PSIHOLOGIJA, 2017, Vol. 50(4), 503–520

UDC 811.131.1'37
81'23
DOI: https://doi.org/10.2298/PSI161208011M

Word-Embeddings Italian Semantic Spaces: A semantic model for psycholinguistic research

Marco Marelli

Department of Psychology, University of Milano-Bicocca, Milano, Italy

Distributional semantics has been for long a source of successful models in psycholinguistics, permitting to obtain semantic estimates for a large number of words in an automatic and fast way. However, resources in this respect remain scarce or limitedly accessible for languages different from English. The present paper describes WEISS (Word-Embeddings Italian Semantic Space), a distributional semantic model based on Italian. WEISS includes models of semantic representations that are trained adopting state-of-the-art word-embeddings methods, applying neural networks to induce distributed representations for lexical meanings. The resource is evaluated against two test sets, demonstrating that WEISS obtains a better performance with respect to a baseline encoding word associations. Moreover, an extensive qualitative analysis of the WEISS output provides examples of the model potentialities in capturing several semantic phenomena. Two variants of WEISS are released and made easily accessible via web through the SNAUT graphic interface.

Keywords: distributional semantics; word embeddings; Italian; semantic similarity; psycholinguistic resources

Highlights:

- Distributional semantics provides valid computational models in psycholinguistics
- For many languages, such models are unavailable in easy-to-access formats
- A model is proposed for Italian, based on state-of-the-art computational methods
- The model is validated against datasets obtained in psycholinguistic studies

Corresponding author: marco.marelli@unimib.it

Acknowledgement. The research reported in this paper was partially done while I was working at Ghent University. I would like to thank Paweł Mandera for technical support concerning SNAUT and useful discussions about model development, and Cristina Burani and Francesca Peressotti for sharing the data used for model validation.

Vector space modeling has been widely applied in psychology (Landauer & Dumais, 1997) and computational linguistics (Turney & Pantel, 2010) to the purpose of representing word meanings, with several proposals being advanced under the general label of distributional semantic models (hence, DSMs). The principle at the basis of this approach is the distributional hypothesis (Harris, 1954; Wittgenstein, 1953): the meaning of a word is (or can be approximated by) the contexts in which that word appears. Computationally speaking, this typically amounted to extract lexical co-occurrences from large collections of texts. The distributional hypothesis certainly fits well with our intuitions as language speakers, since we are able to easily (if roughly) estimate the meaning of a word we have never heard before from the way it is used in a sentence: if we hear *The small wug flew towards the tree* we easily come up with the idea of the *wug* as a bird. Indeed, as experimentally demonstrated by McDonald and Ramscar (2001), DSMs can well capture how human speakers learn novel-word meanings (for a related work using state-of-the-art multimodal techniques, see Lazaridou, Marelli, & Baroni, 2017). Indeed, a number of DSM proposals (such as LSA or HAL; Landauer & Dumais, 1997; Lund & Burgess, 1996) became popular in cognitive-oriented literature.

However, DSMs are more than simple collections of word co-occurrences. They may rely on basic co-occurrence counts, but focus on extracting the underlying statistical structure from them. In the literature, this was achieved through several methods, including typically dimensionality reduction techniques (for a review, see Turney & Pantel, 2010), but also predictive systems such as neural networks (Levy & Goldberg, 2014; Mikolov, Chen, Corrado, & Dean, 2013). The architecture resulting from these operations is a collection of vectors encoding activations across a set of (sub-symbolic) nodes, in turn defining coordinates in a high-dimensional semantic space. As a result, similar meanings are found in the same portion of the space – their corresponding vectors are close together. The lower-dimension vector representations are relatively far from the original word-co-occurence counts, representing to all intents and purposes an abstraction from these latter low-level data, and providing a profitable computational way to model the conceptual system. Landauer and Dumais themselves, in their paper on Latent Semantic Analysis (1997), saw dimensionality-reduction techniques as a way to explain how we can derive general, high-level concepts from multiple everyday experiences. Griffiths, Steyvers, and Tenenbaum (2007) further characterized meaning as a distribution across topics, defined as probabilistic characterizations of the reduced dimensions. More in general, the vector approach is believed to represent a viable way to model how the human brain stores and processes information (e.g., Eliasmith et al., 2012). These models can also explain how symbolic representations can emerge from distributed, continuous experiences, reconciling perspectives that are apparently at odd with each other.

The computational and theoretical appeal of DSMs was mirrored on the empirical ground. Extensive testing of DSMs indicated that the approach can

generate reliable predictions concerning behavioral effects. These included explicit intuitions concerning word meanings (e.g., Griffiths et al., 2007; Landauer & Dumais, 1997), as well as processing measures such as response times (e.g., Buchanan, Westbury, & Burgess, 2001), priming effects (e.g., Jones, Kintsch, & Mewhort, 2006), and fixation durations in reading (e.g., Griffiths et al., 2007). More recently DSMs were also validated using neuroscientific methods. Vector representations obtained through DSMs were shown to closely resemble brain activations, as measured through neuroimaging experiments (e.g., Mitchell et al., 2008), as well as electric response on the scalp (e.g., Murphy, Baroni, & Poesio, 2009) during language processing.

