Consciousness and Cognition 56 (2017) 178–187

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text

Edgar Altszyler[a,*], Sidarta Ribeiro[b], Mariano Sigman[c], Diego Fernández Slezak[a]

[a] Depto. de Computación, Universidad de Buenos Aires, Ciudad universitaria, CONICET, Pabellon 1, C1428EGA, Argentina
[b] Instituto do Cérebro, Universidade Federal do Rio Grande do Norte, Natal, Brazil
[c] Universidad Torcuato Di Tella – CONICET, Argentina

ARTICLE INFO

Keywords:
Dream content analysis
Word2vec
Latent Semantic Analysis

ABSTRACT

Computer-based dreams content analysis relies on word frequencies within predefined categories in order to identify different elements in text. As a complementary approach, we explored the capabilities and limitations of word-embedding techniques to identify word usage patterns among dream reports. These tools allow us to quantify words associations in text and to identify the meaning of target words. Word-embeddings have been extensively studied in large datasets, but only a few studies analyze semantic representations in small corpora. To fill this gap, we compared Skip-gram and Latent Semantic Analysis (LSA) capabilities to extract semantic associations from dream reports. LSA showed better performance than Skip-gram in small size corpora in two tests. Furthermore, LSA captured relevant word associations in dream collection, even in cases with low-frequency words or small numbers of dreams. Word associations in dreams reports can thus be quantified by LSA, which opens new avenues for dream interpretation and decoding.

1. Introduction

The analysis of dream contents marked the dawn of Psychology (Freud, 1900; Kraepelin, 1906). Dream contents show gender and cultural differences, consistency over time, and concordance with waking-life experiences, such as activity and emotions (Bell & Hall, 2011; Domhoff, 2002; Domhoff & Schneider, 2008a). Dream contents change after drug treatment (Kirschner, 1999) or due to psychiatric disorders (Domhoff, 2000). Along this line, recently, Mota, Furtado, Maia, Copelli, and Ribeiro (2014) have shown that the graph analysis of dreams reports is quite informative about psychosis, being useful to predict the Schizophrenia diagnosis (Mota, Copelli, & Ribeiro, 2017).

Dream content analysis has been employed to infer the mechanisms that shaped the evolution of dreaming. Threat Simulation Theory (TST) has been particularly influential. It describes the function of dreaming in terms of an evolutionarily selected mechanism, which provides a *world simulation* where we can train responses to threatening experiences (Revonsuo, 2000). This theory brings the analysis of dream contents centerstage, as a natural approach for the investigation of their functionality. For example, Valli et al. (2005) tested the Threat Simulation Theory by comparing the content of the dreams reported by traumatized and non-traumatized children. Their results show an increased number of threatening dream events in traumatized population, thus giving support to the Threat Simulation Theory. Despite the large empirical evidence supporting this theory, some contradictory evidence has been reported against TST (Malcolm-Smith, Koopowitz, Pantelis, & Solms, 2012; Malcolm-Smith, Solms, Turnbull, & Tredoux, 2008). In

* Corresponding author.
E-mail address: ealtszyler@dc.uba.ar (E. Altszyler).

http://dx.doi.org/10.1016/j.concog.2017.09.004
Received 6 March 2017; Received in revised form 25 August 2017; Accepted 10 September 2017
Available online 21 September 2017

this sense, authors have suggested more structured and data-driven tests to be developed (Revonsuo, Tuominen, & Valli, 2015; Revonsuo & Valli, 2008; Valli, 2011). Thus, computational approaches could provide a quantitative framework to tests hypothesis derived from dreams theories.

Most of computational dream content analysis methods are based on frequency word-counting of predefined categories in dreams reports (Bulkeley, 2009; Domhoff & Schneider, 2008b). For example, these techniques have been successfully used to quantify the presence of emotions, sexual content and references to cognitive activity (Bulkeley, 2014). This approach is focused on the salience of words without identifying the context in which they appear. For instance, the occurrence of the word *fall* in a dream report may be used in different contexts, such as *falling* from a cliff, teeth *falling* out or *falling* sick. Moreover, since language is inherently polysemic (Sigman & Cecchi, 2002), semantic ambiguity due to polysemic word associations could hinder the automated analysis of dream reports. In this context, we set out to study the capabilities of word embeddings to capture relevant word associations in dream reports. We believe that word embeddings can be useful not only in extracting words meaning but also in establishing relationships between elements present in dreams.

Corpus-based semantic representations (i.e. embeddings) exploit statistical properties of textual structure to embed words in a vector space. In this space, terms with similar meanings tend to be located close to each other. These methods rely on the idea that words with similar meanings tend to occur in similar contexts (Harris, 1954). This proposition is called *distributional hypothesis* and provides a practical framework to understand and compute semantic relationship between words. Word embeddings have been used in a wide variety of applications such as sentiment analysis (Socher, Huval, Manning, & Ng, 2012), psychiatry (Bedi et al., 2015), psychology (Elias Costa, Bonomo, & Sigman, 2009; Sagi, Diermeier, & Kaufmann, 2013), philology (Diuk, Slezak, Raskovsky, Sigman, & Cecchi, 2012), literature (Altszyler & Brusco, 2015), cognitive science (Landauer, 2007), finance (Galvez & Gravano, 2017) and social science (Carrillo, Cecchi, Sigman, & Analysis, 2015; Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015).

Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007; Landauer & Dumais, 1997), is one of the most used methods for word meaning representation. LSA takes as input a training corpus, i.e. a collection of documents. A word by document co-occurrence matrix is constructed. Typically, tf-idf transformation is applied to reduce the weight of uninformative high-frequency words in the words-documents matrix (Dumais, 1991). The output of the tf-idf transformation is a matrix W where the element $w_{i,j}$ is the weight of word i in the document j,

$$w_{i,j} = tf_{i,j} \cdot \log_2\left(\frac{D}{df_i}\right), \tag{1}$$

where $tf_{i,j}$ is the frequency of the word i in the document j,D is the number of documents in the training corpus and df_i is the number of documents in which word i appears. Then, each document weight is normalized to unit length and a dimensionality reduction is implemented by a *truncated Singular Value Decomposition*, where only the k largest singular values are selected. This method, provides a low-dimensional vectorial representation of every word present in the trained corpus. The success of LSA in capturing the latent meaning of words comes from this low-dimensional mapping (Turney & Pantel, 2010).

More recently, neural-network language embeddings have received increasing attention (Collobert & Weston, 2008; Mikolov, Chen, Corrado, & Dean, 2013), leaving aside classical word representation methods such as LSA. In particular, Word2vec models (Mikolov, Chen, et al., 2013; Mikolov et al., 2013) have become especially popular in embeddings generation.

Word2vec consists of two neural network language models, Continuous Bag of Words (CBOW) and Skip-gram. In both models, a window of predefined length is moved along the corpus, and in each step the network is trained with the words inside the window. Whereas the CBOW model is trained to predict the word in the center of the window based on the context words (the surrounding words), the Skip-gram model is trained to predict the context words based on the central word. In the present study, we use a Skip-gram model, which shows better performance in Mikolov, Corrado, et al. (2013) and in Asr, Willits, and Jones (2016) semantic tasks.

For each training example, the Skip-gram model tends to maximize the log probability of observed context words w_k given the center word w_j,

$$\log p(w_k|w_j) = \log\frac{\exp(\mathbf{c_k}\cdot\mathbf{v_j})}{\sum_i \exp(\mathbf{c_i}\cdot\mathbf{v_j})}, \tag{2}$$

where $\mathbf{c_k}$ and $\mathbf{v_j}$ are the vectorial representations of the context and central words, respectively, and the index i is spanning along all words in the vocabulary. Once the neural network has been trained, the average between both learned vectorial representations is taken as the final word representation. In order to increase the efficiency of the method, Mikolov, Chen, et al. (2013) propose a different version of the technique, the *negative sampling* method. In this variant, for each training example (w_k,w_j), the model is feed with a predefined number of words sampled from the vocabulary, as examples of words that did not appear in the context of w_j (for a more detailed explanation see (Jurafsky and Martin, 2014; Mikolov, Chen, et al., 2013)).

An intrinsic difference between LSA and Word2vec is that while LSA is a counter-based model, Word2vec is a prediction-based model. Although prediction-based models have strongly increased in popularity, it is not clear whether they outperform classical counter-based models (Baroni, Dinu, & Kruszewski, 2014; Levy & Goldberg, 2014; Levy, Goldberg, & Dagan, 2015).

In particular, Word2vec methods have a distinct advantage in handling large datasets, since they do not consume as much memory as some classic methods like LSA and, as part of the Big Data revolution, Word2vec has been trained with large datasets of about billions of tokens. However, only a few studies analyze semantic representations of small corpora, such as the typical dream collection. In a recent study, Asr et al. (2016) show that a co-occurrence model, like LSA, outperforms Skip-gram model in a semantic classification task over a medium size corpus (8 million words). In the same line, Sahlgren and Lenci (2016) compare the performance

of different embeddings techniques in several semantic task when the corpus size is varied. They showed that the Skip-gram model outperforms other embeddings in the case of large corpus size (1 billion words), while in the case of a smaller corpus size (1 million words) LSA beats other models. Since this was a large scale analysis, the authors avoided the parameter optimization, which may be relevant as it is known that larger corpora require more dimensions for their optimal vectorial representation (Fernandes, Artífice, & Fonseca, 2011).

In a first experiment, we will revisit the optimality of the methods to achieve reliable semantic mappings at different corpus size, varying the number of the embeddings dimensions. Then, we will test the capabilities of LSA and Skip-gram models to identify patterns in the usage of words among dreams reports. In particular, we set out to analyze the semantic neighborhood of the word *run* present in different collection of dream reports. We chose this word because of its high frequency in dreams, and the great variety of contexts in which it can be used. For example, *run* may be associated to sports activities or with the chase/escape situation, which is reported to be one of the most typical dreams (Griffith, Miyagi, & Tago, 1958; Malcolm-Smith et al., 2012; Nielsen et al., 2003) and is a central element in the Threat Simulation Theory (Revonsuo, 2000; Valli et al., 2005).

The final goal of this work is to analyze the capabilities of embeddings models to identify word associations in dreams reports. It is worth noting that training the embeddings in auxiliary corpus and using context vectors could be an alternative approach (Sagi et al., 2013), nevertheless it could provide spurious relations coming from the auxiliary corpus.

2. Methods

2.1. Semantic representations

Both, LSA and Skip-gram semantic representations were generated with the Gensim Python library (Rehuek & Sojka, 2010). In LSA implementation, a tf-idf transformation was applied before the truncated Singular Value Decomposition. The number of dimensions of LSA's representation was varied among a wide range of values (7, 15, 25, 50, 100, 200, 400), selecting the dimensionality that produces the best performance in each experiment. In Skip-gram implementations no minimum frequency threshold was used, and the window size (noted as *win*) and the number of negative samples (noted as *neg*) were varied among different values to analyze their dependency and select the best option. All other Skip-gram parameters were set to default Gensim values.

