Picking buttercups and eating butter cups: spelling alternations, semantic relatedness and their consequences for compound processing

Marco Marelli, Georgiana Dinu,
Roberto Zamparelli, and Marco Baroni
Center for Mind/Brain Sciences, University of Trento, Italy

Abstract

Semantic transparency (ST) is a measure quantifying the strength of meaning association between a compound word (*buttercup*) and its constituents (*butter*, *cup*). Borrowing ideas from computational semantics, we characterize ST in terms of the degree to which a compound and its constituents tend to share the same contexts in everyday usage, and we collect separate measures for different orthographic realizations (solid vs. open) of the same compound. We can thus compare the effects of semantic association in cases in which direct semantic access is likely to take place (*buttercup*), vis-a-vis forms that encourage combinatorial procedures (*butter cup*). ST effects are investigated in an analysis of lexical decision latencies. The results indicate that distributionally-based ST variables are most predictive of RTs when extracted from contexts presenting the compounds as open forms, suggesting that compound processing involves a conceptual combination procedure focusing on the merger of the constituent meanings.

Keywords: semantic transparency, distributional semantic models, compound processing, orthographic variability

In this paper we investigate the role of semantic transparency in English compound processing. In order to pursue this aim, we formalize semantic transparency using a distributional-semantics approach, focusing on different degrees of similarity between a compound and its constituents and the orthographic variability in the usage of English compounds.

The role of semantic transparency (ST) has been a central issue in the literature on compound word processing. This variable measures the degree to which the meaning of a compound is associated to the meanings of its constituents: *carwash* is semantically trans-

This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES). Correspondence concerning this 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

parent, since it denotes something associated to both cars and washing, whereas *fleabag* is not transparent, since it does not denote an actual bag containing fleas. To what extent ST influences the processing of a compound word is far from being clear. The seminal study of Sandra (1990) suggests that compound ST modulates the involvement of constituent representations in word processing. Sandra has shown that semantic priming emerges only for semantically transparent compounds (e.g., priming effect for *truck–carwash*), while the constituents of opaque compounds cannot trigger the same facilitation effect (e.g., no priming effect for *moon–Sunday*). This result suggests that constituents are helpful in accessing a compound only when their meaning is sufficiently related to the compound concept, and hence compound processing is ultimately semantically driven.

The results of later studies have only partially confirmed this view: even if a ST effect has been often reported, at least in the limited sense that transparent compounds are recognized faster than their opaque counterparts (e.g., Libben, Gibson, Yoon, & Sandra, 2003; Ji, Gagné, & Spalding, 2011). However, the semantic properties of a compound have been hardly proven to modulate constituent access: constituent priming effects are present in both opaque and transparent compounds (e.g., priming effect for both *car-carwash* and *sun-Sunday*; Libben et al., 2003), and constituent frequency is a reliable predictor of processing time irrespective of ST (Pollatsek & Hyönä, 2005). Moreover, some results indicate that even the semantic representations of constituents are routinely accessed independently of ST (e.g., Dohmes, Zwitserlood, & Bölte, 2004; Ji et al., 2011). Indeed, current evidence hedges in favor of the idea that constituent representations are *always* accessed at the lexical level, and only late semantic processing of these representations generates the reported ST effects (Libben, 1998).

Recently, it has been proposed that a variety of results associated with ST may depend on how the measure is conceptualized (Marelli & Luzzatti, 2012), and hence operationalized. ST is most often modeled in terms of *semantic relatedness*, focusing on the semantic similarity between a compound and its constituents. That is, access to constituent representations helps recognizing transparent compounds because their meanings activate the compound meaning. This view is efficiently formalized within a localist framework (Libben, 1998), assuming that constituents and compounds are discrete units in a network representing word meanings as symbolic nodes: in the case of high ST, the constituent nodes are linked to the compound, whereas such links are rescinded in the case of opaque compounds. However, ST can also be interpreted as the degree of *semantic compositionality* of a compound, measuring how well the combination of the constituents represents the compound meaning, independently of the degree to which the components, when treated as independent words, are related to the meaning of the whole. This approach assumes that the concepts associated to the first (hence modifier) and second (hence head) constituent are combined into a structured representation during compound processing: the meaning of *swordfish* is not related to the meaning of *sword*; nevertheless, when *sword* and *fish* are considered together, it becomes apparent that *sword* underlines features which highly characterize the combined concept *swordfish*, hence *swordfish* is semantically compositional to a certain degree. Although this approach can also be defined in localist terms (Gagné & Shoben, 1997; Devereux & Costello, 2006), it is probably more profitably characterized by a distributed view of lexical meanings (Lynott & Ramscar, 2001), under which a (transparent) compound is seen as the result of a merger between the constituents' subsymbolic

nodes. The compositional view of ST has recently found empirical support. Indeed, an interaction between constituent frequency measures and ST emerges when the conceptual combination between constituents is encouraged by experimental manipulations (e.g., by presenting the two constituents in different colors, or employing compound-like nonword fillers; Ji et al., 2011). Moreover, if ST ratings are quantified by asking subjects about the extent to which compound meaning is predicted from the meaning of the underlying constituents, the variable significantly modulates constituent frequency effects (Marelli & Luzzatti, 2012), whereas the same modulation does not emerge when subjects are directly asked for the degree of semantic similarity between the components and the whole word. In conclusion, conceptualizing ST in terms of semantic compositionality is a promising approach when trying to understand the role of ST in compound processing.

