To appear in *Psychological Review*

Affixation in Semantic Space: Modeling Morpheme Meanings with Compositional Distributional Semantics

Marco Marelli, Marco Baroni
Center for Mind/Brain Sciences, University of Trento, Italy

Abstract

The present work proposes a computational model of morpheme combination at the meaning level. The model moves from the tenets of distributional semantics, and assumes that word meanings can be effectively represented by vectors recording their co-occurrence with other words in a large text corpus. Given this assumption, affixes are modeled as functions (matrices) mapping stems onto derived forms. Derived-form meanings can thus be thought of as the result of a combinatorial procedure which transforms the stem vector on the basis of the affix matrix (e.g., the meaning of *nameless* is obtained by multiplying the vector of *name* with the matrix of *-less*). We show that this architecture accounts for the remarkable human capacity of generating new words that denote novel meanings, correctly predicting semantic intuitions about novel derived forms. Moreover, the proposed compositional approach, once paired with a whole-word route, provides a new interpretative framework for semantic transparency, that is here partially explained in terms of ease of the combinatorial procedure and strength of the transformation brought about by the affix. Model-based predictions are in line with the modulation of semantic transparency on explicit intuitions about existing words, response times in lexical decision, and morphological priming. In conclusion, we introduce a computational model to account for morpheme combination at the meaning level. The model is data-driven, theoretically sound, and empirically supported, and it makes predictions that open new research avenues in the domain of semantic processing.

Keywords: distributional semantic models, compositionality, word formation, derivational morphology, semantic transparency, novel words

Large-scale statistical models induced from text corpora play an increasingly central role in computational simulations of various aspects of human language processing and acquisition (see, e.g., Dubey, Keller, & Sturt, 2013; Brent & Cartwright, 1996; Hay & Baayen, 2005, for just a few examples). Within this general trend, the last few decades witnessed widespread interest in using methods from *distributional semantics* to obtain quantitative estimates of important but hard-to-operationalize semantic variables such as the degree of conceptual or topical similarity between two words. These methods (whose most famous implementations might be Latent Semantic Analysis, HAL and Topic Models) approximate lexical meanings with vectors that summarize the contexts in which words appear, under the hypothesis that similar words will occur in similar contexts.

However, words are not the smallest meaning-bearing units in a language. Most words are composed by smaller elements consistently associated to specific semantic aspects: *nullify* contains *null* and *-ify*, and means *to make something null*; *driver* contains *drive* and *-er*, and means *someone who drives*. These elements, called morphemes (Bloomfield, 1933), are at the base of the lexical productivity of human languages, that is, their capacity to generate endless novel words that are immediately understandable by native speakers. The examples above fall, in particular, within the domain of *derivational* morphology, where a free-standing morpheme, or word (the stem, e.g., *null*), is combined with a bound element (the affix, e.g., *-ify*) to generate the derived form, that is perceived as a separate lexical item. *Inflectional* morphology generates instead inflected variants of the same item, as in *sing/sings*.

Distributional semantics has already been used in studies of morphological processing, where distributional similarity between a derived form (e.g., *nullify*) and its stem (e.g., *null*) can be employed to estimate the degree of semantic transparency of the complex form, under the assumption that opaque forms should be semantically far apart from their constituents. While this is a promising approach to quantify the

We are very grateful to Roberto Zamparelli and Angeliki Lazaridou, who developed the main ideas presented in this paper together with us. Special thanks to Angeliki for implementing the first version of the FRACSS model we use here. We had many fruitful discussions with the other members of the COMPOSES team, especially Georgiana Dinu, and we received useful feedback from Simona Amenta, Harald Baayen, Sebastian Padó and the audiences at ACL 2013, MOPROC 2013, AIP 2014, and The Mental Lexicon 2014. We also thank John Anderson, David Plaut, Kathy Rastle, and Erik Reichle for constructive criticism. This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

Correspondence concerning the article should be addressed to Marco Marelli, Center for Mind/Brain Sciences, University of Trento, Corso Bettini 31, 38068 Rovereto (TN), Italy. Tel: +39 0464 80 8620. E-mail address: marco.marelli@unitn.it.

degree of semantic relatedness between the starting and end points of a morphological process, the field is still missing an empirical method to characterize the semantic side of the process itself. Thanks to distributional semantics, we have an objective way to measure that, say, *redo* is highly related to its stem *do* whereas *recycle* is not so similar to *cycle*, but we are still missing a model that, given an appropriate meaning representation for an affix, say *re-*, and a stem, say *do* or *cycle*, generates a meaning representation for the corresponding derived forms (*redo* and *recycle*).[1]

This is a big gap. Without an explicit account of how *morphological composition* works on the semantic side, our models of morphological processing are ignoring, paradoxically, one of the core reasons why morphological processes exist, that is, to express new meanings by combining existing morphemes. More concretely, fundamental debates in the literature (e.g., Di Sciullo & Williams, 1987; Sandra, 1994), for example on the extent to which complex words must be listed in the lexicon as "semantically unpredictable", are bound to remain purely theoretical in lack of an objective model of how semantically predictable meanings of complex words should look like.

In this article, we purport to fill the gap. We exploit recent advances in distributional semantics to develop a fully automated and data-induced morphological composition component that, given distributional representations of stems and affixes, produces a distributional representation for the corresponding derived form.

The proposed model can generate distributional representations for the meanings of novel derived words, hence tapping into one of the core functions of derivational morphology, that is, lexical productivity. Therefore, in the first set of experiments we focus on the predictions that our model makes about novel forms. We show that certain quantitative properties of compositionally-obtained semantic representations of nonce forms obtained with our models are significant predictors of subject intuitions about their semantic meaningfulness (*harassable* and *windowist* are equally unattested in a very large corpus, but participants found the first highly acceptable, the second meaningless). We show moreover that words that our model automatically picks as highly related to composed nonce forms ("nearest neighbours") are indeed closer in meaning to the nonce forms than to other terms, including their stems, according to subject judgments.

Next, we apply our compositional model to existing derived forms in three experimental case studies. We use the compositional model to account for behavioral patterns influenced by the degree of *semantic transparency* between a derived form and its stem. More specifically, we let our model predict explicit semantic relatedness

[1]Of course, it is fairly standard in the theoretical morphological literature to use place-holders or even carefully crafted feature structures to represent the meaning of stems and affixes (e.g., Lieber, 2004). For example, the meaning of *re-* might be equated to a [+iterate] feature. However, these featural representations are not detailed and flexible enough to make quantitative predictions about the morphological phenomena studied in the experimental literature. Moreover, hand-crafting feature structures is only feasible for closed-class affixes, as there are relatively few of them and their meanings are very general, and leaves stem representations largely unspecified.

intuitions, modulate stem frequency effects in a lexical decision task, and account for morphological priming results. These successful experiments demonstrate that, when equipped with an appropriate semantic combination module, an approach in which complex words are derived compositionally can predict effects associated to different degrees of semantic transparency that are interestingly complementary to those that are best captured by relying on full-form representations for the meanings of opaque words. Overall, a more nuanced picture of semantic transparency effects emerges from our experimental results.

Taken together, the evidence presented in this paper indicates that our compositional distributional semantics framework provides an effective meaning layer for simulations of morphological processing. The richer, more flexible meaning composition rules that our system learns from data capture a wider range of composition patterns than just fully transparent ones, and have important theoretical implications for the development of models of word meanings. The model can be profitably used to obtain data-driven, quantitatively-defined semantic representations for complex forms, irrespective of them being well-known or never heard before.

Semantic aspects of morphological processing

The psycholinguistic literature has long investigated the role of morphology in word recognition (Taft & Forster, 1975). This line of research suggests that morphological information influences the way a word is processed beyond pure semantic and form similarity. Priming experiments (Feldman, 2000; Rastle, Davis, Marslen-Wilson, & Tyler, 2000) showed that presenting a morphological related prime before the target (e.g., *cattish-cat*) leads to larger facilitations in response times compared to using semantically (*dog-cat*) or form-related primes (*cattle-cat*).

Although these results indicate that morphology cannot be reduced to a by-product of semantic similarity, the semantic component appears to be important in many morphological effects. Indeed, the degree of semantic transparency of words modulates the mentioned priming effect (W. Marslen-Wilson, Tyler, Waksler, & Older, 1994; Feldman & Soltano, 1999; Rastle et al., 2000): if the meaning of a complex word is associated to the meaning of its constituents (*dealer-deal*), the priming effect will be larger than that observed for opaque pairs (*courteous-court*). This modulation on morphological priming, although it might differ in magnitude, can be observed across several languages and experimental manipulations (e.g., Diependaele, Sandra, & Grainger, 2005; Rueckl & Aicher, 2008; Feldman, O'Connor, & Moscoso del Prado Martín, 2009; Diependaele, Sandra, & Grainger, 2009; Kazanina, 2011; Järvikivi & Pyykkönen, 2011; Feldman, Kostić, Gvozdenović, O'Connor, & Prado Martín, 2012; Marelli, Amenta, Morone, & Crepaldi, 2013, but see Frost, Forster, & Deutsch, 1997). Later in the paper, we will come back to the semantic transparency issue and discuss it in detail, since it will play a central role in the empirical assessment of our model. For the present discussion, it is sufficient to conclude from the relevant experimental evidence that the processing of a complex word is influenced by the semantic

properties of the elements it is made of.

This notion is also supported by the literature on family size effects. The family size of a complex word is computed as the count of the distinct words that contain the same stem. The variable has a facilitatory effect on word recognition for both complex (Bertram, Baayen, & Schreuder, 2000) and simple (Schreuder & Baayen, 1997) word processing. Crucially, the nature of the family size effect is essentially semantic. First, it emerges only late (i.e., at central processing levels) in a progressive demasking condition (Schreuder & Baayen, 1997). Second, it works better as a predictor for response latencies if opaque forms (e.g., *cryptic* when considering the family size of *crypt*) are excluded from counting (Bertram et al., 2000; Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004). Third, the effect can be dissociated from the influence of family frequency (e.g., the cumulative corpus counts of morphological relatives), which is believed to be associated to visual familiarity (De Jong, Schreuder, & Baayen, 2000). Fourth, irregular relatives crucially contribute to the effect in virtue of their semantic connection and despite their orthographic dissimilarity with the word (De Jong et al., 2000). Fifth, the effect of family size interacts with other semantic dimensions (e.g., concreteness) of the target word (Feldman, Basnight-Brown, & Pastizzo, 2006). Sixth, family size is predictive of monolinguals' lexical decision latencies across unrelated languages (Moscoso Del Prado Martín et al., 2005). Taken together, these results indicate that the morphological relations entertained by a word play a role when that word is read, and this happens on the basis of meaning-mediated associations. Data from Finnish (Laine, 1999; Lehtonen, Harrer, Wande, & Laine, 2014) further suggest that stems and affixes may be differentially represented at the semantic level: in a semantic decision task on inflected words, violations affecting the affix were more difficult to reject than violations affecting the stem, indicating that suffix-related information is secondary to stem meaning.

Additional supporting evidence for the importance of semantics in morphology comes from studies of compound processing. Recent work has shown that the semantic properties of the individual constituents (*year* and *book*) influence the recognition of the whole compound (*yearbook*), either because of activation/interference from their core meanings (Ji, Gagné, & Spalding, 2011; Marelli & Luzzatti, 2012) or for an effect associated with their emotional valence (Kuperman, 2013).

In conclusion, there is plenty of evidence that morpheme semantics plays a role in the processing of complex words: morphological effects are not simply dependent on superficial, formal similarities between morphologically related words, but also involve access to morpheme meanings. Even more importantly, the ecological purpose of word processing is *comprehension*: a crucial question that any complete model of morphological processing should address is thus how we understand the meaning of a morphologically complex word, and as a consequence how morpheme meanings are represented in the semantic system.

These empirical and theoretical considerations notwithstanding, morphological processing models often lack a detailed description of how morphemes are repre-

sented at the semantic level, mostly focusing on early, orthographic-lexical levels of word recognition. In some cases (e.g., Crepaldi, Rastle, Coltheart, & Nickels, 2010) the architecture of the semantic system is left purposely underspecified. A similar approach is adopted by Taft (2004): the meaning level is generically described as containing "semantic information", that is in turn activated by lemma representations assumed to be holistic for opaque words (*cryptic*) and morpheme-based for transparent words (*re-do*). Interestingly, an explicit (as well as essential for model building) assumption about semantic representations was made within the connectionist framework. This class of models explains word recognition by using distributed hidden layers interfacing orthographic and semantic information (Plaut & Gonnerman, 2000). In these architectures, the semantic level is populated by subsymbolic nodes representing semantic features; word meanings are then represented as activation distributions across these nodes. Indeed, since such models do not conceive a lexical system specifying representations for morphological units, they explain morphological effects as a by-product of the large overlap, in terms of both form and meaning, between a derived form and its stem: morphology is seemingly important because *read* and *reader* have similar distributed representations, but *read* is not actively used to construct the meaning of *reader* online. In opposition to this, some models of word processing (Caramazza, Laudanna, & Romani, 1988; Baayen & Schreuder, 1996; Baayen, Milin, Durdević, Hendrix, & Marelli, 2011) postulate stored morpheme representations in the semantic system. These models assume that stems and affixes are eventually combined to obtain the whole-word meaning, but how this procedure unfolds is left unspecified, at least from a computational point of view. Meaning composition has been instead central to the study of novel compounds, that is, research on conceptual combination between content words (e.g., Gagné & Spalding, 2009; Costello & Keane, 2000). Although these models might provide some insight as to how also affixes are processed, combining two content words (e.g., *stone+squirrel*) and combining a root and an affix (e.g., *stone+ful*) are relatively different operations, each of them subtending its own procedures and posing its own problems.

In conclusion, although different assumptions have been made time after time, the psycholinguistic literature is generally lacking detailed descriptions of how affixed words are represented and processed at the meaning level. This gap is puzzling, all the more so when considering that the time course of the semantic influence on morpheme processing is one of the central issues of current psycholinguistic research on lexical morphology (e.g., Rastle & Davis, 2008; Feldman et al., 2009).

Distributional semantic models

Distributional semantic models (DSMs) automatically extract word meaning representations from large collections of text, or *corpora*. Recent surveys of these models include Clark (2015), Erk (2012), Lenci (2008) and Turney and Pantel (2010).

DSMs rely on the idea, known as the *distributional hypothesis* (Firth, 1957; Harris, 1954; Miller & Charles, 1991), that if two words are similar in meaning they will

	runs	*wheels*
automobile	1	4
car	2	4
horse	4	1

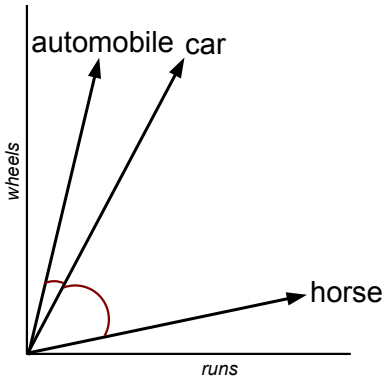

Figure 1. In this toy DSM example, the target words *automobile*, *car* and *horse* are represented by the vectors on the left, recording the number of times they co-occur with the context terms *runs* and *wheels* in a hypothetical corpus. The vectors are represented geometrically on the right, where we see that those for *automobile* and *car*, that share similar contextual patterns, form a narrower angle (in real DSMs, the vectors would have hundreds or thousands of dimensions).

have similar distributions in texts, that is, they will tend to occur in similar linguistic contexts. But then, by inverting the postulated dependency between meaning and context, we can use context similarity to infer meaning similarity. In concrete, most DSMs represent the meaning of a word with a vector that keeps track of how many times the word has occurred in various contexts in a corpus (where contexts are, for example, documents or other words co-occurring with the target in a short passage). Thanks to this representation of word meaning, DSMs can quantify semantic relatedness with geometric methods, in particular by measuring the width of the angle formed by the vectors associated to the words of interest. A toy example illustrating the idea is sketched in Figure 1.

DSMs differ widely in terms of what counts as context, how raw co-occurrence counts are weighted and whether dimensionality reduction techniques are applied. Indeed, some famous DSMs in cognitive science are derived by fixing some parameters in the construction of semantic spaces. For example, LSA (Landauer & Dumais, 1997) is a semantic space based on a word-document co-occurrence matrix, to which Singular Value Decomposition is applied for dimensionality reduction; on the other hand, HAL (Lund & Burgess, 1996) is built from word-to-word co-occurrences, whose collection can be delimited by different window sizes. More recently, Neural Language Models (Collobert et al., 2011; Mikolov, Chen, Corrado, & Dean, 2013) induce vectors trained to *predict* contextual patterns, rather than directly encoding them.

DSMs have several desirable properties as computational models of meaning for cognitive simulations. First, they induce meaning from large amounts of naturally occurring linguistic data (the source corpora), not unlike what children and teenagers must do in order to acquire the huge vocabularies that adults command (of course, while the input might be similar, nobody claims that the mechanics of DSM induction are plausible acquisition models; Landauer & Dumais, 1997). Second, DSMs can easily induce and encode meaning representations for thousands or even millions

of words, making them very practical in the design of experiments and simulations. Third, DSMs naturally provide a graded notion of meaning (via the continuous similarity scores they produce), in accordance with mainstream "prototype" views of lexical and conceptual meaning (Murphy, 2002).

Indeed, DSMs have been found to be extremely effective in simulating an increasingly sophisticated array of psycholinguistic and lexical-semantic tasks, such as predicting similarity judgments and semantic priming, categorizing basic-level nominal concepts or modeling the selectional preferences of verbs (e.g., Baroni, Barbu, Murphy, & Poesio, 2010; Landauer & Dumais, 1997; Lund & Burgess, 1996; McDonald & Brew, 2004; Padó & Lapata, 2007; Erk, Padó, & Padó, 2010). As this aspect will be relevant to our nonce form acceptability experiments below, we remark that, although DSMs are mostly used to measure vector similarity as a proxy to semantic relatedness, intrinsic properties of distributional vectors, e.g., their length and entropy, have also recently been shown to be of linguistic interest (Kochmar & Briscoe, 2013; Lazaridou, Vecchi, & Baroni, 2013; Vecchi, Baroni, & Zamparelli, 2011).

Not surprisingly, morphological processing scholars have seized the opportunity offered by distributional methods, and it has become almost standard to use DSMs to quantify the degree of semantic transparency of derived or compound words in terms of geometric distance of the morphologically complex form from its stem (in derivation) or its constituents (in compounding). For example, Rastle et al. (2000) used LSA to quantify the degree of semantic relatedness between morphologically related primes and targets in a study of visual word recognition (e.g., *depart* and *departure* are semantically close, *apart* and *apartment* are far). Other studies using LSA in similar ways include Rastle, Davis, and New (2004); Moscoso Del Prado Martín et al. (2005); Milin, Kuperman, Kostić, and Baayen (2009); Feldman et al. (2009); Gagné and Spalding (2009); Diependaele, Duñabeitia, Morris, and Keuleers (2011). Heylen and De Hertog (2012) used DSMs together with other distributional cues to predict the degree of semantic transparency of Dutch compounds. Working with English and Chinese, Wang, Hsu, Tien, and Pomplun (2013) found good correlations between constituent-to-compound similarities as measured by LSA and human transparency judgments. As a final example, Kuperman (2009) found that operationalizing constituent-to-compound semantic similarity in terms of LSA scores led to reliable transparency effects in lexical decision and eye-movement latencies.

While the studies we just reviewed provide evidence for the usefulness of DSMs in morphology, a crucial ingredient is missing. Standard DSMs provide representations for the words that constitute the input and output of a morphological process (*consider* and *reconsider*, *contain* and *containment*, etc.), but they have nothing to say about the process itself, and the meaning of the morphemes that trigger it (how does *re-* contribute to the meaning of *reconsider*). Without a way to model meaning composition in word formation, traditional DSMs are rather poor semantic surrogates for the study of morphology.