Given a vectorial representation, the semantic similarity (S) of two words was calculated using the cosine similarity measure between their respective vectorial representation ($\mathbf{v_1}$,$\mathbf{v_2}$),

$$S(\mathbf{v_1},\mathbf{v_2}) = \cos(\mathbf{v_1},\mathbf{v_2}) = \frac{\mathbf{v_1}.\ \mathbf{v_2}}{\|\mathbf{v_1}\|.\ \|\mathbf{v_2}\|} \tag{3}$$

where $\|\mathbf{v_i}\|$ refers to the euclidean norm of the vector $\mathbf{v_i}$. The semantic distances between two words $d(\mathbf{v_1},\mathbf{v_2})$ was calculated as 1 minus the semantic similarity ($d(\mathbf{v_1},\mathbf{v_2}) = 1-S(\mathbf{v_1},\mathbf{v_2})$).

2.2. Semantic tests

To compare LSA and Skip-gram semantic representation quality, we performed two tests: (1) a semantic categorization test and (2) a word-pairs similarity test. For each test, we studied how the performance of LSA and Skip-gram embeddings depend on the corpus size using two different corpora (TASA and UkWaC). To do this, we took 6 nested sub-samples of both training corpora, following the procedure described in Bullinaria and Levy (2007, 2012). In each case, we started with the whole corpus and we progressively discarded random documents to produce 6 sub-corpus with decreasing corpus size. We varied the corpora size with equally-spaced sizes in a logarithmic scale. After the documents reduction step, if any of the test words did not appear at least once in the sub-corpus, a random document was replaced with one of the discarded ones containing the missing word. The minimum sub-corpus size contained only 600 documents.

2.2.1. Semantic categorization test

In this test we measured the capabilities of the model to represent semantic categories (Bullinaria & Levy, 2007; Patel, Bullinaria, & Levy, 1997) (such as, drinks, countries, tools and clothes). The dataset is composed by 53 categories with 10 words each and it was used first by Patel et al. (1997). In order to measure how well the word i is grouped vis-à-vis the other words in its semantic category we used the Silhouette Coefficients, $s(i)$ (Rousseeuw, 1987),

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \tag{4}$$

where $a(i)$ is the mean distance of word i to all other words within the same category, and $b(i)$ is the minimum mean distance of word i to any words within another category (i.e. the mean distance to the neighboring category). In other words, Silhouette Coefficients measure how close a word is to other words within the same category compared to words of the closest category. The Silhouette Score is computed as the mean value of all Silhouette Coefficients. The score takes values between -1 and 1, higher values reporting localized categories with larger distances between categories, representing better clustering.

2.2.2. Word-pairs similarity test

This test measures the capabilities of the model to capture semantic similarity between concepts. We used the well established

WordSim353 test collection (Finkelstein et al., 2001), which consist of 353 word-pairs (such as Maradona-football or physics-chemistry) associated with a mean human-assigned similarity score. Each word-pair is rated on a scale ranging from 0 (highly dissimilar words) to 10 (highly similar words). The evaluation score is computed as the Spearman correlation between the human scores and the model semantic similarities.

2.3. Case study: Semantic association in dreams reports

In this case study, we analyzed the capabilities of the models to capture semantic word associations, testing whether the embedding models can capture the semantic neighborhood of a target word in a single dream collection (a collection of dream reports produced by the same person or by a certain group). In particular, we selected the word *run* as the target word, and we focused on the detection of its distance to escape/chase contexts. To calculate the distance of a given word "w" with respect to the target word *run*, we order all words in the vocabulary according their similarity to *run* in a decreasing order and we use the rank (position on the list) as a measure of the distance.

The rank distance of a given word "w" with respect to *run* was measured as the rank of "w" among the cosine similarity between *run* and all other words in the vocabulary.

For example, if a word has a rank of *20*, it means that, among all words in the vocabulary, it is the 20th closest word using a cosine similarity metric. Finally, we will define the rank distance of escape/chase concept as the minimum value within the rank of the words *escape*, *escapes*, *escaping*, *escaped*, *chase*, *chases*, *chasing* and *chased* with respect to *run*.

For each dream collection, two independent annotators from our team read all the dreams in which the word *run* appears, and labeled whether they refer to an escape/chase situation or not.

Escape/chase situations were defined as those in which (1) someone is being chased or is under the impression of being chased or (2) someone is escaping from a real or imaginary threat. Also, annotators do not count escape/chase situations associated to a clear positive emotional valence, thus discarding, for instance, escape/chase situations related to games or sports. With these criteria, for each dream collection the annotators calculate the fraction of times the word *run* appears in an escape/chase context. In order to show the annotators agreement in the fractions calculation, a Pearson correlation over dream series between the two annotators was computed, obtaining a correlation coefficient of 0.98. For each dream collection, we use the average over the fraction measured by each annotator as the ground truth, and we will refer to this value as the *escape/chase fraction*.

We used the *escape/chase fraction* as a ground truth to test the embeddings quality. Good representations should produce low rank distance in collections with high *escape/chase fraction* and high rank distance in collections with low *escape/chase fraction*. Thus, not only do we expect negative correlations between the escape/chase rank distance and the ground truth, but we also expect the differences in rank distances to be large when the models are trained with low and high *escape/chase fraction*.

In order to quantify these differences, we computed the linear regression of $\log_{10}(rank\ distance)$ vs the *escape/chase fraction*, and we used the log-linear slope as a measurement of performance. Thus, the more negative the slope is, the better the performance. It should be noted that in this analysis the collections in which the word *run* appears less than 5 times were excluded.