The two hypotheses described above are obviously not mutually exclusive, and ST, as usually measured, may encompass different latent variables. Indeed, most transparent compounds are both compositional and similar to their constituents, making it difficult to disentangle effects associated to either interpretations of ST. This study attempts to paint a clearer picture by exploiting quantitative measures developed under the tenets of distributional semantics. The central idea of distributional semantic models (DSMs) is that the meaning of words and phrases can be approximated by vectors summarizing their patterns of co-occurrence in corpora. The similarity between two word meanings can hence be modeled as the proximity of their corresponding vectors. In other words, these models work under the assumption that the more similar two meanings are, the more often they will be found in the same contexts. DSMs have already been successfully applied in cognitive science. Popular models such as LSA (Landauer & Dumais, 1997), HAL (Lund & Burgess, 1996) and Topic Models (Griffiths, Steyvers, & Tenenbaum, 2007) have been shown to predict human performance in a number of tasks involving lexical semantic knowledge. Kuperman (2009) and Wang, Tien, Hsu, and Pomplun (2012) already observed that compound ST can be characterized in terms of proximity of the compound vector to the head and modifier vectors. However, these approaches do not address the issue we discussed above, namely the confusion between semantic relatedness and semantic compositionality. By focusing on similarity between lexical meanings, they are hardly informative about compositionality.

Nevertheless, DSMs permit to take compositionality into account from a different point of view, i.e., by taking notice of the spelling alternations of written compounds. English compound words (Bauer, 1998) can be written either as unique orthographic strings (hence *solid form*) or with a blank space separating the constituents (hence *open form*).[1] Recently, Kuperman and Bertram (2013) have extensively studied this phenomenon, showing that spelling variability is dependent on a number of variables, including semantic, lexical, and orthographical dimensions. In particular, semantic similarity between the two constituents has a significant effect on the preference for either alternatives: compounds with similar constituents are more likely to be written as open forms. The authors associate this effect to a semantic combination procedure: similar constituents are easier to integrate, and

[1] Hyphenated forms are also attested. However, it is not clear what the usage of such forms cues in terms of semantic properties (Kuperman & Bertram, 2013). For this reason, in this paper we focus on contrasting open and solid compounds, and leave the investigation of hyphenated forms to future research.

they are thus written in a form favoring combinatorial processing.[2] This interpretation is in line with the traditional view on spelling variation, that assumes different forms to be associated to different degrees of semantic compositionality (Frisson, Niswander-Klement, & Pollatsek, 2008). The open form would cue a more productive usage of compounding, and thus procedures in which constituent concepts are actively combined. We propose that semantic representations extracted from contexts in which a compound is written in open versus solid form will capture more or less compositional usages of the compound, and that this orthographic cue can thus be used as a proxy for compositionality. The usage of a compound in open form suggests an active, systematic combination between constituent meanings, with the combined concept being computed online. On the contrary, the usage of a solid form should be more often associated to a lexicalised compound meaning, resulting from a certain degree of semantic drift that is unpredictable on the sole basis of the constituents.

In this work we exploit distributionally-based measures of ST and orthographic variability in order to investigate the processing of English compound words. First, we run a crowdsourcing analysis to gather further evidence regarding the semantic dimensions influencing spelling alternations. The purpose of this preliminary study is to bring concrete support to our intuition regarding spelling forms and their informativeness about semantic compositionality. Second, we use the very measures extracted in the first study to predict lexical decision latencies. ST effects were tested in a regression-design study including, along with the distributionally-based ST measures, a series of psycholinguistic variables concerning the compound and its constituents. The results we obtained are informative of the semantic procedures involved in compound reading.

Experiment 1: Semantic compositionality and orthographic variability in compound words

In the present experiment, we look for evidence supporting the hypothesis that different spelling realizations subtend different connotations of the compound meaning. Although this is a common belief in the morphological literature, the research on the issue is (with the noteworthy exception of Kuperman & Bertram, 2013) extremely limited. Further results are needed to strengthen our assumption. Here, we consider compounds appearing in both open and solid form in a large corpus, and evaluate to what extent these orthographic forms are associated to qualitatively different semantic neighborhoods. If open forms subtend more compositional connotations vis-á-vis solid forms, we expect their neighbors to be semantically closer to the constituent meanings. Crucially, this alleged effect would be independent from the actual similarity between a compound and its constituents (or between the constituent themselves), further indicating that compositionality cannot be reduced to the semantic relatedness between independent lexical meanings.

Materials and Methods

A set of 100 compound words were used in the present experiment. These were randomly sampled from an initial set comprising all two-constituent compounds listed in

[2]This is not necessarily the case when considering novel compounds. For example, Downing (1977) found that high semantic similarity between constituents may hinder the interpretability of the novel combination.

the English Lexicon Project database (ELP; Balota et al., 2007).

Sentence internal co-occurrences[3] were collected from the 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 words in total). Compound meanings were approximated by vectors indicating their frequency of co-occurrence with the 10000 most frequent content words in the corpora. Raw counts were reweighted using Positive Pointwise Mutual Information (Turney & Pantel, 2010). Separate vectors were collected for contexts related to open and solid compound forms. Open forms were identified on the basis of the noun-noun dependencies revealed by Maltparser (Nivre, Hall, & Nilsson, 2006). In order to evaluate the semantic connotations of the solid and open forms, we extracted the nearest neighbors of their corresponding vectors. Nearest neighbors are words whose vectors are very close to the target (i.e., the compounds) in the semantic space, and are informative of the semantic area the word belongs to.