A related issue arises when trying to scale DSMs to handle meaning *above* the

word level. DSM proponents have indeed been interested, from the very start, in ways to derive the meaning of phrases, sentences and whole passages from the combination of the distributional representations of their constituent words (Landauer & Dumais, 1997). However, despite the similar compositional challenge, most of the approaches suggested for constituents above the word do not adapt seamlessly to derivational morphology, since they rely on the assumption that the input to composition is a set of word-representing vectors (Guevara, 2010; Mitchell & Lapata, 2010; Socher, Huval, Manning, & Ng, 2012; Zanzotto, Korkontzelos, Falucchi, & Manandhar, 2010). For example, a simple and surprisingly effective approach constructs the vector representing the meaning of a phrase by summing the vectors of its constituent words. The problem with extending this method to morphological derivation should be clear. If we want to derive a representation for *reconsider*, we can use the corpus-harvested vector for *consider*, but how do we get a vector for *re-*, given that the latter never occurs as an independent word? There are of course *ad-hoc* ways around this problem. For example, in Lazaridou, Marelli, Zamparelli, and Baroni (2013), we created a vector for *re-* by simply accumulating co-occurrence information from all words prefixed with *re-* in the corpus.[2] However, such artificial solutions are not required by the theoretically grounded functional model we are about to introduce.

The *functional* approach to DSM composition was proposed by Baroni and Zamparelli (2010) and further developed by Baroni, Bernardi, and Zamparelli (2014) (Clark, 2013, Coecke, Sadrzadeh, & Clark, 2010, and Grefenstette & Sadrzadeh, 2011, present a closely related framework).[3] The approach follows formal semantics in characterizing composition as *function application*. For example, an adjective modifying a noun (*red car*) is treated as a function that takes the noun vector as input, and returns a modified phrase vector as output. This naturally extends to derivation, where we can think of, e.g., the prefix *re-* as a function that takes a verb vector as input (*consider*) and returns another verb vector with an adjusted meaning (*reconsider*) as output. No independent vector representation of affixes is assumed. This is the approach we will pursue in this article.

Combinatorial versus full-form meaning

In the next section we will introduce our implementation of a distributional semantic model equipped with a derivational composition component based on the functional approach. Of course, a purely combinatorial procedure for morphemes is not the only possible solution for the role of morphology in the mental lexicon, and not necessarily the best one. It may be tempting to conceive a semantic system populated by full-form meanings (i.e., separate representations for *run, runner, homerun*)

[2]Luong, Socher, and Manning (2013), more elegantly, learn vector representations of morphemes within a recursive neural network architecture trained to predict word n-gram contexts.

[3]Interestingly, a first sketch of the functional approach was developed by Guevara (2009) in the context of modeling derivational morphology, although Guevara did not evaluate his method in quantitative terms.

and explain alleged morphological effects as by-products of semantic and formal similarity, and/or lexical links between related whole-word representations. This solution permits dealing with the idiosyncratic semantics characterizing (to different degrees) nearly all complex words. It can also handle cases where a complex form contains a reasonably transparent affix meaning but the stem is not a word: *grocer*[4] clearly displays the agentive sense of *-er*, but *to groce* is not a verb, so the noun cannot be derived compositionally.[5] However, holistic meanings by themselves fall short in explaining the surprising productivity of morphological systems. Native speakers of a language are able to build new words by means of existing morphemes, and people in the same linguistic community are immediately able to understand the meanings of these novel constructs: *herringless dish* can be clearly assigned the meaning of *dish without herrings*, even if the word *herringless* does not appear in English dictionaries and it has likely never been heard before by the listener. Any model of the semantic system should be able to explain these phenomena, but in order to do so the cognitive architecture needs some representation for morpheme meanings, as well as a combinatorial procedure operating on them.

Assuming a combinatorial process does not exclude the possibility that holistic meanings may also be stored in the semantic system together with separate morphemic entries, and the phenomena we discussed above indeed suggest that both full-form representations and a combinatorial route are called for. This makes a purely full-form meaning approach and a mixed one difficult to disentangle from an empirical point of view. Still, providing a computationally-defined formalization of the combinatorial mechanism will permit one to assess to what extent the meaning of a complex word can be predicted by systematic processes, and conversely help to determine when a complex word really needs a holistic meaning representation of its own.

A distributional model for morpheme combination

Distributional semantic space

We mentioned in the introductory section on distributional semantics that DSMs greatly vary in terms of how co-occurrences are defined and which mathematical transformations are applied to the co-occurrence matrix (Turney & Pantel, 2010). We adopt here a set of parameters that led to top performance in previous empirical tests (e.g., Boleda, Baroni, McNally, & Pham, 2013; Bullinaria & Levy, 2007, 2012). For model implementation we relied on the freely available DISSECT toolkit (Dinu, Pham, & Baroni, 2013a).

[4]We owe the example to David Plaut.

[5]Although representing the whole-word meaning of *grocer* is out of its scope, a compositional approach can still capture the general semantic properties associated to the *-er* affix appearing in this form. For more on this, see the general discussion in the novel words section.

We extracted co-occurrence data from the concatenation of the widely used ukWaC (`http://wacky.sslmit.unibo.it/`), English Wikipedia (`http://en.wikipedia.org/`), and BNC (`http://www.natcorp.ox.ac.uk/`) corpora (about 2.8 billion tokens in total). The words in these corpora have been automatically mapped to dictionary forms and annotated with their parts of speech. As a consequence, in the resulting DSM, (a) separate vector representations are stored for homographs with different grammatical class (e.g., a vector for the noun *run* and a vector for the verb *run*), and (b) different inflectional forms are represented by the same vector (e.g., the occurrences of *speak, speaks, spoke* are all used to construct a single *speak* vector). In model building, we considered the top 20,000 most frequent content words (adjectives, adverbs, nouns, verbs), along with any lexical items used during the affix function training phase (described in the next subsection).

Word-to-word *co-occurrence counts* were collected by imposing a 5-word context window, that is, each target word was considered to co-occur with the two (content) words preceding and following it. Window-based lexical co-occurrences, as in HAL, have proven to be optimally-performing in a number of semantic tasks, and are also attractive for their simplicity in comparison to collocates based on syntax-based links (e.g., Sahlgren, 2008; Bruni, Boleda, Baroni, & Tran, 2012). Narrow-window collocates (as opposed to the word-document co-occurrences used by LSA) usually entail very close semantic and syntactic relations, and this approach is hence expected to capture a more locally-based kind of semantic similarity, such as the one found in close taxonomic relations (Sahlgren, 2006). This granular property of the narrow-window approach is all the more attractive for our purpose, since morphological derivation will typically change meaning in rather subtle ways that might be missed by a coarse context representation (e.g., *-ly* simply transforms an adjective into an adverb, *-er* adds an agentive role, denoting someone doing the action the verb stem describes).

Weighting schemes are usually applied to raw co-occurrence counts in order to best capture the information they carry by down-playing the role of chance co-occurrence. In the present study, we adopted (nonnegative) Pointwise Mutual Information (PMI, Church & Hanks, 1990), an information-theoretic measure of association widely used in computational linguistics. Given target word t and context word c, PMI is computed as follows:

$$PMI(t, c) = \log \frac{p(t, c)}{p(t)p(c)}$$

The measure compares the probability of co-occurrence of two words estimated directly from the corpus with the probability of those two words co-occurring by chance, and hence quantifies the extent to which their co-occurrence is not random. Consider, for example, the word pairs *the+dog* and *dog+barks*. Even if *the+dog* is likely much more frequent than *dog+barks*, the PMI of *the+dog* is much lower, since the association between the two words is not meaningful: simply, *the* is so fre-

quent that it is likely to co-occur with any noun in the corpus. On the other hand, *dog+barks* will have a high PMI score, since their co-occurrence is far from being random, being based on the semantic and syntactic association between the two words. The nonnegative version of PMI we apply here (negative values are replaced by zeros) was shown to lead to high-performance models by Bullinaria and Levy (2007). Landauer and Dumais (1997) have speculatively related such information-based association measures to the Rescorla-Wagner formalization of discriminative learning (Rescorla & Wagner, 1972). Indeed, the higher the PMI, the more informative a context word will be, and informativeness of a cue (in this case, the contextual collocate) is strongly associated to its discriminative power: *the*, being associated to a large number of different words, is not a good discriminative cue for any of them; on the other hand, where *barks* occurs, the presence of *dog* is also expected. Therefore, the information-weighted encoding of word co-occurrences in DSMs is intuitively similar to the way organisms create simple associations between phenomena. This similarity could explain why DSMs perform so well as models for the human semantic system.

Dimensionality-reduction techniques perform a mapping of the data to a lower dimensional space while trying to preserve certain properties of the original full-space representations (e.g., variance). This procedure makes the data matrix easier to handle, but its purpose is not purely practical. Landauer and Dumais (1997) consider the reduced dimensions as an analogue to abstract semantic features emerging from the co-occurrence of superficial elements. This hypothesis was further developed by Griffiths, Steyvers, and Tenenbaum (2007) with Topic Models. In their proposal, dimensionality reduction techniques identify a set of topics emerging from word distributions; word meanings (or *gist*) can in turn be modeled as probability distributions across topics. In place of the better-known Singular Value Decomposition (Landauer & Dumais, 1997) and Latent Dirichlet Allocation (Griffiths et al., 2007) methods, in the present study we performed dimensionality reduction by Nonnegative Matrix Factorization (NMF). This technique leads to a significant improvement in model performance (Arora, Ge, & Moitra, 2012; Boleda et al., 2013), and the dimensions it produces have been shown to be comparable to the interpretable topics of Topic Models (Dinu & Lapata, 2010). On the basis of recent empirical results and without our own tuning, we set the number of dimensions to 350.

The semantic space resulting from these operations is a set of approximately 20,000 350-dimensional vectors, each representing a word meaning. These define a multidimensional space in which geometric proximity can be treated as a proxy for contextual, and hence semantic similarity. In concrete, semantic similarity is measured as the width of the angle formed by two vectors. More technically, following standard DSM practice, we quantify the angular distance between vectors by the *cosine* of the angle they form (the narrower the angle, the higher the cosine, that is, the more similar the words being compared are expected to be). Given two vectors $\vec{a}$ and $\vec{b}$, their cosine is computed as follows:

$$\cos(\vec{a},\vec{b}) = \frac{\sum_{i=1}^{i=n} a_i \times b_i}{\sqrt{\sum_{i=1}^{i=n} a_i^2} \times \sqrt{\sum_{i=1}^{i=n} b_i^2}}$$

When all vector components are nonnegative, as in our case, the cosine is also nonnegative, and it ranges from 0 for perpendicular vectors to 1 for parallel vectors.

Inducing functional representations of affixes

Using the distributional semantic space described above as our starting point, we now proceed to build affix representations. Following the functional approach in compositional distributional semantics (Baroni & Zamparelli, 2010), affixes can be seen as functions modifying the semantics of word stems to obtain new meanings.

Specifically, Baroni and Zamparelli, for reasons of elegance, interpretability and computational tractability, restrict composition functions to the class of linear transformations (but see Appendix B on the nature of this restriction), so that words or affixes encoding functions can be represented by coefficient matrices, and function application corresponds to vector by matrix multiplication (Strang, 2003). When an n-dimensional (row) vector is multiplied by a $n \times n$ matrix, the output is another n-dimensional vector.[6]

The value in the i-th dimension of the output is a weighted sum of all input vector dimensions, each multiplied by the corresponding coefficients in the i-th column of the matrix. Thus, the matrix representing an affix encodes how much each input vector dimension affects each dimension of the derived output representation.

Our affix-specific coefficient matrices constitute Functional Representations of Affixes in Compositional Semantic Space (FRACSSs), and suit derivational morphology particularly well: not only are they in line with the view of affixes as functional elements from descriptive and theoretical linguistics, but they are also in accordance with psycholinguistic results indicating that, at the semantic level, stems are accessed first and affix meaning enters the picture only subsequently (e.g., Laine, 1999).

Let's clarify how FRACSSs operate with a toy example. Assume that when the prefix *re-* attaches to an activity verb *V* it has an iterative meaning, more or less *to V again* (cf. *sing* and *re-sing*). When it attaches to an accomplishment verb such as *open*, the meaning is instead restitutive: *to re-open (the door)* does not imply that the door was ever opened before, just that it is no longer closed (this account of the semantics of *re-* is so enormously simplified as to be flawed; see, e.g., Lieber, 2004, for a more nuanced story). Let's assume moreover that *continuously* is a typical contextual feature of activity verbs, with high scores in their distributional vectors, and similarly for *completely* in accomplishment verbs. We take the contextual feature

[6]For simplicity, we ignore in this exposition the intercept row, that actually makes our matrices $(n+1) \times n$-dimensional. Nothing substantial changes, but see the discussion at the end of the novel words section on how the intercept might be interpreted as capturing the "average" meaning of the derived forms sharing the same affix.

Table 1
Toy FRACSS matrix representing the prefix re-*. Each column contains the weights determining the impact of each input dimension (associated to the corresponding row label) on the value of the output dimension corresponding to the column label.*

	d1	*d2*	*completely*	*continuously*	*back*	*again*
d1	1	0	0	0	0	0
d2	0	1	0	0	0	0
completely	0	0	1	0	2	0
continuously	0	0	0	1	0	2
back	0	0	0	0	1	0
again	0	0	0	0	0	1

Table 2
Toy distributional vectors before and after multiplication by the re- *FRACSS matrix in Table 1 (words associated to vectors in the row labels, contextual dimensions in the column labels).*

	d1	*d2*	*completely*	*continuously*	*back*	*again*
sing	3	2	0	2	0	0
re-sing	3	2	0	2	0	4
open	1	3	1	0	1	0
re-open	1	3	1	0	3	0

again to be very salient in verbs expressing iterative meanings, and *back* in restitutive readings. Finally, let's assume that verbs live in a very narrow 6-dimensional space, where dimensions *d1* and *d2* pertain to characteristics that are not affected by *re-* (e.g., how the actions denoted by verbs are performed). Then, the *re-* FRACSS might look as in Table 1. Each cell of this matrix states how much the input dimension corresponding to a row label will affect the output dimension in a column label: for example, the 0 in the second cell of the first row tells us that the input *d1* dimension has no effect on the output *d2* dimension. Importantly, the fifth and sixth columns of the table contain the weights that the input vector values will be multiplied by to obtain the output *back* and *again* dimension values, respectively. In the case of *back*, the output value will be magnified by summing to the input *back* value twice the input *completely* value, and similarly for *again* with respect to *continuously*. Table 2 shows how the FRACSS operates on hypothetical *sing* and *open* vectors, illustrating how the same matrix multiplication operation (equivalently: linear function application) correctly emphasizes the iterative dimension of the first verb, the restitutive dimension of the second. Realistic distributional vectors and matrices will contain, of course, hundreds or thousands of cells (in our semantic space, vectors have 350 dimensions, FRACSS matrices 350×350 cells), allowing a much richer multivariate representation of factors such as iterativity, and a much more nuanced treatment of how input and output dimensions interact in semantic transformations. It is also worth remarking that FRACSS matrices can also be seen as vectors in a higher dimensional space, and possess meaningful semantic properties in their own right, e.g., similar affixes should have similar matrix representations (Baroni & Zamparelli, 2010; Baroni et al., 2014).

The weights to fill the actual FRACSS cells are estimated from corpus-extracted examples of input-output pairs of the relevant function application using standard least-squares methods (Dinu, Pham, & Baroni, 2013b). The intuition is that an affix is the carrier of a transformation, so we want to learn its representation from pairs that illustrate the transformation it carries through. To estimate the *re-* FRACSS, for example, we might use corpus-extracted distributional vectors of pairs such as ⟨*do*, *re-do*⟩, ⟨*think*, *re-think*⟩, etc. The *re-* FRACSS coefficients are set so that, on average, when the example input vectors are multiplied by them, they produce output vectors that are geometrically close to their corpus-extracted equivalents (in the running example, weights are set so that multiplying the *do* vector by the *re-* matrix will produce a vector that is close to the corpus-extracted *redo* vector, etc.). Once the FRACSS is estimated, it can of course be applied to arbitrary vectors that were not part of the training examples to generate new derived forms (for example, the matrix in Table 1 might have been estimated on examples such as ⟨*do*, *re-do*⟩, ⟨*think*, *re-think*⟩, but once its weights have been fixed it can be applied to the morphologically simple vectors of Table 2 – that were not used as training data – to generate the corresponding prefixed forms).

If vector representations indicate how the usage of a certain word is distributed over the set of contexts, FRACSSs, because of the way they are estimated, will capture systematic patterns linking two separate context distributions. For example, for the agentive *-er*, FRACSS will represent the association between contextual representations of certain actions (e.g., *deal, run, drink, drive*) and contextual representations of entities able to perform (or usually performing) those actions (e.g., *dealer, runner, drinker, driver*). In the previous section, we proposed that a simple learning process resulting in the storage of word-to-word associations is at the base of the induction of distributional models (Landauer & Dumais, 1997). The same principle of development of association between phenomena on the basis of statistically reliable co-occurrence patterns is at the basis of FRACSS learning. However, whereas in the case of word associations the processed phenomena are words in context, FRACSSs capture association between the "meanings" (distributions over contexts) encoded in distributional vectors. In other terms, in the present proposal, affixes are to be considered as high-order associations between the distributional semantics of different words (or, better, word sets). These associations are hidden within natural language usage, but may emerge when the apt statistical learning procedure is applied.

We trained FRACSSs for 34 affixes using the DISSECT toolkit (Dinu et al., 2013a). Each affix was associated to a training set composed by at least 50 stem/derived-form pairs, obtained by exploiting the morphological annotation from the CELEX English Lexical Database (Baayen, Piepenbrock, & Gulikers, 1995). The pair elements matched the most common part-of-speech signature for the corresponding affix (e.g., *-er* pairs had verbal stems and nominal derived forms). For each pair, both elements occurred at least 20 times in our source corpus. The Appendix reports a list of the affixes together with information on the associated morpho-syntactic

transformations and number of training data used.

In principle, picking training examples from CELEX could have reduced the naturalness of the experimental setup, by favouring examples of productive, transparent, and/or synchronic affixation. In practice, this was not the case. The CELEX morphological segmentation was performed semi-automatically, and it is extremely liberal in parsing words as morphologically complex. Indeed, most of the words typically included as opaque items in priming experiments are tagged as morphologically complex in CELEX (e.g., 61% of the opaque words in Rastle et al., 2000, and 71% of the opaque words in W. D. Marslen-Wilson, Bozic, & Randall, 2008). As a consequence, words like *listless, department, corny, seedy, whisker, audition, awful, virtual, archer* are included in our training set. These forms show that picking training examples from CELEX does not add much unintended supervision to the setup. Essentially, it is equivalent to picking training examples by using simple surface-based distributional heuristics that should not be beyond the reach of human language learners, and have been shown in many computational simulations to suffice in providing reasonable parses of complex forms (see Goldsmith, 2010).

Examples of FRACSS-derived representations

Before we delve into the experiments that will bring quantitative support to our approach, it is interesting to inspect, qualitatively, the semantic representations of affixed forms composed with FRACSS. The examples discussed here show how such representations mostly reflect the meanings that we would assign to the corresponding derived forms, sometimes capturing surprisingly nuanced details. We will see that the linear matrix-multiplication approach we take, thanks to the flexibility afforded by representing each separate affix with a different matrix and the interplay of input vector dimensions and matrix weights, provides enough room to learn representations that can capture affix polysemy, pick the right sense of the stem, and handle different sub-classes of stems differently.