Recent developments of DSMs moved from techniques based on co-occurrence counts (such as those implemented in LSA and HAL) to implement predictive systems based on neural networks (Mikolov et al., 2013). In this approach, nodes in the input and output layers represent words, and the system learns to predict a target from the surrounding context words by incrementally adjusting the network weights on the basis of the training text. These proposals, often referred as *word-embeddings* models, have been extremely successful at the empirical validation, outperforming traditional approaches in a number of natural-language-processing tasks (Baroni, Dinu, & Kruszewski, 2014) as well as psycholinguistic experiments (Mandera, Keuleers, & Brysbaert, 2017). Remarkably, they were also claimed to represent a more sound model of how humans learn word meanings. In fact, Mandera et al. (2017) highlighted the close relation between the delta-rule and psychologically-grounded learning systems (in the form of the Rescorla-Wagner equations; Rescorla & Wagner, 1972), in order to establish word embeddings as a plausible cognitive proposal at both the computational and the algorithmic level.

However, distributional semantics is not only a theoretical and computational approach concerning how meaning is structured in the human mind. It is also a useful resource for psychological studies, and in particular for the investigation of language processing. Indeed, DSMs can conveniently produce semantic-relatedness estimates in terms of cosines between vectors: the more related two words, the closer their corresponding vectors, the higher the cosine of the angle between them. As a result, DSMs have been often used as a shortcut to quantify semantic relations in a number of experiments, both for the purpose of matching stimuli across different conditions (e.g., Jones, 2010) and of properly defining experimental manipulations (e.g., Kreher, Holcomb, & Kuperberg, 2006). Moreover, in the morphological processing domain, DSM estimates were used to capture semantic transparency by contrasting the meaning of a complex form and that of its stem (*corner-corn* vs. *teacher-teach*, e.g., Rastle, Davis, & New, 2004). Indeed, estimates from DSMs represent a precious instrument for language scientists. It must be noted, though, that most of the studies exploiting estimates from DSMs focused on English. In fact, despite their usefulness, DSMs remain largely unavailable for many languages. This is not to say that DSMs cannot be developed for languages different from English:

in principle, open-source toolboxes (e.g., Dinu, The Pham & Baroni, 2013; Wild, 2011) and text corpora (e.g., http://wacky.sslmit.unibo.it/doku.php) make it now possible to automatically obtain semantic measures for many languages. However, this theoretical possibility is not necessarily mirrored on the practical side. Limitations in programming knowledge or in computational instruments make it difficult for these data to be produced by many research groups, which would rather benefit by out-of-the-box resources in this respect. The importance of data accessibility should not be underestimated: the popularity of LSA as an instrument to obtain semantic estimates has been certainly boosted by the easy-to-use website of Colorado University (http://lsa.colorado.edu/).

The present paper describes a new resource in this respect, namely a set of semantic spaces for Italian to be used for psycholinguistic research: Word-Embeddings Italian Semantic Spaces (WEISSs). On the technical side, these models adopt a state-of-the-art word-embeddings approach, which is not only more reliable in terms of predictive power (Baroni et al., 2014) but also more sound from a cognitive perspective (Hollis & Westbury, 2016; Mandera et al., 2017). On the practical side, the data are released through the SNAUT website (http://meshugga.ugent.be/snaut//), ensuring a high level of accessibility by means of an easy-to-use graphic-user-interface.

Method

I trained a series of WEISSs on ItWaC, a freely available text corpus based on web-collected data and consisting of about 1.9 billion tokens (Baroni & Kilgarriff, 2006). The raw corpus was tokenized and all characters were converted to lower case. Special characters were removed, with the exception of vowels with orthographically-marked stresses (à, è, é, ì, ò, ù).

In order to induce semantic spaces, I applied the freely available word2vec tool (Mikolov et al., 2013) that exploits neural network techniques to automatically obtain vector representations for word meanings. I obtained such meaning approximations for all the words in the corpus with a minimum frequency of 100, for a total of 180,110 different forms. Since vector size and co-occurrence window size are known to affect the quality of the obtained space (see Bullinaria & Levy, 2007; Mandera et al., 2017), I trained separate models by systematically varying these parameters, for a total of 20 semantic spaces. As possible vector sizes I considered 100, 200, 300, 400, and 500 dimensions. In a word-embeddings approach, the number of these dimensions is related to the number of nodes considered in the hidden layer of the neural network on which the training is based. As maximum co-occurrence window sizes I considered 3, 5, 7, and 9 words. That is, the predictive ability of the context words with respect to the target is considered within a maximum window of 3, 5, 7, and 9 words, respectively. How these two parameters should be set in order to obtain the best performing model is tested in the next sections. The other parameters were kept fixed: I applied the CBOW method, as it is computationally more efficient (Mikolov, Le, & Sutskever, 2013) as opposed to the skipgram approach, and, following Mandera et al. (2017) (and in line with the empirical results of Baroni et al., 2014) I set negative sampling to k = 10, and subsampling to t = 1e-5.

The obtained WEISSs were validated against two datasets obtained by previous experiments. These sets consist of lists of word pairs associated with semantic information. The

former set was obtained from the conditions in the semantic-priming experiments described in Burani, Tabossi, Silveri and Monteleone (1989). The latter was obtained from the semantic association norms of Peressotti, Pesciarelli, and Job (2002). Since there is no previously established benchmark for these models and tasks on Italian, in order to assess the model performances I computed a baseline based on Pointwise Mutual Information (hence PMI). PMI is a popular association measure in information theory and computational linguistics (Church & Hanks, 1990). Given two words, PMI quantifies the discrepancy between their probability to co-occur (i.e., their joint distribution) and the probability of them to co-occur by chance (based on their individual distributions).

$$PMI(word1,word2) = log\frac{p(word1,word2)}{p(word1)*p(word2)}$$

As a result, larger PMI values will indicate a stronger, non-random association between the two words. PMI has the convenient property of estimating lexical associations net of the frequency of the individual elements. This is crucial, since words that are used often (e.g., function words) will co-occur with many others simply in virtue of their frequency. However, these latter co-occurrences are not an index of actual association: *the* and *dog* are not particularly informative of each other, whereas *dog* and *bark* are, irrespective of the former pair having higher co-occurrence frequency than the latter. PMI, by also taking into account the frequency of the individual elements, conveniently assigns larger association estimates to the latter pair.