2.4. Corpora

In both tests, we used as training corpora the TASA corpus (Zeno, Ivens, Millard, & Duvvuri, 1995) and a random subsample of ukWaC corpus (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). The TASA corpus is a commonly used linguistic corpus consisting of 37k educational texts with a corpus size of 5M words in its cleaned form. UkWaC consists of web pages material from .uk domain. The random subsample has 140k documents with a corpus size of 57M words in its cleaned form. This corpus was used in Sahlgren and Lenci (2016) work.

For the case study we used the Dreambank reports corpus (Domhoff & Schneider, 2008b; Schneider & Domhoff, 2016). The DreamBank corpus consists of 58 collections of dream reports in English, containing about 19k dreams with a total of 1.3M words in its cleaned form (the names of the 58 collections where detailed in the appendix).

To clean the corpora, we performed a word tokenization, discarding punctuation marks and symbols. Then, we transformed each word to lowercase and eliminated stopwords, using the stoplist in the NLTK Python package (Bird, Klein, & Loper, 2009). Also, all numbers were replaced with the string "NUM".

3. Results

3.1. Corpus size analysis in semantic test

As a first step for all analyses, we carried out the Skip-gram parameter optimization for both tests (see Table 1). The parameters that produces the best performance, in each test and corpus, were selected to perform the corpus size analysis. In the semantic categorization test, in the case of TASA corpus, the negative sampling parameter was chosen as 15 ($neg = 15$) given that it showed a slightly better performance.

To compare LSA and Skip-gram embeddings quality in different size corpora, we tested both methods in random nested subsamples of TASA and ukWaC corpus (see Fig. 1). Given that the appropriate embeddings dimensions depends on the corpus size (Fernandes et al., 2011), for each sub-corpus, we ran the models with a wide range of dimension values (7, 15, 25, 50, 100, 200, 400), using in each case the dimension that produces the best performance.

Table 1
Skip-gram's parameter selection. Silhouette scores for the categorization test and correlations for the WordSim353 test. In all cases the embedding dimensions were set to 100 and the number of negative samples (*neg*) and the window size (*win*) were varied. Best scores are shown in bold for each test and corpus.

	win\neg	5	10	15
Silhouette score				
TASA	5	0.107	0.107	0.109
	10	0.110	0.117	0.119
	15	0.115	**0.121**	**0.121**
ukWaC	5	0.150	0.151	**0.155**
	10	0.146	0.149	0.151
	15	0.141	0.145	0.145
Correlation				
TASA	5	0.603	0.592	0.589
	10	0.615	0.610	0.602
	15	0.623	0.618	**0.626**
ukWaC	5	0.643	0.633	0.638
	10	0.644	0.643	0.637
	15	**0.647**	0.640	0.642

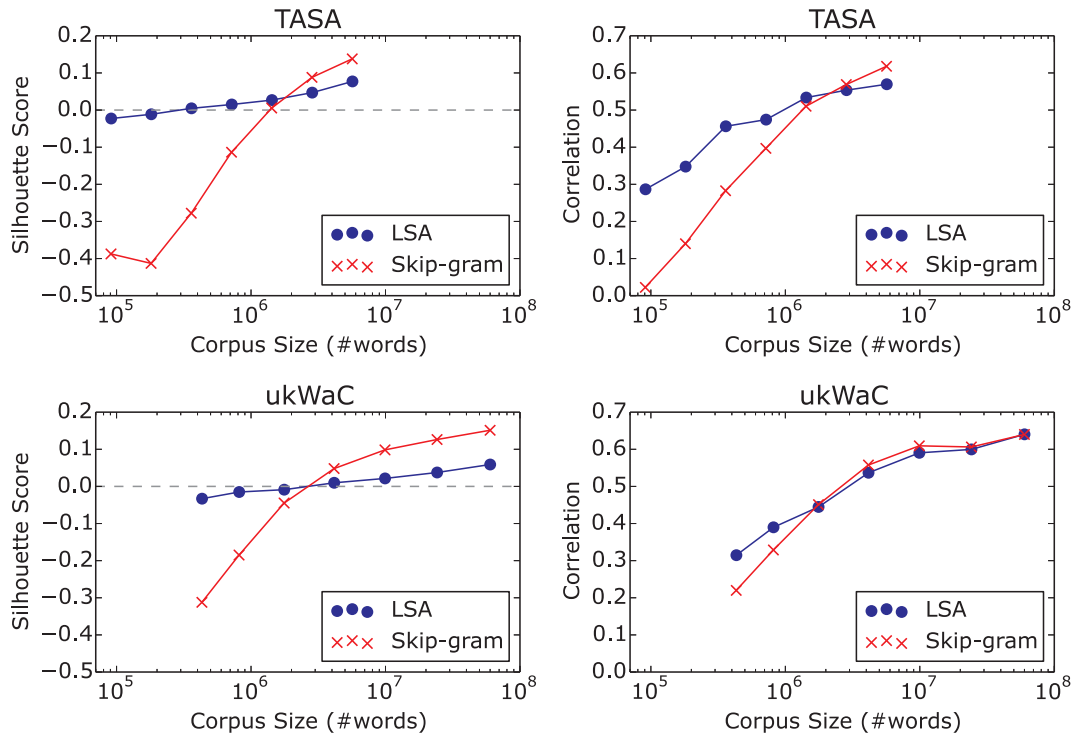

Fig. 1. Semantic categorization test performance (Silhuouette score), in left graphs, and WordSim353 test performance (correlation), in right graphs, in function of the corpus size for LSA and Skip-gram model. The size of the different corpus are considered as the number of tokens that they contain.