We conduct the qualitative analysis by inspecting derived-form vectors constructed via multiplication of a stem vector by a FRACSS matrix. For example, the *cellist* vector we will discuss results from multiplying the *cello* vector by the *-ist* matrix. We then assess what is the meaning that the model has produced by looking at the nearest neighbors of the composed vector, that is, its closest vectors (cosine-wise) in our distributional semantic space. In particular, all the neighbors we discuss here are among the nearest 20 to each composed form of interest, in a space containing more than 20,000 items.[7] The cases discussed here were picked as good representatives of various phenomena, but they are by no means exceptional with respect to the

[7]The semantic space where we search for neighbors contains only vectors directly extracted from the corpus, also for derived forms. For example, when we say that *flutist* is a close neighbor of (composed) *cellist*, we mean that the vector we constructed from the corpus contexts of *flutist* (as a whole word) is one of the nearest – the nearest one, actually – to the vector we constructed multiplying the *cello* vector by the *-ist* matrix.

larger set of about 4,000 composed items (taken from the stimuli of the experiments below) that we scanned when looking for examples.

We start with some cases illustrating how FRACSS representations capture different senses of the same affix. The *-er* matrix, for example, produces an agent (*V-er = X who Vs*) meaning from *carve* but an instrument one (*V-er = X used for Ving*) when the input is *broil*. Consequently, among the neighbors of *carver* we find a number of other craftsmen performing related activities, e.g., *potter*, *engraver*, *goldsmith*. On the other hand, the *broiler* neighbors are tools such as *oven*, *stove*, as well as other words related to the function of *broilers*: *to cook*, *kebab*, *done*.

While many forms in *-ment* are ambiguous between a process and a result interpretation (*achievement* as the act of achieving vs. what has been achieved), with some stems one of the two readings is much more likely. The FRACSS for *-ment* appears to have captured the difference: For *interment*, the neighbors strongly cue a process reading. There are verbs such as *inter*, *cremate* and *disinter*, as well as other nouns with a dominant event meaning: *burial*, *entombment*, *disinterment*, *funeral*... On the other hand, for *equipment*, where the result reading is more prominent, we find neighbors that are clearly related to *equipment* as a set of physical tools: *maintenance*, *servicing*, *transportable*, *deploy*.

Marchand (1969), in his classic treatment of English morphology, distinguishes between *full of N* and *causing N* senses of *-ful*, that are indeed reflected in the *-ful* FRACSS. The neighbors of *careful* cue the *full of care* sense: *judicious*, *attentive*, *compassionate*... Those of *dreadful* imply instead the correct *causing dread* sense: *frightful*, *horrible*, *unbearable*, etc.

While it's hard, based on nearest neighbor evidence alone, to decide if the *re-* FRACSS is capturing the difference between the main iterative and restitutive senses of the prefix, the difference between the default iterative sense and some more marked intepretations does emerge. So, for iterative *reissue* we clearly detect the reference to a former *issuing* event in neighbors such as *original*, *expanded* and *long-awaited*. On the other hand, the neighbors of *retouch* cue the strong presence of a "correction" sense: *repair*, *refashion*, *reconfigure*... The unusual intensifying sense of the prefix is also captured, as shown by the neighbors of *resound*: *reverberate*, *clangorous*, *echo*, etc.

Other cases highlight how a FRACSS can pick up the right sense of a stem even when the latter is unusual. For example, the nearest neighbors of the noun *type* indicate the prominence of the computational and "typing" senses: *keyword*, *subtype*, *parse*... However, *-ify* correctly selects the "characteristic example" sense in *typify*, cf. neigbours such as *embody*, *characterize*, *essentially* (more distant neighbors such as *codify* and *subsume* suggest that the, now secondary, computational sense has also been preserved). The architectural and mathematical senses dominate the neighbors of *column*: *arch*, *pillar*, *bracket*, *numeric*... However, *-ist* correctly emphasizes the journalistic sense in *columnist*, cf. nearest neighbors such as *publicist*, *journalist*, *correspondent*.

Beyond *columnist*, the *-ist* suffix possesses many nuances that are accurately captured by its FRACSS. So, a *cellist* is someone who plays the *cello* (neighbors: *flutist*, *virtuoso*, *quintet*...). An *entomologist* on the other hand is an expert of *entomology*, which is quite near the disciplines of her/his neighbors: *zoologist*, *biologist*, *botanist*... For *propagandist*, we get the right connotation of devotion to a political cause, that the suffix carries when combined with the relevant class of stems (nearest neighbors: *left-wing*, *agitator*, *dissident*, etc.). As a final example, a *rapist* belongs to the felon class of *-ist* derivatives, with nearest neighbors such as *extortionist*, *bigamist* and *arsonist* (and, among the neighbors that do not contain *-ist*, *pornographer*, *criminal*, *pimp*).

The same stem vector might produce quite different derived vectors when multiplied by different FRACSS. Among the nearest neighbors of *industrial*, for example, we find *environmental*, *land-use* and *agriculture*, whereas among those of *industrious* we see *frugal*, *studious* and *hard-working*.

In all the examples above, the patterns induced by FRACSS might be very specific, but they still have some degree of systematicity: For example, the need to account for the corpus-observed contexts of terms such as *essayist*, *journalist* and *novelist* during the matrix estimation phase (see section on training FRACSS above) must have led to the *-ist* FRACSS matrix encoding the correct generalization for *columnist*. These semantic sub-regularities resemble the *islands of reliability* described by Pinker and Prince (1988) and Albright and Hayes (2002) for the phonological side of morphological combination. In their models, very general, "regular" morphological rules (such as "append *-d* to form past participle") are accompanied by rules that capture more specific, yet still reliable sub-regularities for the very same change (such as "change root *i* to *u* if word ends in *-ng*" – cf. *sting/stung, sing/sung*, etc.). Similar ideas are also being explored in syntax, where certain constructions (e.g., the English resultative, cf. Goldberg & Jackendoff, 2004) are treated as families of sub-regularities. Clearly, when moving to the semantic level, the boundaries of these islands/families are fuzzier and difficult to define, since they are represented as distributions across different semantic dimensions. Still, FRACSSs seem flexible enough to capture such subtle semantic regularities across words.

Interestingly, when the dominant meaning of a derived form is heavily lexicalized and not part of a (semi-)systematic pattern, FRACSS composition will produce an alternative, more semantically transparent interpretation of the same form. For example, among the nearest neighbors of the *nervous* vector directly extracted from the corpus, we find *anxious*, *excitability* and *panicky*. On the other hand, the nearest neighbors of *nervous* composed by multiplying the *nerve* vector with the *-ous* matrix include *bronchial*, *nasal* and *intestinal*. We find this duplicity a desirable aspect of our model, since, for humans as well, it is likely that the dominant metaphorical meaning of *nervous* is also learned holistically and stored with the whole word, whereas the medical sense can be generated compositionally from the *nerve* stem (Amenta, Marelli, & Crepaldi, in press). We will see below how the possibility of generating

compositional meanings for lexicalized forms might play a role in explaining semantic transparency effects on priming.

The examples in the present section speak for the flexibility of the functional approach in capturing a wide range of phenomena, including affix and stem polysemy, affix-stem interactions and -to a certain degree- opaque derivations. This large degree of flexibility is ensured by the affixation procedure being modelled as vector-by-matrix multiplication: each single dimensions of the derived-form is the result of a multipliticative combination of a set of affix-specific weights with the *whole dimension distribution* of the stem. It is therefore possible that, given different stems, the same dimension in the corresponding derived forms will be at times emphasized, at times inhibited (remember that each dimension, in the current approach, encodes a semantic trait). This implies that, for example, the different possible "senses" that an affix can express in a derived form will be crucially determined by the stem it combines with (see the toy example in Tables 1 and 2). In other words, semantic distinctions in the affixation process can depend on (more or less nuanced) distributional patterns in the stems. This very system can also explain (some) cases traditionally considered opaque: *fruitless* and *heartless* have "opaque" meanings because *fruit* and *heart* have the relevant secondary senses encoded in their distributional representations to begin with (e.g., "the fruits of their labor", "that man has a big heart").

Distributional representations of novel derived words

Our model constructs the semantic representation of a derived word by means of a compositional process that transforms the meaning of its stem through FRACSS application. The model can thus build semantic representations for *novel* words (or *nonce formations*), expressing meanings that are not yet lexicalized in the language. Since the generation of new words must be one of the core reasons why morphological derivation exists, and the new word creation process must be largely compositional, nonce formations represent a natural testing benchmark for our model. Previous psycholinguistic studies of novel complex words have mostly focused on the factors underlying their acquisition (e.g., Tyler & Nagy, 1989) or quantitative aspects associated with morphological productivity (e.g., Hay & Baayen, 2002). We present here two investigations aimed at modeling the degree of meaningfulness of novel words and assessing the quality of the vector representations our method generates for them.

Meaningfulness of novel forms

We focus first on a fundamental but essentially unexplored characteristic of new complex words, namely whether they are *meaningful* or not. While semantics is not the only factor constraining word derivation (nonce forms might also be unacceptable, for example, due to morphophonological or lexical strata constraints), it certainly plays an important role by imposing selectional restrictions on the stems an affix can combine with. For example, since the prefix *re-* conveys an idea of iteration

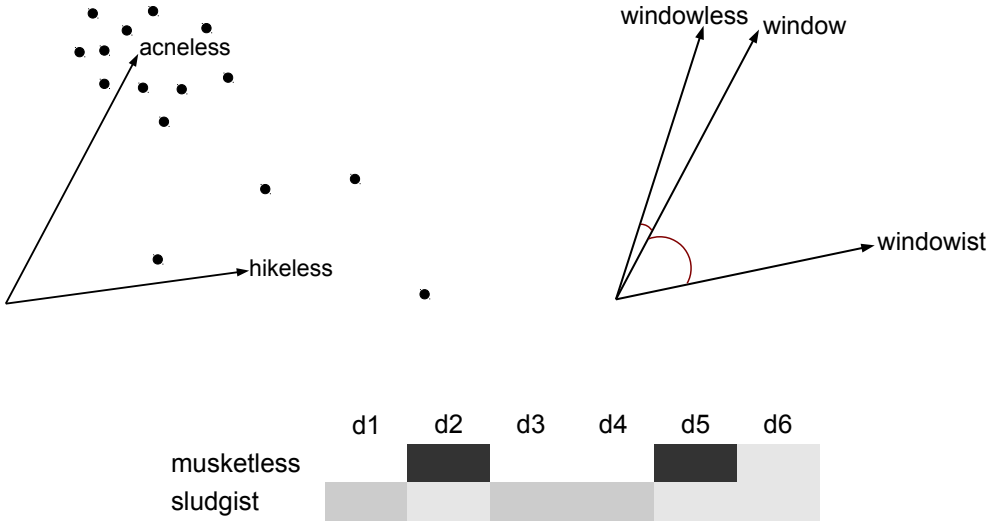

Figure 2. Visual intuitions for the meaningfulness measures. **Top left** (density): the vector of a meaningful derived form such as *acneless* has closer neighbors than a less sensible form such as *hikeless*; **top right** (stem proximity): the vector of a sensible derived form such as *windowless* is closer to the stem than that of a less sensible form such as *windowist*; **bottom** (entropy): looking at distributional vectors as probability distributions (darker shades = higher probabilities), a good form such as *musketless* has most of its probability mass concentrated on few dimensions (low entropy), the probability mass of a less sensible form such as *sludgist* is uniformly spread across many dimensions (high entropy).

or going back to a previous stage, *to re-die* sounds rather strange. Often such restrictions do not lead to sharp judgments, but rather to intuitions laying on a scale of relative acceptability. For example, in the survey we describe below, subjects found nonce forms such as *re-browse* and *re-append* perfectly meaningful, they assigned intermediate ratings to forms such as *re-provoke* or *re-wonder*, and strongly disliked *re-decease* and *re-matter*. Modeling meaningfulness should thus be a task well-suited for compositional DSMs, that provide continuous scores predicting degrees of meaningfulness as we will describe next.

Previous studies have shown that the meaningfulness of novel word combinations (phrases) is efficiently captured by quantitative properties of their distributional representations. We apply the properties that were used by Lazaridou, Vecchi, and Baroni (2013) to quantify the semantic acceptability of phrases (in turn adapted from Vecchi et al., 2011) to novel derived forms. Specifically, the measures proposed by Lazaridou and colleagues were computed here on novel word vectors derived by FRACSS application to the corresponding stem representations.

Neighborhood density measures how close, on average, a word vector is to the vectors of its nearest neighbors in distributional semantic space. The rationale for this measure is that a vector encoding a meaningful new concept should live in a region of semantic space that is densely populated by the vectors of many related concepts that

have already been lexicalized, whereas a vector denoting something that makes no sense should be quite far from the vector of any concept meaningful enough to have been lexicalized. It's easy to think of concepts related to the nonce word *acneless* (a form deemed meaningful by our participants), such as *pimple*, *teenager*, *lotion*, etc. On the other hand, it's hard to assign a precise, fixed sense to *hikeless* (a form that received low ratings), and consequently no related concepts spring to mind. This intuition is illustrated graphically at the top left of Figure 2. Formally, if $n_i(t)$ is the i-th nearest neighbor of a target nonce form t, then $density(t)$ is computed as follows:

$$density(t) = \frac{\sum_{i=1}^{i=N} \cos(\vec{t}, \vec{n}_i(t))}{N}$$

where N, the number of top nearest neighbors to be considered, is a free parameter. Following Lazaridou and colleagues, we set it to 10 without tuning (that is, density is operationalized as the average cosine of the target item with its 10 nearest neighbors).[8]

Stem proximity is the cosine of the derived-form vector with the vector of its stem. This measure captures the intuition that, in productive word formation, if the derived form has a radically different meaning from that of its stem, something went wrong, because a nonce derived word should never be semantically opaque. If I produce the complex word *windowless* (high ratings), I do it because I expect my reader/listener to be able to transparently recover its meaning from that of *window*. Consider instead *windowist* (low ratings): here, it is difficult to see in which way the meaning of *window* should contribute to the meaning of the derived form. This intuition is illustrated graphically at the top right of Figure 2. Note that stem proximity has also been employed to estimate the degree of semantic transparency of existing derived forms, and we will extensively use it for that purpose below. Here, however, the interpretation changes: whereas an *existing* derived form that is far from its stem is likely to have developed a different meaning through its usage, a nonce derived form cannot have an idiosyncratic meaning. Thus, if it is semantically far from its stem, this is a likely cue that derivation broke down. Formally, given distributional representations of a target derived form t and its stem s (such that $t = d(s)$ for some derivation process $d()$), stem proximity is simply the cosine:

$$proximity(t) = \cos(\vec{t}, \vec{s})$$

Finally, the *entropy* of a vector is lower when it has a skewed distribution with just few dimensions having large values, higher when the distribution tends to be uniform (see, e.g., Cover & Thomas, 2006). Since, as discussed in the section on distributional semantic space construction above, the dimensions of distributional vectors, alone or in clusters, cue different semantic domains or "topics", a high-entropy (that

[8] We considered the top 20,000 most frequent content word lemmas in our corpus as potential neighbors. Virtually the same results were obtained when the candidate neighbor set was expanded to include all derived forms used in the experiments.

is, uniform) vector does not carry any specific meaning. Thus, we expect an inverse correlation between entropy and meaningfulness. If there is little doubt that the highly rated nonce word *musketless* pertains to the domain of military matters, it's hard to associate *sludgist* (low ratings) to any specific semantic domain, as we don't know what this word is about. This intuition is illustrated graphically at the bottom of Figure 2. Independent evidence that entropy should correlate with acceptability comes from the observation that attested derived words (that are all, presumably, meaningful to a certain degree) have much lower entropy than derived nonce forms (that are likely to contain a mixture of sensible and meaningless formations). Specifically, the entropy range in a sample of 900 existing derived words (taken from the materials of the semantic transparency experiments we will discuss below) is 2.01-4.51 ($mean = 3.27; SD = .39$), whereas the nonce forms of the present experiment have an entropy range of 4.77-5.58 ($mean = 5.34; SD = .18$), with no overlap between the two sets.

Formally, if $t_1, \ldots, t_k$ are the values in the K components of the distributional vector of a target nonce form t, its entropy $H(t)$ is computed as follows:

$$H(t) = \log K - \frac{1}{K} \sum_{i=1}^{i=K} t_i \log t_i$$

Note that entropy is defined for vectors encoding probability distributions, that cannot contain negative values. Our vectors, obtained by Nonnegative Matrix Factorization of the co-occurrence matrix, satisfy this condition.

Materials and methods. We focused on the 4 suffixes and 2 prefixes presented in Table 3. The affixes were selected among those for which we trained FRACSSs, as described above. These affixes are all reasonably productive according to the quantitative indices reported by Hay and Baayen (2002), and they are not subject to strict morphophonological or lexical strata constraints according to Marchand (1969). We observe in passing that none of the quantitative productivity indices reported in Hay and Baayen (2002) correlates significantly with the average meaningfulness scores for our affixes, indicating that semantic acceptability (a property of specific derived words) cannot be reduced to productivity (a property of word formation processes).

For each of the target affixes, we automatically generated derived forms by attaching the affix to a stem of the appropriate syntactic category (e.g., *-able* was only attached to verbs). Stems had to occur at least 2,500 times in our reference corpora. Orthographic rules where semiautomatically applied where appropriate (e.g., the final *-e* of *demote* was deleted before appending *-er*).

We randomly sampled 100 derived forms per affix, manually excluding those that might sound strange for reasons independent of semantics (e.g., *rereprocess* because of the repeated prefix). We checked, moreover, that the candidate nonce forms never occurred in our very large corpus, also considering spelling (dash/no dash, English/American, etc.) and inflectional variants. We cannot of course guarantee that all forms were absolutely novel to all our subjects, but it is highly unlikely that any

Table 3
Affixes in the novel word meaningfulness data set, with syntactic categories of input stems and output derived forms (A: adjective, N: noun, V: verb), mean acceptability scores across derived forms containing the affixes, and examples of forms with high and low meaningfulness scores.

Affix	Stem→Derived	Mean acceptability (*SD*)	Examples
-able	V→A	3.82 (0.51)	*high:* sketchable, harassable *low:* dawnable, happenable
-er	V→N	3.82 (0.42)	*high:* surpasser, nicknamer *low:* relenter, pertainer
-ist	N→N	3.18 (0.52)	*high:* hologramist, liaisonist *low:* windowist, rasterist
-less	N→A	3.64 (0.44)	*high:* acneless, musketless *low:* eaterless, rinkless
re-	V→V	3.52 (0.45)	*high:* reappend, reinsult *low:* relinger, rematter
un-	A→A	3.46 (0.46)	*high:* undiligent, unheartfelt *low:* unthird, unmessianic

of them would have heard or produced more than a few times a derived form that never occurs in a corpus of 2.8 billion words.

The resulting set of 600 derived nonce forms was annotated to mark the degree of meaningfulness of each item by means of a crowdsourcing study. Crowdsourcing is an online survey method increasingly used in the cognitive sciences to collect large amounts of data (Schnoebelen & Kuperman, 2010). Here, crowdsourcing was used to reach a larger and more diverse population than the one usually taking part in psycholinguistic experiments. Participants were recruited from Amazon Mechanical Turk through the CrowdFlower platform (`http://www.crowdflower.com`). Participants were asked to rate each item on the basis of how easy it was to assign a meaning to it, using a 5-point scale ranging from "almost impossible" to "extremely easy". In the instructions, we specified that there was no right answer, and we stressed that we were specifically interested in the *meaning* of the new words; participants were invited to ignore spelling considerations, and consider alternate spellings if they made the words appear more natural. Each novel words was evaluated by 10 different participants (that had to declare to be native speakers of English). Average ratings were then computed for each item.[9]

Ten items were excluded from the analysis due to technical reasons. The third column of Table 3 reports average meaningfulness ratings across the forms with each affix and the corresponding standard deviations. We observe relative high ratings (associated with a general rightward skewness in the distribution), and some limited variation across affixes. However, we also observe some significant variance *around*

[9]The data collected on nonce words can be downloaded from `http://clic.cimec.unitn.it/composes/FRACSS/`. We hope these data sets will foster further research on the factors determining semantic acceptability of derived forms.