For each compound we considered the three closest neighbors to the open form and the three closest neighbors to the solid form. The resulting six words were paired with both compound constituents, hence obtaining 12 word pairs associated to each compound. Participants were asked to rate each pair for the relatedness between the meanings of the two words, using a 5-point rating scale ranging from 'completely unrelated' (1) to 'almost the same meaning' (5). Each pair was evaluated by ten different raters. Participants were recruited in a crowdsourcing study using the *Crowdflower* platform (`www.crowdflower.com`). Crowdsourcing is a widespread method in economics and social sciences aimed at the fast collection of large amounts of data by means of online surveys. Recently it has been shown (Schnoebelen & Kuperman, 2010) that, as far as linguistic materials are tested, crowdsourcing provides as reliable data for psychological experiments as those collected in traditional, pen-and-paper tests. A total of 55 participants were recruited in the experiment, each evaluating 218 pairs on average. Only native speakers of English were admitted.

The collected ratings were evaluated in a mixed-effects analysis including orthographic form as predictor. In other words, the analysis was aimed at testing whether the perceived semantic relatedness between the extracted neighbors and the constituent meanings was higher for open than for solid forms. Compounds, constituents, neighbors and participants were included as random intercepts. Participants' ratings were log-transformed for their distribution to be more Gaussian-shaped. Atypical outliers were removed (employing 2.5 *SD* of the residual errors as criterion) in order to ensure that results were not overly influenced by them.

Results

Table 1 reports examples of 'nearest neighbors' for different forms of the same compounds.

The reported examples suggest that the different spelling forms are used to express different meaning connotations of the same compound word. When written as open forms, compounds tend to be related to the literal constituent meanings, displaying neighbors that are very close to its constituent lexemes. On the contrary, solid realizations have neighbors related to an extended (if not metaphorical) meaning of the compound word, often at

[3]Focusing on sentence co-occurrences leads to measures reflecting local dependencies, vis-á-vis the topical relations derived from word-document matrices (e.g., as those used in LSA models)

Compound	Solid form	Open form
moonlight	dream, love, wonder	shine, light, dark
homework	teacher, classroom, lesson	job, home, help
football	coach, soccer, team	kick, throw, round
paintbox	collage, artwork, creative	ink, paint, brush
pitfall	knowledge, understand, process	trap, cave, bat

Table 1: Examples from the semantic neighborhoods of the same compounds in the two different forms. The example words are chosen from the top ten neighbors of each compound form in the described DSM.

an abstract level. This qualitative observation is confirmed by the crowdsourcing study. According to the participants' intuitions, semantic neighbors are closer to the constituent meanings when extracted from open forms ($mean = 2.25; SEM = .06$) as opposed to solid forms ($mean = 2.13; SEM = .06$). The difference, although small, is significant in the mixed-effects analysis ($pMCMC = .0006$).

Discussion

In this preliminary experiment we investigated the different meanings associated to open versus solid compound forms. We exploited methods from distributional semantics to extract the semantic neighbors of each form, and had them judged by participants in a crowdsourcing study. Participants' intuitions indicated that vectors associated to open forms populate a semantic area more associated to the meanings of their constituents, as opposed to the vectors associated to solid-form.

This piece of evidence is in line with the results by Kuperman and Bertram (2013) and Frisson et al. (2008), and confirms our assumption regarding spelling alternations: open forms reflect productive, constituent-based combinatorial procedures, as opposed to solid forms reflecting a more lexicalised interpretation of the compound. As a consequence, the semantic relationship between a compound and its constituents can change on the basis of the spelling form considered, leading to quite different predictions in psycholinguistic terms. In the following experiment, we will use this orthographic variability as a proxy for semantic combination to test the consequences of the latter on compound-word processing.

Experiment 2: Semantic processing in the recognition of compound words

In this experiment we exploit distributionally-based measures of ST and orthographic variability in order to investigate the processing of English compound words. We extract distributionally-derived semantic measures for open and solid forms of the same compounds. On the basis of the results of the preliminary study, we assume that the distributional semantic similarity between a compound written in solid form and its parts provides an estimate of the relatedness of the compound to its constituents when the compound is used non-compositionally, whereas the same measure for the compound written in open form

captures a more compositional usage of the compound itself. These measures were used to predict lexical decision latencies in a regression-design study. If ST effects are purely dependent on the semantic relatedness between a compound and its constituents, and compositionality plays only a limited role, we should find very similar effects on response times for measures associated to open and solid compounds. On the contrary, the conceptual-composition hypothesis would predict that semantic similarity will be more reliable as a ST measure when calculated in contexts where the compound is used in an actively compositional way (i.e., *open compounds*), in comparison to contexts in which the compound is more lexicalized (i.e., *solid compounds*).

Materials and Methods

Two-constituent compounds listed in the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995) were selected as the initial pool of stimuli. Those for which also lexical decision latencies were available in the English Lexicon Project database (ELP; Balota et al., 2007) were chosen. In order to obtain reliable measures from the DSMs, only those compounds that occur more than 50 times in both open and solid forms in the source corpora were eventually used as experimental stimuli. The final set comprised 1176 two-constituent compound words. Response times (RTs) in lexical decision from the ELP (Balota et al., 2007) were employed as dependent variable. RTs were logarithmically transformed in order to obtain a more Gaussian-shaped distribution. Note that compounds in the ELP are presented as unique orthographic strings, i.e., we investigated lexical decision latencies for compounds in *solid forms*; still, ST predictors were considered for both solid and open forms. Frequencies of both the compounds and their constituents were collected from the CELEX database (Baayen et al., 1995). The bias towards concatenated spelling (*BiasC*), proposed by Kuperman and Bertram (2013), was also considered. *BiasC* (the proportion of times the compound is written in solid form) measures the preference for writing the compound as a unique word, as opposed to separate constituents, and it was shown to be an important predictor in lexical decision.