Fig. 1 shows that Skip-gram word-knowledge acquisition rate tends to be larger than LSA's. While Skip-gram tends to produce better embeddings than LSA when they are trained with larger corpora, under training with smaller corpora Skip-gram performance is considerably lower than LSA's. In accordance with Sahlgren and Lenci (2016) results, the threshold in corpus size below which LSA outperform Skip-gram is around the million of words.

3.2. Case study: semantic association in dreams report

In order to check the expected differences between the associations of the word *run* in dreams and waking life, we built LSA and Skip-gram embeddings trained each corpora, and we extracted the 25 words most similar to *run* in each case (see Table 2). Infrequent words which appear less than 15 times were excluded. As is expected we found that word embeddings are capable of identifying

Table 2
Most similar words to *run* for LSA and Skip-gram (S-G) embeddings trained in Dreambank, TASA and UkWac corpus.

Dreambank	
LSA	chase, running, scream, chasing, escape, runs, chases, grab, screaming, nazi, hide, chased, yells, safety, wolf, devil, stairwell, evil, away, attacking, killing, slam, yell, crouch, marathon
S-G	running, escape, catch, chase, chasing, follow, ran, sight, coming, runs, dangerous, guards, robbers, hide, toward, hiding, escaped, safely, firemen, fence, chute, safe, shoot, protect, tornado
TASA	
LSA	running, runs, ran, go, operate, organise, compete, start, break, install, operated, gone, move, set, managed, jump, walk, organised, connect, pull, perform, rowing, ride, put, operating
S-G	drive, ride, running, stay, go, haul, walk, jump, throw, staying, get, carry, move, stop, cut, driving, send, pass, climb, reach, steer, travel, runs, pull, take
UkWac	
LSA	running, runs, marathon, bash, start, rlogin, runners, starts, jump, loaded, weekend, vms, startx, marathons, mkdir, clocking, program, syslog, redhat, novice, filesystem, executable, startup, stall, nfs
S-G	running, dash, jumping, jump, yell, workouts, kick, jogging, workout, stretch, tiring, fun, fast, throw, repetitions, calisthenics, stretching, runs, overload, runner, exercises, yelled, cycling, runners, bicycling

differences in usage patterns of word between dreams and waking life. In TASA and ukWaC corpora, *run* is linked with words associated with a big variety of contexts, such as sports, means of transport and programming, while in dreams, *run* is directly related with words associated with chase/escape situations. These over-representation of chase/escape situations in dreams content is consistent with the *Threat Simulation Theory*, which propose that dream production mechanism evolve to simulate threatening events (Revonsuo, 2000; Valli et al., 2005).

Then, we tested the ability of both models to extract semantic tendencies in single dreams sets following the method described in Section 2.3. The sensitivity of the method to detect escape/chase situation in the context of the term *run* rely on the steepness of the slope between the escape/chase rank distance and the escape/chase fraction. Good models will produce steep negative slopes (see Methods section for details). A parameter selection was made, obtaining the best performance for LSA in 200 dimensions and for Skip-gram in win = 15 and neg = 10 (Tables 3 and 4).

While both models present a downward trend, the LSA outperforms Skip-gram, with a negative log-linear slope of -1.99 ± 0.06 with LSA and -1.12 ± 0.06 with Skip-gram. The distribution of slopes obtained with the 10 repetitions with each methods show a significant difference, with a *p-value* $< 3 \times 10^{-4}$, in a Kolmogorov-Smirnoff test. In Fig. 2 we plot the calculated distance vs the ground truth for each dreams collection in the selected parameters.

In order to test to what extent we can use these methods to explore the usage pattern of a target word in individual dream collections, we compare two groups of dreams collections: (1) dreams collections where *run* is highly associated with escape/chase situation (escape/chase fraction >0.5) (2) dreams collections where *run* is weakly associated with escape/chase situation (escape/chase fraction <0.1). Under these conditions both groups has 8 dream collections. In Fig. 3 we compare the rank distance distributions between these groups with both models, in a logarithmic scale. Both, LSA and Skip-gram models show significant difference between the median rank distance of the groups with *p-values* of 0.0033 and 0.0046, and a statistic of 8.65 and 8.04 respectively in a Kruskal-Wallis test.

Additionally, we show in Table 5 the 25 closest words of *run* in 3 different dreams collections, using the same parameter set as in Fig. 2. Collections 1 and 2, are the two collections with the highest escape/chase fraction, while collection 3 has no escape/chase situations in dreams that contain the word *run*. In the first two collections, we observe that *run* neighborhood in LSA embedding contains words highly related with escape/chase situations, such as *chased* and *hide* in collection 1 and *chasing* and *chases* in collection 2. Conversely, Skip-gram embeddings do not succeed in identifying escape/chase contexts in these dream collections. As a control case, it can be seen that collection 3 do not show escape/chase related words.