Table 4
Fixed effects in the analysis of novel word meaningfulness.

Predictor	Estimate	Std. Error	t	p
Intercept	3.62	0.13	28.43	.0001
Stem frequency	0.05	0.02	2.22	.0801
Stem proximity (linear)	0.78	0.29	2.65	.0388
Stem proximity (quadratic)	-3.39	0.89	-3.78	.0002
Vector entropy	-0.91	0.13	-3.75	.0008

each affix average, confirming that semantic acceptability cannot be explained away by properties of the affixes, without considering the specific stems they attach to.

Results. The judgments were analyzed in a mixed-effects model (Baayen, Davidson, & Bates, 2008) using the measures described above (neighborhood density, stem proximity and vector entropy) as predictors. In addition, we introduced (log-transformed) stem frequency as a covariate, in order to account for the influence of stem familiarity on participants' intuitions. Affix-associated random intercepts and random slopes (for all predictors) were also introduced, to partial out affix-specific effects. We considered moreover quadratic terms, finding that only for stem proximity this form of non-linear modeling improved the fit. All predictors were mean-centered in order to ensure more reliable parameter estimation.

Table 4 presents the results for the fixed effects in the regression analysis. The parameter associated with density was removed because it did not significantly contribute to the model goodness-of-fit. P-values were computed adopting the Satterthwaite approximation for degrees of freedom (Satterthwaite, 1946) as implemented in the lmerTest R package (Kuznetsova, Brockhoff, & Christensen, 2013).

The effects of entropy, stem proximity, and stem frequency are represented in Figure 3. Entropy has a negative effect on meaningfulness: the more entropic the vector, the less easy it is to understand a novel word. Stem proximity predicts the highest semantic acceptability at intermediate scores (about .4), and progressively lower ratings for more extreme proximity values. A trend for the effect of stem frequency also emerged, but failed to reach significance.

Discussion. As expected, vector entropy has a negative effect on meaningfulness judgments. High-entropy (that is, more uniform) vectors fail to identify specific meanings, making the corresponding novel words harder to interpret than their low-entropy counterparts.

The non-linear effect of stem proximity is more surprising, but it makes sense when considering that novel words are meant to carry new meanings: affixation is thus expected to modify the core meaning of the stem enough for the new word not to be superfluous. Among the forms with the highest proximity values in our data we find *opticianist* and *scholarist*, both receiving low ratings. Arguably, the problem with these forms is that the *-ist* prefix attached to a profession name is redundant: *opticianists* and *scholarists* probably would do exactly what *opticians* and *scholars* do, which our model captures by high stem proximity. At the other extreme of the

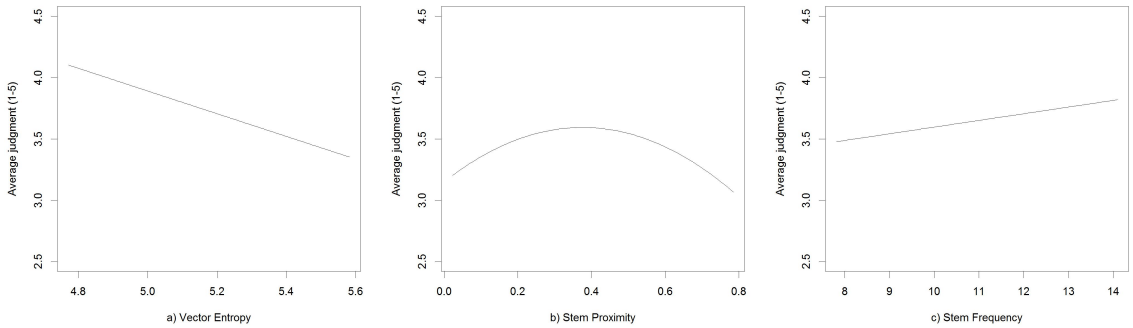

Figure 3. Partialized effects of vector entropy (a), stem proximity (b), and stem frequency (c) on meaningfulness judgments of novel words.

proximity scale, we find two more forms in *-ist* with low subject ratings, namely *sludgist* and *windowist*. The problem here is not redundancy, but that it's not clear how *sludgists* and *windowists* would "specialize" in *sludge* and *windows*, respectively. The model accounts for this by assigning very low proximity to these forms, as if to mark the fact that their relation to the stems is obscure. In conclusion, a novel word should be far enough from its stem to avoid redundancy, but not so distant that the new meaning is no longer interpretable. This pattern is captured by the non-linear effect of stem proximity on meaningfulness.

Interestingly, proximity and entropy are not statistically associated, and are thus probably capturing different aspects of novel word meaning: whereas entropy is diagnostic of topic-specificity of the novel concept, proximity indicates to what extent the derived meaning differentiates itself from the original stem.

Regarding the lack of a density effect, we observe first that this measure has a relatively high correlation with entropy ($r = .36$). In post-hoc analyses, we regressed entropy on density and *vice versa*, entering the residuals of one measure (residualized entropy or density) together with the other as uncorrelated predictors. While entropy was consistently a significant predictor, density only reached significance when entropy was residualized. The overall pattern thus suggests that density does not account for acceptability judgments beyond what is already explained by entropy, and the latter is a stronger predictor. Note that our current operationalization of density might fail to capture the intuition we gave for this measure. In particular, the current implementation only takes into account the distance of the derived form to its nearest neighbors, but the relation of these neighbors to each other (are they also close, making the neighborhood truly "dense"?) is not taken into account.

As it is not the focus of the current study, we leave a more in-depth investigation of the specific measures we picked, and in particular density, to further studies. For our current purposes, the important result presented in this section is that quantitative properties of our compositionally-derived vectors are able to capture a significant portion of the variance in semantic intuitions about nonce derived forms, even when

other factors such as stem frequency and affix type are considered.

Quality of novel form vector representations

In the previous experiment, we have tested to what extent *quantitative* aspects of the vector representation of a novel word are able to predict its perceived meaningfulness. As an added benefit, the survey we ran produced a list of novel derived forms that participants rated as highly meaningful. We can now use this subset to look into the *quality* of composed novel word representations more directly.

In particular, following Lazaridou, Marelli, et al. (2013), we assume that the semantic quality of a distributional vector is reflected in its semantic neighborhood. A good vector representation of a word should live in a region of the semantic space populated by the vectors of intuitively related words; for example, a vector can more convincingly be said to have correctly captured the meaning of *car* if it places it closer to *automobile* than *potato*. We extend this approach to *novel words*, asking participants for relatedness judgments about the neighbors of the corresponding vector representations.

Materials and methods. Neighbors were extracted from our reference semantic space (described in the section on model development) on the basis of the vector representations for novel words obtained using FRACSSs. We focused on 236 novel forms, that received an average meaningfulness rating of at least 4 in the previous study, and their top 10 nearest neighbors. From this set, we had previously filtered out words that were overly represented across neighborhoods (i.e., occurring in the top neighbor lists of more than 10% of the considered forms; Radovanović, Nanopoulos, & Ivanović, 2010). These shared elements were poorly informative of the specific novel forms under analysis, since they were primarily connected to affix meaning, independently of the stem: they were affixed words found in the neighborhoods of many words with the same affix, and some of them were nearly synonymous to it (e.g., *unlike* for *un-* forms, *manageable* for *-able* forms).

We randomly selected up to 5 neighbors for each nonce form, resulting in a set of 853 neighbors, each contrasted with the corresponding novel form. Each neighbor was also assigned two control items, namely the stem of the nonce form and an unrelated baseline word. Baseline words were randomly chosen from the whole set of words in the semantic space, pending that they were reasonably different from both the stem and the nonce form they were assigned to (cosine similarity less than .30). Therefore, the final item set included 2,559 pairs, organized into three conditions: in the *nonce form* condition, nonce-form neighbors were contrasted with the corresponding nonce forms (*blameworthy-apologizable*); in the *stem* condition, nonce-form neighbors were contrasted with the stem of the corresponding nonce forms (*blameworthy-apologize*); in the *random* condition, nonce-form neighbors were contrasted with unrelated random words (*blameworthy-blazer*). Stimulus examples are given in Table 5.

The resulting set was annotated in a crowdsourcing study to mark the degree of semantic relatedness for each pair. Participants were recruited from Amazon Me-

chanical Turk through CrowdFlower and asked to rate each pair using a 7-point scale ranging from "completely unrelated" (1) to "almost the same meaning" (7). In the instructions, we warned participants that some of the words could be unknown to them, but pointed out that those very terms were made up of portions of existing English words (e.g., *quickify=quick+ify*), and thus conveyed a meaning that could be evaluated. Each pair was rated by 10 different participants (requested to be native speakers of English). Average ratings were then computed for each pair.

Results. The judgments were analyzed in a mixed-effects model using the experimental condition as predictor. The condition of interest (*nonce-form*) was modeled as reference level. Random intercepts for both terms of each pair were included in the model. Average ratings in the *nonce form* condition ($mean = 2.41; SEM = 0.03$) are significantly larger than those in the *random* ($mean = 1.87; SEM = 0.02; t = -14.44; p = .0001$) and *stem* ($mean = 2.19; SEM = 0.02; t = -5.81; p = .0001$) conditions.

Discussion. We evaluated the quality of the vector representations of meaningful novel words constructed by FRACSS application. We focused on the semantic-space neighbors of these vectors, and had them judged by participants in a crowdsourcing study. Crucially, similarity ratings between nonce forms and the produced neighbors were higher than those in two control conditions.[10]

First, the neighbors of a nonce form were deemed to be closer to the nonce form itself than to an unrelated control word. This confirms that that the region in semantic space individuated by the nonce form vector is far from random, being populated by words that native speakers consider semantically related to the unfamiliar derived term. Second and more importantly, the neighbors of a nonce form were deemed to be closer to the nonce form itself than to its stem; this result further supports the reliability of the obtained vectors as representations for novel derived words, indicating that the compositional procedure does not produce simple replicas of stem meanings, but can capture the specific semantic connotations springing from the derivational process. Table 5 reports some examples of neighbors that were rated more similar to nonce forms than to their stems. We observe a variety of patterns that link the novel derivation to its neighbor, forming a tighter semantic connection than with the stem. The negation effect of *-less* in *pastureless* brings this form near *barren*. In a case like *disagreer/doubter*, both derived-form and neighbor are agents rather than events. For *soakable*, we get the antonym *waterproof*, and so on.

General discussion of the novel word experiments

As first benchmark test for the proposed compositional model, we used FRACSSs to generate distributed semantic representations for novel derived words, namely

[10]The average ratings were somewhat low in absolute terms, but this only reflects the difficulty of producing judgments about unfamiliar words. Indeed, the values are similar to those obtained by Lazaridou, Marelli, et al. (2013) for similarity judgments between low-frequency existing words and their neighbors.

Table 5
Examples of novel-form neighbors that were rated more similar to the novel forms than to the corresponding stems.

neighbor	Nonce form	Stem	Random
redo	refinalise	finalise	sip
curable	counteractable	counteract	wedding
reprehensible	insultable	insult	meat
waterproof	soakable	soak	email
propagandist	provocationist	provocation	joystick
doubter	disagreer	disagree	palsy
accountant	leverager	leverage	ulceration
defenceless	garrisonless	garrison	qualitative
barren	pastureless	pasture	authenticate
greyness	sunlightless	sunlight	incitement
flawed	unsuperior	superior	headstone

stem-affix combinations that are unattested in a very large corpus and are hence likely to be unfamiliar to most speakers. The representations that we obtained provide a compact computational model of what happens in the semantic system when a speaker has to understand the meaning of an unknown complex word. The procedure can be summarized as follows: the distributional pattern associated to the meaning of a familiar word (the stem) is modified through the application of a function, in the form of the affix FRACSS; the FRACSS acts on the basis of systematic statistical relations that the affix entertains in language usage (as experienced by the speaker and here captured by corpus counts); because of the nature of FRACSS representations and function application, each dimension of the resulting distributional pattern will be influenced by *all* the dimensions of the original stem vector, with different FRACSS weights for each affix and unit in the output distribution, thus granting highly flexible results. Indeed, the newly obtained distributional pattern has a series of properties that can be meaningfully quantitatively characterized, and can be compared to those of existing, familiar words to identify its semantic connotation.

In the first experiment, we showed that the perceived meaningfulness of a novel word is predicted by the properties of FRACSS-generated distributions. More meaningful forms have less entropic representations, that is, the distribution they display is less uniform, with a few dimensions being particularly active with respect to the others. Since each dimension in a vector representation can be associated to a semantic domain (Griffiths et al., 2007), a less entropic distribution cues a novel word with a more specific meaning. More informally, an unknown derived word is considered more meaningful if it elicits a clear sense in the mind of the speaker, and entropy computed on the FRACSS-generated vector is a good predictor of this sense clarity. Perceived meaningfulness was also predicted by the proximity of the newly obtained representations to the familiar stem, a measure of the extent to which the distributional semantic representation of the stem is transformed by affixation. If this

transformation is too extreme, the relation to the stem is lost, and it becomes difficult to assign a meaning to the new word. On the other hand, low meaningfulness judgments are also obtained in cases in which FRACSS application is nearly transparent, hardly affecting stem representation. Arguably, in this case, the affixation procedure is perceived as redundant. Meaningful novel words must be distant enough from their stems to usefully encode a different meaning, but not so distant that their meanings can no longer be recovered.

It is worth emphasizing that the training sample on which FRACSSs are based does not provide direct examples of different degrees of meaningfulness: it includes familiar and relatively frequent words, all expected to be highly meaningful. Indeed, meaningfulness becomes a sensible testing ground only after we obtain new word representations through the compositional procedure. In other words, meaningfulness is predicted by properties of the newly obtained complex words, that crucially emerge as a by-product of the *combination* between a stem vector and an affix matrix (neither of them intrinsically informative about word meaningfulness).

Through the survey conducted in the first experiment, we obtained a list of nonce derived forms that were deemed meaningful by participants. In the second experiment, hence, we could directly evaluate the quality of their FRACSS-based representations by considering which familiar (existing) words "resonate" with their distributional patterns, that is, have vectors that are close to the distributional representations of the novel forms. If the compositionally obtained novel representations capture the meanings that speakers assigned to the corresponding novel forms adequately, we expect their neighbors to be words that are semantically related to them also according to speakers' judgments. Indeed, participants found automatically extracted neighbors of novel forms closer in meaning to the novel forms themselves than to random control words or to the novel word stems. This latter evidence is particularly important because it indicates that the compositional procedure based on FRACSS is generating representations that capture the peculiarities of novel derived forms over and above the meaning of their stems.

To sum up, FRACSSs provide a good model for the semantic processing of novel morphologically complex words, paving the way to a thorough understanding of the main determinants of meaning construction in new morpheme combinations.

We conclude this section with preliminary data that show how our system might produce broadly sensible guesses about the meaning of a novel affixed form even when the stem of the form itself is novel, simulating the "wug" test (Berko, 1958), in which children or adults are asked to productively attach affixes to non-lexical stems. Typically, morphophonological or orthographic properties of the derived form are investigated ("This is a wug. Now there is another one. There are two of them. There are two... *wug[z]*"). However, we expect wug words to also come with a certain degree of semantic expectation. Even if we do not know what *zibbing* is, we can guess that a *zibber* is either a person who zibs or a zibbing tool. Essentially, in this case, a speaker must resort to the general semantic properties of the affix to deduce part of

the meaning of the derived form – the part associated to the affix.

As a result of the standard least-squares estimation procedure, a FRACSS matrix contains an intercept vector encoding the averaged contextual distribution (hence, the distributional meaning) of all derived forms that were used for training. This intercept vector should be a reasonable proxy of "zibber" words, since all we can deduce about zibbers is that they must do what, on average, *-er* derivations do. Indeed, we find that the FRACSS intercepts of productive affixes are associated to semantic neighbors that contain the relevant affixes. By using such intercepts to represent zibber words, we naturally capture the fact that, if all we know about the meaning of a form is that it contains an affix, we can only guess that its meaning will be related to that of other forms containing the affix. Considering the most productive affixes in Hay and Baayen (2002) (i.e., those with type frequency above 400), the affix *-y* is found in 19 of the top 20 neighbors of the *-y* intercept; the affix *-er* is found in 11 of the top 20 neighbors of the *-er* intercept; the affix *-ness* is found in 19 of the top 20 neighbors of the *-ness* intercept and the affix *-ly* is found in all of the top 20 neighbors of the *-ly* intercept. Moreover, even those neighbors that do not directly contain the affix, typically are associated to it at the meaning level: many neighbors of the *-er* intercept, for example, denote professions (*compositor, salesman, projectionist*), suggesting that the agentive meaning of the suffix dominates its semantic neighborhood.

This pilot experiment suggests how FRACSSs could capture aspects of the affixes they represent also when applied to semantically void stems. Besides making predictions about wug derivation, a similar approach might be used for existing words, such as *grocer*, that have an active affix attached to a non-lexical stem. Although the *grocer* whole-word meaning should be represented holistically, a combinatorial procedure would still be able to capture the affix traits through the FRACSS intercept.

Modeling semantic transparency effects

In the previous section, we have shown how the composition-based approach to distributional semantics can be profitably used to generate meaning representations for novel words. The question arises whether the same compositional methods also have a role to play when familiar derived words are processed. From a theoretical point of view, if the compositional procedure works for accessing the meaning of novel forms, it is not clear why it should be blocked when processing other derived forms. Such conjectured routine application of a combinatorial semantic procedure presupposes a systematic activation of morphemic units: for composition to operate at the semantic level, morphemes need to have been previously activated (Rastle & Davis, 2003, 2008). Moreover, in order for composition to be applied to any string potentially containing morphemes, morphemic parsing needs to proceed in a semantically blind manner. Crucially, empirical results consistently show that any parsable orthographic string is associated to the activation of morphemic information, irrespective of effective morphological structure (*corner* vs. *darkness*, Longtin, Segui,

& Hallé, 2003), semantic transparency (*courteous* vs. *darkness*, Rastle et al., 2004) and familiarity (*quickify* vs. *darkness*, Longtin & Meunier, 2005). We can thus build on extensive evidence that morphemes are automatically accessed when processing any word that is potentially complex (Rastle & Davis, 2008).

Assuming that a compositional procedure is always applied when morphological information is available does not imply that this operation will always be successful at retrieving the full-fledged semantic denotation of the derived form. Because of the way they are obtained, FRACSSs reflect *statistical systematicities* in the relations between stem and derived word meanings. Our qualitative analysis above showed that this systematicity encompasses a larger degree of semantic variations than usually assumed, but the latter are still limited to (semi-)regular, predictable, synchronic operations. Such procedures, that are effective when building the meaning of novel derived forms, are bound to miss a certain amount of information when dealing with some existing words. The lexicon of a language is continuously evolving through time: complex words becomes progressively more lexicalized, and many of them are subject to a certain amount of semantic drift. In order to fully explain the semantic processing of morphologically complex words, the compositional procedure must be paired with a way to directly access the meaning of the derived form as a whole. This alternative and complementary *whole-word route* should capture meaning shifts caused by diachronic phenomena (as well as morphemic accidents of the *corner* type).