PMI can be ideally employed as baseline for evaluating the quality of semantic spaces. In fact, PMI has proven very reliable in predicting lexical intuitions in human participants (Paperno, Marelli, Tentori, & Baroni, 2014), and it can even outperform traditional DSMs in widely used test sets (e.g., Budiu, Royer, & Pirolli, 2007; Bullinaria & Levy, 2006). However, PMI only captures syntagmatic relations, that is, associations that emerge from actual co-occurrences between elements. But semantic relations are also expressed in paradigmatic terms, that is, through the degree of substitutability of two given elements. Consider synonyms, as an example. Their meanings are by definition extremely associated, but they are not usually found together in a sentence, since it is rarely needed to express the very same meaning more than once in such a limited context. Paradigmatic relationships are exactly the ones that DSMs, with their focus on the extent to which words appear in similar contexts, are supposed to capture. These considerations make PMI an ideal baseline for the purposes of the present study: on the one hand, it is a good predictor of intuitions on word associations, making it a difficult-to-beat competitor; on the other hand, it ignores an important source of semantic information, leaving thus room for improvement.

Results and Discussion

Evaluation on priming datasets

In Evaluation 1, I test the obtained 20 WEISSs against the word pairs used in Burani et al. (1989). In these semantic priming experiments, a related condition (*argento – oro*, silver – gold) were contrasted with an unrelated condition (*libro – veleno*, book – venom). The related pairs were further divided into pairs having a coordinative association (*argento – oro*) and pairs having an attributive association (*sangue – rosso*, blood – red). I considered items from both the experiments described in Burani et al. (1989), for a total of 69 word pairs.

First, I evaluated to what extent WEISSs were able to distinguish between unrelated and related pairs by considering the average model-predicted similarity in the two conditions. Second, I zoomed in into the related subset (46 pairs) and evaluated how WEISSs performed in capturing differences within more nuanced categories, namely the correlative-associative contrast. In both tests, for each word pair I computed the cosine proximity between the corresponding semantic vectors and evaluated to what extent these were different between the conditions individuated by Burani et al (1989). Differences were measured in terms of t-scores. Results are reported in Figure 1 and Figure 2.

In both tests, all WEISS estimates are significantly different between conditions ($t>1.96$) and outperform the PMI baseline. In the former test, focused on the related-unrelated contrast (Figure 1), better performances are achieved by considering smaller vector sizes (100–300 dimensions) and mid-size co-occurrence windows (5–7 words). In the latter test, focused on the correlative-associative contrast (Figure 2), better performances are achieved with small co-occurrence windows (3–5 words). The contribution of WEISS estimates with respect to the PMI baseline was tested by means of logistic regressions. First, a baseline model was fitted in which priming conditions were predicted by PMI values only. Then, separate models were tested in which estimates from each of the considered semantic spaces were included over and above the PMI baseline. The WEISS performance, in terms of improvement in explained variance, was evaluated by means of goodness-of-fit tests comparing each of the model including WEISS estimates with the baseline model. All the semantic spaces considered significantly improved predictions of human-based data in both priming sets (lowest performance in the relatedness test: 3-word window and 500 dimensions, $F(1) = 28.09$, $p = .0001$; lowest performance in the correlative-associative test: 9-word window and 400 dimensions, $F(1) = 9.39$, $p = .0022$). These results indicate that WEISSs capture variance in human data that cannot be explained by simpler PMI associations.

Taken together, these results speak for the usefulness of the proposed models to characterize cognitively-relevant semantic categories and for their applicability to semantic priming experiments. Indeed, a good performance is observed not only for the "extreme" contrast between related and unrelated pairs, but also when considering more nuanced, finer-grained semantic relations (correlative vs. attributive associations).

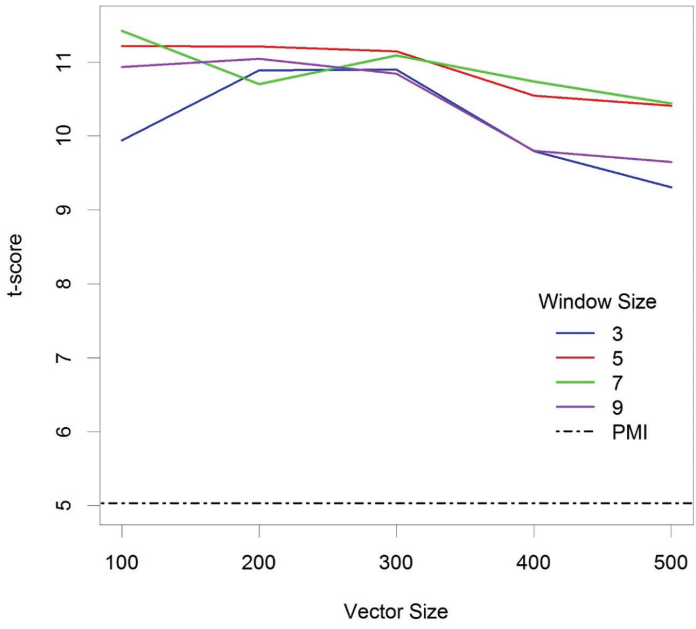

Figure 1. Results of the model validation against the unrelated-vs.-related word pairs by Burani et al. (1989).