Sentence internal word-to-word co-occurrences were collected following the same procedure described for Experiment 1. The reference corpus was a concatenation of the 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 words in total). Constituent meanings were approximated by vectors indicating their frequency of within-sentence co-occurrence with the 10000 most frequent content words in the corpora. Compound vectors were obtained following the same procedure; separate vectors were collected for contexts related to open and solid compound forms. Raw counts were reweighted using Positive Pointwise Mutual Information (Turney & Pantel, 2010). Two ST measures were obtained: ST relative to the modifier constituent and the head constituent was modeled as the proximity (measured by cosine of angle) between the compound vector and modifier and head vector, respectively. Since in our set of stimuli the very same compounds appeared in both orthographic forms, each compound had different measures for the open and the solid form. Figure 1 reports the density distributions of the ST measures. Table 2 summarizes the distributions of all the considered predictors.

Generalized Additive Models (GAMs; Wood, 2006) were employed as primary statistical tool. First, the effects of a series of well-studied lexical variables were considered, in

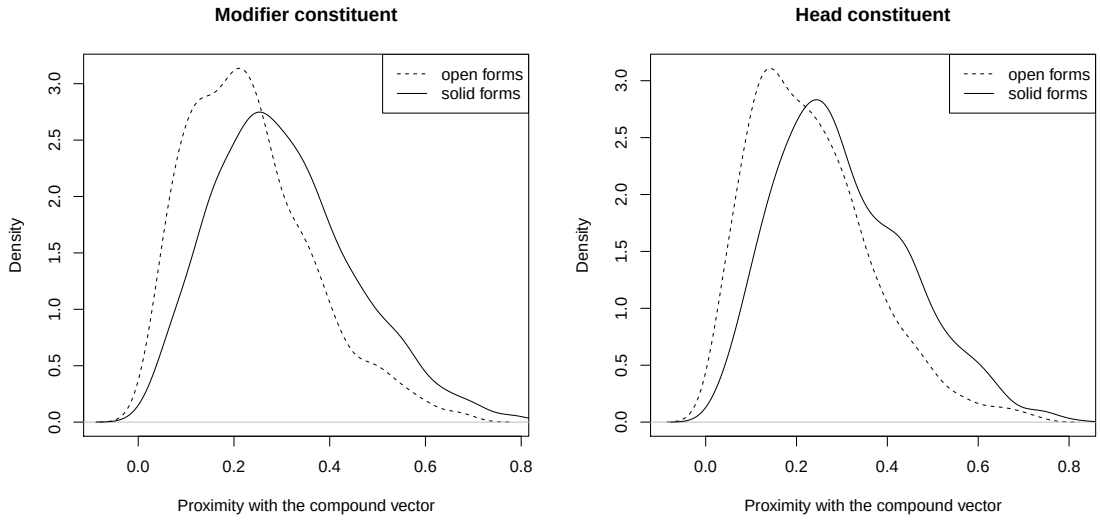

Figure 1. Density distributions of the ST measures, quantified as the proximities between constituent and compound vectors.

Predictor	Minimum	Maximum	Mean	SD
Compound Frequency	1	2177	37.95	116.74
Modifier Frequency	1	72370	3854	5660.51
Head Frequency	1	80740	5077	8064.27
Compound Length	5	14	8.57	1.34
BiasC	.02	.99	.82	.22
Modifier ST (solid form)	.01	.79	.30	.15
Head ST (solid form)	.02	.81	.30	.15
Modifier ST (open form)	.01	.69	.23	.13
Head ST (open form)	.01	.73	.23	.13

Table 2: Summary of the independent variables. Frequency measures are extracted from the CELEX database (18.6 million words). BiasC and ST measures were based on a concatenation of of the ukWaC, English Wikipedia, and BNC corpora (about 2.8 billion words in total).

order to partial out their contributions. Log-transformed frequencies of both the compound and its constituents,[4] as well as string length (i.e., number of letters), were introduced in the model, and potential interactions (as reported by, e.g., Bertram & Hyönä, 2003; Kuperman, Schreuder, Bertram, & Baayen, 2009; Marelli & Luzzatti, 2012) were also tested. *BiasC* was introduced as an additional covariate. The model was progressively simplified by removing

[4]Since in this dataset constituent family sizes (De Jong, Schreuder, & Baayen, 2000) were highly correlated with the corresponding frequency measures ($r = .59$), and disentangling family-size from frequency effects is out of the scope of this work, we focused on the latter measure only.

parameters that did not significantly contribute to the model fit; effects were evaluated one by one on the basis of likelihood ratio tests.

Once this lexical baseline was established, we investigated whether the introduction of the ST measures improved the model fit. Effects associated to head- and modifier-ST were tested, along with a dummy variable specifying whether the semantic measure is derived from solid or open forms. Since the correlation between ST measures was relatively high (.51), modifier ST was regressed on head ST, and the latter was replaced by the residuals of the resulting model (Kuperman et al., 2009). Once the final model was identified, atypical outliers were removed (employing 2.5 *SD* of the residual errors as criterion). The model was then refitted to ensure that the results did not depend on these few outliers. Reported results are consistent with those of the refitted model.