When dealing with existing words the questions that have to be addressed are hence rather different from those explored for novel forms. First, we want to evaluate to what extent a compositional procedure can explain the semantic variations present in familiar morphological constructs; or, in other terms, to what extent (semi-)systematic semantic relations can account for morpho-semantic phenomena. Second, we aim at assessing the relative efficiency of the two semantic procedures (compositional vs. whole-word) in different tasks and experimental conditions.

A natural domain to test the relative role of the compositional and whole-word routes to complex word meanings is the empirical study of the degree of *semantic transparency* of a derived form with respect to its stem. The semantic transparency of a complex word indicates how easily the whole-word meaning can be inferred from the meaning of its parts. For example, the meaning of the transparent derived word *rename* is largely predictable from the meaning of its stem (*name*) and its affix (*re-*), whereas the same cannot be said for the opaque *remark*, whose meaning is not (easily) understood given the meaning of its parts.

The role of semantic transparency has been a central theme in the literature on complex word processing, with most research revolving around the hypothesis that transparent words can be accessed through the representations of their morphemes, whereas opaque words have to be represented on their own, and thus accessed directly. This assumption has been largely investigated by means of priming paradigms (e.g., Feldman & Soltano, 1999), where typically the derived form (e.g., *dealer*) is used as the priming stimulus and participants are asked to recognize the following

stem target (e.g., *deal*). The priming effect corresponds to the amount of facilitation in the response latencies as compared to a control unrelated-prime condition (e.g., *speaker-deal*), whereas possible confounds (e.g., orthography, semantics) are excluded in a series of control conditions. Prime duration (usually measured through the Stimulus Onset Asynchrony, SOA) can be manipulated in order to investigate different processing levels. The alleged modulation of semantic transparency on morpheme access is supported by results associated to long-SOA primes: a significant priming effect is found only for transparent words, whereas with opaque primes it does not emerge (Rastle et al., 2000; Rueckl & Aicher, 2008; Meunier & Longtin, 2007). Results are less clear-cut at short SOAs (in particular, in the masked priming condition): a priming effect is observed for both word types, indicating that, as mentioned above, at early processing stages morphemes are routinely accessed irrespective of semantic transparency (Rastle et al., 2004). Still, even at short SOAs some studies report significant differences in priming effect sizes, with more facilitation for transparent than opaque words (Diependaele et al., 2005; Feldman et al., 2009; Diependaele et al., 2009; Järvikivi & Pyykkönen, 2011; Kazanina, 2011; Feldman et al., 2012; Marelli et al., 2013). Moreover, recent results from both priming (Tsang & Chen, 2014) and eye-tracking studies (Amenta et al., in press), although confirming that morphological parsing proceeds in a semantically blind manner, also suggest that the morpheme meanings are accessed straight away after word decomposition. In conclusion, empirical evidence shows that semantic transparency plays a role in complex word recognition, although its effect can be more or less prominent depending on the processing stage under examination.

Prima facie, opaque words, being traditionally defined by the property of having a meaning that is not predictable from their parts, may be seen as outside the possibilities of our compositional model. Yet we think that this conclusion is far from granted, as it depends on a series of assumptions regarding the nature of semantic transparency that are common in the psycholinguistic literature, but not necessarily warranted. First, semantic transparency is often conveniently operationalized in terms of meaning similarity between a derived form and its stem. This approach hence focuses on the meaning of the shared stem across the two forms, overlooking the crucial role played by the affixes. In fact, the latter are often active and meaningful also in opaque words (Baayen et al., 2011), since they carry the correct morphosyntactic information and respect the grammatical constraints of the combination (*-ous* marks the adjective class in *courteous*, *-ic* combines with a noun to generate the adjective *cryptic*), and often more (*-less* marks the absence of something in *fruitless*). This important role, totally missed when we focus on the derived-form stem only, is a crucial aspect of the compositional approach that represents affixes as FRACSSs, i.e., the functional elements of morphological composition. Second, in most studies the opaque test set is populated by highly heterogeneous elements, ranging from pseudo-derived words such as *corner* to genuinely derived and not entirely opaque ones such as *fruitful*. Certainly, in the former case, the correct meaning cannot be obtained

through combinatorial processes. However, some semantically-opaque derived words may still show a certain degree of compositional systematicity (Plaut & Gonnerman, 2000; Royle, Drury, Bourguignon, & Steinhauer, 2012), provided the combinatorial procedure is flexible enough to account for the fact that the affixation process should select only some specific features that the stem carries (e.g., the metaphorical meaning of *fruit* in *fruitful*, the *crypt* quality of being dark and difficult to access in *cryptic*). The distributional representations we adopted can arguably encode these separate facets of meaning as specific dimensions of the stem vector (Griffiths et al., 2007), and FRACSSs should be flexible enough to highlight different features of the input vectors when generating the derived form (see the examples discussed above). Third, semantic transparency should not be confused with the degree of *systematicity* of derivation: a privative suffix such as *-less* will in general alter the meaning of the stem quite a lot, even in forms where the meaning shift is largely predictable: arguably, such forms are systematic but not fully transparent. We think that much of the earlier literature has mixed up systematicity with the strength of the effect that the affix-driven transformation has on the meaning of the stem. The results we are about to report, where our compositional model makes good predictions about semantic transparency effects, suggest that semantic transparency can, at least in part, be dissociated from systematicity.[11]

In conclusion, opacity is not, *a priori*, a theoretical limit for morpheme combination at the meaning level, but it rather represents a good empirical benchmark for the corresponding model, testing its nuance and flexibility.

Quantifying semantic transparency

Following a long tradition in psycholinguistic research (that mostly exploited LSA-based measures, e.g., Rastle et al., 2000; Milin et al., 2009; Gagné & Spalding, 2009), we operationalized semantic transparency as the proximity between the vector associated to a target derived form t and its stem s (such that $t = d(s)$, at least potentially, for some derivation process $d()$):

$$ST = \cos(\vec{s}, \vec{t})$$

We use this mathematical formulation to describe semantic transparency as predicted by either the composition procedure or the traditional direct comparison between stem and derived form. The only difference lies in how $\vec{t}$ (the representation of the target derived form) is obtained. Under the *composition* approach, $\vec{t}$ is obtained by multiplying the stem vector by the FRACSS, hence without relying on an explicit representation of the derived form. In the latter *whole-word* approach, $\vec{t}$ is a vector directly extracted from the corpus-based co-occurrence profile of the derived form,

[11] As a consequence of our findings, we might argue that transparency and opacity are somewhat misleading terms for the phenomenon we are trying to model. Still, we stick to them for terminological coherence with the earlier literature.

hence ST depends on the similarity between two separate, explicit representations. To further clarify: the composition approach also yields a single vector representing the whole-word meaning of the derived form, however the latter is obtained by stem+affix composition, instead of being induced directly from the corpus contexts in which the derived word occurs.

Although we eventually compute the same transparency score (that is, the cosine between stem and derived-form vectors) under both approaches, the theoretical implications associated to how $\vec{t}$ is obtained are crucial, and can be connected to the morpho-semantic routes we proposed for the semantic processing of existing words. Whole-word-based ST is not truly a morphological measure, as it rather quantifies the association between two independent meanings. Under this approach, ST indicates how much a derived form is related to its stem in the same way in which the cosine measure might tell us how much *dog* is related to *cat*. Therefore, the whole-word model explores a semantic system populated by holistic, encapsulated meanings. On the contrary, composition-based ST quantifies how much the meaning of a stem word is altered by the affixation process, that is, to what extent a FRACSS changes the stem meaning, where this change should be, to a certain extent, systematic (as opposed to unpredictable). This approach is more easily connected to a morpheme-based semantics, since it does not assume explicit representations for the derived forms, which are generated on-line by combining the morphemes they are made of.

In the following experiments, the ST measures resulting from the whole-word and composition-based approaches will be assessed by considering a series of behavioral effects in morphological processing. The phenomena under examination include explicit judgments of semantic transparency, facilitation in morphological priming, and morpheme-frequency effects on lexical decision latencies.

Explicit intuitions about affixed words

Human ratings about the semantic properties of words are the traditional benchmark for the reliability of distributional semantic measures, and the case of semantic transparency of complex forms makes no exception (e.g., Kuperman, 2009). The assumption is that cosine similarity between stem and derived-form vectors should correlate with explicit intuitions about the degrees of semantic transparency of the derived form. Certainly, this alleged correlation does not necessarily mean that distributional measures would be effective predictors of language *processing* (e.g., Baayen, 2013). Still, distributional semantic measures are expected to explain, at least partially, the variability observed in human judgments about the semantic transparency of derived forms.

Materials and Methods. A set of 900 word pairs, each including a derived form and its stem, were included in the experiment. Stimuli were chosen by randomly sampling 50 derived forms from each of the 18 affixes (15 suffixes and 3 prefixes) with the highest number of training examples (i.e., largest family sizes) from our FRACSS set (see Table A1).

Semantic transparency ratings were collected by means of a crowdsourcing study. Participants were again recruited from Amazon Mechanical Turk through CrowdFlower. Only (self-declared) native speakers of English were admitted. Participants were asked to rate the pairs for how strongly related the meanings of their component words were on a 7-point scale, ranging from "completely unrelated" (1) to "almost the same meaning" (7). 7 judgments were collected for each pair. In order to ensure that participants were committed to the task and exclude non-proficient English speakers, we used 60 control pairs as verification items, consisting of pairs either including highly transparent derived forms (*singer-sing*) or pseudo-derived words whose apparent complexity is just orthographic happenstance (*corner-corn*). Participants who gave obviously wrong answers to these control pairs (at the opposite of the expected end of the transparency scale) were automatically excluded from the experiment. By-item average scores were used as dependent variable. The resulting dataset has already been employed in Lazaridou, Marelli, et al. (2013), and can be downloaded from `http://clic.cimec.unitn.it/composes/FRACSS/`. Six pairs were excluded from the analysis for technical reasons.

As described above, semantic transparency was operationalized as the cosine similarity between stem and derived-form vector. For semantic composition, the derived-form vector was induced by applying the relevant FRACSS to the stem vector, whereas for the whole-word approach we extracted the derived-form vector directly form the corpus. These distributional measures were separately tested as predictors of participants' ratings.

Results. Collected judgments had an inter-rater agreement of 60%. The distribution of the average ratings was negatively skewed (mean rating: 5.52; standard deviation: 1.26). Figure 4 compares this distribution with those produced by the models (human ratings were rescaled to the 0−1 range for direct comparability). Although the model-based scores are clearly more Gaussian-distributed, their rank-based correlations with participants' ratings are significant (composition: $rho = .32$, $p = .0001$; whole-word: $rho = .36$, $p = .0001$).

In order to rule out the possibility that the performance of a method depends on particularly effective representations of only few, more regular, affixes, we tested the ST measures in mixed-effects models including random intercepts and slopes associated to the affixes (p-values were computed adopting the Satterthwaite approximation for degrees of freedom). The effects are indeed confirmed for both whole-word ST ($t = 5.75, p = .0001$) and composition ST ($t = 2.57, p = .0221$), with the former approach clearly outperforming the latter ($\Delta AIC = 81$, see Wagenmakers & Farrell, 2004).

Results are consistent in analyses using ranks in place of average judgments (whole-word: $t = 6.97, p = .0001$; composition: $t = 3.57, p = .0004$), indicating that they are not overly influenced by the skewness of the dependent variable. Still, to exclude the possibility that good performance is due to transparent words being overrepresented, we median-split the items on the basis of their transparency ratings.

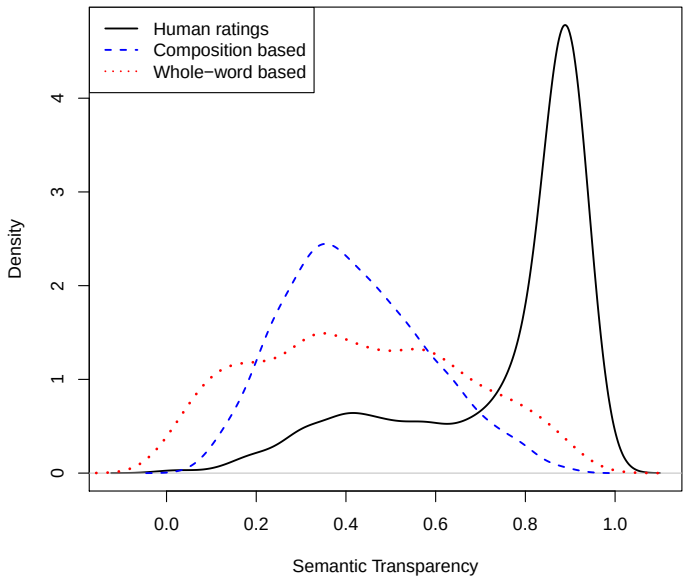

Figure 4. Distribution of Semantic Transparency values in human ratings and in the model-generated measures.

The effects of the transparency measures hold in both the high-transparency (whole-word: $t = 4.35, p = .0002$; composition: $t = 2.57, p = .0103$) and low-transparency sets (whole-word: $t = 3.41, p = .0027$; composition: $t = 2.17, p = .0339$), confirming the reliability of the results and good performance of the models.

Discussion. Results indicate that distributionally-based semantic transparency measures significantly predict human intuitions. The semantic composition approach does not perform as efficiently as the direct measure of semantic relatedness based on whole-word vectors. Still, the effect of the composition-based transparency variable is significant throughout a series of control analyses, and in particular when focusing on low-transparency words. Semantic opacity can hence be reframed, at least in part, in terms of strong but systematic semantic composition effects on the meaning of the stem, still amenable to a suitably flexible compositional analysis. An opaque word is not necessarily a word subject to unpredictable, lexicalized drift; it can also be a derived form in which the affixation process has a strong effect on the resulting meaning, taking it farther way from the stem than is the case in transparent words.

As suggested in the qualitative-analysis section, this more general and nuanced view of semantic transparency is possible because FRACSSs are flexible enough to capture sub-regularities in the production of new meanings through derivation, thus extending the boundaries of systematicity. For example, *-ful* and *-less* do not always modify the primary meanings of the stems they are attached to, but rather they

apply to a metaphorical or secondary sense when attached to certain words (e.g., *fruit, heart*). Similarly, *-y* often generates adjectives recalling general, less-defined connotations of their stems, rather than their proper meanings (as in *beefy, foxy*). Specific semantic features or alternative facets of meanings are captured by different dimensions in the vector representations we developed (as already discussed, vector dimensions can be assigned an intuitive semantic interpretation, see, e.g., Griffiths et al., 2007). The functional approach we adopted is able to learn which dimensions are more likely to generate more opaque meanings when combined with specific affixes: a word that is particularly characterized by dimensions denoting a metaphorical meaning, for example, may be more likely to generate an opaque form when combined with *-less*.

We are not claiming that the whole spectrum of semantic transparency effects can be explained compositionally. As mentioned, there are phenomena in the lexicon that cannot be predicted in compositional terms, such as lexicalization, semantic drift, and purely etymological relations. To understand the meaning of, e.g., *archer*, whole-word access is needed, since in contemporary English *arch* has lost the relevant *bow* sense. This explains why the whole-word approach outperforms composition in predicting human intuitions: all these non-systematic aspects are part of a speaker's lexical knowledge, and obviously impact the way participants perceive complex words in a task involving explicit judgments. In these particular cases, the compositional approach may even be misleading, as it might generate transparent meanings for the opaque forms. Indeed, the semantic neighborhood of the composed vector for *archer* includes *decorator, craftsman, carver, carpenter*, all words reflecting the present meaning of *arch* as an architectural element.

However, the present results suggest that these latter cases represent only a limited portion of the vectors generated using FRACSS, and that semantic composition captures a much wider range of the opacity continuum than previously thought. Indeed, words like *heartless, fruitful, foxy, courteous* are usually classified as opaque, but the compositional procedure can produce a good approximation of their meanings: we find that the FRACSS-derived vector for *fruitless* includes in its neighborhood *depressingly, monotonous, dreary*; *foxy* includes *sluttish, leggy, dishy*; and *heartless* includes *unfeeling, brutish, callous, pitiless*.

As mentioned in the introduction to this experiment, explicit judgments are not necessarily the best way to assess a model of complex word meaning. In fact, the rating distribution is very skewed: most words are perceived to be extremely transparent by participants, that are arguably missing subtle meaning variations between a derived word and its stem. Vector-based measures, on the contrary, have more Gaussian-shaped distributions. In other words, it seems that these latter measures are better suited to capture the continuous nature of semantic transparency (see Gonnerman, Seidenberg, & Andersen, 2007) than explicit judgments are. It is not surprising, then, that they were found to be better predictors of processing measures (e.g., response latencies, fixation times) than human ratings (Baayen, 2013). In the

next experiments, we will thus turn to predicting response times in lexical decision tasks.

Priming effects at different SOAs

In the present experiment we consider priming effects in lexical decision, focusing on paradigms in which the derived form is used as prime for the corresponding stem. The scope of this empirical analysis is twofold. First, priming paradigms are traditionally used to test the time course of lexical processing: the relation between priming effects and vector-based measures can thus shed light on which processing levels are affected by the semantic operations we are modeling. Second, in the previous section we studied a large random sample of derived forms; priming experiments offer instead the opportunity to focus on small, well-defined test sets, in which the difference between transparent and opaque words is extreme by design, thanks to the selection procedure carried out by expert language researchers.

Materials and Methods. We employed item sets previously used by Rastle et al. (2000) in a series of priming experiments. In that study, priming effects were tested in a number of different conditions, including orthographic, purely semantic, and unrelated pairs. We focused on the morphologically transparent (*dealer-deal*) and opaque conditions (*cryptic-crypt*). The former set included 24 derived-stem pairs, and the latter set included 21 pairs (we excluded the original pair *creature-create*, because *-ure* is not among our FRACSSs, and *apartment-apart*, because *apart* is not in our semantic space). The two sets were originally validated by human ratings on semantic transparency and LSA measures.

The item pairs were used in masked priming experiments adopting different SOAs (43ms vs. 72ms vs. 230ms), where SOAs correspond to the duration of presentation of the prime stimulus, i.e., the derived word. SOA effects are believed to be informative of the involved processing stages: since prime processing is limited by presentation time, the shorter the SOA, the earlier the associated priming effect will occur (although this assumption is questionable, see Tzur & Frost, 2007; Norris & Kinoshita, 2008). We used the average reaction times (RTs) of each pair as reported in the appendix of Rastle et al. (2000).

We employed these stimuli, and the associated RTs, as a test set for our vector-based measures. As in the previous analysis, we used proximity between stem and derived vectors as a proxy for ST, where the derived vector could be constructed through the compositional method or directly extracted from the corpus (whole-word approach). First, we tested whether the measures were able to correctly distinguish between opaque and transparent items. Second, we assessed the association between vector-based measures and RTs at different SOAs.

Results. Figure 5 reports the average proximity for opaque and transparent pairs, as predicted by the composition- and whole-word-based approaches. The composition-based measure predicts derived forms to be more similar to their stems

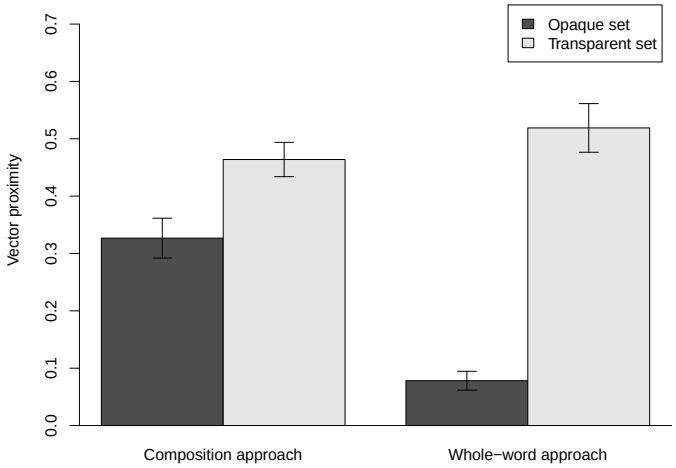

Figure 5. Similarity between derived words and their stems in the opaque and transparent sets, as predicted by the composition- and whole-word-based proximity measures.

than the whole-word-based measure does, but both models correctly distinguish transparent and opaque sets: proximity is significantly higher in transparent than in opaque pairs for both the composition ($t(43) = 3.01$, $p = .0043$) and whole-word measures ($t(43) = 9.19$, $p = .0001$).