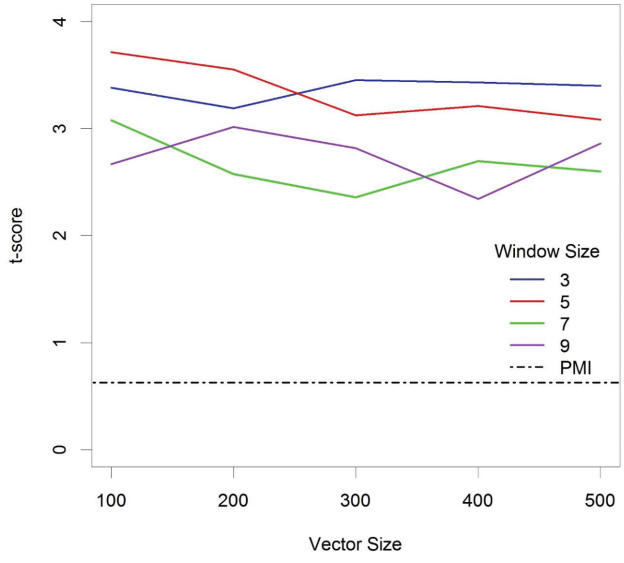

Figure 2. Results of the model validation against the coordinative-vs.-associative word pairs by Burani et al. (1989).

Evaluation on association norms

In the present section, I discuss an additional evaluation of WEISS, focusing on the association norms collected by Peressotti et al. (2002). These data are obtained through free associations, that is, participants are asked to produce the first words that come to their mind when hearing the experimental stimulus. Three responses are requested from each participant. The final dataset lists the frequency of production of each response. Given these premises, these data constitute quite a difficult test set for a DSM: since all the produced words are to a certain degree related to the target, in this task a reliable system is expected to capture small variations within a very homogenous set of highly associated word pairs. This is clearly a tougher evaluation than the one presented in the previous section (focused on extreme examples – related vs. unrelated pairs – or well-defined categories – correlative vs. attributive associations). Therefore, it is an ideal test to further assess the quality of the proposed resource.

I extracted all the pairs from the data released by Peressotti et al. (2002); each pair was formed by an experimental stimulus and a response produced by participants, and was associated with a count value, representing the frequency of production of that response. I then removed responses that occurred only once in the participants' productions, in order to limit the skewness of the distribution and removing potentially uninformative data (e.g., responses dependent on the personal experience of a single participant, rather than on an established association between stimulus and response). I also removed pairs with elements not included as vectors in the semantic spaces (usually productions related to multiword expressions: *succo di frutta*, fruit juice, as a response for *albicocca*, apricot) and with words tagged as strictly related to the context where the data were collected (e.g., names of professors at the University of Padua). As a result, I obtained a set of 3754 pairs. The associated count value was taken as an index of the association between the two words of the pair: a response that is very frequently produced for a stimulus would indicate that that response is strongly associated with the word stimulus; on the contrary, lower counts would indicate that the response, although still related, have a weaker association with the stimulus. In other words, I used association norms as a proxy for relatedness within semantically highly homogenous groups.

For each of the word pairs I computed the cosine similarities between the corresponding semantic vectors on the basis of each of the 20 WEISSs. Model performance was then quantified as the Pearson correlation between the produced cosine estimates and the pair-associated counts indicating how frequently a given word was produced in the human responses. Mandera et al. (2017) followed a different approach for the evaluation of DSMs against association norms. Namely, they considered the average by-target relative entropy between the distribution of human responses and the distribution of model predictions (i.e., semantic neighborhood). I opted for a different

approach for two reasons. First, human responses in the available Italian dataset are much fewer than the ones found in the corresponding English resource used by Mandera et al. (2017); as a result, the present test set is characterized by very skewed distributions that are difficult to approximate, at the target level, in terms of automatically obtained semantic neighbors. Second, in order to provide a convincing validation of WEISS I wanted to compare their predictions to a strong baseline that is theoretically and empirically motivated, namely the PMI association scores. In Mandera at al.'s (2017) method, conversely, a simple random baseline was employed.

Results are represented in Figure 3. A general trend is observed in association with vector size: model performances tend to be better when considering higher-dimensionality vectors. However, the effect of window size is even more evident: the larger the window size, the better the model performance, with the smaller window size not beating the PMI baseline. Overall, the results confirm the usefulness of the obtained resource for capturing semantic associations; indeed, WEISS can provide reliable estimates even when tested on nuanced differences in meaning association: in the present analysis, models are expected to capture different degrees of relatedness within an item set of highly related pairs. The difficulty of the test indeed affects model performance: in comparison with the previous analysis, the model estimates are much closer to the baseline, and correlation scores are relatively low. Moreover, models are more affected by variations in the considered parameters (vector size and window size), with a number of parameter combinations leading to performances that are lower than the one obtained by the baseline.