Results

In the analysis dedicated to the lexical variables, significant main effects were found for all the considered predictors. In particular, length led to longer RTs (the longer the compound, the slower the response). Moreover, interactions between frequency measures were also included in the final model, resulting in 36.4% of explained deviance. The introduction of ST measures significantly increased the overall goodness-of-fit (38.9%, $p = .0003$). In particular, an interaction between *BiasC*, head ST, modifier ST, and spelling form was eventually included. Outlier removal further increased model fit to 42.4%.

	Estimate	Std. Error	t	p
Intercept	6.55	.059	112.43	.0001
Compound Length	.022	.002	13.18	.0001
Compound Frequency	-.085	.012	7.36	.0001
Modifier Frequency	.013	.007	1.75	.0807
Head Frequency	.025	.007	3.41	.0007
Comp. Freq. BY Mod. Freq.	.005	.001	4.86	.0001
Comp. Freq. BY Head Freq.	.002	.001	2.09	.0363
Mod. Freq. BY Head Freq.	-.006	.001	5.63	.0001
	edf	Ref.df	F	p
Open forms: Mod. ST BY Head ST BY BiasC	15.31	17.89	6.96	.0001
Solid forms: Mod. ST BY Head ST BY BiasC	13.37	14.62	7.91	.0001

Table 3: Parameters included in the final model and corresponding significance tests. In the bottom part of the Table the approximate significance of the smoothing terms are reported.

Table 3 summarizes the effects in the final model. ST measures resulted in an interaction best described in non linear terms (the introduction of the smoothing term led to significant fit improvement; $p = .0001$). The modulation of the dummy variable (coding for the orthographic form) was also justified ($p = .0034$), indicating that the interaction between the ST measures is different between the two forms. These effects were further qualified by an interaction with *BiasC* ($p = .0001$). Figure 2 represents the resulting 4-way

interaction. The heat-maps report modifier ST on the x-axis, and head ST on the y-axis; the dependent variable (log-transformed RTs) is reported on the z-axis, with the color variation ranging from red (slower RTs) to green (faster RTs). Contour lines delimit areas with significantly different RT values. As a consequence, the number of contour lines is an index of effect size. Left-hand panels represent the effects for solid forms, right-hand panels represent the effects for open forms. Progressively increasing *BiasC* levels are represented by different panels in the vertical dimension: the lower the panel, the larger the *BiasC* values. The top and bottom panels, corresponding to extreme levels of *BiasC*, indicate that at these levels the interaction between ST measures is minimal. Mostly, only a main effect of *BiasC* can be observed, with fastest RTs for large *BiasC* (lowest panels) and slowest RTs for small *BiasC*. On the contrary, central panels indicate an interaction between ST measures, as RTs become progressively shorter as long as both ST measures become higher. In other words, RTs are faster when both modifier and head are strongly associated to the compound meaning (upper right corners of the four central panels); however, it is sufficient that either constituent is unrelated to observe slower responses. A further piece of evidence comes from the difference between compound forms. Comparing the right-hand to left-hand panels, it is clear that the effect is much larger when associated to open forms, vis-á-vis solid spelling: ST measures that are collected in contexts presenting open forms predict a boost in RTs for related constituents that is hardly observed when considering solid compounds.

In addition, linear interactions between frequency measures were found (see Figure 3). The lower the frequency of the compound, the larger the effect of constituent frequencies (left panel and central panel of Figure 3). In the left portion of these panels (corresponding to lower compound frequency), contour lines bend and become more horizontal: this indicates that the frequency of the constituents helps compound recognition more when the compound frequency is low. The effect is observed for both the modifier and the head constituent, although it is larger for the former. Moreover, constituent frequency measures interact with each other (right panel of Figure 3): the shortest RTs are observed when both constituents have very high frequency (upper-right corner), and compound recognition is most difficult when either constituent has low frequency, regardless of its position.

Discussion

In the present study we tested distributionally-based ST measures as predictors of RTs in lexical decision (Balota et al., 2007). Different measures were extracted on the basis of the different usage of compounds, capitalizing on the orthographic variability that these constructs show in English (open vs. solid forms). Indeed, the measures we adopted had significant effects on response latencies, and in particular turned out to improve the fitting of a baseline model composed of lexical variables.

The main result of this experiment, however, is that the effects of the ST measures are mostly observed when these are based on the open form of the compound, i.e., when the co-occurrence vectors refer to contexts in which the compound is written as two separate constituents. On the basis of previously reported evidence and the very results emerging from Experiment 1, the orthographic variability of compounds may be thought of as a reflection of compositionality in compound production. We can thus assume that the open form is used in contexts in which the two constituents are actively combined in order to create a new, composed meaning, and the associated effect can be interpreted in terms of