Figure 6 reports priming effects at different SOAs, when derived forms are used as primes and the corresponding stems as targets (Rastle et al., 2000). As suggested by a visual comparison of figures 5 and 6, the proximities produced by the composition approach pattern very well with results at the shortest SOA, whereas whole-word-based predictions are more in line with data from longer SOAs. Indeed, when using the composition-based approach, the vector-based proximity measure is correlated with RTs at $SOA = 43ms$ ($r = -.38$, $p = .0104$), but neither at $SOA = 72ms$ ($r = -.24$, $p = .1114$) nor at $SOA = 230ms$ ($r = -.22$, $p = .1419$). The opposite pattern is found with the whole-word approach: vector similarity is not correlated with RTs at $SOA = 43ms$ ($r = -.27$, $p = .0735$), but it's correlated with results at both $SOA = 72ms$ ($r = -.53$, $p = .0001$) and $SOA = 230ms$ ($r = -.54$, $p = .0001$). Results are confirmed in a series of mixed-effects analyses[12] including affix-associated random effects.

One reviewer suggested that the latter results might be accounted for in terms of orthography-semantics dissociation: whereas whole-word ST is a measure of seman-

[12]Composition ST: $SOA = 43ms, t = 2.68, p = .0104$; $SOA = 72ms, t = 0.88, p = .3801$; $SOA = 230ms, t = 1.05, p = .3010$
Whole-word ST: $SOA = 43ms, t = 1.75, p = .0889$; $SOA = 72ms, t = 4.09, p = .0002$; $SOA = 230ms, t = 4.27, p = .0002$

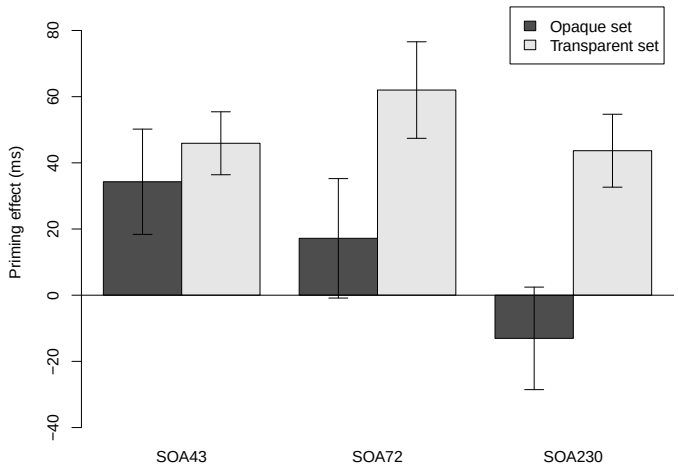

Figure 6. Priming effects of derived forms on recognition of the corresponding stems at different SOAs. Adapted from the results of Rastle et al. (2000).

tic relatedness, compositional ST would mainly capture the orthographic similarity between stems and derived words, and it would be this form similarity to explain short-SOA effects. However, this hypothesis does not hold against empirical evidence. If the compositional approach is actually capturing systematic orthographic relations, the composed derived representations should mainly encode orthographic information. It would follow that the neighbors of a composed derived form should be orthographically similar words, and significantly more so than the neighbors of the whole-word vector representing the same form. This is not the case. The Levenshtein distance (Yarkoni, Balota, & Yap, 2008) between a derived form and its top ten nearest neighbors is not significantly different when the derived-form vector is obtained compositionally ($mean = 6.82; SD = 2.73$) as opposed to being directly constructed from co-occurrence counts ($mean = 6.64; SD = 3.16$). This is confirmed by a mixed-effects analysis including random intercepts and slopes of target words ($t = 0.93; p = .3571$). The reported results cannot thus be explained in orthographic terms.

Discussion. The present results corroborate those we obtained on transparency ratings, indicating that, even when a dichotomized set of transparent vs. opaque forms are considered, distributionally-based measures are able to effectively distinguish the groups. This is not surprising for the whole-word approach, as LSA measures were used, in the very first instance, to construct the two sets (Rastle et al., 2000). However, results indicate that the same holds for composition-based similarity estimates, confirming the hypothesis that even opaque words manifest a certain degree of compositionality that is effectively captured in the proposed model.

Moreover, the present results indicate that whole-word and compositional approaches dissociate with respect to their quantitative predictions about the item sets and that these predictions pattern quite well with results at different SOAs. The compositional approach predicts opaque and transparent items to be more similar, in terms of ST, than the whole-word approach does. This prediction is mirrored by priming effects at very short SOAs, indicating facilitation for both transparent and opaque words, with a slight advantage for the former (e.g., Diependaele et al., 2005; Feldman et al., 2009). On the other hand, the large difference found between transparent and opaque items by the whole-word-based measure resembles quite faithfully the pattern of results at longer SOAs (e.g., Rastle et al., 2000; Rueckl & Aicher, 2008), where the priming effect is found in transparent pairs only. This dissociation between the distributional measures indicates that the two approaches we described do not necessarily exclude one other, that is, the compositional approach is not simply a full-parsing attempt to recreate the corpus-extracted distributional vectors of derived forms. Rather, from a cognitive point of view, they might constitute models of different semantic processes. The compositional approach captures an early, automatic procedure that capitalizes on regularities and sub-regularities in the semantic system to attempt to combine the meanings of the observed morphemes. The whole-word approach captures instead late procedures based on the semantic similarity between lexical items (including stored representations of derived forms); this similarity is not only determined by systematic aspects, but also by unpredictable lexicalization processes that fall beyond the possibilities of the compositional approach. The late semantic procedure taps into stored knowledge about word meanings that cannot arguably be accessed during the early, fast composition procedure.

As previously mentioned, short-SOA morphological priming experiments mainly indicate a purely form-based decomposition (e.g., Rastle et al., 2004): early in processing, words are morphologically parsed solely on the basis of their apparent morphological complexity, irrespective of actual morphological structure. Indeed, priming is found for pairs like *number-numb* and *dealer-deal*, and not for pairs such as *dialog-dial* (where *-og* is not a potential morpheme). The present results do not challenge the role of purely orthographic segmentation in short-SOA priming effects. To the contrary, as we discussed in the introduction to the semantic transparency experiments, our proposal presupposes this automatic parsing of superficially complex strings, so that the compositional procedure can apply to any potentially complex word (Rastle & Davis, 2003, 2008). However, the present simulations also suggest that an automatic combinatorial procedure of *morpheme meanings* would build upon this orthography-based segmentation (in line with Tsang & Chen, 2014; Amenta et al., in press). This additional semantic procedure, consequent to a semantically-blind parsing, would explains the asymmetry that we observe, at short SOAs, between transparent and opaque forms. The present results confirm that a combinatorial view is crucial in understanding this early semantic processing. Suppose that, after semantically-blind parsing of complex words, the resulting morphemes are system-

atically recombined in an early semantic operation. As suggested by the examples discussed in the previous section, the combinatorial procedure will be, in some cases, able to generate the proper "opaque" meaning of some words (e.g., *fruitless* as "unproductive"). As this meaning is quite different from that of the stem, the semantic contribution to priming effects will be absent or weak. On the other hand, it will produce a transparent version of the meanings of very idiosyncratic combinations (e.g., *archer* as"'builder of arches", *corner* as "corn grower"), which paradoxically will trigger semantic priming (over and above the form facilitation). Most opaque words will fall somewhere between these extremes, resulting, at the aggregated level, in the pattern represented in Figure 5, that is, a priming effect that is not as strong as that observed for transparent words (where the compositional procedure always results in a meaning close to the one of the stem).

This interpretation may help explaining the brittleness of the (small) advantage for transparent vs. opaque priming at short SOAs (e.g., Diependaele et al., 2005; Feldman et al., 2009; Kazanina, 2011; Järvikivi & Pyykkönen, 2011). The variability in this much discussed effect could depend on the different makeup of the item lists. The semantic effect would emerge in studies whose opaque items have meanings mostly obtainable compositionally; on the other hand, the semantic effect would not emerge in cases where most items have highly idiosyncratic meanings, for which the compositional procedure would only generate transparent alternatives. In the former scenario, a priming effect at the semantic level will be (somewhat) smaller for the opaque set, since the corresponding item stems are more heavily modified by their affixes. In the latter case, the compositional procedure will generate transparent meanings for both the transparent and the opaque sets, leading to no difference between the two conditions. More generally, the semantic contribution to the priming effect will be limited (and thus the ST modulation small) because it builds on the stronger systematic influence of the orthographic form at short SOAs.

Certainly, the present considerations seem at odds with the traditional take on masked priming studies, that are usually assumed to reflect pre-semantic processing. However, more recent results are in line with our interpretation: in masked priming conditions, Tsang and Chen (2014) found that opaque words significanlty prime targets that are semantically associated to their stems (e.g., *butterfly-bread*). The authors further showed (Tsang & Chen, 2013) that semantic properties of *morphemes* are activated in a masked priming paradigm. Although we must be cautios in drawing strong methodological conclusions from our simulations (proposing a new theory of priming is not the purpose of the present paper), we believe that the present model may help reconciling this apparent inconsistency in masked priming results. The traditional priming experiments have investigated semantic associative relations that the present model ascribes to the whole-word route. This latter would be based on long-term, stored lexical knowledge, hence explaining why the associated priming effect can only appear in long-SOA conditions, where the association between prime and target can be explicitly appreciated. On the other hand, the more recent results

have focused their attention to *morpheme meanings*. The results they obtained fit well the predictions of the combinatorial route, that rapidly computes the whole-word semantics by relying on the representations of morpheme meanings. The combinatorial take on semantics would thus be crucial in explaining semantic effects at short SOAs. Interestingly, this hypothesis is also consistent with results from eye-tracking literature, indicating that early semantic transparency effects in complex word processing are compositionally connoted (Marelli & Luzzatti, 2012; Amenta et al., in press). Further research on the issue is certainly needed, but we find the converging evidence between the present approach and the results of these recent studies very promising.

As a final note, it is worth stressing again that the present proposal implies an early processing of complex words centered on orthography. As a consequence, it does not preclude the possibility to isolate this pre-semantic processing stage through experimental manipulations. For example, Gold and Rastle (2007) found no overlap between neural priming effects for semantic pairs and neural priming effects for opaque morphological pairs, and the areas associated with morphological effects (left anterior middle occipital girus) were quite unlikely candidates for a semantic procedure. The very short SOA (30ms) used in the study may have helped evidencing a purely morpho-orthographic procedure. As proposed by Gold and Rastle (2007) themselves, further studies manipulating SOAs (e.g., incremental masked priming) would be helpful for a better understanding of the influence of prime duration.

Modulation of frequency effects in lexical decision

Frequency effects have been traditionally seen as diagnostic of the involvement of the corresponding word (or morpheme) representations in lexical processing. If, when reading a derived word, participants' performance is influenced by stem frequency, the corresponding stem representation must contribute in some way to the processing of the derived word. Although many studies have exploited frequency effects to investigate derived-word processing (e.g., Taft, 2004; Baayen, Wurm, & Aycock, 2008; Traficante, Marelli, Luzzatti, & Burani, 2014), surprisingly this has not been done in conjunction with semantic transparency measures. This tradition is instead well-established in the compound domain, where many studies investigated how ST modulates constituent frequency effects (Pollatsek & Hyönä, 2005; Frisson, Niswander-Klement, & Pollatsek, 2008; Marelli & Luzzatti, 2012). In the present experiment we take inspiration from this research line to test our model: the impact of distributionally-defined ST measures will be assessed in a lexical decision task by evaluating how they modulate stem and whole-word frequency effects.

Materials and Methods. A set of 3,806 affixed words and the corresponding lexical decision latencies were extracted from the English Lexicon Project database (ELP; Balota et al., 2007). All selected stimuli contained one of the trained affixes (see Table A1) and were considered morphologically complex on the basis of the morphological annotation provided by CELEX (Baayen et al., 1995). Response times (RTs)

Table 6
Summary of the fixed effects in the analysis of lexical decision latencies, using either composition-based ST or whole-word-based ST as fixed predictor.

Predictor	Composition ST			Whole-word ST		
	Estimate	t	p	Estimate	t	p
Intercept	6.635	1060.18	.0001	6.636	1048.49	.0001
Stimulus length (RCS 1)	0.015	6.69	.0001	0.015	6.81	.0001
Stimulus length (RCS 2)	0.017	5.73	.0001	0.017	5.67	.0001
Derived-word frequency	-0.043	-37.11	.0001	-0.043	-36.52	.0001
Stem frequency	-0.011	9.83	.0001	-0.011	-9.68	.0001
ST	-0.018	-1.11	.2809	-0.007	-0.59	.5615
Derived-word frequency * ST	0.021	2.84	.0045	0.009	1.79	.0732
Stem frequency * ST	-0.014	-2.23	.0261	-0.010	-2.29	.0216

in lexical decision were employed as dependent variable. RTs were logarithmically transformed in order to obtain a more Gaussian-like distribution.

Word frequency of derived forms and stems were collected from the CELEX database (Baayen et al., 1995). Again, semantic transparency was modeled as the proximity (measured by cosine of angle) between the stem vector and either the composed vector of the derived form (composition approach) or its corpus-extracted (whole-word) vector. The interactions between these vector-based semantic measures and the log-transformed frequency variables were tested through mixed-effects analyses, including stimulus length (in letters) as an additional covariate.

Results. Table 6 reports the results of the analyses employing either composition- or whole-word-based ST measures. We also included per-affix random effects on the intercept and the slope of the ST measures, in order to account for affix-associated variance. P-values were computed adopting the Satterthwaite approximation for degrees of freedom. All predictors were mean-centered in order to ensure more reliable parameter estimation. A non-linear length effect improved the model fit; it was computed using restricted cubic splines with three nodes. The interactions between ST and both stem frequency and derived-word frequency were significant for the semantic composition model, whereas only the interaction with stem frequency was evident in the whole-word-based analysis. In the present results, the composition approach provides a better fit to the data with respect to the whole-word approach ($\Delta AIC = 8$). This difference is not negligible: following the approach proposed by Wagenmakers and Farrell (2004), $\Delta AIC = 8$ would indicate here that the composition model is 54.6 times more likely to be a better model (in terms of Kullback-Leibler distance from the distribution generated by the "true" model) than the whole-word approach.

The interactions involving stem frequency in the two analyses are represented in Figure 7. The effects have very similar patterns: the higher the similarity between the stem and the derived-form vectors (either composed or extracted from the corpus), the more facilitatory the effect of stem frequency. In other words, having frequent stems

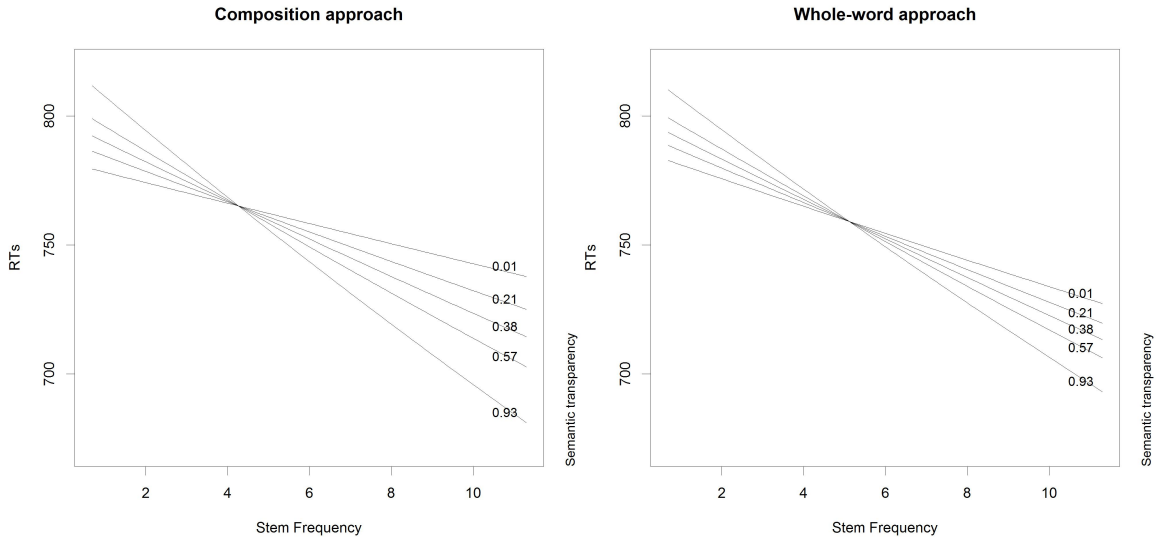

Figure 7. Interactions between stem frequency and ST measures in the composition- (left panel) and whole-word-based (right panel) analyses.

is most helpful when these are similar in meaning to the corresponding derived forms. The interaction involving derived-form frequency in the composition-based analysis is represented in Figure 8: the lower the similarity between the stem vector and the composed vector of the derived form, the more facilitatory the effect of derived-form frequency is.

The superiority of the compositional approach in the present task is further supported by a follow-up analysis on the residual RTs of the statistical model employing whole-word ST. These latter data capture the variance in response times that is not explained when applying whole-word ST. Indeed, a significant interaction between compositional ST and derived-form frequency emerges in this control test as well ($t = 2.071, p = .0384$), indicating that the compositional approach is able to explain a portion of variance in RTs that is crucially missed when using whole-word ST.

Discussion. In the present section we have shown that the ST measures extracted through our model significantly interact with frequency effects in visual word recognition. Frequency effects arguably reflect the ease of access to the concepts subtending the corresponding words (Baayen, Feldman, & Schreuder, 2006). Hence, these interactions are highly informative of the interplay between morpheme meanings during the processing of morphologically complex words.