Despite these shortcomings related to the difficult test employed, the good performance of the models is supported by statistical tests. Indeed, all correlations between WEISS estimates and human-based responses are significant (lowest performance: 3-word window, 100 dimensions, $r = .18$; $t(3752) = 11.11$, $p = .0001$). Moreover, as in the previous analyses, I employed a regression approach to evaluate the WEISS contribution in comparison to the PMI baseline. A mixed-effects linear regression was used (Baayen, Davidson, & Bates, 2008), in order to account for the non-independency of observations by means of by-target random intercepts. Also in this case, all the considered semantic spaces resulted to significantly improve the goodness-of-fit as opposed to the models with PMI scores only, including those spaces whose estimates have lower correlations with human data than the one provided by PMI (lowest performance: 3-word window, 100 dimensions, $\chi^2(1) = 54.71$, $p = .0001$). These results further confirm that WEISS can capture critical variance in human data, and its estimates cannot be reduced to word-to-word statistical associations.

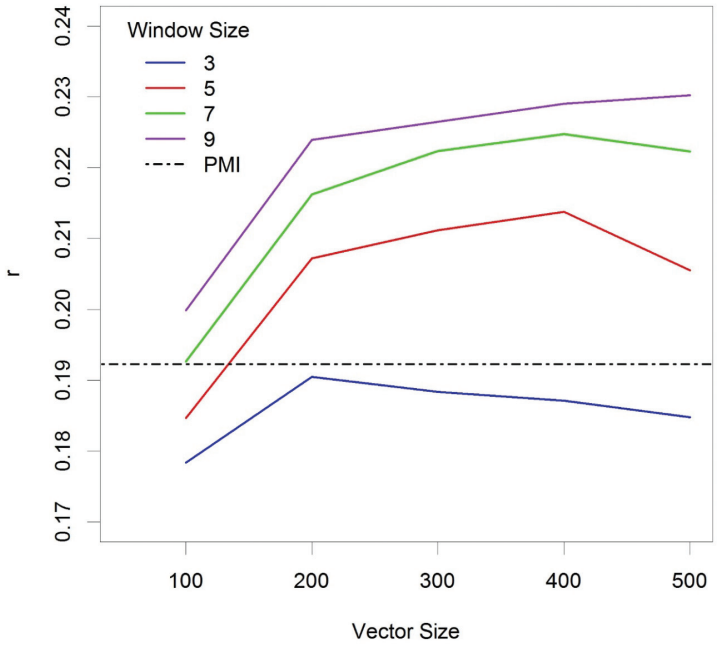

Figure 3. Results of the model validation against the association norms by Peressotti et al. (2002).

Qualitative analysis

The present section describes qualitative predictions about semantic relationships that are automatically generated by WEISS. All the reported examples are obtained from the semantic space with 400 dimensions and 9-word co-occurrence window. DSMs are well suited for this kind of qualitative analyses. By using vector similarity to extract words that are close to a given target, the approach permits to automatically extract semantic neighbors. These provide intuitive cases that speak for the model quality, giving a hint about the meaning structure represented by the vectors.

For example, in WEISS, *casa* (house) has in its top neighbors[14] a series of clearly related words: *abitazione* (residence), *casetta* (small house), *famiglia* (family), *villetta* (cottage), *appartamento* (apartment), *stanza* (room). The neighborhood of *violino* (violin) mostly includes other musical instruments: *violoncello* (cello), *pianoforte* (piano), *flauto* (flute), *liuto* (lute), *clarinetto* (clarinet), *clavicembalo* (harpsicord). However, also morphologically related words are found (*violini* – violins and *violinisti* – violinists). Other targets present more varied (but still sensible) neighborhoods. For example, for *operaio*

14 All the examples in the present section are taken from the top 10 semantic neighbors of the target vector.

(laborer) the neighbors indicate first inflectional variants (*operai* – laborers and *operaia* – female laborer), then specifications (*manovale* – manual laborer and *bracciante* – day laborer), and finally other kinds of workers (*magazziniere* – warehouse worker, *elettricista* – electrician, *contadino* –farmer). *gambe* (legs) has in its neighborhood other human body parts (*ginocchia* – knees, *caviglie* – ankles, *cosce* – thighs, *braccia* –arms), but also the functionally-related *zampe* (paws) and the attributes *muscolose* (feminine plural form of muscular) and *affusolate* (feminine plural form of long/tapered). On the other hand, the neighbors of *lavoro* (work) are most often morphologically related words: *lavorativo, lavorativa, lavorative* (all inflectional variants of work as an adjective), *lavoratore, lavoratori* (singular and plural forms of worker). The impact of morphology in the model predictions is even more evident with verbs, some of which have neighbors dominated by their inflectional paradigm (which, in Italian, is particularly rich): the extracted neighbors of *correre* (to run), for example, are almost exclusively its inflectional variants, with the only exception of *camminare* (to walk).