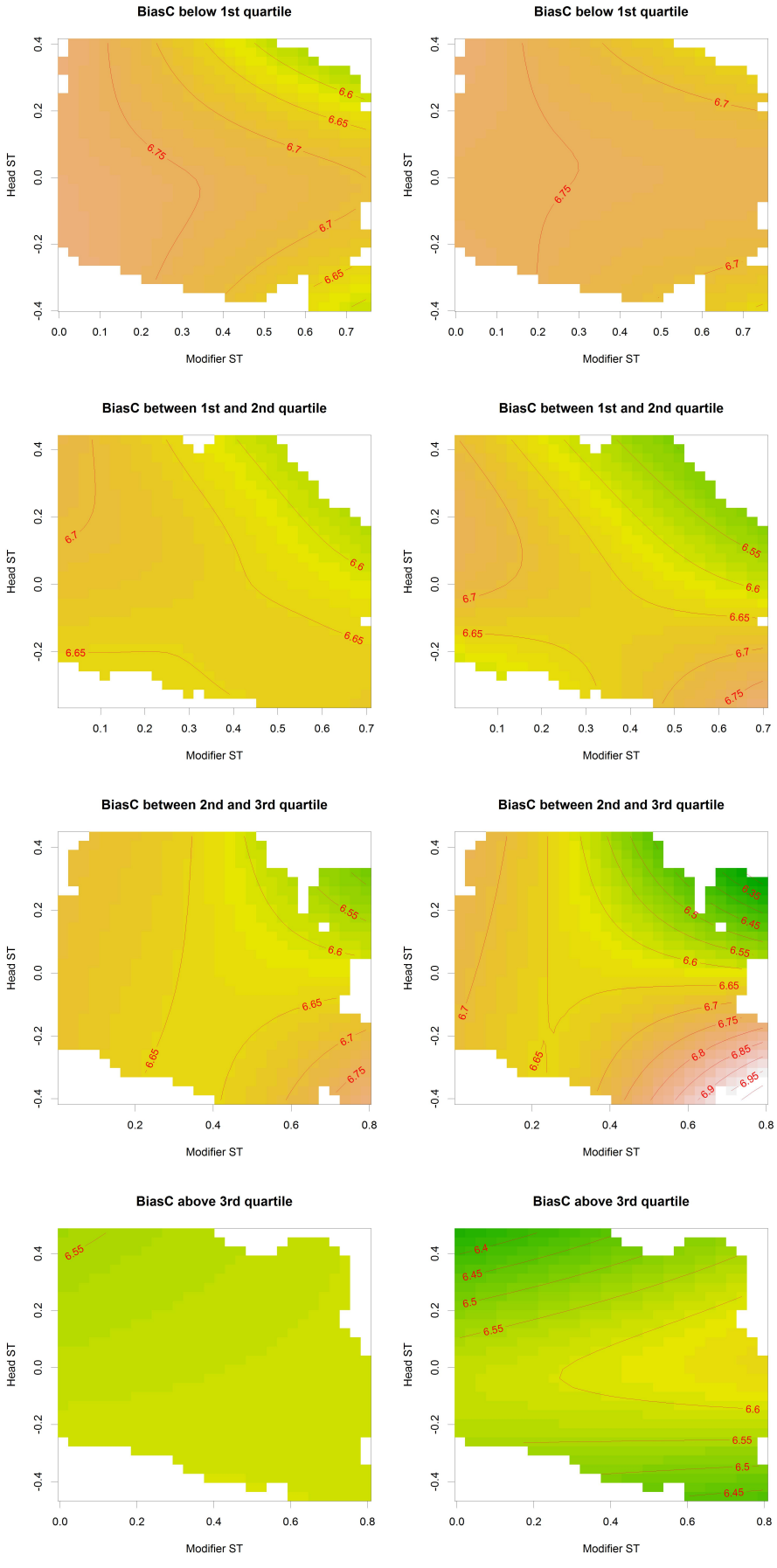

Figure 2. Interaction between ST measures, extracted from solid (left panel) and open (right panel)

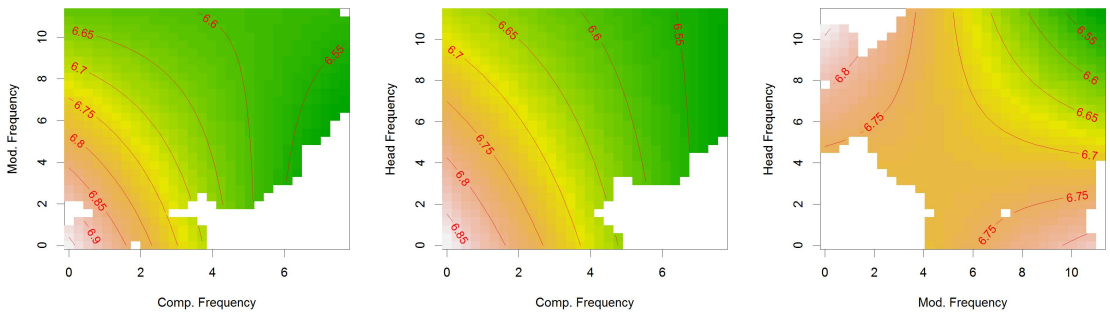

Figure 3. Linear effects of the frequency measures: interaction between compound and modifier frequency (left panel), interaction between compound and head frequency (central panel), interaction between modifier and head frequency (right panel). Response times are represented by the color variation on the z-axis: longer latencies are represented by red areas, and shorter latencies are represented by green areas.

conceptual combination: when processing a compound, the reader attempts to combine the constituent meanings, and does not only access directly the meaning of the whole word. This interpretation is confirmed by inspecting the interaction between ST measures (central-right panels of Figure 2): a boost in compound recognition is observed when both constituents are strongly related to the open-compound meaning, indicating that compound processing is easier when the semantic contribution of both constituents is important. Either constituent being far from the compound meaning suffices to make the processing more difficult. The possibility to integrate both constituents is thus crucial for the semantic processing of compounds, an effect that is difficult to explain with a pure relatedness-based model (for which *fly* should be helpful in recognizing *butterfly*, irrespective of the unrelated constituent *butter*). A combinatorial procedure, on the other hand, would underline the importance of both constituent meanings, in line with the reported interaction between constituent-based ST measures.