On data from a straightforward lexical decision task, the measure from the composition approach (a) outperforms the corresponding measure from the whole-word framework in terms of fit and (b) it is able to capture a wider range of phenomena associated to ST. The better performance of the composition-based measure can be explained by considering what we have discovered so far about the cognitive processes underlying it. First, the composition procedure encompasses a wide set of semantic

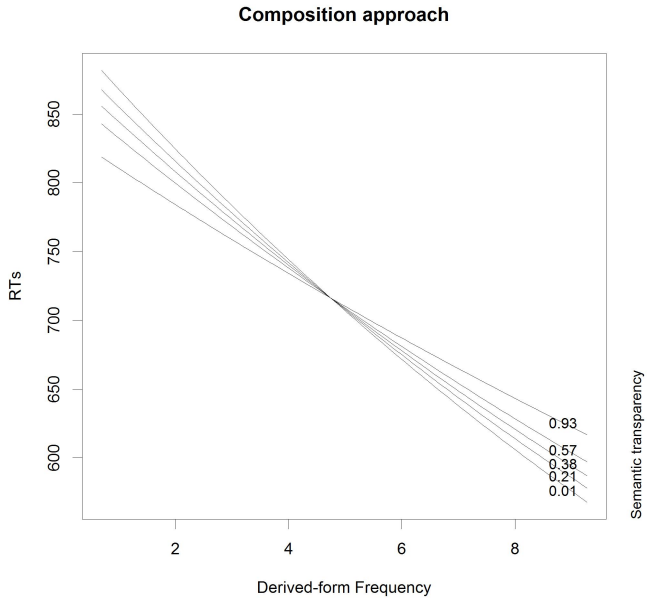

Figure 8. Interaction between derived-form frequency and ST in the composition-based analysis.

regularities and sub-regularities in the derivational process, and it is thus able to produce words on a wide range of the semantic transparency scale (as emerging from the analysis of explicit ST judgments as well as the qualitative analysis of FRACSS-based vectors of derived words). Second, the compositional procedure is fast and automatic, and builds over very early access to constituent morphemes (as suggested by the section on morphological priming). These properties are particularly useful in a lexical decision task, where participants are asked to evaluate as fast as possible the stimulus lexical status, rather than accessing the whole range of semantic properties of the target word. The composition process is arguably able to produce a "familiar enough" meaning to efficiently perform this task, even if it cannot account for a a series of semantic aspects (resulting from semantic drift, diachronic lexicalization, etc.) that we established to be outside the scope of composition, and rather captured by whole-word semantics (and hence whole-word ST). This explanation is in line with a number of results in the literature indicating the importance of stem meaning when performing lexical decision of derived words (e.g., family size effects; De Jong et al., 2000). Also, it fits well with previous results on the pervasiveness of morphological combination, showing that it occurs even when experimental manipulations make composition much less efficient (Taft, 2004).

Over and above the good overall performance of the compositional approach, the associated ST measures resulted in interactions that are also informative of the dynamics involved in the composition process. First, we found an interaction between ST and stem frequency. The effect indicates that, as expected, the familiarity with the concept subtending the stem is more important when the latter is less strongly

affected by the combination with the affix (i.e., when ST is higher). In other words, the ease of access to the stem meaning is important when that meaning is not drastically changed by the FRACSS (transparent words), but not very helpful when that meaning is not maintained through the combination process (opaque words). Second, we found an interaction between ST and derived-form frequency: the lower the ST, the stronger the facilitatory effect of frequency. Under the traditional view of ST, this effect could be easily explained as whole-word access for opaque words. However, in the compositional approach (the only one that shows this effect) there is no stored representation of whole-word meaning. Hence, these results suggest an alternative explanation, following the hypothesis that whole-word frequency effects would reflect stored combinatorial knowledge about morphemes (i.e., their joint probability), rather than being evidence for whole-word lexical representations (Baayen, Wurm, & Aycock, 2008). On the basis of the present results, this stored combinatorial knowledge would be more helpful for opaque words. In these cases, the composition process radically changes the stem meaning, and will thus be particularly demanding in terms of cognitive resources; these words will hence benefit more from whole-word frequency since the possibility to rely on stored information will be much more helpful in cases where the underlying process is more difficult.

General discussion of the semantic transparency experiments

In this section we have tested measures generated from our model in tasks involving existing derived words. In particular, we focused on semantic transparency, operationalized as the proximity between the vector associated to the word stem and the vector associated to the derived form. This latter distributional representation could be either directly extracted from corpus co-occurrences of the derived form, treated as a standalone item (*whole-word* approach), or generated through our data-induced compositional procedure (*composition* approach). The two approaches do not constitute alternative explanations for the same process; rather, they appear to be models of cognitively different and behaviorally distinguishable procedures. Indeed, in a series of three benchmark tests we observed a clear dissociation between composition- and whole-word-based representations. Composition is most predictive of lexical decision latencies and short-SOA priming effects. It can thus be described as an early, fast procedure, that builds on automatically accessed morphemes (Rastle et al., 2004) and generates derived-word meanings on the basis of systematic semantic (sub-)regularities. The whole-word-based measure is a good predictor for explicit judgments on semantic transparency and long-term priming effects. These results suggest a procedure that emerges late during word processing, capitalizes on the similarity between different meanings, captures non-systematic, unpredictable phenomena, and is at least partially based on stored knowledge.

The architecture described is that of a dual procedure system similar to those often proposed in morphological processing (Chialant & Caramazza, 1995; Schreuder & Baayen, 1995; Clahsen, 1999). However, these models mostly focused on the lexical

processing of complex forms. In the present study, the dual route architecture is applied to semantic computation. On the one hand, the meaning of a derived form can be accessed directly as an activation pattern throughout a series of semantic nodes; this distributed representation would include the full extent of the meaning information holistically associated to the word, including non-systematic aspects depending on lexicalization processes. This procedure would model ST as a by-product of the similarity between the meaning of the derived form and the co-activated representation of the stem, in a "network resonance" process similar to the one proposed to explain family size effects (De Jong et al., 2000). On the other hand, the composition route would capitalize on a series of semantic nodes activated by the stem, that is in turn transformed through the FRACSS application. The resulting activation pattern would approximate the derived word meaning on the basis of statistical regularities in the affix semantics. In this procedure, ST will capture the amount of meaning modification that the stem undergoes following affix application, independently of the degree of predictability of the transformation (up to a limit).

We hypothesize that the two routes apply to any word that is (apparently) complex, irrespective of its actual morphological complexity. In other words, both words traditionally considered transparent (e.g., *builder, happiness*) and words traditionally considered opaque (e.g., *fruitless, archer*) would undergo the same dual procedure. Whereas the whole-word route would obviously efficiently retrieve all the semantic aspects of the derived word, irrespective of its ST, one may reasonably doubt of the effectiveness of the composition procedure when dealing with opaque words. Surprisingly, we have shown that a composition approach can explain a wider range of phenomena that one may expect: (many) opaque words present a certain degree of systematicity, that FRACSSs are able to capture. As a consequence, the meanings of words like *fruitless, foxy,* or *heartless* can be obtained compositionally, making the corresponding route reliable for most complex words. At the same time, the proximity of the stem to the obtained derived form serves as an effective cue of (certain aspects of) semantic transparency.

Certainly, in some cases (*archer, corner*) there is no systematic or synchronic relation between the derived form and its (pseudo-)morphemes to rely on. These cases represent an obvious limitation for the compositional route that, we have shown, ends up generating "transparent" alternatives for the meaning of very opaque words (e.g., *archer* as an artisan who builds arches). Given these limitations, one may wonder why a suboptimal system should be applied at all, given the reliability of the alternative whole-word route. Many reasons support the compositional conjecture. First, the empirical results we reported indicate that, in specific tasks, the composition approach generates ST scores that are better at predicting human performance than the whole-word ones. When semantic access is constrained by the experimental setting (short SOA priming) or not fully required to perform the task (lexical decision), composition provides a faster alternative to the whole-word route.

Second, from a theoretical point of view, the processes of a biological system

need not to be optimal, but rather satisficing (Simon, 1956; for a thorough discussion on the issue, see Bowers & Davis, 2012). The composition procedure seems indeed to be "good enough", being effective on the majority of complex words (all the transparent ones and many of the opaque ones). Moreover, the assumption of multiple procedures for the same purpose is in line with the principle of maximization of opportunity (Libben, 1998), that has provided theoretical backing to successful models of morphological processing (e.g., Kuperman, Schreuder, Bertram, & Baayen, 2009).

Third, there are many cases in which a compositional procedure *is required* to obtain the correct meaning. These cases do not only include novel words, but also opaque words that can have alternative transparent readings: words like *chopper* and *ruler*, despite having dominant opaque meanings, carry also transparent senses that can be obtained compositionally. Indeed, it is possible to imagine contexts in which even the most opaque word can be used compositionally: *forty* indicates a number, but one could imagine a group of ancient Romans spotting a piece of land and saying "That area looks quite forty" (i.e., a good place to build their fort)[13]. Recent results confirm that context can be used to prime a transparent reading of opaque words (Amenta et al., in press), and that a compositional procedure is used to retrieve the alternative meaning.

The proposal of a dual-route system opens new research questions related to the relative efficiency of the two procedures. These do not only include the variability in performance across tasks and word types, investigated here by focusing on the ST scores produced by either procedure. It also gives the chance to investigate the extent to which the two routes produce consistent results, and how this affects word processing. Such consistency can be easily quantified by the cosine similarity between vectors representing derived forms obtained compositionally and derived-word vectors represented holistically from corpus co-occurrences. This index will measure the degree of systematicity of a derived-word meaning, that is, it would indicate to what extent the meaning of the word is computable through semantic (sub-)regularities. The cognitive process underlying the measure would be a stage at which the semantic information from the composition and the whole-word route are integrated into a unique representation. A reasonable prediction is that the closer the representations generated by either route, the easier the integration will be, corresponding to shorter processing times observed at the behavioral level.

Finally, the experiments reported in this section suggests that semantic transparency is a more nuanced phenomenon than usually assumed. Specifically, it encompasses both the traditional dichotomy between forms whose meaning can be predicted from their stems and idiosyncratic ones, but also the amount of transformation of stem meaning that is brought about by an ultimately systematic affixation process. This latter effect only becomes clear thanks to our compositional framework, that allows us, for the first time, to go beyond intuitive arguments about systematicity, making precise predictions about which affix-triggered meaning transformation patterns

[13]We are grateful to Kathy Rastle for this example

are statistically robust enough to be captured by a suitably flexible compositional process.

An important point that will require further investigation is the relative speed of the processing routes. In fact, why the composition route should be faster than its counterpart remains an open question. The reasons may rest on the properties of the routes themselves: the whole-word procedure has to retrieve the semantic representations of two independent words (stem and derived form), whereas composition relies on stem meaning only, that is then transformed using one of a limited set of functions. The latter procedure, in our model, can thus be seen as the update of a single semantic pattern (the stem), rather than the actual combination of two independent meaning representations. This hypothesis is in line with approaches positing qualitative differences between the meanings of stems and affixes (Laine, 1999; Lehtonen et al., 2014). On the other hand, it seems at odds with the model by Grainger and Ziegler (2011), in which the coarse-grained (global) route accesses semantics faster than its fine-grained counterpart. We believe this inconsistency is only apparent. Grainger and Ziegler model how orthographic information can activate semantics, whereas the present approach simulates operations within the semantic system itself (and, in particular, how a ST effect can emerge). Therefore, the two approaches focus on different processing levels, and address very different theoretical questions. The present model assumes that the operations postulated by Grainger and Ziegler (2011) have already taken place.

Indeed, one may also hypothesize that the two semantic routes we described build on information from different pre-semantic stages. Let's consider the work by Crepaldi et al. (2010) as reference. This model assumes an early morpho-orthographic stage, at which morphemes are accessed in a semantically-blind fashion, followed by later lexical processing, in which full forms of words (including both stem and derived form) are activated. This contrast fits well the characterization of the two routes of the present model. On the one hand, composition would proceed from the earlier morpho-orthographic stage, exploiting the activated morphemes to generate a semantic representation. On the other hand, the whole word route would concern later lexical stages, with ST effects emerging from the degree of relatedness between independent semantic entries.

In conclusion, most of the literature on morphological processing is focused on lexical aspects, and how these influence semantic activation, rather than the representation of semantics per se. For this reason, future studies will need to delve into the interplay between the present, semantic-centered, system, and previous model capturing form-based aspects of word processing.

General discussion

Semantics has been for many decades the skeleton in the closet of scientific approaches to language analysis (Tomalin, 2006). On the one hand, conveying meaning is arguably the very reason language exists; on the other, the latent nature of meaning

in the linguistic signal makes it hard to study objectively. Distributional semantics offers a way out of the conundrum by suggesting that meaning can be captured with the same distributional approach that has been a core part of linguistic analysis at least since structuralism. Not by chance, it was Zellig Harris, a structuralist deeply concerned with sound methodological procedures, who pioneered distributional semantics already in the fifties (Harris, 1954). If at the time this was just a theoretical program, in the last few decades distributional semantics has become a very concrete proposition, offering empirically effective corpus-induced meaning representations for thousands of words.

The usefulness of distributional semantics has not escaped the attention of the morphological processing community, where it has become fairly standard to use distributional semantic models for quantitative estimates of the relation between stems and derived forms (or compounds and their constituents). But standard distributional semantic models are models of *whole-word* meaning. They might be useful to assess after-the-fact similarity between a derived form and its stem, but they are of no help in modeling the process of derivation at the semantic level.

In contrast, by building on recent research in *compositional* distributional semantics, we introduced here a model of morphological derivation that can account for the dynamic process of meaning construction in word formation. In the FRACSS model, stems are represented by standard distributional vectors, whereas affixes are linear functions that act on stem vectors, modifying them to produce vector representations of the output forms. A qualitative analysis of the semantic neighbours of FRACSS-derived forms confirmed that the transformations encoded in the FRACSS matrices have enough nuance to capture systematic and semi-systematic variations in affix meanings and how they affect stems. Still, FRACSSs do not have enough capacity to capture very idiosyncratic meanings, that must thus be stored holistically. Future research should investigate empirically the predictions we make on the divide between meaning patterns that are systematic enough to be captured by FRACSSs (and could, for example, be productively extended) and what must be left unanalyzed.

By deriving meanings through a compositional process, FRACSSs allowed us, for the first time, to run computational simulations of the all-important phenomenon of novel word derivation. We just started exploring this new field of investigation with our attempt to model nonce-form sensicality and similarity judgments, but of course these explicit judgments and the studied properties are only the tip of the iceberg.

Interesting, however, FRACSSs also led to new insights when we turned our attention to widely studied semantic transparency effects on morphological processing. Equipped with an explicit model for the semantic side of morphological combination, we found that there are important aspects of semantic transparency that had until now been ignored.

In particular, the issue of transparency must be kept clearly distinguished from that of whole-word storage. The changes in stem meaning that FRACSS representations bring about are rich and nuanced enough that the approach can produce *com-*

positionally derived forms that are *opaque* (in the sense of being far away from their stem meaning), without requiring storage of whole-word information (put in other terms, *opaque* does not entail *unsystematic*, if your model of systematicity is flexible enough). We do not claim that *all* derived forms can be obtained compositionally: there are certainly plenty of highly idiomatic complex words whose meanings must be stored holistically. Indeed, the picture emerging from our semantic transparency experiments suggests a place for both compositional and whole-word meanings. Still, the FRACSS model makes concrete predictions about *which* words must be stored in full form due to semantic considerations, and it paves the way to a new line of interesting empirical studies, as well as to more explicit modeling of competition and integration between composition and a direct meaning-retrieval route.

It is interesting, to conclude, to look at how our approach to morpheme semantics fits within the more general picture of morphology and psycholinguistics. There is a long line of research in theoretical morphology that treats (more or less explicitly) affix meanings as feature bundles that affect the meaning of stems (also represented as feature structures) (see Jackendoff, 2002; Lieber, 2004; Scalise, 1984, among many others). In this line of research, feature structures are typically manually specified for just a few affixes and (partially) for a few stems, they contain categorical (unary or binary) features and the combination operations are very simple. Distributional representations for stems and FRACSSs for affixes can be seen as an extension of the featural approach, with much larger, automatically-induced real-valued feature structures (the distributional vectors and matrices) and a general operation to combine them in composition. Interestingly, if recent trends in morphology (e.g., Booij, 2010) tackle the paucity of fully systematic morphological processes and richness of sub-regularities by emphasizing lexicalized schemas over productive rule-based composition processes, our approach suggests an alternative model, that leaves more room to compositional word formation, but assumes richer underlying representations, and makes the process of composition more flexible and nuanced, so that it can also capture patterns that would appear, on first sight, to be only partially predictable.

The FRACSS approach offers a new perspective on the rules vs. analogy and rules vs. similarity debates, as in the famous English-past-tense controversy (e.g., McClelland & Patterson, 2002; Pinker & Ullman, 2002) or in more general discussions (e.g. Hahn & Chater, 1998). The system we propose is pervasively characterized by systematic composition function application, which can be seen as a rule-based process (indeed, according to the criteria of Hahn and Chater, we are proposing a rule-based system). However, on the one hand, the content of the rules (corresponding to FRACSS morpheme representations) are learned via an analogical process in which the FRACSS matrix weights are set so as to provide the best approximation to examples of the composite meanings they should produce. On the other hand, the rules we learn do not operate in terms in discrete terms, but as continuous transformations of real-valued vectors. As such, they are rich and nuanced enough to capture

a good portion of that grey area of half-systematic generalizations that have been traditionally seen as the domain of analogy. Under our proposal, processes are *triggered* in a discrete manner by all-or-nothing formal properties of morphological derivation (affix x being discretely attached to stem y), but they operate in a (superficially) fuzzy manner over continuous lexico-semantic representations: we believe that this distinction between categorical syntactic rules (such as affix concatenation) and less clear-cut lexico-semantic operations (such as affix-triggered stem meaning alteration) has a strong intuitive appeal. On the other hand, we have at the time little to say about processes that appear to be fuzzy on the syntactic side as well, as in the partial formal compositionality of forms such as *grocer* (but see the general discussion of the novel word experiments on how our model could capture the fact that such form might contain semantic features of the affix).

Our approach is very close to connectionist research not only in its representation of word meaning as distributed patterns (e.g., Plaut & Gonnerman, 2000), but also in its emphasis on the need for compositional operations that act over them (Smolensky, 1990). Our proposal is fully in the spirit of Smolensky's early approach to composition in terms of operations on vectors, or more generally tensors (see also Landauer and Dumais (1997) for an interpretation of DSMs as neural networks). We bring two main innovations with respect to this line of research. First, unlike most traditional connectionist networks whose input and training data were de-facto hand-coded, our corpus-induced distributional representations allow us to run real-life, large scale simulations of linguistic phenomena, and provide a natural story for learning. Second, by recognizing the strong asymmetry between argument (free stems) and functional elements (affixes), we propose a view of composition where input vectors are transformed into other vectors that belong to the same space of the input vectors. Under Smolensky's original tensor product proposal, instead, the output of composition is a tensor of much higher dimensionality than the input elements, which is both problematic in computational terms and, more seriously, implausible from a linguistic point of view (an affixed form, for example, is no longer comparable to its stem).

On a more technical level, FRACSS matrices can be seen as fully-connected feed-forward one-layer networks without non-linearities. This makes them easy and efficient to induce from data, and directly interpretable (as illustrated in the toy example in the section on training FRACSS). We stress that the simplicity of the model is offset by the fact that *each* affix is represented by a separate matrix/network, and indeed when composition is seen as a function of affix (matrix) and stem (vector) representations, their relation is *not* linear, as we will show in Appendix B. More importantly, both the qualitative examples and the experimental evidence we reported suggest that the relatively simple FRACSS approach can account for an important portion of the (semi-)systematic semantic patterns encountered in derivational morphology. Further research should ascertain whether extending FRACSS with multiple layers and non-linearities brings about empirical improvements that justify the trade-off in added model complexity.

More generally, the FRACSS approach follows the same path of connectionism and cognitively-oriented distributional semantic models in reproducing the remarkable human capacity of extracting systematic knowledge from complex statistical patterns. These architectures focus on learning associations between levels or linguistic elements, e.g., orthographic features and semantic features (in connectionist models of morphology: Plaut & Gonnerman, 2000), words and documents (in LSA and Topic Models: Landauer & Dumais, 1997; Griffiths et al., 2007), or words and other words (in HAL: Lund & Burgess, 1996). We extend this capacity to second-order associations, in the sense that FRACSSs capture systematic relations between two sets of distributional vectors (stems and derived forms), that are in turn encoding associations between words and their contexts.

With regards to theoretical, "box-and-arrows" models of morphological processing, the FRACSS approach fills a long-standing gap in the definition of morphemes at the meaning level. The model is in line with all those frameworks that assume, more or less explicitly, a combinatorial step between morpheme meanings, and in particular the proposals conceiving qualitatively different representations for stems and affixes (e.g., the single-route decomposition model by Stockall & Marantz, 2006, based on rule-governed concatenations of stems and affixes).