Model estimates do not only capture lexical semantics and morphology. Some examples suggest that the distributional system can also extract from texts a certain degree of world-knowledge (if not grounded) information, in line with the symbol-interdependence proposal by Louwerse (2011). For example, the neighborhood of *vino* (wine) is populated by different types of (Italian) wines: *aleatico, passito, marzemino, spumante, grignolino, sciacchetrà. parco* (park) produces as semantic neighbors many names of Italian national parks: *Mercurago, Gennargentu, Beigua, Orsiera* and *Rocciavrè*. Moreover, *rocca* (stronghold) is associated with names of famous Italian castles, like *Calascio, Sanvitale*, and *Torrechiara*; however, since Rocca is also the surname of the Alpine skier Giorgio Rocca, it also produces as neighbors names of other athletes, like *Raich* (Benjamin Raich, also Alpine skier), *Schoenfelder* (Olivier Schoenfelder, ice dancer), and *Rivellino* (Roberto Rivellino, football player). To a certain extent, also geographical information is captured in the model (in line with the results by Louwerse & Benesh, 2012). The neighbors of *Belgio* (Belgium) are either surrounding countries (*Francia* – France, *Olanda* – Netherlands, *Germania* – Germany) or Belgian cities (*Namur, Charleroi*). The neighborhood of *Norvegia* (Norway) is populated by the other Scandinavian countries (*Svezia* – Sweden, *Danimarca* – Denmark, *Islanda* – Iceland, *Finlandia* – Finland). The vector of *Laos* is close to the vectors of *Birmania* (Burma), *Cambogia* (Cambodia), *Vietnam*, and *Myanmar*. Finally, in line with the observations by Hutchinson and Louwerse (2014), a certain degree of numerical information seems to be encoded in these lexical-based models, especially concerning small numbers that are found in a relatively defined area of the semantic space. For example, the closest neighbors of *due* (two) are *tre* (three), *quattro* (four), *cinque* (five), in this order; the closest neighbors of *tre* are *quattro*, *cinque*, *due*, *sette* (seven), *otto* (eight), *nove* (nine), in this order; the closest neighbors of *quattro* are *tre, cinque, due, sette, otto,* in this order.

Semantic neighborhoods are not the only way to assess the quality of semantic spaces (and in particular semantic spaces based on word embeddings). Neighborhoods simply capture one-to-one associations between meanings. However, Mikolov et al. (2013) have shown that distributional models are also able to approximate relationships between several meanings: if we subtract the vector of *male* from the vector of *king*, and we add the vector of *female* we obtain a new vector that has between its closest neighbors the vector of *queen*. This analogy test works for many examples in different domains, and indeed I observed similar phenomena in the present Italian models. For example, if I subtract from *calciatore* (football player) the vector of *calciare* (to kick) and add the vector of *suonare* (to play, referred to a musical instrument) I obtain a vector that is close to *musicista* (musician). These relationships at times mirror lexical relations at the morphological level. For example, the plural inflection is well captured in the following case: *stelle* (stars) minus *stella* (star) plus *casa* (house) returns *case* (houses). Gender-related aspects are evident in words denoting professions: *professore* (male professor) minus *uomo* (man) plus *donna* (woman) produces *professoressa* (female professor), that was not a neighbor of *professore* to begin with. Certainly, the observed lexical effects are semantically driven, as other examples in the profession-word group suggest: *dottore* (male doctor) minus *uomo* plus *donna* produces *dottoressa* (female doctor), but only after *ginecologa* (female gynecologist) and at the same distance of *infermiera* (female nurse), and at a similar distance of less expected items (e.g. *bimba*, little girl, and *signora*, lady).

Moreover, in line with previous examples from semantic neighborhoods, there are analogies that are geographically connoted: *Italia* (Italy) minus *Roma* (Rome) plus *Parigi* (Paris) returns *Francia* (France); *Cina* (China) minus *Pechino* (Beijin) plus *Mosca* (Moscow) returns *Russia*. These examples also work at a more local level: *Milano* (Milan) minus *Lombardia* (Lombardy) plus *Lazio* returns *Roma* (Rome), where Rome and Milan are the region capitals of Lazio and Lombardy, respectively. Even if capital cities are typically in the neighborhood of the corresponding country or region, it must be noted that the analogy operation brings them closer, rank-wise, to the newly obtained vectors (e.g., the vector of *Francia* is the fourth closest neighbor of *Parigi*, and the second closest neighbor to the vector resulting from *Italia* minus *Roma* plus *Parigi*).

World knowledge in vector relationships is also exemplified by stereotypical foods and products: *Germania* (Germany) is the result of both i) *Italia* (Italy) minus *vino* (wine) plus *birra* (beer) and ii) *Italia* minus *pizza* plus *wurstel* (frankfurter); *Italia* minus *pasta* plus *riso* (rice) returns *Cina* (China). In a different domain, *FIAT* minus *Italia* plus *Germania* returns *Volkswagen*, and *FIAT* minus *Italia* plus *Francia* returns *Renault*. Also, WEISS seems to capture associations between countries and some of their famous intellectuals and artists: *Galileo* minus *Italia* plus *Germania* returns *Copernico* (Copernicus), *Fellini* minus *Italia* plus *Spagna* (Spain) returns *Bunuel*, *Beccaria* minus *Italia*

plus *Francia* returns *Voltaire*, *Boccioni* minus *Italia* plus *Spagna* returns *Picasso*, *Pirandello* minus *Italia* plus *Russia* returns *Gogol*, *Leopardi* minus *Italia* plus *Inghilterra* (England) returns *Wordsworth*.

Similarly, WEISS seems also to be able to capture a certain degree of historical information. When computing *Mussolini* minus *Italia* plus a different country, I obtain *Roosevelt* for *America*, *Stalin* for *Russia*, *Hitler* for *Germania*, *Hirohito* for *Giappone* (Japan). Although not as precisely, the same analogy leads to vectors close to *Churchill* and *Petain* (leader of Vichy France) for *Inghilterra* and *Francia*, respectively. However, in this domain the time-defined historical relation seems to be mixed with more general-level associations between countries and their infamous dictators: *Mussolini* minus *Italia* plus *Cile* (Chile) returns *Pinochet*, whereas *Mussolini* minus *Italia* plus *Iraq* returns *Saddam*.