3.3. Conclusions

Psychology has an early interest in the study of dream contents, addressing questions such as “what do we dream about?” and “how our cultural background and waking life experiences shape our dream contents? (Bell & Hall, 2011; Domhoff, 2000, 2002; Domhoff & Schneider, 2008b). The *Threat Simulation Theory* describes the function of dreaming in terms of a evolutionarily selected mechanism, which provides a *world simulation* where we can train responses to threatening experiences (Revonsuo, 2000). Current

Table 3
Skip-gram’s parameter selection in the dream reports analysis. The scores are the slopes in the log-linear regression of the escape/chase rank distance vs escape/chase fraction. The shown values are the mean score among 10 repetitions. The embedding dimensions were set to 50. The best score is shown in bold.

win\neg	5	10	15
5	−0.63	−0.96	−0.99
10	−0.95	−1.06	−1.01
15	−0.95	**−1.12**	−1.02

Table 4
LSA's parameter selection in the dream reports analysis. The scores are the slopes in the log-linear regression of the escape/chase rank distance vs escape/chase fraction. The shown values are the mean score among 10 repetitions. The best score is shown in bold.

dim	30	50	100	200	300	400
Slope	−1.65	−1.67	−1.96	**−1.99**	−1.73	−1.77

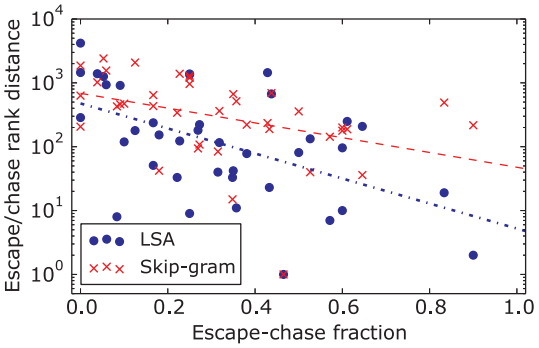

Fig. 2. Escape/chase rank distance vs the escape/chase fraction for each individual dream collection. A log-linear regression was performed for one sample of LSA and Skip-gram models (blue dash-dotted line and red dashed line respectively). LSA measurements present a log-linear slope of −2.10, while the Skip-gram model has a slope of −1.12. Also, the LSA measures of log(rank distances) show a correlation of −0.57 with the escape/chase fraction while the Skip-gram measures of log(rank distances) show a correlation of −0.42. The correlations show a *p-value* of 0.0001 and 0.007, for LSA and Skip-gram measures respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

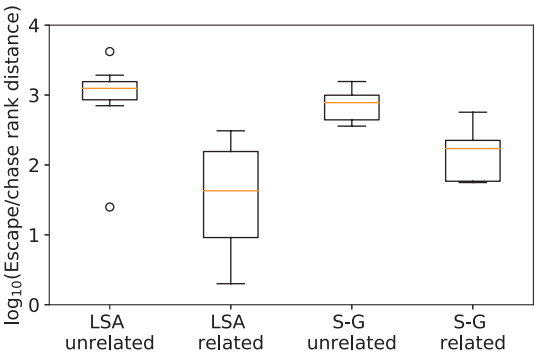

Fig. 3. Box plot comparing between the log(rank distance) distributions of dreams collections where *run* is highly associated with escape/chase situation (related) and dreams collections where *run* is not associated with escape/chase context (unrelated). This comparison was made training the collections with both, LSA and Skip-gram (S-G) models. In the box plot, the bottom and top of the boxes are the first and third quartiles, the orange line inside the box is the median, and the whiskers show the lowest/highest value that lies within an extension of the box of 1.5 times the interquartile range. Values outside the whiskers are plotted as outliers with a dot. Each group consists of 8 dream collections, with a mean and standard deviation of the log(rank distance) distribution of 2.93 ± 0.62,1.55 ± 0.75,2.86 ± 0.22 and 2.17 ± 0.37, for LSA-unrelated, LSA-related, SG-unrelated and SG-related respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

computational methods are based on frequency word-counting for dream content analysis (Bulkeley, 2009; Domhoff & Schneider, 2008b). This approach is focused on the salience of words without identifying the context in which they appear, which can be determinant to capture words associations and semantic ambiguity due to polysemy.

To fill this gap, we compared the LSA and Skip-gram capabilities to extract semantic word associations in dream reports collections. Given that semantic word representations have not been extensively tested and tuned in small dataset such as dreams collections, we started analyzing the models performance and parameter dependencies in two semantic tests when they were trained with small datasets. In accordance with Sahlgren and Lenci (2016) we found that LSA outperforms the Skip-gram model when they are trained in corpora smaller than 1 million words.

Then, to test whether word embeddings are capable to identify differences in usage patterns of words between dreams and waking life, we compared the differences in the semantic neighborhood of the word *run* when word embeddings are trained with different corpus. We found that in waking life corpora *run* is linked with words associated with a big variety of contexts, such as sports, means of transport and programming, while in dreams, *run* is directly related with words associated with threatening events and chase/escape situations. This result is consistent with the *Threat Simulation Theory*, which propose that dream production mechanism evolve to simulate threatening events (Revonsuo, 2000; Valli et al., 2005).

Table 5
Semantic neighborhood of the word *run* in 3 different dreams collections for LSA and Skip-gram's (S-G) word embeddings. Collections 1 and 2 are the two collections with the highest escape/chase fraction, while collection 3 is a control collection, which has no escape/chase situations in dreams that contain the word *run*. Collection 1 is the "*Seventh Grade girls*" collection, in which only 5 of its 69 dream reports contain the word *run* and on average 90% of these dreams refer to chase/escape situations (escape/chase fraction of 0.9). Set 2 is the "*Bay Area girls: Grades 7–9*" collection, in which 6 of its 154 dream reports contain the word *run*, 83% of which refer to a chase/escape situation (escape/chase fraction of 0.833). Set 3 is the "*Madeline3: Off-Campus*" collection, in which 13 of its 348 dream reports contain the word *run* and none of them refers to a chase/escape situation (escape/chase fraction of 0).