The analysis involving ST measures based on solid compounds led to different results. In this case, the interaction between ST predictors has a much smaller (if not negligible) effect size. It seems that this condition is not ideal to describe compound properties associated to semantic combination. Indeed, this is not surprising: solid forms tend to be used in contexts in which the compound meaning is most crystallized, therefore reflecting situations in which the role of constituents is not crucial (see examples in Table 1). In other words, the semantic properties of the compounds are reflected in the way these are used in context, and the corresponding measures are the most informative when built out of contexts where presumably some conceptual composition took place (namely, open-form contexts). Note however that compound stimuli are presented in their solid forms in the ELP. The properties associated to the everyday usage of a compound can thus dissociate from its actual form, and arguably represent information stored in the mental lexicon. In other words, the properties observed for open forms are associated to the compound representation itself, and play a role during processing irrespective of the way the compound

is actually presented (obviously, how a compound is presented in the experimental setting can still influence the relative importance of the combination procedure; see Juhasz, Inhoff, & Rayner, 2005).

The routine access to both (open-form and solid-form related) connotations of a compound is also supported by the modulation of *BiasC*. Indeed, the interaction between ST measures is evident when both open and solid forms are familiar to the reader (i.e. at mid-levels of *BiasC*, central panels of Figure 2), but it is not found when extreme *BiasC* levels are considered. Very large *BiasC* values (i.e. cases where the compound is almost always written as a unique word) are associated to fast RTs and no interplay between constituent ST measures, suggesting that in the processing of these compounds constituent meanings play only a limited role. On the contrary, low *BiasC* values, although also showing limited interactions between ST measures, are associated to very slow RTs, probably due to the interference emerging from presenting a compound that often appears in open form as a unique word. This result is in line with the interpretation in terms of the familiarity of a form over the other, suggested by Kuperman and Bertram (2013), but also implies that both forms (and information associated to both forms) are accessed during processing.

In conclusion, the present results support the hypothesis of a routinely combinatorial procedure (Ji et al., 2011; Marelli & Luzzatti, 2012), even for known compounds presented in concatenated form. As proposed by Marelli and Luzzatti (2012), this may be thought of as complementary semantic processing in a multiple route framework (Kuperman et al., 2009). The present data fit well with this proposal also when considering the effects of the lexical covariates that we included in the analysis: the reported frequency effects are a direct prediction of the model, and reflect trade-offs in the access to the different lexical representations involved. In particular, the interactions between constituent and compound frequency indicate that constituent cues are more involved in processing when the compound is difficult to recognize directly (Kuperman et al., 2009), and the interaction between constituent frequencies is typical of lexical decision, when constituents are arguably processed in parallel and the participants' choice capitalizes on the properties of both (Marelli & Luzzatti, 2012). In both cases, the phenomena are easily explained by conceiving parallel and interacting routes in compound processing (Kuperman, Bertram, & Baayen, 2008), in which different variables are exploited in order to maximize the efficiency of word recognition.

General discussion

In this paper we studied semantic effects in compound processing by borrowing methods from computational (distributional) semantics. We measured ST by examining the similarity between co–occurrence vectors representing the compound and its constituents, but different measures were extracted on the basis of the different usages of compounds, capitalizing on their orthographic variability in English (open vs. solid forms). These different orthographic forms were taken as a proxy measure of semantic compositionality. Indeed, solid and open forms subtend rather different meaning connotations of the same compounds: open forms reflect a productive, constituent-based combinatorial procedure, whereas solid forms have more lexicalised meanings, associated to the whole compound, often at an abstract level. This assumption was confirmed in Experiment 1, showing that semantic neighbors of the compounds were rated to be closer to constituent meanings when extracted from open forms, as opposed to solid ones. In Experiment 2, the extracted ST

measures were tested in a lexical decision task, and turned out to improve the fitting of a baseline model of lexical variables. Results indicate that distributionally-based ST variables are most predictive of RTs when extracted from contexts presenting the compounds as open forms, and suggest that compound processing involves a conceptual combination procedure focusing on the merger of constituent meanings.

DSMs has already been successfully employed in psycholinguistic experiments, but usually they were aimed at predicting off–line measures or, at most, quantifying word association (e.g., to replicate semantic priming effects; Lund, Burgess, & Atchley, 1995). The present study exploits vector space modelling to characterize the internal properties of compound words. Although it represents one of the first steps (see also Kuperman & Bertram, 2013) in applying this interdisciplinary approach to the study of isolated word processing, and its nature is thus exploratory, the results are clear-cut: not only the proposed measures are good predictors of the processing time, and can be profitably employed in psycholinguistic studies, but the results offer theoretical insights that are consistent with recent perspectives on compound processing.

A central issue of the present manuscript concerned how to best describe the role of constituent meanings in compound processing: either the contribution of the constituent meanings depends on their semantic similarity to the separate, holistic meaning of the compound, or the constituent meanings are combined in an active process to obtain the compositional concept. In previous studies, these conditions has been effectively dissociated by means of experimental manipulations (Ji et al., 2011; Marelli & Luzzatti, 2012). In the present work, we aimed at replicating these alternative accounts by exploiting corpus-based measures focusing on different orthographic realizations of the same compounds (open vs. solid forms). Under these conditions of natural language usage, characterizing the constituent-contribution issue in strictly binary terms may be misleading: we cannot expect a 100% consistent association between orthographic realizations and semantic processing, and a certain amount of noise in the data is unavoidable. In other words, contexts where the orthographic form is not informative of a specific semantic connotation (compositional vs. lexicalised) are to be expected, as well as examples in which the form choice depends on non-semantic aspects of the compound (Kuperman & Bertram, 2013). The consequent partial overlap between open and solid form variables notwithstanding, measures extracted from either realization still resulted to be informative of semantic processing: open forms generated representations that are perceived to be closer to the constituent meanings (Experiment 1) and ST measures that are better predictors of the constituent roles in compound access (Experiment 2).