From the point of view of distributional semantics, our research program addresses an important weakness of classic DSMs, namely that they are static, word-based models of the lexicon, providing meaning representations only for (simple and derived) words that are sufficiently frequent in the source corpus. FRACSSs enrich the distributional semantic lexicon with dynamic, word-and-affix processes that allow us to create representations of new words from existing primitive or derived elements. Interestingly, the functional approach has been first developed to account for syntactic composition above the word level. By extending it *below* the word to handle morphological phenomena, we blur the boundary between morphological and syntactic derivation, proposing a unified account for semantic composition at both levels. As the morphology-syntax boundary is far from sharp, we see this as a very promising development.

Note that we have here only experimented with "vanilla" DSM representations. An interesting direction for future research is to experiment with FRACSSs induced on different spaces (e.g., spaces more akin to LSA, Topic Models or Neural Language Models), to see if they capture complementary aspects of semantic derivation.

Many questions are still open, and they will have to be investigated in other studies. However, we believe that the results we presented here demonstrate that the functional approach to distributional semantics can lead to important new insights into the semantic structures and processes of derivational morphology.

References

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological learning - Volume 6* (pp. 58–69).

Amenta, S., Marelli, M., & Crepaldi, D. (in press). The fruitless effort of growing a fruitless tree: Early morpho-orthographic and morpho-semantic effects in sentence reading. *Joural of Experimental Psychology. Learning, Memory, and Cognition*.

Arora, S., Ge, R., & Moitra, A. (2012). Learning Topic Models – going beyond SVD. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium* (pp. 1–10).

Baayen, R. H. (2013). Decomposition makes things worse: Letter bigram troughs in compound reading are not segmentation cues but the critical cues for compound meaning. In *8th International Morphological Processing Conference.* Cambridge, UK.

Baayen, R. H., Davidson, D. J., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 390–412.

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Baayen, R. H., Milin, P., Durdević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM).* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., & Schreuder, R. (1996). Modeling the processing of morphologically complex words. In T. Dijkstra & K. d. Smedt (Eds.), *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing* (p. 166-191). London: Taylor and Francis.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2008). Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities. *The Mental Lexicon*, *2*, 419–463.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Baroni, M., Barbu, E., Murphy, B., & Poesio, M. (2010). Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, *34*(2), 222–254.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for

compositional distributional semantics. *Linguistic Issues in Language Technology*, *6*(6), 5–110.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP* (pp. 1183–1193). Boston, MA.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150-177.

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*, 390-405.

Bloomfield, L. (1933). *Language*. London: Allen and Unwin.

Boleda, G., Baroni, M., McNally, L., & Pham, N. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS* (p. 35-46). Potsdam, Germany.

Booij, G. (2010). *Construction Morphology*. Oxford, UK: Oxford University Press.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in Technicolor. In *Proceedings of ACL* (pp. 136–145). Jeju Island, Korea.

Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

Bullinaria, J., & Levy, J. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, *44*, 890–907.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*, 297-332.

Chialant, D., & Caramazza, A. (1995). Where is morphology and how is it processed? the case of written word recognition. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (p. 55-78). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, *16*, 22-29.

Clahsen, H. (1999). Lexical entries and rules of language: a multi-disciplinary study of German inflection. *Behavioral and Brain Sciences*, *22*, 991–1060.

Clark, S. (2013). Type-driven syntax and semantics for composing meaning vectors. In C. Heunen, M. Sadrzadeh, & E. Grefenstette (Eds.), *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse* (pp. 359–377). Oxford, UK: Oxford University Press.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *Handbook of Contemporary Semantics, 2nd ed.* Malden, MA: Blackwell. (In press; `http://www.cl.cam.ac.uk/~sc609/pubs/sem_handbook.pdf`)

Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, *36*, 345–384.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: constraint-guided conceptual combination. *Cognitive Science*, *24*(2), 299 - 349.

Cover, T., & Thomas, J. (2006). *Elements of Information Theory, 2nd ed.* Hoboken, NJ: Wiley.

Crepaldi, D., Rastle, K., Coltheart, M., & Nickels, L. (2010). 'Fell' primes 'fall', but does 'bell' prime 'ball'? Masked priming with irregularly-inflected primes. *Journal of Memory and Language*, *63*(1), 83–99.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*, 329-365.

Diependaele, K., Duñabeitia, J. A., Morris, J., & Keuleers, E. (2011). Fast morphological effects in first and second language word recognition. *Journal of Memory and Language*, *64*, 344-358.

Diependaele, K., Sandra, D., & Grainger, J. (2005). Masked cross-modal morphological priming: Unravelling morpho-orthographic and morpho-semantic influences in early word recognition. *Language and Cognitive Processes*, *20*, 75–114.

Diependaele, K., Sandra, D., & Grainger, J. (2009). Semantic transparency and masked morphological priming: The case of prefixed words. *Memory & Cognition*, *37*(6), 895-908.

Dinu, G., & Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of EMNLP* (pp. 1162–1172). Cambridge, MA.

Dinu, G., Pham, N., & Baroni, M. (2013a). DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL (System Demonstrations)* (pp. 31–36). Sofia, Bulgaria.

Dinu, G., Pham, N., & Baroni, M. (2013b). General estimation and evaluation of compositional distributional semantic models. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality* (pp. 50–58). Sofia, Bulgaria.

Di Sciullo, A.-M., & Williams, E. (1987). *On the definition of word* (Vol. 14). Springer.

Dubey, A., Keller, F., & Sturt, P. (2013). Probabilistic modeling of discourse-aware sentence processing. *Topics in Cognitive Science*, *5*(3), 425–451.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Erk, K., Padó, S., & Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, *36*(4), 723–763.

Feldman, L. B. (2000). Are morphological effects distinguishable from the effects

of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1431-1444.

Feldman, L. B., Basnight-Brown, D., & Pastizzo, M. J. (2006). Semantic influences on morphological facilitation: Concreteness and family size. *The Mental Lexicon*, *1*(1), 59-84.

Feldman, L. B., Kostić, A., Gvozdenović, V., O'Connor, P. A., & Prado Martín, F. Moscoso del. (2012). Semantic similarity influences early morphological priming in Serbian: A challenge to form-then-meaning accounts of word recognition. *Psychonomic Bulletin & Review*, *19*(4), 668-676.

Feldman, L. B., O'Connor, P. A., & Moscoso del Prado Martín, F. (2009). Early morphological processing is morphosemantic and not simply morpho-orthographic: A violation of form-then-meaning accounts of word recognition. *Psychonomic Bulletin & Review*, *16*(4), 684–691.

Feldman, L. B., & Soltano, E. G. (1999). Morphological priming: The role of prime duration, semantic transparency, and affix position. *Brain and Language*, *68*(1-2), 33–39.

Firth, J. R. (1957). *Papers in Linguistics, 1934-1951.* Oxford, UK: Oxford University Press.

Frisson, S., Niswander-Klement, E., & Pollatsek, A. (2008). The role of semantic transparency in the processing of english compound words. *British Journal of Psychology*, *99*(1), 87–107.

Frost, R., Forster, K. I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 829-856.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*, 20-35.

Gold, B., & Rastle, K. (2007). Neural correlates of morphological decomposition during visual word recognition. *Journal of Cognitive Neuroscience*, *19*(12), 1983–1993.

Goldberg, A., & Jackendoff, R. (2004). The English resultative as a family of constructions. *Language*, *80*(3), 532–568.

Goldsmith, J. A. (2010). Segmentation and morphology. In *The Handbook of Computational Linguistics and Natural Language Processing* (pp. 364–393). Wiley-Blackwell.

Gonnerman, L. M., Seidenberg, M. S., & Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General*, *136*(2), 323.

Grainger, J., & Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Frontiers in psychology*, *2*.

Grefenstette, E., & Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP* (pp. 1394–1404). Edinburgh, UK.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Guevara, E. (2009). Compositionality in distributional semantics: Derivational affixes. In *Proceedings of the Words in Action Workshop.* Pisa, Italy.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of GEMS* (pp. 33–37). Uppsala, Sweden.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, *65*, 197–230.

Harris, Z. (1954). Distributional structure. *Word*, *10*(2-3), 1456–1162.

Hay, J., & Baayen, H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences*, *9*, 342–348.

Hay, J., & Baayen, R. H. (2002). Parsing and productivity. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 2001* (p. 203-235). Dordrecht: Kluwer Academic Publishers.

Heylen, K., & De Hertog, D. (2012). *A distributional corpus analysis of the degree of semantic compositionality of Dutch compounds.* Poster presented at LSD 2012. Leuven, Belgium.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford, UK: Oxford University Press.

Järvikivi, J., & Pyykkönen, P. (2011). Sub- and supralexical information in early phases of lexical access. *Frontiers in Psychology*, *2*, 282.

Ji, H., Gagné, C. L., & Spalding, T. L. (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language*, *65*(4), 406 - 430.

Kazanina, N. (2011). Decomposition of prefixed words in russian. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1371-1390.

Kochmar, E., & Briscoe, T. (2013). Capturing anomalies in the choice of content words in compositional distributional semantic space. In *Proceedings of RANLP* (pp. 365–372). Hissar, Bulgaria.

Kuperman, V. (2009). Revisiting semantic transparency in english compound words. In *6th International Morphological Processing Conference.* Turku, Finland.

Kuperman, V. (2013). Accentuate the positive: Semantic access in English compounds. *Frontiers in Psychology*, *4*(203).

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: towards a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 876-895.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of

lme4 package). [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lmerTest` (R package version 2.0-3)

Laine, M. (1999). Meaning analysis of inflected words. *The Quarterly Journal of Experimental Psychology Section A*, *52*(1), 253-259.

Landauer, T., & Dumais, S. (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.

Lazaridou, A., Marelli, M., Zamparelli, R., & Baroni, M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL* (pp. 1517–1526). Sofia, Bulgaria.

Lazaridou, A., Vecchi, E., & Baroni, M. (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of EMNLP.* Seattle, WA. (In press)

Lehtonen, M., Harrer, G., Wande, E., & Laine, M. (2014). Testing the stem dominance hypothesis: Meaning analysis of inflected words and prepositional phrases. *PLoS ONE*, *9(3).*

Lenci, A. (2008). Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

Libben, G. (1998). Semantic transparency and processing of compounds: consequences for representation, processing and impairment. *Brain and Language*, *61*, 30–44.

Lieber, R. (2004). *Morphology and Lexical Semantics.* Cambridge, UK: Cambridge University Press.

Longtin, C.-M., & Meunier, F. (2005). Morphological decomposition in early visual word processing. *Journal of Memory and Language*, *53*(1), 26 - 41.

Longtin, C.-M., Segui, J., & Hallé, P. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *18*, 313-334.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, *28*, 203–208.

Luong, T., Socher, R., & Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL* (pp. 104–113). Sofia, Bulgaria.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word Formation. A Synchronic-Diachronic Approach.* München: Beck'sche Verlagsbuchhandlung.

Marelli, M., Amenta, S., Morone, E. A., & Crepaldi, D. (2013). Meaning is in the beholder's eye: Morpho-semantic effects in masked priming. *Psychonomic Bulletin & Review*, *20*(3), 534-541.

Marelli, M., & Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, *66(4)*, 644-664.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and

meaning in the English mental lexicon. *Psychological Review*, *101*, 3-33.

Marslen-Wilson, W. D., Bozic, M., & Randall, B. (2008). Early decomposition in visual word recognition: Dissociating morphology, form, and meaning. *Language and Cognitive Processes*, *23*(3), 394–421.

McClelland, J., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*, 465–472.

McDonald, S., & Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of ACL* (pp. 17–24). Barcelona, Spain.

Meunier, F., & Longtin, C.-M. (2007). Morphological decomposition and semantic integration in word processing. *Journal of Memory and Language*, *56*(4), 457 - 471.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* `http://arxiv.org/abs/1301.3781/`.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. (2009). Words and paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition.* Oxford: Oxford University Press.

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429.

Moscoso Del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., & Baayen, R. H. (2005). Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch. *Journal of Memory and Language*, *53*, 496–512.

Murphy, G. (2002). *The Big Book of Concepts.* Cambridge, MA: MIT Press.

Norris, D., & Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: insights from masked priming. *Journal of Experimental Psychology: General*, *137*, 434-455.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193.

Pinker, S., & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456-462.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4/5), 445-485.

Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes*, *20*(1-2), 261-290.

Prado Martín, F. Moscoso del, Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1271–1278.

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2010). On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 186–193).

Rastle, K., & Davis, M. (2003). Reading morphologically-complex words: Some thoughts from masked priming. In S. Kinoshita & S. Lupker (Eds.), *Masked priming: State of the art.* Psychology Press.

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*, 942-971.

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*, 507-537.

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

Rescorla, R. A., & Wagner, A. W. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Royle, P., Drury, J. E., Bourguignon, N., & Steinhauer, K. (2012). The temporal dynamics of inflected word recognition: A masked ERP priming study of French verbs. *Neuropsychologia*, *50*(14), 3542 - 3553.

Rueckl, J. G., & Aicher, K. A. (2008). Are CORNER and BROTHER morphologically complex? Not in the long term. *Language and Cognitive Processes*, *23*, 972-1001.

Sahlgren, M. (2006). *The Word-Space Model*. Ph.D dissertation, Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33–54.

Sandra, D. (1994). The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and cognitive processes*, *9*(3), 227–269.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, *2*(6), 110–114.

Scalise, S. (1984). *Generative morphology*. Dordrecht: Foris.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In

L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (p. 131-154). Hillsdale, New Jersey: Lawrence Erlbaum.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129–138.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence*, *46*, 159–216.

Socher, R., Huval, B., Manning, C., & Ng, A. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP* (pp. 1201–1211). Jeju Island, Korea.

Stockall, L., & Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *The Mental Lexicon*, *1*(1), 85–123.

Strang, G. (2003). *Introduction to linear algebra, 3d edition*. Wellesley, MA: Wellesley-Cambridge Press.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, *57A*, 745–765.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638-647.

Tomalin, M. (2006). *Linguistics and the Formal Sciences*. Cambridge, UK: Cambridge University Press.

Traficante, D., Marelli, M., Luzzatti, C., & Burani, C. (2014). Influence of verb and noun bases on reading aloud derived nouns: evidence from children with good and poor reading skills. *Reading and Writing*, *27*(7), 1303–1326.

Tsang, Y.-K., & Chen, H.-C. (2013). Early morphological processing is sensitive to morphemic meanings: Evidence from processing ambiguous morphemes. *Journal of Memory and Language*, *68*(3), 223–239.

Tsang, Y.-K., & Chen, H.-C. (2014). Activation of morphemic meanings in processing opaque words. *Psychonomic Bulletin & Review*, *21*(5), 1281-1286.

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

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, *28*(6), 649 - 667.

Tzur, B., & Frost, R. (2007). SOA does not reveal the absolute time course of cognitive processing in fast priming experiments. *Journal of Memory and Language*, *56*(3), 321 - 335.

Vecchi, E. M., Baroni, M., & Zamparelli, R. (2011). (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality* (pp. 1–9). Portland, OR.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights.

Psychonomic Bulletin & Review, *11*(1), 192–196.

Wang, H.-C., Hsu, L.-C., Tien, Y.-M., & Pomplun, M. (2013). Predicting raters' transparency judgments of English and Chinese morphological constituents using latent semantic analysis. *Behavior Research Methods*. (In press)

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Zanzotto, F., Korkontzelos, I., Falucchi, F., & Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of COLING* (pp. 1263–1271). Beijing, China.

Appendix A
Learned FRACSSs

Table A1
Complete affix set used in our experiments.

Affix	Type	Stem POS	Derived Form POS	Training Examples
-able	suffix	verb	adjective	284
-al	suffix	noun	adjective	341
-ance	suffix	verb	noun	56
-ant	suffix	verb	adjective	105
-ary	suffix	noun	adjective	89
-ate	suffix	noun	verb	118
-en	suffix	adjective	verb	74
-ence	suffix	adjective	noun	177
-ent	suffix	verb	adjective	75
-er	suffix	verb	noun	1074
-ery	suffix	noun	noun	95
-ful	suffix	noun	adjective	148
-ic	suffix	noun	adjective	386
-ify	suffix	noun	verb	50
-ion	suffix	verb	noun	764
-ish	suffix	noun	adjective	127
-ism	suffix	adjective	noun	108
-ist	suffix	noun	noun	341
-ity	suffix	adjective	noun	495
-ize	suffix	noun	verb	210
-less	suffix	noun	adjective	206
-ly	suffix	adjective	adverb	2884
-ment	suffix	verb	noun	241
-ness	suffix	adjective	noun	1270
-or	suffix	verb	noun	164
-ous	suffix	noun	adjective	234
-y	suffix	noun	adjective	596
de-	prefix	verb	verb	74
dis-	prefix	verb	verb	122
en-	prefix	noun	verb	56
in-	prefix	adjective	adjective	238
mis-	prefix	verb	verb	61
re-	prefix	verb	verb	159
un-	prefix	adjective	adjective	329

Appendix B
On the (non-)linearity of FRACSSs

While in the system we presented each affix corresponds to a linear operation, the relation between the stem (vector) and affix (matrix) representations involved in a composition is actually not linear.

In traditional distributed models of composition (e.g., classic work by Smolensky, 1990, but also more recent work by Mitchell & Lapata, 2010, Guevara, 2010, Zanzotto et al., 2010, Luong et al., 2013), composition is seen as a function $f()$ operating on the concatenation of two input vectors representing the morphemes to be composed. Many of these models are based on matrix multiplication, where the concatenated input vectors are multiplied by the *single matrix* representing the composition function. This is of course a linear operation (each dimension of the output vector is a weighted sum of the input dimensions).

In the FRACSS approach, we associate instead *different* composition functions to certain classes of linguistic expressions (i.e., affixes). In particular, we assign a matrix to each affix, and perform composition by multiplying it by the vector representing the stem. We thus have a separate composition function $f_{affix}()$ for each affix. This function applies to stem vectors, not to concatenations of affix and stem vectors. In this perspective, the approach is linear. However, it has much more power than the linear approaches briefly outlined above, because each affix is represented by a matrix instead of a vector. This implies, first of all, that (if stem vectors are d-dimensional) we have $d \times d$ weights to represent the affix meaning, instead of d.[14] Second, because the affix corresponds directly to the matrix, the output dimensions are no longer weighted sums of input vectors, but sums of dimension-wise products of affix and matrix dimensions (which should capture their interaction beyond additive effects: for example, 0 in an affix cell can cancel out the corresponding stem component).

We can thus re-interpret our model from a different perspective. Suppose that, as in the traditional approaches, we look at composition as a single function $f()$ that applies to the concatenation of distributed representations of the affix and the stem. The affix matrix can of course be unfolded into a $(d \times d)$-dimensional vector, and so the concatenation will be a vector with $(d \times d) + d$ dimensions.

Given the vector v resulting from the concatenation of the affix and the stem, when the composition function $f()$ is applied to v, the k-th dimension o_k (for k ranging between 1 and d) of o, the output vector, is given by:

$$o_k = \sum_{i=1}^{i=d} v_{((k-1)\times d)+i} \times v_{(d\times d)+i}$$

This is no longer a linear function. For example, given a constant a:

[14] We ignore the intercept dimensions here for simplicity, and since they do not affect our main point.

$$af(v) \neq f(a \times v)$$

Compared to a composition model assuming a single or a limited set of linear functions operating on concatenated stem and affix representations (such as those proposed by Guevara, 2010, and Zanzotto et al., 2010), the functional approach we adopted possesses a lot more flexibility thanks to the choice to encode each affix as a separate function, and to the interactions captured by the multiplicative relation between stem and affix dimensions.