Serie 1 ↑chase/esc.	
LSA	metal, **chased**, **hide**, suitcases, expensive, climb, fashioned, diamond, worked, apartment, stairs, playground, janitor, jeans, dusty, led, rooms, things, empty, shirts, wake, building, everywhere, running, open
S-G	get, turned, color, wake, another, started, saw, like, left, away, time, window, black, behind, dad, stopped, walking, car, older, moved, right, thought, hi, back, would
Serie 2 ↑chase/esc.	
LSA	footsteps, alleyway, move, nervous, thinking, trapped, feet, behind, satin, trash, jump, turns, sits, winds, forth, alley, anymore, yells, **chases**, farther, drill, fights, stuck, **chasing**, heard
S-G	color, mouth, wake, watched, whitney, moved, supposed, left, recognize, eventually, black, turned, liked, drove, staff, except, windy, stopped, brian, cookies, pretty, sandwiches, fun, older, principal
Serie 3 ↓chase/esc.	
LSA	loop, surfboards, thumb, flip, laugh, motorcycle, matte, chotto, kudasai, corrected, polite, squealed, potbellied, petted, wild, paintball, written, housemates, thin, pinky, tried, foam, stuck, follow, throwing
S-G	color, afraid, reason, zebra, day, city, turned, watched, realizing, closer, food, mouth, course, main, moving, milk, last, riding, moved, safe, ready, rip, kept, town, anyway

Finally, we test word embeddings capabilities to identify word associations in dream reports collections. In particular, we test whether these tools were able to capture accurately, in different dream reports collections, the semantic neighborhood of the word *run*.

We found that LSA can effectively differentiate different word usage pattern even in cases of collections with low number of dreams and low frequency of target words. This is a step forward in the application of word embeddings to the analysis of dream content. We propose that LSA can be used to explore word associations in dreams reports, which could bring new insight into this classic field of psychological research. On one hand, the validation of semantic metrics to analyze word associations in dream reports promises a much more accurate quantification of socially-shared meaning in dream reports, with great potential application in psychiatric diagnosis (Mota et al., 2014, 2017), dreams theories testing (Revonsuo & Valli, 2008; Revonsuo et al., 2015) and dream decoding research (Horikawa & Kamitani, 2017; Horikawa, Tamaki, Miyawaki, & Kamitani, 2013).

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

We want to thank the teams behind the TASA (Zeno et al., 1995), WaCky (Baroni et al., 2009) and Dreambank (Domhoff & Schneider, 2008b) projects for providing us the corpora. This research was supported by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad de Buenos Aires, and Agencia Nacional de Promoción Científica y Tecnológica.

Appendix A

The Dreambank dataset contains the following dream collections:

Alta, Angie, Arlie, Barb Sanders, Barb Sanders2, Bay Area girls: Grades 4–6, Bay Area girls: Grades 7–9, Blind dreamers (F), Blind dreamers (M), Bosnak, Chris, Chuck, Dahlia, David, Dorothea, Ed, Edna, Emma, Emmas husband, Esther, Hall female, Jeff, Joan, Kenneth, Mack, Madeline 1, Madeline 2, Madeline 3, Madeline 4, Mark, Melissa, Melora, Melvin, Merri, miami-home, miami-lab, Midwest teenagers (F), Midwest teenagers (M), Nancy, Natural Scientist, Norman, Norms (F), Norms (M), Pegasus, Peruvian women, Peruvian men, Phil 1, Phil 2, Phil 3, Physiologist, Ringo, Samantha, Seventh Graders, Toby, Tom, UCSC women, Vickie, West Coast teens.

For a detailed description of the dream collections see http://dreambank.net/.

References

Altszyler, E., & Brusco, P. (2015). Análisis de la dinámica del contenido semántico de textos. In *Argentine symposium on artificial intelligence (ASAI 2015)-JAIIO 44 (Rosario, 2015)*.

Asr, F. T., Willits, J. A., & Jones, M. N. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In *Proceedings of CogSci*.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language*