All the examples reported speak for the possibilities offered by the WEISS resource, showing how word embeddings can capture sometimes surprisingly nuanced details concerning meaning relations (in line with evidence from other languages: Mikolov et al., 2013) The cases discussed were picked as good representatives of various phenomena, but they are by no means exceptional with respect to the semantic space in general. On the other hand, it remains that they constitute exploratory examples, and cannot be taken as strong evidence about what it is possible to capture with distributional methods. Indeed, there are scenarios in which the model output is not ideal, expressing patterns of similarity relations that does not fit common human intuitions. This is already evident in some of the cases above: for example, when describing how numerical information is encoded in WEISS, *sei* (six) and *uno* (one) are not found because these forms are also used as an inflectional variant of the verb *essere* (to be) and as an indefinite article (a/an), respectively; as a result, their associated vectors also encode these additional meanings, leading to less precise estimates for the corresponding words. In fact, polysemy is an obvious source of noise in this kind of vanilla, general-purpose distributional models, as different meanings ends to be packed together in the same distributional vector (Sahlgren, 2002). More sophisticated techniques need to be applied to deal with this issue (e.g., Schütze, 1998). Moreover, model predictions seem to be affected by stereotypes and social biases, as expressed through the source corpus. This is clearly exemplified by the case *dottore* minus *uomo* plus *donna*, described above and producing as results *infermiera* and *ginecologa*. Although these cases can be object of investigation per se (e.g., Bhatia, 2017), they constitute an additional source of ambiguity when the purpose is generating norms that strictly capture semantic aspects (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). More generally, it must be noted that vector algebra rarely produces a representation in which the desired word is captured by the very first neighbor – in most cases, the obtained vector is still very close to one of those initially introduced in the analogy triplet. For example, when computing *calciatore* minus *calciare* plus *suonare,* the top neighbors of the obtained vector are *suonare* and *calciatore*, and the desired outcome *musicista* is lower in the neighbor ranking. Even if this observation

does not invalidate the model (*musicista* was not a neighbor of either *suonare* or *calciatore*, so it is remarkable that it is found at all between the top neighbors of the obtained vector), it indicates that there is room for improvement with respect to the analogy test.

Apart from this rather systematic issues, there are more specific examples of model predictions that may not fit human intuitions. These are in most cases related to neighbor ranking. For example, between the top ten neighbors of *villaggio*, village, the model includes *kfar* (the Hebrew word for village) and *Watamu*, a small town in Kenya and popular tourist destination with many holiday villages; *rana*, frog, has as second closest neighbor *triturus*, the scientific term for a genus of newts commonly called *tritoni* in Italian; the top neighbors for *concerto*, concert, are *palastampa* and *filaforum*, arenas located in Turin and Milan, respectively, and often used for concerts. Similarly to the *vino* and *parco* examples above, strictly speaking these predictions are not wrong; nevertheless, they are overly specific and certainly not the responses one would expect from an average speaker.

Conclusions

In this paper I described a new DSM for Italian, namely WEISS (Word-Embeddings Italian Semantic Space). The resource is meant to provide an easy-to-use and automatic way to obtain semantic estimates concerning Italian words, by building on techniques from distributional semantics. WEISSs are based on state-of-the-art word-embeddings approaches and are released through the SNAUT website, providing a user-friendly interface to explore semantic spaces. Through SNAUT, it is possible to extract semantic neighbors for a given word, compute analogies such as the ones described in the qualitative analysis section, and obtain semantic-similarity estimates for lists of word pairs.

On the basis of the evaluation reported in the present paper, I release two semantic spaces: WEISS1 (http://meshugga.ugent.be/snaut-italian/), based on a CBOW model with 400 dimensions and a 9-word window, trained on Italian forms from ItWaC; and WEISS2 (http://meshugga.ugent.be/snaut-italian-2/), based on a CBOW model with 200 dimensions and a 5-word window, trained on Italian forms from ItWaC. These two models dissociate in their performance with respect to the proposed evaluations, with WEISS1 being better at predicting spontaneous word associations, and WEISS2 being better at capturing different conditions in semantic priming. This dissociation is in line with previous results (e.g., Mandera et al. 2017), and interpretations concerning window-size effect such as the one proposed by Sahlgren (2008). Releasing both models will permit future users to opt for the solution that is best suited for their research.

As a cautionary note, it must be stressed that the present evaluation is based only on two relatively limited datasets, due to the lack of large-scale resources for the Italian language. Hopefully, with the development and release of new shared datasets, in the future large-scale validations of WEISS (and other Italian

DSMs) will be possible. The present results are anyway an indication of the quality of the models that can be obtained through word-embeddings systems in Italian. In particular, the priming-based analysis shows that WEISS predictions are well apt at capturing differences in terms of semantic relatedness (usually quantified by collecting human ratings): the *t*-value of the best performing model corresponds to a point-biserial correlation of $r = .74$. These results suggest that the WEISS estimates could be used as easy-to-obtain replacement for semantic-relatedness ratings. The significant impact of WEISS estimates in explaining word associations, notwithstanding the apparently low correlations obtained, further strengthen this conclusion: the test set considered in this case is in fact particularly tough for DSMs, as it requires to capture nuanced differences between equally related word pairs. The significant impact of the WEISS estimates, even against a strong baseline condition (represented by the PMI scores), is indeed a further testament to the model quality.

In conclusion, WEISS will be a precious resource for language scientists interested in Italian, that have lacked for long an access to instruments based on DSMs (like the popular LSA website for researches on English). The estimates obtained through WEISS will be useful in a number of domains. For example, WEISS will permit to easily obtain semantic norms for variable matching and for controlling for potential semantic confoundings in studies on word processing or sentence reading. The same norms can be used to computationally operationalize the experimental conditions of priming studies (as shown by the evaluations in the present paper). More generally, the norms could serve as the basis to develop new psycholinguistic measures (e.g., Marelli, Amenta, & Crepaldi, 2015), or to quantify difficult-to-formalize semantic predictors (e.g., semantic transparency in morphological processing: Marelli, Dinu, Zamparelli, & Baroni, 2015; Marelli & Baroni, 2015). The qualitative analysis also suggests potential application aimed at evaluating the degree of grounded information that can be captured by text-based approaches (Louwerse, 2011), as well as the investigation of social biases and stereotypes (e.g., Bhatia, 2017; Caliskan, Bryson, & Narayanan, 2017). Thanks to the accessibility granted by the SNAUT interface, all these applications are now readily available to a large psycholinguistic community.

References

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

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL* (pp. 238–247). Associations for Computational Linguistics. http://anthology.aclweb.org/P/P14/P14–1023.pdf

Baroni, M., & Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of ACL* (pp. 87–90). Association for Computational Linguistics. http://www.aclweb.org/anthology/E/E06/E06–2001.pdf

Bhatia, S. (2017). The semantic representation of prejudice and stereotypes. *Cognition, 164*, 46–60. https://doi.org/10.1016/j.cognition.2017.03.016

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and Reducing Stereotypes in Word Embeddings. *arXiv preprint arXiv:1606.06121*.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8*, 531–544. doi:10.3758/BF03196189

Budiu, R., Royer, C., & Pirolli, P. (2007). Modeling information scent: A comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *Proceedings of RIAO 2007: Large scale semantic access to content (text, image, video, and sound)* (pp. 314–332).

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510–526. doi:10.3758/BF03193020

Burani, C., Tabossi, P., Silveri, M. C., & Monteleone, D. (1989). Relazioni semantiche associative e non associative: Effetti di priming. *Giornale Italiano di Psicologia*, *4*, 617–636.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356(6334),* 183–186. doi:10.1126/science.aal4230

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29. http://www.aclweb.org/anthology/J90–1003

Dinu, G., The Pham, N., & Baroni, M. (2013). DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL* (pp. 31–36). Association for Computational Linguistics. https://aclweb.org/anthology/P/P13/P13–4006.pdf

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*, 1202–1205. doi:10.1126/science.1225266

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. http://dx.doi.org/10.1037/0033–295X.114.2.211

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146–162. http://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, 1–13. doi:10.3758/s13423–016–1053–2

Hutchinson, S., & Louwerse, M. M. (2014). Language statistics explain the spatial–numerical association of response codes. *Psychonomic bulletin & review*, *21*(2), 470–478. doi:10.3758/s13423–013–0492–2

Jones, L. L. (2010). Pure mediated priming: a retrospective semantic matching model. *Journal of experimental psychology. Learning, memory, and cognition*, *36*(1), 135–146. http://dx.doi.org/10.1037/a0017517

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. http://dx.doi.org/10.1016/j.jml.2006.07.003

Kreher, D. A., Holcomb, P. J., & Kuperberg, G. R. (2006). An electrophysiological investigation of indirect semantic priming. *Psychophysiology*, *43*(6), 550–563. doi:10.1111/j.1469–8986.2006.00460.x

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

Lazaridou, A, Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science, 41*(4), 677–705. doi:10.1111/cogs.12481

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems* (pp. 2177–2185). http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. doi:10.1111/j.1756–8765.2010.01106.x

Louwerse, M. M., & Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive science*, *36*(8), 1556–1569. doi:10.1111/cogs.12000

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. doi:10.3758/BF03204766

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. http://dx.doi.org/10.1016/j.jml.2016.04.001

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

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

Marelli, M., Dinu, G., Zamparelli, R., & Baroni, M. (2015). Picking buttercups and eating butter cups: Spelling alternations, semantic relatedness, and their consequences for compound processing. *Applied Psycholinguistics*, *36*(06), 1421–1439. https://doi.org/10.1017/S0142716414000332

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 611–616). http://conferences.inf.ed.ac.uk/cogsci2001/pdf-files/0611.pdf

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

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–1195. doi:10.1126/science.1152876

Murphy, B., Baroni, M., & Poesio, M. (2009). EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the Conference on EMNLP* (pp. 619–627). Association for Computational Linguistics. http://www.aclweb.org/anthology/D09–1#page=657

Paperno, D., Marelli, M., Tentori, K., & Baroni, M. (2014). Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. *Cognitive psychology*, *74*, 66–83. http://dx.doi.org/10.1016/j.cogpsych.2014.07.001

Peressotti, F., Pesciarelli, F., & Job, R. (2002). Le associazioni verbali PD-DPSS: norme per 294 parole. *Giornale italiano di Psicologia, 29(1),* 153–172.

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*(6), 1090–1098. doi:10.3758/BF03196742

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: Current research and theory*.

Sahlgren, M. (2002). Towards a flexible model of word meaning. In *AAAI Spring Symposium* (pp. 25–27). http://www.aaai.org/Papers/Symposia/Spring/2002/SS-02–09/SS02–09–005.pdf

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33–54. http://www.italian-journal-linguistics.com/wp-content/uploads/Sahlgren.pdf

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics, 24(1),* 97–123.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37(1),* 141–188. http://www.aaai.org/Papers/JAIR/Vol37/JAIR-3705.pdf

Wild, F. (2011). lsa: Latent Semantic Analysis [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lsa (R package version 0.73).

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

RECEIVED 08.12.2016.
REVISION RECEIVED 11.05.2017.
ACCEPTED 13.07.2017.