The effect of the ST measures indicate that, as far as constituents are concerned, a combinatorial procedure is more apt at characterizing their role in the semantic processing of the compound, and this combinatorial procedure can in turn be captured by looking at contexts in which the compound is produced in its open form. The central role played by compositional processes in accessing a compound meaning is a long-standing proposal in the conceptual-combination literature. In fact, although this research line stems from the study of novel combinations (e.g., Gagné & Shoben, 1997), more recent results indicate that also for known, lexicalised compounds a procedure capitalizing on the active combination of constituent meanings is routinely in place (Gagné & Spalding, 2006; Gagné & Spalding, 2009). In line with this view, the results by Ji et al. (2011) and Marelli and Luzzatti (2012)

indicate that compound ST and conceptual combination are strongly associated phenomena, and that the latter may provide the ideal theoretical framework to characterize the former.

Clearly, the ST measures we employed are meant to test the role of constituent meanings, and the present data cannot be taken as suggesting that a lexicalised holistic meaning is not (also) accessed when reading a compound word. On the contrary, the very effect of *BiasC* (Kuperman & Bertram, 2013) suggests that both a compositional approach (captured by open forms) and the access to the lexicalised compound meaning (captured by solid forms) may take place during processing (although the relevance of either procedure is influenced by the relative familiarity with the orthographic variants). The hypothesis of multiple processing streams suits well the proposal of a the multi-route model (Kuperman et al., 2009). The principle at the base of this model is that, when processing a complex word, every possible lexical cue is considered in order to maximize the efficiency in accessing the word meaning. Indeed, both a direct access to the lexicalized meaning of the compound and a combinatorial procedure involving the constituent meanings are potentially useful sources of information regarding the compound semantics, and could be interactively at work during compound processing. The access to the holistic meaning is crucial to activate those semantic aspects that are secondary to lexicalization phenomena, and cannot thus be computed online on the basis of the constituent meanings. The combinatorial procedure, capitalizing even on partial information regarding the constituents (Marelli & Luzzatti, 2012), has a very early onset and can thus provide a boost to compound access before information regarding the whole-word is available.

This kind of architecture suits well a distributional approach to semantics. Indeed, if compound meaning is represented as a distribution across semantic features (Landauer & Dumais, 1997) or topics (Griffiths et al., 2007), separate procedures (direct vs. combinatorial) are expected to activate very similar patterns, and hence be both informative with regards to the same compound representation (at least when both procedures lead to consistent results, that is, in compounds that are reasonably transparent). The computational details of how this should be formalized are still unclear, but the promising development of compositional DSMs (that derive a distributional meaning for a composite expression from the distributional representations of its parts, e.g., Mitchell & Lapata, 2010; Baroni & Zamparelli, 2010) will hopefully provide tools to pursue this aim.

References

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

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

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (p. 1183-1193).

Bauer, L. (1998). When is a sequence of noun + noun a compound in english? *English Language and Linguistics*, *2*, 65-86.

Bertram, R., & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language*, *615-634*, 48.

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

Devereux, B., & Costello, F. (2006). Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (p. 184-189).

Dohmes, P., Zwitserlood, P., & Bölte, J. (2004). The impact of semantic transparency of morphologically complex words on picture naming. *Brain and Language*, *90*, 203 - 212.

Downing, P. (1977). On the creation and use of english compound nouns. *Language*, 810–842.

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.

Gagné, C. L., & Shoben, E. J. (1997). The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 71–87.

Gagné, C. L., & Spalding, T. L. (2006). Conceptual combination: Implications for the mental lexicon. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 145–168). Oxford, UK: Oxford University Press.

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.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. T. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-214.

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.

Juhasz, B. J., Inhoff, A. W., & Rayner, K. (2005). The role of interword spaces in the processing of english compound words. *Language and Cognitive Processes*, *20*(1-2), 291-316.

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

Kuperman, V., & Bertram, R. (2013). Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes*, *28*, 939-966.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*, 1089–1132.

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.

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.

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

Libben, G., Gibson, M., Yoon, Y., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, *84*, 50–64.

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

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Meeting of the Cognitive Science Society* (p. 660-665).

Lynott, D., & Ramscar, M. J. A. (2001). Can we model conceptual combination using distributional measures? In *Proceedings of the 12th Irish Conference on Artificial Intelligence and Cognitive Science* (p. 1-10).

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.

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

Nivre, J., Hall, J., & Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of lrec* (Vol. 6, pp. 2216–2219).

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.

Sandra, D. (1990). On the representation and processing of compound words: automatic access to constituent morphemes does not occur. *Quarterly Journal of Experimental Psychology*, *42A*, 529-567.

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

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

Wang, H. C., Tien, Y. M., Hsu, L. C., & Pomplun, M. (2012). Estimating semantic transparency of constituents of english compounds and two-character chinese words using latent semantic analysis. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.*

Wood, S. N. (2006). *Generalized additive models.* New York: Chapman & Hall/CRC.