Resources and Evaluation, 43, 209–226.
Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 238–247). http://dx.doi.org/10.3115/v1/P14-1023.
Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., ... Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia, 1*. http://dx.doi.org/10.1038/npjschz.2015.30 <http://www.nature.com/articles/npjschz201530> .
Bell, A. P., & Hall, C. S. (2011). *The personality of a child molester: An analysis of dreams.* Transaction Publishers.
Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python.* O'Reilly Media, Inc.
Bulkeley, K. (2009). Seeking patterns in dream content: A systematic approach to word searches. *Consciousness and Cognition, 18*, 905–916.
Bulkeley, K. (2014). Digital dream analysis: A revised method. *Consciousness and Cognition, 29*, 159–170.
Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*, 510–526. http://dx.doi.org/10.3758/s13428-011-0183-8.
Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and svd. *Behavior Research Methods, 44*, 890–907.
Carrillo, F., Cecchi, G. A., Sigman, M., & Analysis, L. S. (2015). Fast distributed dynamics of semantic networks via social media (Supplementary Material). http://dx.doi.org/10.1155/2015/712835.
Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Architecture, 20*, 160–167. http://dx.doi.org/10.1145/1390156.1390177 <http://portal.acm.org/citation.cfm?id=1390177> .
Deerwester, S., Dumais, S. T., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *JAsIs,* 41.
Diuk, C. G., Slezak, D. F., Raskovsky, I., Sigman, M., & Cecchi, G.a. (2012). A quantitative philology of introspection. *Frontiers in Integrative Neuroscience, 6*, 1–12. http://dx.doi.org/10.3389/fnint.2012.00080.
Domhoff, G. W. (2000). Methods and measures for the study of dream content. *Principles and Practices of Sleep Medicine, 3*, 463–471.
Domhoff, G. W. (2002). *Using content analysis to study dreams: Applications and implications for the humanities*. New York: Palgrave.: Bulkeley (Ed.).
Domhoff, G. W., & Schneider, A. (2008a). Similarities and differences in dream content at the cross-cultural, gender, and individual levels. *Consciousness and Cognition, 17*, 1257–1265.
Domhoff, G. W., & Schneider, A. (2008b). Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition, 17*, 1238–1247. http://dx.doi.org/10.1016/j.concog.2008.06.010.
Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers, 23*, 229–236. http://dx.doi.org/10.3758/BF03203370.
Elias Costa, M., Bonomo, F., & Sigman, M. (2009). Scale-invariant transition probabilities in free word association trajectories. *Frontiers in Integrative Neuroscience, 3*, 19.
Fernandes, J., Artífice, A., & Fonseca, M. J. (2011). Automatic estimation of the lsa dimension. In *KDIR* (pp. 309–313).
Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web* (pp. 406–414). ACM.
Freud, S. (1900). The Interpretation of Dreams, SE 4-5.
Galvez, R. H., & Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science, 19*, 43–56. http://dx.doi.org/10.1016/j.jocs.2017.01.001 <http://www.sciencedirect.com/science/article/pii/S1877750317300091> .
Griffith, R., Miyagi, O., & Tago, A. (1958). The universality of typical dreams: Japanese vs. Americans. *American Anthropologist, 60*, 1173–1179. http://dx.doi.org/10.1525/aa.1958.60.6.02a00110.
Harris, Z. (1954). Distributional structure. *Word, 23*, 146–162.
Horikawa, T., & Kamitani, Y. (2017). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience,* 11.
Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science, 340*, 639–642.
Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A., & McNamara, D. (2007). Strengths, limitations, and extensions of LSA. In *Handbook of latent semantic analysis* (pp. 401–426). http://dx.doi.org/10.1164/rccm.201012-2079ED. < http://141.225.42.135:8080/Wpal/CSEP/pdf/xhuChapOct6.pdf > .
Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing, Vol. 3*. Pearson.
Kirschner, N. T. (1999). Medication and dreams: Changes in dream content after drug treatment. *Dreaming, 9*, 195.
Kraepelin, E. (1906). *Über Sprachstörungen im Traume*. Engelmann.
Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on World Wide Web (WWW '15)* (pp. 625–635). http://dx.doi.org/10.1145/2736277.2741627. <http://dl.acm.org/citation.cfm?id=2741627> . Available from 1411.3315.
Landauer, T. K. (2007). Lsa as a theory of meaning. In *Handbook of latent semantic analysis* (pp. 3–34).
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240. http://dx.doi.org/10.1037/0033-295X.104.2.211.
Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems (NIPS)* (pp. 2177–2185). < http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization > .
Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211–225 <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570> .
Malcolm-Smith, S., Koopowitz, S., Pantelis, E., & Solms, M. (2012). Approach/avoidance in dreams. *Consciousness and Cognition, 21*, 408–412.
Malcolm-Smith, S., Solms, M., Turnbull, O., & Tredoux, C. (2008). Threat in dreams: An adaptation? *Consciousness and Cognition, 17*, 1281–1291.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In *Nips* (pp. 1–9). http://dx.doi.org/10.1162/jmlr.2003.3.4-5.951. Available from 1310.4546.
Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013b). Efficient estimation of word representations in vector space. In *Proceedings of the international conference on learning representations (ICLR 2013)* (pp. 1–12). http://dx.doi.org/10.1162/153244303322533223. <http://arxiv.org/pdf/1301.3781v3.pdf>. Available from 1301.3781v3.
Mota, N. B., Copelli, M., & Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia, 3*, 18.
Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., & Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Scientific Reports, 4*, 3691.
Nielsen, T. A., Zadra, A. L., Simard, V., Saucier, S., Stenstrom, P., Smith, C., & Kuiken, D. (2003). The typical dreams of Canadian University students. *Dreaming, 13*, 211–234. http://dx.doi.org/10.1023/B:DREM.0000003144.40929.0b.
Patel, M., Bullinaria, J. A., & Levy, J. P. (1997). Extracting semantic representations from large text corpora. In *Proceedings of the 4th neural computation and psychology workshop* (pp. 199–212).
Rehuek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta, Malta: ELRA.
Revonsuo, A. (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences, 23*, 877–901.
Revonsuo, A., Tuominen, J., & Valli, K. (2015). The simulation theories of dreaming: How to make theoretical progress in dream science. In *Open MIND*. Open MIND. Frankfurt am Main: MIND Group.
Revonsuo, A., & Valli, K. (2008). How to test the threat-simulation theory. *Consciousness and Cognition, 17*, 1292–1296.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7 <http://www.sciencedirect.com/science/article/pii/0377042787901257> .

Sagi, E., Diermeier, D., & Kaufmann, S. (2013). Identifying issue frames in text. *PLoS ONE, 8*, 1–9. http://dx.doi.org/10.1371/journal.pone.0069185.
Sahlgren, M., & Lenci, A. (2016). The effects of data size and frequency range on distributional semantic models. *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 975–980). ACL.
Schneider, A., & Domhoff, G. W. (2016). Dreambank. < http://www.dreambank.net/ > Last accessed: Sep. 12, 2016.
Sigman, M., & Cecchi, G. A. (2002). Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences, 99*, 1742–1747.
Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201–1211).
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.
Valli, K. (2011). Dreaming in the multilevel framework. *Consciousness and Cognition, 20*, 1084–1090.
Valli, K., Revonsuo, A., Pälkäs, O., Ismail, K. H., Ali, K. J., & Punamäki, R.-L. (2005). The threat simulation theory of the evolutionary function of dreaming: Evidence from dreams of traumatized children. *Consciousness and Cognition, 14*, 188–218.
Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster.