Philosophy & Technology (2021) 34:149–214
https://doi.org/10.1007/s13347-020-00393-9

RESEARCH ARTICLE

# Why Can Computers Understand Natural Language?

## The Structuralist Image of Language Behind Word Embeddings

**Juan Luis Gastaldi**[1]

Received: 16 November 2018 / Accepted: 15 January 2020 / Published online 14 May 2020


**Abstract**

The present paper intends to draw the conception of language implied in the technique of word embeddings that supported the recent development of deep neural network models in computational linguistics. After a preliminary presentation of the basic functioning of elementary artificial neural networks, we introduce the motivations and capabilities of word embeddings through one of its pioneering models, word2vec. To assess the remarkable results of the latter, we inspect the nature of its underlying mechanisms, which have been characterized as the implicit factorization of a word-context matrix. We then discuss the ordinary association of the "distributional hypothesis" with a "use theory of meaning," often justifying the theoretical basis of word embeddings, and contrast them to the theory of meaning stemming from those mechanisms through the lens of matrix models (such as vector space models and distributional semantic models). Finally, we trace back the principles of their possible consistency through Harris's original distributionalism up to the structuralist conception of language of Saussure and Hjelmslev. Other than giving access to the technical literature and state of the art in the field of natural language processing to non-specialist readers, the paper seeks to reveal the conceptual and philosophical stakes involved in the recent application of new neural network techniques to the computational treatment of language.

**Keywords** Word embeddings · Natural language processing · word2vec · Neural networks · Philosophy of language · Matrix models · Distributional hypothesis · Structuralism

✉ Juan Luis Gastaldi
juan.luis.gastaldi@gess.ethz.ch

[1] ETH Zürich (HPM, D-GESS), Zürich, Switzerland

*Il n'y a pas de "philosophie" du langage.*
*Il n'y a que la linguistique.*
Louis Hjelmslev
Principes de Grammaire Générale, 1928

## 1 Introduction

Whoever relies ever so little on the regular use of computational treatment of natural language, such as automatic translation, speech recognition, or spelling and grammar correction, cannot have failed to notice that, in the last few years, even the most elementary tools available to the general public have experienced a remarkable increase in performance. At the origin of this phenomenon is a significant technological shift brought about by the landing of the new generation of artificial neural network (NN) architectures in the field of computational linguistics or Natural Language Processing (NLP). Yet the rapid adoption of new (i.e., *deep*) NN techniques for the treatment of everyday language has more profound consequences than a backdoor technical improvement, a substantial gain in performance, and a handful of novel applications. The success of NN linguistic models brings to the surface a whole new *image of language*.

By image of language[1] we should understand a pre-theoretical and natural notion of what language is, often attributed more to common sense than to sophisticated conceptual elaboration, that acts as an implicit presupposition even for the most advanced scientific inquiry. To continue to paraphrase Deleuze's words (1994, p. 131), it is in terms of this image that everybody knows and is presumed to know what it means to speak or write.

To a large extent, over the past century, the image of language in most scientific and philosophical fields of research has been attached to what is usually referred to as "philosophy of language," a specific region of philosophical inquiry that tends to tie the problem of language and its meaning to those of representation, reference, truth, and reality.[2] In most of the cases, that image also shows us that language is the manifestation of the competence of individual subjects somehow rooted in their biological constitution—and in their brain as privileged biological substratum—as an instrumental component of their cognitive relation with their environment. In its pre-scientific version, that image is so widely established that it often pervades scientific research in multiple fields, by discretely indicating the places where the truth about language should be looked for. And even when scientific results from the specific field of linguistics happen to challenge any of the fundamental features of that image, it is most likely that the former rather than the latter will be viewed with suspicion.

Recent NN models for NLP are not the exception in this respect. Their results are nevertheless so appealing that suspicion in this case often turns into sheer awe. To

[1] I borrow this expression from Maniglier (2016, p. 359), who in turn takes inspiration from Deleuze's notion of "image of thought" (1994, ch. III).

[2] See for instance Hale and Wright (1997).

such an extent that, to the unavoidable question of if and how computers can "understand" natural language, it has become customary to speak about the "mysteries" and "magic" of NN models.[3] As if it were easier to attribute the grounds of scientific results to some obscure cause than to accept the challenge of critically reassessing our received image of language in such a way that its adequacy with the facts diminishes our state of bewilderment.

The following pages are an attempt to start depicting the image of language that those new computational models offer to us. As a way, of course, of contributing to the demystification of the magical power of new technologies, but also of acknowledging their full philosophical import. It will turn out that such image is not so unfamiliar to us after all: not only can most of the underlying mechanisms of NN models for NLP be explained by well-known techniques in computational linguistics, but an early outline of the image of language they convey can also be recognized in the structuralist background of the original distributional hypothesis, often invoked as justification in the development of the field in present days. Among the multiple aspects of those NN models, we will focus exclusively on *word embeddings*, a general representation of linguistic units through vectors in continuous space which became the standard input form for most NN architectures dealing with NLP tasks. After a preliminary presentation of the basic functioning of elementary NN models, we shall introduce the motivations and capabilities of word embeddings through one of its pioneering models, word2vec (Section 2), and, relying on existing literature, we will relate the mechanisms of word embeddings to already existing NLP models based on term-context matrix analysis (Section 3). We will then assess the image of language stemming from those mechanisms by contrasting it to the one usually conveyed by the association between the distributional hypothesis and a use theory of meaning (Section 4) and we will finally trace that image back up to the structuralist background of distributionalism (Section 5). We will conclude with a detailed summary of our inquiry and some final comments, to which the reader can refer to have an overall reconstruction of the arguments advanced in the next pages.

## 2 Word Embeddings: the Ground Layer of Natural Language

### 2.1 Preliminaries: Neural Network Models[4]

In their most elementary form, artificial NNs can be seen as ways of transforming a vector (i.e., a list of numbers) into another vector, through successive parametrized

[3]See for instance Christopher Manning and Richard Socher's tutorial "Deep Learning for Natural Language Processing (without Magic)" at https://nlp.stanford.edu/courses/NAACL2013/, implicitly denouncing such attitudes in the reception of field.

[4]The following presentation is deliberately concise and intends to give a background idea only. The reader already acquainted with NN architectures may skip this section. A complete presentation of deep NNs can be found in Goodfellow et al. (2016). See Schmidhuber (2015) for a historical overview of deep learning techniques in neural networks.

transformations,[5] each of which takes a vector as an input and yields a vector as an output that will be fed as the input vector of the next transformation. An input vector is thus fed into the network that will be transformed into another vector, which will be in turn transformed into a new one again and again until a final vector, which will be taken as the output of the procedure. Each one of those successive vectors is considered as a *layer*. The existence of multiple layers is what makes neural networks *deep*.

Insofar as different kinds of contents (in the form of images, sounds, texts, etc.) can be represented as lists of numbers (which is trivially the case if those contents exist in digital form), the hope is that, given a significant number of pairs of such vector representations corresponding to meaningful relations between the respective contents (e.g., an image and its textual description, a sound and the image of the object to which it corresponds, a written sentence and its translation into another language, etc.), the parameters of the transformations leading from one layer to the next can be so adjusted that, for any new input vector not included in the set of given pairs, the network produces a meaningful output vector bearing the same relation with the former as the pairs of that set. To that end, NNs are *trained* using the initial set of pairs of input-output vectors by means of an algorithm known as *backpropagation*.[6] At the beginning of the training, the parameters are initialized randomly, and the input vectors of the training set are fed into the network. The difference (also called *error* or *loss*) between the output vector thus computed by the network and the expected one given by the training set is then measured by a *loss function*, and such measure is used to progressively adjust the parameters in order to minimize the error, until minimization stabilizes, hopefully at a negligible level. Once the NN reaches such a state, it can be expected to successfully treat new unseen inputs.[7]

The unexpected efficacy attained by NN models in the last decades could, nonetheless, never overshadow a fundamental drawback of that method: since the adjustment of the parameters is performed automatically and the states of the model (i.e., the particular values of the intermediate vectors or *hidden layers*) are uninterpretable in principle, the reasons behind such efficacy are opaque at best. Hence, reliance on such methods within their multiple fields of application remains an open question. Moreover, from an epistemological viewpoint, the high efficacy levels of NNs contrast with the practically non-existent intelligibility they provide upon the phenomenon they are able to reproduce.

[5] Each one of those transformations consists typically of $n$ weighted sums $s_i = \sum w_{ij} x_j$ of the components $x_j$ of the input vector (where $1 \leq i \leq n$, $1 \leq j \leq m$, with $n$ the dimension of the output vector and $m$ the dimension of the input vector), plus a bias term $b_i$ such that $z_i = s_i + b_i$. A non-linear transformation $a_i = f(z_i)$ is computed on top of that, such as a sigmoid function which "squeezes" the result of the linear transformation $z$ between $-1$ and 1. If the weights $w_{ij}$ are expressed as a matrix $W$ of dimensions $n \times m$ and the biases $b_i$ as an $n$-dimensional vector, the entire transformation can be expressed as $a = f(Wx+b)$. Each layer of the network (i.e., each successive transformation) has a comparable form, and the parameters to be adjusted correspond to the weights collected in the matrices $W$ and the biases of the vectors $b$ of each layer.

[6] Implementing (different versions of) stochastic gradient descent.

[7] The whole system can thus be seen as a procedure to approximate any kind of function. For an accessible presentation of universal approximation theorems, see Nielsen (2015, ch. 4).

### 2.2 Word2vec: the Triumph of Word Vector Representations

Compared to other tasks, such as image or sound processing, deep NN models have been rather marginal among computational models for natural language until recent years. Early applications of deep NNs to natural language were mostly focused on specific tasks, and in speech recognition in particular, for which they could show some significant results,[8] while more general models found it difficult to compete with well-established statistical inference approaches, like *n-gram* models, in which a word is predicted by computing the conditional probability of that word given a fixed number of immediately preceding words.[9] However, during the 2000s, an alternative conception of the representation of words grew stronger among the community of NN NLP which would eventually lead to a reversal of that situation.[10]

The idea gradually emerging from those works was the following. If we think of an application of NN techniques to NLP, it is easy to see that in the vast majority of the cases, the input of the model takes the form of (one or several) linguistic units, typically words. In accordance with classic statistical language models, NN models would then represent those words as input vectors in the form of indexes over a given vocabulary. Thus, if the word "house" is the 542nd word of a vocabulary $V$ containing a total of 3 million words and phrases,[11] that word would be represented by a large 3-million-dimensional "one-hot" vector, i.e., composed of 0 everywhere, except for a 1 at the 542nd position:

$$v_{house} = \underbrace{(0, 0, 0, 0, 0, 0, 0, 0, \ldots, 0, \overbrace{1}^{\text{542nd position}}, 0, \ldots, 0, 0, 0, 0, 0, 0, 0, 0)}_{\text{3 million dimensions}}$$

Given this input, the output of the model depends, in principle, on the specific task the network is supposed to perform. For instance, in the case of automatic translation, this could be a one-hot vector (or several concatenated one-hot vectors) indicating the corresponding word (or words) in a given vocabulary of the target language; or in sentiment analysis, a one or two-dimensional vector indicating if the input is positive or negative. Finally, the loss function would be designed according to the task, and the network would be trained on relevant corpora.

However, researchers around NN models progressively realized that the layer resulting from the first transformation of the input vector (the "projection layer") had a special significance. Indeed, the vectors resulting from that transformation once the whole network was trained on one specific task could be used as input vectors

[8] See for instance Dahl et al. 2012.

[9] See Manning and Schütze 1999 for a detailed presentation.

[10] Cf. Bengio et al. 2003; Schwenk and luc Gauvain 2002; Blitzer et al. 2005; Mnih and Hinton 2007; Collobert and Weston 2008; Turian et al. 2010; see Bengio 2008 for an overview.

[11] As it is the case for the vocabulary of pre-trained word vectors available for download at the official word2vec website (Google inc. 2013). This vocabulary contains all the word forms seen in the corpus (for instance, "house" and "houses" are two different words of the vocabulary) as well as proper names and phrases such as "college grads", "geographically dispersed", "Volga river", or "Chief Executive Steve Ballmer". The vocabulary is ordered by the frequency of words and phrases in the training corpus. The models preceding word2vec were however of a much smaller scale.

for networks designed for other NLP tasks, with a significant increase in performance on the latter. In other terms, those transformed vectors could be considered as *generic representations of words* capturing some of their essential linguistic features, in opposition to the localist, atomistic and purely indexical one-hot representation suggested by traditional methods. On the basis of that remarkable fact, researchers figured out that one could then train those vectors independently of any specific task, and replace traditional indexical representations of words with those trained vectors for practically all NLP tasks.

Yet, if, as we have seen, tasks are essential to train NN models, since they provide the loss function that permits to adjust their parameters, how could we then train those new word vector representations *independently of any specific task*? In other terms, what could be the relation between input and output vectors that can orient the training of perfectly generic word representations? The answer was to base the training on *the relation between a word and its context words within a given corpus*.

The idea of a distributed representation of words was not new at that time, constituting, for instance, an essential part of the connectionist program laid out by Rumelhart and McClelland (1986) (cf. also Hinton 1986; Elman 1996). Elman (1990) had, in turn, proposed the idea of producing such representations using NN models.[12] The new advances in deep NN during the 2000s provided however a new general setting in which word vector representations could be thought, designed, and tested as generic representations for the vast majority of the tasks relevant to computational linguistics. Training such vector representations remained nonetheless computationally expensive and the high requirements in terms of training data made them unsuitable for real-life applications.

A turning point occurred in 2013, when Tomáš Mikolov and colleagues at Google,[13] released a set of pre-trained vectors and a software package distribution implementing two significantly efficient models, which have since been popularized under the name of their software package distribution: *word2vec* (Google inc. 2013). The models were introduced in a series of papers (Mikolov et al. 2013a, c, e) and extend the main idea that emerged from the work of the previous decade: in the *Skip-gram* model, a network is trained to predict the context words of a given focus or center word; conversely, the continuous bag of words (CBOW) model aims at predicting the center word given the set of context words around it.

Take, for instance, the Skip-gram model. One-hot vector representations are used for both the input center word chosen in the corpus, and the output context words surrounding it,[14] and only one intermediate or hidden layer is trained.[15] The input one-hot vector is then transformed through randomly initialized parameters into a

[12] The study of the connectionist image of language falls out the scope of the present paper. For an analysis convergent with the one we elaborate here, see Maniglier (2016).

[13] Google had officially adopted deep NN technology for speech recognition the year before (Jaitly et al. 2012).

[14] A window size is defined in advance, determining the number of words to be taken as context words immediately to the left and to the right of the chosen word.

[15] In this sense, the NN implementing the model is "shallow," i.e., not "deep."

low-dimensional vector[16] (the hidden layer), which is in turn similarly transformed into an output vector of as many dimensions as the size of the vocabulary, and finally normalized so as its components can be interpreted as a probability distribution over the vocabulary.[17] The error between this output vector and each one of the one-hot vectors corresponding to the context words is then used to adjust the parameters of the network through backpropagation. After that process is finished for one particular word with respect to its context, the network is fed with the next word in the corpus, trying to predict its own context, and the process is repeated with every word of the corpus until the minimization of the error reaches a stable state, starting over from the beginning of the corpus, if necessary, when the last word is attained.

After the training process is finished, the set of intermediate low-dimensional vectors corresponding to each of the input words provide the sought dense vector representations. Thus, while the word "house" was previously represented by a one-hot (very) large and sparse vector indexing its place in a vocabulary, now the same word will be represented by a dense low-dimensional vector given by the hidden layer of this network, whose first and last components, in the case of this particular model,[18] are as follows:

$$v_{house} = (\underbrace{0.157227, -0.0708008, 0.0539551, \ldots, -0.041748, 0.00982666, -0.00494385, -0.032959}_{300 \text{ dimensions}})$$

Although its main objective was to provide an efficient algorithm to train NN models for NLP, the success and popularity of word2vec[19] marked the triumph of distributed vector representations over traditional methods. But more importantly, the predominant position thus obtained by NN models of NLP within those two respective fields provided the opportunity for the development of novel perspectives concerning the relation between the formal objects inherent to NNs and the nature of linguistic objects. In other terms, as we will see in the next sections, the technical achievement of word2vec turned the use of NNs for the treatment of natural language into the unexpected occasion of mutual intelligibility.

### 2.3 What Can Word Embeddings Do? The Analogy Behind the Similarity

As any other vector, ($d$-dimensional) word vectors can be represented as points in (a $d$-dimensional) space. The coordinates of those points are given by the components of the corresponding vectors. While atomic representations are discrete (their vector components are either 0 or 1), the distributed vectors produced by NN models live

[16] Typically of dimension $d = 300$, also defined in advance.

[17] Standard normalization function is softmax. In the case of word2vec, more efficient versions are used, namely hierarchical softmax and negative sampling.

[18] As provided by the word2vec package (Google inc. 2013). It is important to bear in mind that models resulting from different training procedures as well as from different training corpora may differ substantially and that there is no unique linguistic model for a language.

[19] Relying significantly on the exceptional computational and data resources of a company such as Google, which were not available to most researchers at the time.

in continuous space (typically $\mathbb{R}^d$), since their components are real numbers (usually between $-1$ and 1). Moreover, index vectors are, by definition, orthogonal to each other (i.e., each corresponding point lies in a different dimension with respect to all the others). This implies that no relation whatsoever between the words they represent can be derived from their spatial configuration. In contrast, the density of the new vectors provides the means for their spatial "embedding" to capture general linguistic features associated with the relations between the words they represent. It is this mapping of words into real space that motivates the generic denomination "*word embeddings*" for dense word vector representations, in particular those produced by NN models.

Drawing from those spatial properties, word embeddings are able to grasp relevant linguistic features, starting with *word similarity*. Indeed, based on a distance measure over $\mathbb{R}^d$, such as the distance between two points (Euclidean distance), or the angle between the lines that connect those points to the origin point of the space (cosine distance), the relative distance between word vectors can be measured. Once those vectors have been fully trained on a raw corpus as explained, it appears that their relative distance in the embedding space correlates with the similarity or "relatedness" between the *contents* of the corresponding words. For instance, in the Skip-gram model referred to above, the 10 closest word vectors to the vector representation of the word "house" using cosine distance[20] are "houses" (cosine distance = 0.292761), "bungalow" (0.312144), "apartment" (0.3371), "bedroom" (0.350306), "townhouse" (0.361592), "residence" (0.380158), "mansion" (0.394181), "farmhouse" (0.414243), "duplex" (0.424206), "homes" (0.43802).

This is a remarkable fact, since in a classical NLP representation of words (such as index representation), every word is at the same distance from each of the others, and their relatedness must rely on features *external to the word representation itself*, such as annotations or handcrafted categorization. Word embeddings have thus somewhat internalized the similarity relation between words in their very representation. More generally, through relative distance, relevant groups of words can be identified, clustered by their content within that vector space, and different kinds of word categorization can take place based on those spatial clusterings (see, for instance, Senel et al. 2017). A common method to explore and visualize such clusters in a 300-dimensional space is to plot a projection of word vectors onto the most relevant directions of variation of the space.[21] Figure 1 presents one of such projections of a selection of word2vec vectors showing that words with similar content dispose in fairly identifiable clusters (e.g., numerals, number words, temporal nouns, but also pronouns, verbs, etc.).[22]

Relatedness and, in particular, semantic similarity between words play a decisive role in most NLP tasks, since it allows the model to rely on implicit content rather

[20] The cosine distance between the vectors $v$ and $u$ is defined as follows: cosine distance$(u, v) = 1 - \frac{v.u}{\|v\|\|u\|}$, where $x.y$ is the dot product between the vectors $x$ and $y$ and $\|\cdot\|$ is the norm of the vector.

[21] Computed through techniques such as principal component analysis (PCA) or T-distributed stochastic neighbor embedding (t-SNE).

[22] An online tool for visualizing such projections can be found at https://projector.tensorflow.org.

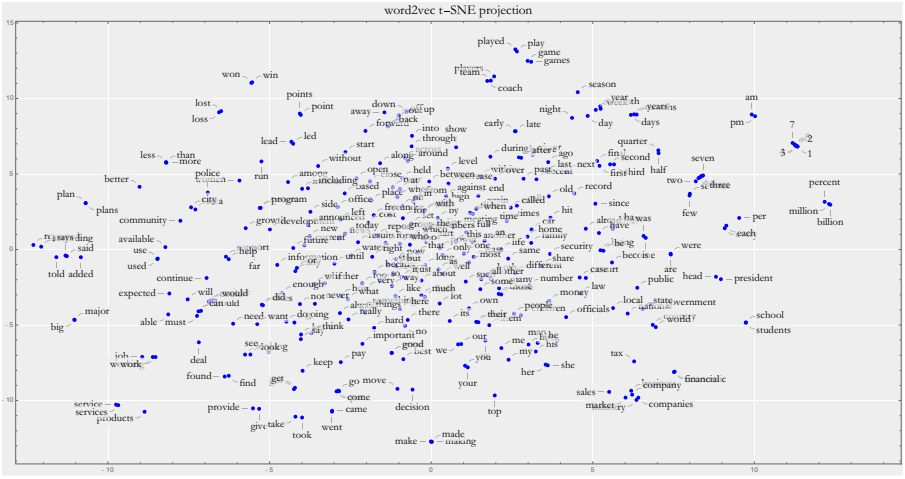


**Fig. 1** T-SNE two-dimensional projection of the word2vec vectors representing a selection of the most frequent words in its training corpus

than on rigid explicit expressions. From the point of view of information retrieval, machine translation, classification, or even part-of-speech tagging, it is crucial to know, for instance, that the word "house" is similar to the word "bungalow" (even if the corresponding written or spoken expressions are not), or the word "catch" is similar to the word "caught" (while "natch" is not similar to "naught"). More generally, capturing different aspects of similarity between words is evidence of the capability of the model to grasp the meaning of the fundamental building blocks of language. For this reason, word similarity has been regularly used as a privileged measure to assess the quality of NLP models in the past. Somewhat unexpectedly, predictive word embedding models such as word2vec were capable of attaining state-of-the-art levels at the moment of their introduction, and even outperforming previous probabilistic models in practically all similarity tasks, including semantic relatedness, synonym detection, and concept categorization (cf. Baroni et al. 2014; Schnabel et al. 2015; Levy et al. 2015). Those results are surprising not only because of the capacity of a brand new technique to surpass well-established ones, but also most of all because of the frugality of the means by which the new approach was able to do it. Indeed, while traditional models had reached high levels of performance only after several decades of careful design and fine-tuning over specific tasks, relying in most of the cases on external knowledge, handmade rules and supervised learning out of manually annotated corpora, NN word embeddings are merely based on a simple unsupervised predictive task requiring only word contexts automatically extracted from almost raw linguistic corpora with remarkable computational efficiency.

The capacity of models like word2vec to outperform existing models at traditional tasks was a proof of the fertility of distributed word vector representations and confirmed that NN models could indeed have a place in the landscape of computational linguistics. Still, this success could hardly be more surprising than one of the many technical improvements to which such a field is familiar with. Yet, there is a sense

in which word2vec was capable of going beyond its predecessors, including the NN tradition. Indeed, other than remarkable performance at already existing tasks, the newcomer was able to reveal a deeper linguistic significance of metric relations in the embedding space in what has come to be known as the *analogy* task.

The idea of analogy within this framework relies on the fact that, like any other vector, word vectors can be the object of arithmetical operations; in other terms, they can be added, subtracted, or multiplied. Of course, there is no reason, in principle, that arithmetic operations between vectors represent any linguistic relation whatsoever between the words those vectors represent (not to mention semantic relations); at least not any more than does the alphabetical order of words in a dictionary. However, Mikolov and his colleagues noticed that if, for instance, the word vector for "man" is subtracted from that of "king", and the result is then added to that of "woman", the closest vector to the resulting vector is the one representing the word "queen". In other terms, we have that

$$v_{king} - v_{man} + v_{woman} \approx v_{queen}$$

Or, what is practically the same

$$v_{king} - v_{queen} \approx v_{man} - v_{woman}$$

Far from being an isolated event, the same arithmetic relation holds for a great number of analogous cases. Thus, we also have that the vectors corresponding to the couples *prince : princess*, *landlord : landlady*, *bull : cow*, etc. bear approximately the same linear relation as those of *king : queen* and *man : woman*. Geometrically, such relation appears in the embedding space as a regular "offset" between pairs of words, in such a way that the common relation that holds between those two series of words (in this case, the relation between male and female, or gender) could be itself represented by a constant vector along the direction of that offset (Fig. 2).

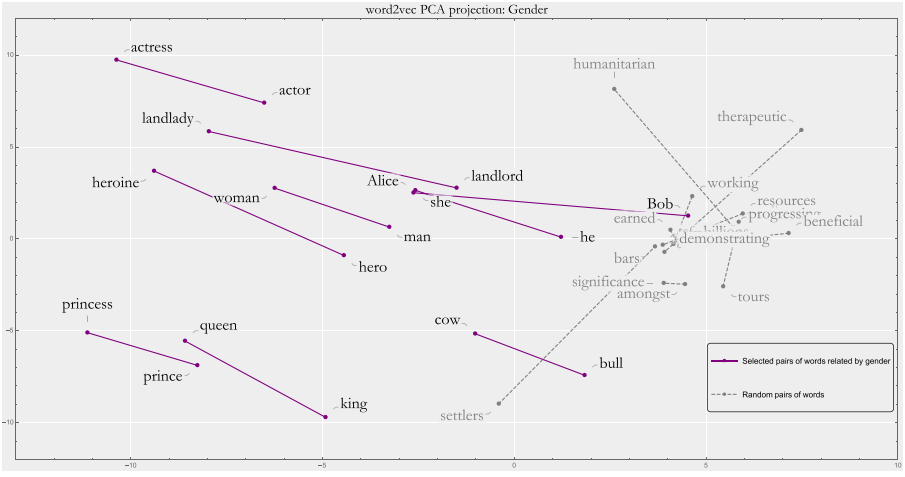


**Fig. 2** Offset representing the gender relation, revealed in the embedding space (word2vec) by a PCA projection

This surprising analogical property of word vectors has been shown to hold for a remarkable number of semantic relations. If we come back to our example of the vector representation of "house", we can find, for instance, that

$$v_{house} - v_{city} + v_{countryside} \approx v_{farmhouse}$$

Typical analogies attested in word2vec are of semantic nature, such as capital cities (*China : Beijing*, *Russia : Moscow*, *Poland : Warsaw*, ...), professions (*Einstein : scientist*, *Messi : midfielder*, *Picasso : painter*, ...), chemical nomenclature (*copper : Cu*, *gold : Au*, ...), company CEOs (*Microsoft : Balmer*, *Apple : Jobs*, ...) or typical national dishes (*Japan : Sushi*, *Germany : Bratwurst*, ...). However, without any modification of the training objective, the same vectors are capable of accounting for a large series of syntactic forms as well. Word2vec was thus able to grasp analogies between superlative forms (*big : biggest*, *good : best*, ...), verb tenses (*fall : fell*, *take : took*, ...), oppositions (*possibly : impossibly*, *ethical : unethical*, ...), plural nouns (*mouse : mice*, *dollar : dollars*, ...), and many others.[23]

The analogical relations in continuous vector space revealed by word2vec represented such a conceptual novelty for the field[24] that the corresponding analogy task became almost instantly a crucial internal evaluation method for all NLP models. Since then, the efficacy of the new NN word embeddings has been proved in most NLP tasks (cf. Pennington et al. 2014a; Baroni et al. 2014; Schnabel et al. 2015) and a great number of improved models directly based upon, built on top of or inspired by the mechanisms and results of word2vec have been proposed (Pennington et al. 2014a; Levy et al. 2015; Bojanowski et al. 2016; Nickel and Kiela 2017; Peters et al. 2018; Howard and Ruder 2018; Devlin et al. 2018).[25]

The intensive research the success of word embeddings motivated could not but push the original Skip-gram and CBOW models to progressive obsolescence. Yet, if we are more interested in the conceptual and philosophical consequences of the adoption of new techniques, rather than the mere improvement of efficiency standards, we can observe that by raising the conception of distributed vector representations to the state-of-the-art level, word2vec brought back to the forefront of linguistic thought the combined idea that the implicit organization of language can be mirrored not only by a space of local semantic similarities, but more fundamentally, by an *emergent*

[23] Details of the different kinds of semantic and syntactic analogy relations can be found in Mikolov et al. (2013c) and Schnabel et al. (2015).

[24] As pointed out in Schnabel et al. (2015), the analogy task was previously conceived as a classification problem, for instance in Turney (2008).

[25] Other than increasing the performance of linguistic tools, such a profusion of models brought about important analytical insights into the mechanisms of NLP and of NN models in particular. Incidentally, they revealed a fundamental fact of linguistic models that should not be overlooked: in spite of a relative equivalence or convergence on elementary results, *no two models are identical*. In the specific case of NN word embeddings, the non-deterministic nature of the training procedure can yield diverging models even with identical training parameters. Such a disparity is not only related to the irregular behavior of those models, but also to the deep problem of the unity and homogeneity of language itself. For this reason, the facts revealed by word embeddings which will interest us in the following pages should neither be ascribed to a single model, nor to the nature of a unified and homogeneous language, but to a common property of a family of related but differing models with respect to the possibility of general yet partial reconstructions of the underlying mechanisms of specific language practices.

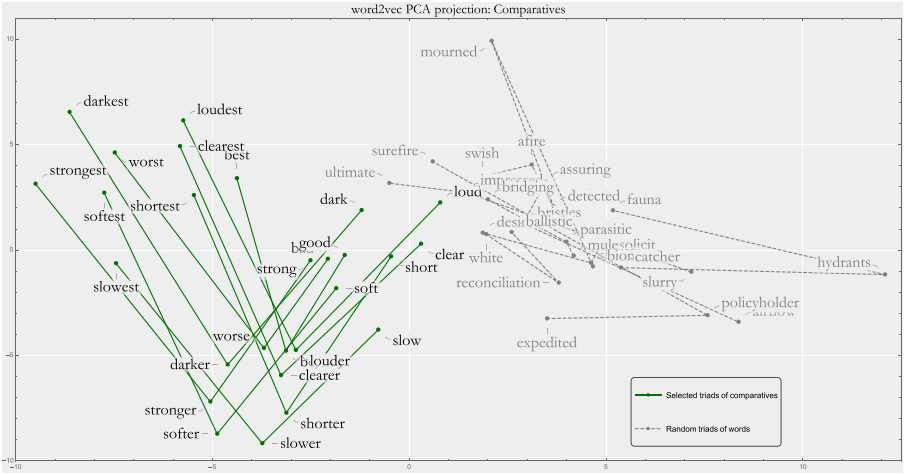


**Fig. 3** Pattern in the embedding space (word2vec) corresponding to the comparative category (base, comparative and superlative forms)

*structure* underlying that space constructed in a simple and unsupervised way. From a conceptual viewpoint, it is this idea that gives word embeddings their full significance and power.[26] Indeed, one only has to take a quick look at the present state of the field to realize that, let alone internal evaluation methods and constant increase in performance on traditional tasks, the most original, promising and stimulating results since the introduction of NN word embeddings concern global configurations within the embedding space rather than local applications of the similarity and analogy task as such. Starting with Mikolov et al. (2013c), who indicate that not only analogical relations but also compositional ones can take place as a linear relation in embedding spaces. For instance,

$$v_{Russia} + v_{river} \approx v_{VolgaRiver}$$

or

$$v_{Germany} + v_{capital} \approx v_{Berlin}.$$

Moreover, the work of Pennington et al. (2014a) suggests that the entire comparative category of English (i.e., the base or positive, the comparative and the superlative forms of adjectives) defines a pattern or even a two-dimensional subspace within the word vector space (see Fig. 3 for an illustration of this fact using a PCA projection of word2vec). Similarly, Mikolov et al. (2013d) suggested that the conjugation of verbs, even if irregular, also defines a coherent subspace (see Fig. 4 for an analogous illustration). Even more strikingly, several works (Mikolov et al. 2013b; Luong

[26]This idea is not entirely new. Accompanied by a significant increase in computational capabilities, the resurgence of empiricist, frequentist, corpus-based and emergentist trends in linguistics in the last decades has sufficiently shown the strengths of all kinds of data-driven approaches to language structure. For an overview of those different trends, see Elman (1996), Bybee and Hopper (2001), McEnery and Wilson (2001), MacWhinney (1999), and Chater et al. (2015). The novel means by which word2vec and alike models connect unsupervised treatment of corpora and derivation of a global implicit structure are however endowed with their own originality.

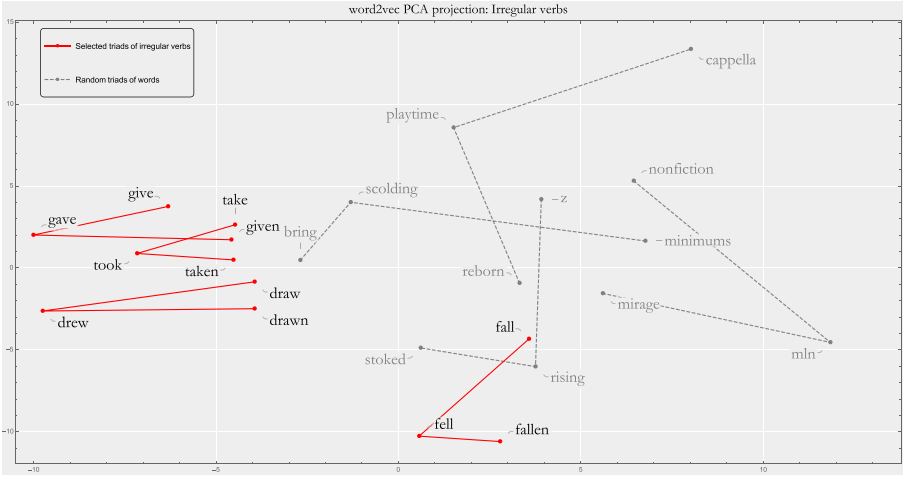


**Fig. 4** Pattern in the embedding space (word2vec) corresponding to conjugation of irregular verbs

et al. 2015; Jansen 2017) explored the possible correspondence between embedding spaces of different languages, suggesting a certain invariance with respect to relative positions of words across multiple word vector spaces (see Fig. 5). Moreover, recent works on semantic change (Kulkarni et al. 2014; Hamilton et al. 2016) were able to trace the alteration of meaning of words over the last centuries through the modification of the relative position of their vector representations in the embedding space with respect to corpora of different epochs (see Fig. 6). Finally, Hewitt and Manning (2019) provide evidence of the fact that syntax trees are encoded, modulo a linear transformation, in the embedding space produced by recent models.

Certainly, as thought-provoking as those research orientations might be, they should not make us underestimate the inherent difficulties of word vector representations to grasp the most intimate mechanism of natural language. Any attempt to extend the operations of word embeddings outside local, specific, and well-controlled conditions is likely to encounter innumerable obstacles, and hence can barely be relied upon in its present state for concrete real-world applications. Several critical

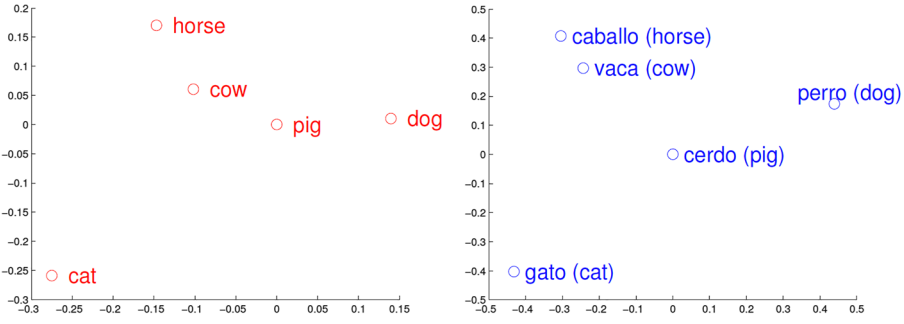


**Fig. 5** Correspondence between word embeddings of different languages (credits: Mikolov et al. 2013b)


**Fig. 6** Visualization of semantic change based on word2vec (credits: Hamilton et al. 2016)

assessments of NN word embeddings have pointed out important limitations in this sense, such as their significant sensibility to different similarity measures, the non-reversibility of the analogy task (e.g., $v_{farmhouse} - v_{countryside} + v_{city} \approx v_{bungalow} \neq v_{house}$), their task dependency, or their lack of consistent performance across downstream tasks (cf. Schnabel et al. 2015; Linzen 2016; Gladkova et al. 2016). Moreover, studies such as Levy and Goldberg (2014a, b, 2015) convincingly show that most of the results and performances achieved by NN models can be obtained by previously existing models if properly adjusted following implicit parameters of the former.

However, from a conceptual or philosophical perspective, the implications of word embeddings outreach their practical outcomes. Even if their most suggestive results are not entirely reliable in their current state, even if they can be proved, in hindsight, not to depend on the specific technology of NN, the success of word vector models made evident in the wake of word2vec brought to the surface a novel combination of three aspects of linguistic analysis that was practically absent as such among the principal research orientations of the field before their arrival. First and foremost, the renewed idea that the reconstruction of the underlying organization of language requires virtually no more human intervention than the one implied in the most ordinary use of language as recorded in an *almost raw linguistic corpus*. Second, the evidence that in the reconstruction of that organization both *semantic and syntactic contents* of words are determined *at the same time and as the result of the same procedure*, which requires that any clear-cut distinction between syntax and semantic be reconsidered and warns, in turn, against any reduction of word relatedness to semantic similarity. Finally, the perspective that word vector representations are not simply disposed or organized in the vector space in such a way that more or less stable neighborhood relations capture more or less significant similarities, but that the vector space itself is *structured following relatively precise directions* at the crossroads of which those syntactic and semantic contents are established.

For that reason, if the uncontroversial improvement word embeddings brought about as standard input for most NN models has already earned them the place of the bedrock in the new epoch of NLP, it is possible that the strength and the freshness of the image of language that they carry with them hold the capacity to disquiet the one that has governed our philosophical approach to language in most areas of knowledge for almost a century.

## 3 Why Does it Work? The Implicit Mechanisms of Word Embeddings

And yet, the attempts to shed some clear light on that image have been relatively infrequent compared to the profusion of technical discussions around word embeddings within this new NN framework. Indeed, epistemological and philosophical reflections are scarce, at best, in the literature of the field. That lack of comprehensive conceptual insight is accentuated by the fact that, at first sight, it is not at all clear why word embeddings resulting from NN models work the way they do. This is not to say that something in the model is unknown. On the contrary, models such as word2vec are, in a way, almost too simple. Yet there seems to be a disproportion between the simplicity of the models and the complexity of the results attained that opens too wide a gap to be spontaneously covered by merely technical explanations.

Given the difficulty of the task, most of the elucidating efforts have been then oriented to answer the question "why does it work?" by identifying the action of already known models under the mechanisms of the new approach. A glance at those results constitutes therefore a first step toward the theory of meaning that lies behind NN word embeddings.

### 3.1 The Figure in the Carpet: the Factorization of Word-Context Matrices

In the last few years, a number of studies, such as Li et al. (2016), Smilkov et al. (2016), Park et al. (2018), Liu et al. (2018), and Chen et al. (2018), developed several visualization tools intending to provide a more intuitive understanding of the way in which word2vec training progressively organizes vectors in the embedding space, with a special focus on the emergence of linear substructures inducing word analogy. In spite of their undeniable heuristic value and utility, those tools put the accent on the optimization of the model, rather than on the intelligibility of its procedures with respect to linguistic structures as such. Less intuitive approaches turn out to provide better insight into this case.

If we come back to Mikolov's et al. original papers, the only explanatory attempt one can find in those pages, other than external references, concerns specifically the technical aspects of the additive compositionality (i.e., $v_{Russia} + v_{river} \approx v_{VolgaRiver}$). Mikolov finds an explanation of the linear properties of vectors—qualified as "surprising" by the author on multiple occasions—by referring it to the specific training objective of his model (Mikolov et al. 2013c, §5). For instance, in the Skip-gram model,[27] each word vector is trained to capture the distribution of its context words (i.e., the frequency in which all the other words actually appear within its context throughout the corpus), and a probability distribution is computed out of it (i.e., the probability for every word in the vocabulary to appear in the context of the given word). Now, if $x$ and $y$ appear frequently in the context of $a$, as well as $y$ and $z$ in the context of $b$, the conjunction of $a$ and $b$ will yield the intersection of their respective contexts, which will thus contain $y$ and exclude $x$ and $z$. By well-known probability properties, such conjunction corresponds to the multiplication of

[27] A similar argument holds for CBOW.

their respective probabilities. Since word vectors representing $a$ and $b$ are in a logarithmic relation to those probabilities (through the softmax layer), such conjunction will then correspond to the addition of the vectors. Adding $a$ and $b$ will then result in a vector similar to $y$. It follows that "if 'Volga River' appears frequently in the same sentence together with the words 'Russian' and 'river', the sum of these two word vectors will result in such a feature vector that is close to the vector of 'Volga River'" (Mikolov et al. 2013c, p. 7).

Other than technical aspects concerning the way traditional probabilities are encoded in NN models, this explanation does little more than associate the capacity of vectors to capture linguistic meaning with the statistics of contexts within a corpus, which was already fairly evident from the basic procedures of the model. Instead, better insight can be obtained by considering an alternative word embedding model called GloVe (for "Global Vectors"), proposed by the Stanford NLP team, in the wake of word2vec (Pennington et al. 2014a). This work associated the fundamental idea of word2vec with well-known non-NN NLP models based on *the analysis of co-occurrence matrices*. In this way, the team was able to obtain a set of word vector representations with comparable results by setting the objective of the NN as that of predicting, not the context words out of a given word (or vice versa), but directly the values of a matrix containing the global frequencies of co-occurrence of words (within a context window of a given size) in a corpus. In other terms, given the $i$th and the $j$th words of a vocabulary as input, the NN is asked to predict, not another word in their context, but their total number of co-occurrences within a corpus, previously computed in a co-occurrence matrix where such number occupies the cell at the intersection of the $i$th row and the $j$th column. The fact that the resulting vectors, constructed in such a different way, are capable of performing the same tasks as word2vec with comparable performance reveals two noteworthy situations. First, that the ultimate reason for vectors to behave the way they do, and capture meaningful linguistic structure, relies on the *global* statistics of word contexts within a corpus.[28] And second, that *NN architectures play only a secondary role* in this relation between linguistic structure and global statistics. They seem to provide nothing more than an efficient way of encoding the information already present in the co-occurrence matrix.

The idea that essential aspects of linguistic structure and meaning can be obtained by probabilistic analysis is far from being new, and plainly trivial in the field of NLP. Yet, what the analysis of GloVe suggests is that the remarkable results of word embeddings with respect to other existing models hinge upon a specific data structure, namely that of *co-occurrence matrices*, rather than, say, plain $n$-grams or Markov models, longtime privileged by classical probabilistic models.

Levy and Goldberg (2014b) provided a more direct and detailed confirmation of this view. Through a formal analysis of the Skip-gram model, the authors showed that word2vec is *implicitly factorizing a word-context matrix*, in other terms a matrix

[28] Schnabel et al. (2015) noted that NN word embeddings encode a great amount of information about word frequencies.

of a similar type as that of GloVe, in which rows correspond to words and columns to contexts with respect to those words. Furthermore, the authors revealed that the cells of the implicit matrix of the Skip-gram model do not correspond to the raw count of the corresponding word-context pairs, but to a certain measure of the strength of their association. The authors identify that measure as being essentially the *pointwise mutual information* (PMI), a common statistical function reflecting the disparity between the probability of the joint distribution between two variables (in this case, word and context) with respect to their respective individual distributions.[29]

That word2vec is implicitly *factorizing* a PMI matrix means that the resulting word vectors can be understood as the rows of a low-dimensional matrix, which, multiplied by another matrix, results in (a close to optimal approximation of) that PMI matrix. Because of the way in which such a factorization takes place, that low-dimensional matrix can function as a substitute for the general matrix, since the former encodes the most relevant information of the latter. Incidentally, the authors conjecture that the better results word2vec exhibits in the analogy tasks, compared to other well-known matrix factorization techniques, can be explained by the fact that, unlike the latter, the factorization performed by the Skip-gram model is weighted, increasing the strength of word-context pairs that are more frequent. What is more, as already mentioned, in Levy and Goldberg (2014a) and Levy et al. (2015), the authors show that, understood as linear combinations of three pairwise word similarities, analogy relations can be also (and even better) captured by explicit (i.e., unfactorized) word-context matrices, and that by adopting some of the implicit parameters of NN models, traditional models based on word-context matrices can achieve comparable performances.

It appears then that the secret of word2vec and alike word embedding models relying on NN architectures resides in the particular way in which the distributions of linguistic units in a corpus are connected to one another through *a word-context relation, which can be properly grasped by the connection between the rows and the columns of a word-context matrix*. As shown by the relevant literature, the components of a word vector are nothing more than an efficient encoding of the global distribution of the contexts of that word throughout a corpus. If words can be adequately represented as dense vectors, and if significant aspects of linguistic structure can be thereby mirrored by the space those vectors define, the reason must then be sought in the relation those word-context matrices maintain with natural language.

### 3.2 Inside the Matrix

PMI matrices are only a special case of word-context matrices. The latter are far from being new in the landscape of NLP: they are the common denominator of a family of linguistic models whose origins go back to the early 1970s. A substantial renewal of that approach during the 1990s earned them the place of state-of-the-art at several NLP tasks, and at information retrieval and word similarity most particularly, until the recent arrival of the new generation of NN models to the field. Significantly, in matrix models, unlike classical statistical ($n$-gram) models, words are also represented by

[29] Formally, $\mathrm{PMI}(x, y) = \log \frac{p(x,y)}{p(x)p(y)}$.

vectors. Hence, the name vector space models (VSMs), generically qualifying the whole family of models.

In a paper that would become influential within the research community of NN word embeddings, Turney and Pantel (2010) draw a fairly comprehensive view of the history and variety of VSMs. As they justly point out (Turney and Pantel 2010, p. 143), vectors had already been widely used as representational tools in the early days of cognitive science and machine learning. Yet an entirely new approach emerged when frequencies of words with respect to contexts within a corpus were used as components of those vectors, an idea that was introduced by the SMART Information Retrieval System (Salton 1971; Salton et al. 1975). Given that the system was designed for information retrieval, the contexts corresponded to documents over which queries could be made. Thus, the terms susceptible to be the object of a query were represented by row vectors in a matrix whose column vectors represented all the documents susceptible to match a query. Each cell $m_{ij}$ of the matrix $M$ registered the number of occurrences of the $i$th term within the $j$th document. Hence, a term in such a system is represented by nothing more than the documents in which it appears (weighted proportionally to the frequency of appearance), and conversely, documents are represented by nothing more than the words they contain. Being a set of terms, a query is then conceived as a sort of document, whose similarity can be computed with respect to all other documents in the embedding space of document vectors (i.e., column vectors). The result is a list, sorted by similarity, of the most relevant documents matching the query.

The success of this approach encouraged the development of a great variety of VSMs during the past decades, in which rows and columns could represent different kinds of pairs of entities: terms-documents, words-contexts, persons-items, pairs of terms-word patterns, etc. In particular, Deerwester et al. (1990) realized that a word-by-document matrix could also be used to determine the similarity between words instead of documents, by comparing the rows rather than the columns of the matrix. Several forms of column contexts were proposed, which proved to be best suited for this new task, going all the way down from documents to individual words as contexts of the initial words, measured within a fixed window of words (Schütze 1992; 1993; Qiu and Frei 1993; Lund and Burgess 1996). Sahlgren (2006) provides a good historical overview of these approaches, which are commonly referred to as *Distributional Semantic Models* (DSMs) to distinguish them from VSMs, mostly oriented to information retrieval. Moreover, he observed (Sahlgren 2006, 2008) that word-by-word matrices have a different behavior concerning word similarity, compared to word-by-document ones, since the latter tend to measure word co-occurrence within the same contexts, while the former is rather oriented to capture their substitutability in the contexts considered.[30]

We will have the opportunity to discuss the relevance of the distinction between word-by-document and word-by-word matrices in the following pages. For the moment, what is important in order to understand the generality of word vector

[30]Lenci (2018) gives a recent account of existing DSMs with special attention on the kind of matrix they employ.

representations put forth by NN word embeddings, is to focus on the most general mechanisms underlying all those models as their common denominator, irrespective of the specific task they are intended to solve. From what precedes, it appears that such mechanisms are essentially determined by the *constitutive relation between terms and contexts as it is expressed by that between the rows and columns of a frequency matrix*. Thus, we will henceforth use the expression "matrix models" (MMs) to refer to those NLP models that rely in a decisive way on that row-column relation, without consideration of other specific properties. For the same reason, unless stated otherwise, we will use the word "terms" to refer generically to the units represented by the rows of those matrices and "contexts" for those represented by the columns, even in the cases in which the latter are of the same type as the former, like in word-by-word matrices.[31]

Turney and Pantel's survey provide a general description of the usual procedure followed in elaborating a MM. After a preliminary linguistic preprocessing that includes tokenization (i.e., division of the corpus in elementary units, such as words), and in some cases also normalization (case folding, stemming, etc.) and annotation (POS tagging, disambiguation, etc.) of the raw text in the corpus, the generic steps of the mathematical processing of MMs are roughly as follows. First, the *frequency matrix* is built, by simply counting the number of times the terms occur within the predefined contexts. Then those frequencies are *weighted* following information-theoretical principles according to the specific goal of the model (in order to give, for instance, higher impact to surprising associations).[32] Matrices resulting from the previous step are usually extremely sparse (i.e., most of its components are zeros), since most of the terms do not appear in most of the contexts. They also tend to be highly noisy, since both rare and very common terms appearing in the corpus might happen to have more impact than they should, due to the limitations of the corpus. This situation entails several difficulties, in particular concerning possible generalizations of the model to data outside the corpus, but also related to high computational costs. The third step in the procedure is therefore the "*smoothing*" of the matrix, in order to achieve a sparsity and noise reduction, as well as a dimensionality reduction. Finally, the resulting vectors—either rows (terms) or columns (contexts), depending on the task—are *compared* using distance or similarity measures, such as the already mentioned cosine or Euclidean distance.

But what exactly gives MMs their analytical power over linguistic units (whether they are terms or contexts)? As already stated, in the initial frequency matrix, the terms (represented by the rows) are only determined by the contexts in which they appear (represented by the columns). If two terms appear in the same contexts, and only in those, then those terms are indistinguishable from the point of view of the model (they are represented by identical row vectors), and the same is true of contexts with respect to terms. As a consequence, any distinction between terms

[31] We follow here the general perspective of Turney and Pantel (2010, p. 148): "In general, we may have a word-context matrix, in which the context is given by words, phrases, sentences, paragraphs, chapters, documents, or more exotic possibilities, such as sequences of characters or patterns.

[32] As we have seen, PMI is one of the usual ways of determining the weights of the frequencies.

(resp. contexts) cannot but result from a difference in the sets of contexts in which they respectively appear (resp. the sets of terms contexts respectively accept). Terms and contexts stand therefore in a relation of co-determination, contexts playing a *discriminating* role over terms, and terms over contexts. It turns out that the *similarity between terms or contexts can be seen as an inverse effect of such discriminating effect*: for a group of terms to be similar in this configuration is nothing more than to be jointly distinguished from other terms.[33]

Sahlgren (2008) correctly recognizes this inverse relation between differences and similarities between terms. However, he seems to imply that both points of view are equivalent.[34] Yet, if similarity can indeed be viewed as an effect of the discriminating action of contexts over terms, the converse is less clear. For the notion of similarity, as it is usually conceived and used, tends to suggest the idea of a loose equivalence relation. Indeed, given a set a of terms, we tend to suppose that every term is similar to itself, if one term is similar to another, the second is similar to the first, and if the second is similar to a third one, so it is the first one. Of course, empirical results often show situations in which any of those three properties could fail; but in the vast majority of the cases, similarity is nonetheless expected to provide a good approximation to a classification of terms of which the partition of the embedding space remains the privileged form (as manifested by the different clustering techniques at work in the field).

Now, if instead of adopting the viewpoint of term similarity, we focus on the discriminating action between terms and contexts as the fundamental mechanism of MMs, we can see that there is much more to it than a simple classification task. For it is evident that not one, but several contexts can contribute to the differentiation of terms and vice versa, and that they can do so at different degrees. Therefore, a group of contexts strongly contributing to the discrimination of a group of terms from the rest can be thought of as an underlying *implicit or latent feature* affecting each explicit context at different degrees (namely, at the degree at which each context contributes to discriminating that group of terms from the rest). Given that different groups of contexts can contribute to discriminating different groups of terms, the underlying features are multiple. As such, they capture the manifold correlations between groups of terms and groups of contexts.

Those correlations have no reason to generate a strict partition of the space of terms or of contexts, since one subgroup of all the strong discriminant contexts of a group of terms will certainly be discriminant for a different group of terms while the remaining contexts will not (and the same is true for terms). Take for instance the example introduced at the beginning of Section 3.1 and suppose, in addition, that $z$

[33] The similarity between terms (resp. contexts) can be measured through the already mentioned similarity measures among the corresponding row (resp. column) vectors. In practice, dot product is the common measure in this case.

[34] For instance, when he argues that "Harris talks about meaning differences, but [...] the distributional hypothesis professes to uncover meaning similarities. There is no contradiction in this, since differences and similarities are, so to speak, two sides of the same coin" (p. 51, note 3); or when he affirms that "the distributional methodology is only concerned with meaning differences, or, expressed in different terms, with semantic similarity" (p. 37).

| | ... | *w* | *x* | *y* | *z* | ... |
|---|---|---|---|---|---|---|
| ... | ... | 0 | 0 | 0 | 0 | ... |
| *a* | ... | 0 | 1 | 1 | 0 | ... |
| *b* | ... | 0 | 0 | 1 | 1 | ... |
| *c* | ... | 1 | 0 | 0 | 1 | ... |
| ... | ... | 0 | 0 | 0 | 0 | ... |

**Fig. 7** Illustration of the co-determination between terms (rows) and contexts (columns) in a term-context matrix

and $w$ appear frequently in the context of $c$, and that, for all the existing terms other than $a$, $b$, or $c$ it so happens that $w$, $x$, $y$, and $z$ do not appear in their contexts (Fig. 7). We will then have that $x$ and $y$ taken together discriminate (strongly) $a$ and (less strongly) $b$ from the rest of the terms, while $w$ and $y$ discriminate (less strongly) the three terms $a$, $b$, and $c$ from the rest. We can thus see that neither the groups of terms $\{a, b\}$ and $\{a, b, c\}$ nor the groups of contexts $\{x, y\}$ and $\{w, y\}$ define a partition in their respective spaces (i.e., the space of terms and of contexts), although their correlations define relevant discriminating dimensions or features underlying both of them.[35]

Latent features are purely formal: their only content is their discriminating effects over all the observable components of the corresponding spaces. Since they are also continuous (terms or contexts can be affected by them at different degrees), the set of all those features can be best thought of as dimensions defining a *latent space*. We can therefore see the latter as a virtual complex of formal (i.e., not necessarily interpretable) features governing the actual articulation of terms and contexts in the use of language. It follows that, if a similarity relation between terms (or contexts) can be certainly drawn from this fundamental mechanism of MMs, the organization of that latent space could provide, in addition, a global underlying *structure* to the embedding space, such as the one manifested by the analogy task of NN word embeddings. Thanks to that structure, terms would not only be disposed in similarity neighborhoods but organized through complex system of differences.

The above example shows that the dimensions of the latent space are not reducible to any single partitioning of terms or contexts. It is in this sense that those dimensions are *latent* with respect to the explicit, manifest, or observable spaces of terms and contexts. Interestingly enough, that virtual discriminating space need not be postulated a priori or handcrafted, since relevant dimensions of it can actually be formally *derived* from the observed practice of language as it is captured by a term-context matrix, in such a way that an explicit representation of those dimensions is produced.

[35] Incidentally, notice that discrimination is not necessarily symmetric, which can be easily seen by observing that $\{x, y\}$ discriminate $\{a, b\}$ but $\{a, b\}$ discriminate $\{x, y, z\}$ (likewise $\{a, b, c\}$ discriminate $\{w, x, y, z\}$, and not just $\{y, z\}$).

A common way of doing so is *matrix factorization*. Through factorization, the original matrix is presented as the multiplication of two or more matrices, one of which (at least) is expected to exhibit strong regularities underlying the original matrix.

To get an idea of how matrix factorization works, we can consider one of the most prominent factorization methods, namely singular value decomposition (SVD). The introduction of SVD within VSMs by Deerwester et al. (1990) in the 1990s resulted in the model known as latent semantic analysis (LSA), which motivated the renewal of the field to which we have referred earlier.[36] Based on linear algebra tools,[37] SVD extracts from the initial co-occurrence matrix $\boldsymbol{M}$ three matrices $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$, such that $\boldsymbol{M}$ is the result of a certain multiplication of those matrices, namely $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}$ (where $\boldsymbol{V}^{\boldsymbol{T}}$ is the transposed of $\boldsymbol{V}$). Intuitively, the rows of the first matrix $\boldsymbol{U}$ correspond to the terms (as in the initial matrix $\boldsymbol{M}$), but the columns now contains orthonormal (i.e., normalized independent) vectors, corresponding to latent or virtual dimensions ordered by decreasing discriminating power. Likewise, the rows of the matrix $\boldsymbol{V}$ (i.e., the columns of the matrix $\boldsymbol{V}^{\boldsymbol{T}}$) correspond to the initial contexts of $\boldsymbol{M}$, and its columns to orthonormal dimensions ordered in the same way. The remaining matrix $\boldsymbol{\Sigma}$ is a diagonal matrix, the values of the diagonal of which can be interpreted as the discriminating strength of the corresponding dimensions, in decreasing order. Given that the dimensions (the columns of the matrices) are ordered by their decreasing discriminating strength, taking only the first $k$ dimensions results in a lower-dimensional approximation of the initial matrix $\boldsymbol{M}$.[38]

Matrix factorization is one of the preferred smoothing techniques of MMs (the third step in the procedure described by Turney and Pantel). As such, it has the effect of transforming particular data corresponding to a limited corpus in such a way that it approximates the most relevant features of language in general. What is at stake here is the possibility of extracting abstract rules of language from purely empirical data. In other words, if there is a place where the structured space underlying word embeddings can emerge in an unsupervised fashion from raw corpora, this is it. As we have seen, Levy and Goldberg showed that the Skip-gram model could be understood as a factorization procedure and compared its results to an SVD factorization of its implicit word-context matrix. Indeed, like word embeddings, the vectors produced through matrix factorization are dense, low-dimensional, computationally efficient and best suited for generalization outside the observed data. It appears then that the remarkable properties exhibited by word2vec and other NN word embedding

[36] LSA terms were taken to be words and contexts were documents or paragraphs. The latent space derived through SVD was therefore singularly suited for the task of information retrieval. But the idea of a latent space as we have presented it is not essentially tied to such a task, no more than to word-by-document matrices, as neither is SVD as a general technique for deriving relevant dimensions of that space. The generality of LSA methods is not unknown to Landauer and his team, as reflected in Landauer et al. (2007), where the relevance of LSA for other models and tasks is assessed.

[37] The technical aspects of SVD procedure fall outside the scope of the present paper. For details, one might consult (Deerwester et al. 1990; Landauer et al. 1998, 2007).

[38] In fact, it can be shown that such $k$-approximation is the best possible approximation, i.e., minimizes the approximation errors with respect to the Frobenius norm.

models are indeed intimately related to the way in which they reconstruct the dimensions of the latent space governing the co-determination between terms and contexts respectively represented by the rows and columns of a frequency matrix.

Yet, we have also seen that NN embeddings exhibited, at least at the moment of their introduction, better performances than classic models based on SVD and other related techniques, in particular in analogy tasks. Since factorization is implicit (i.e., performed by indirect means) in NN models, its mechanisms remain somewhat obscure; hence, it is not easy to tell what can account for that superiority. We have already mentioned Levy and Goldberg's conjecture concerning the implicit strong weight given by the Skip-gram model to frequent co-occurrences. It might also very well be that SVD is not the most adequate way of producing an explicit representation of the latent dimensions of a *linguistic* space.[39] In this sense, Turney and Pantel (2010, p. 160) observe that SVD models presuppose that the elements of the matrix have a Gaussian distribution, which is known not to be the case for words in natural language. However, the authors also observe that PMI is approximately Gaussian, thus explaining the better behavior exhibited by SVD with PMI matrices. As we have seen, Levy and Goldberg have shown that the matrix underlying word2vec was indeed a PMI matrix.

Other matrix smoothing techniques have been developed since the introduction of LSA, such as probabilistic latent semantic analysis (Hofmann 1999, 2001) or latent Dirichlet allocation (Blei et al. 2003), which have proposed original methods for increasing the performance of term-context matrix analysis in ways that are not foreign to the procedures of NN word embeddings. We have also mentioned that Levy et al. (2014a, 2015) convincingly show that most advantages of NN methods rely less on the model itself than on implicit hyperparameters—such as dynamic context windows, context distribution smoothing, or eigenvalue weighting—which can be transferred to previous MMs to obtain comparable performances, even without recourse to factorization techniques or any dimensionality reduction.

The fact that certain models can achieve competitive results without recourse to any explicit representation of the latent space is certainly remarkable. And indeed, this may lead to the impression that in such cases no implicit dimensions are involved in the capacity of those models to reconstruct essential aspects of language. This is specially the case for models based on word-by-word matrices, and one of the reasons they are thought to be substantially different from those built upon word-by-document ones. However, that no explicit representation of implicit dimensions are needed in those particular models does not mean that such dimensions are not at work in the structure of the data to which those models are applied (contributing, then, to their effectiveness). Consider, for instance, the high co-occurrence frequency of the words "he", "she", and "it" on one side, with respect to words like "says", "makes", and "plays" or "talks" on the other. The fact that a model can successfully account for the similarity of the words belonging to one or the other of those two groups without

[39] See, for instance, Caron (2001), Bullinaria and Levy (2012), and Österlund et al. (2015)

constructing an explicit representation corresponding to "third person singular verbs" or "third person singular personal pronouns" does not mean that such features do not play any role in the reasons explaining that success. Indeed, frequency of co-occurrence seems to be practically all those models rely on for establishing word similarity, which in this case has no other source than the existence in the English language of an implicit dimension of the "third person singular" relating, among others, personal pronouns to verbs.

More generally, it seems that the efficacy of MMs to grasp essential aspects of linguistic structure is tied to the way in which the complex co-determination between rows and columns is capable of capturing those latent dimensions of language, whether they construct explicit representations for them or not. However, there is a point in which the properties derivable from an explicit representation of the latent space are irreducible to the ones of models not making any use of them. If we come back to the example presented in Fig. 7, and consider the terms $a$ and $c$, we can see that, from the point of view of observable (i.e., not latent) contexts, these terms are as different as they can be. If similarity was computed based on those vector representations, the result would be that those terms are not similar at all. However, if it so happens that a significant number of terms other than $a$, $b$, and $c$, and only them, frequently appear in the context of all $w$, $x$, $y$, and $z$, whatever those contexts represent (documents or words), and in no other, then those four contexts would naturally define a single latent dimension, in which case the similarity of $a$ and $c$ could be easily established with respect to that single dimension. To get and intuition of this remarkable fact, take for instance $a$ ="she", $c$ ="we", $w$ ="say", $x$ ="says", $y$ ="makes", $z$ ="make". In the case of this example, this means that if the connection between "say", "says", "make," and "makes" can be established by distributional means other than the distribution of "she" and "we" (which is highly probable, since they are all verbs), then the relation between these two terms can in principle be captured by the model, even if their distributions are perfectly disjoint.[40]

That irreducible capacity of latent dimensions to connect terms that do not share any context reveals another decisive property of the latent space MMs are able to grasp, namely the *global* principle guiding the definition of their units. By comparing contexts only within the local scope of paris of terms, vector representations based only on observable co-occurrence frequency are unable to address the possibility of locally disjoint contexts (or terms) being globally linked. This is, of course a remarkable possibility offered by MMs, and a source of their potential strength. Its importance will become more and more evident as we unfold the image of language that stems from the mechanisms of MMs as we have just described them.

[40]The same is true for a possible disjoint distribution of "say" and "make" with respect to that of "says" and "makes" if the similarity of "she" and "we" can be established by their joint distribution outside those contexts.

## 4 Toward a New Image of Language

### 4.1 The Many Uses of the Distributional Hypothesis

In the previous sections, we have seen that NN word embeddings were the result of what can be understood as a particular way of implicitly factorizing a term-context matrix, yielding a set of low-dimensional vectors which define an embedding space whose substructures suggest the capacity of mirroring those of natural language. The analysis of the most elementary mechanisms of MMs showed that such a factorization relates to a complex latent space governing the co-determination between groups of terms and groups of contexts in a given corpus. Moreover, it appeared that the dimensions of that latent space corresponded to nothing more than highly discriminating axes underlying, yet derived from, the discriminating action of observed pairs of contexts and terms.

But if after gaining this insight from MMs we are in a better position to answer to the question "how does it work?," the more philosophical question "why does it work?" is still to be answered. In view of what precedes, this question can now be reformulated as follows: *how is it possible that a latent discriminating space underlying the relations between terms and contexts in a given raw record of linguistic performances succeeds in capturing a surprising amount of linguistic properties, including syntactic regularities and structures, semantic relatedness, isomorphic configurations between different languages and historical change?* As already noted, most explanatory efforts within the recent NN word-embedding literature have mainly focused on the technical specifications that allow any one of those computational models to perform slightly better than others in specific linguistic tasks. Yet if we want to disclose the image of language animating the entire series of those models, we need to consider their success as something more than a purely technical feat with respect to specific aspects of language, and redirect that question to *the nature of language itself*. In other terms, to the question "why can computers understand natural language?" we should direct our attention to natural language rather than to computers, and ask: *what must natural language be for the specific procedures of MMs and word embedding models to succeed in revealing some of its most essential aspects?*

It barely deserves mentioning that this is not the way the community around NN word embeddings tends to approach the problem.[41] However, there is indeed one possible answer that one can find repeatedly wielded throughout the literature, namely "the distributional hypothesis," invariably referred to Zelig Harris's (1970) and condensed into John Firth's quote turned into a motto: "You shall know a word by the

[41] To get an idea of a rather widespread feeling within the NLP community in this respect, it might be worth reminding here the "snappy" version of the words of Fred Jelinek, of the IBM speech group, as reported in Jurafsky and Martin (2008, p. 189): "Every time I fire a linguist the performance of the recognizer improves". Let us better not try to know what NLP computer scientists think about philosophers.

company it keeps!" (Firth 1957, p. 11). Although often restricted to this laconic reference within the NN community, the distributional hypothesis has been the object of abundant discussion accompanying the development of previous MMs. In his detailed treatment of this question, Sahlgren (2008, p. 33-34), for instance, lists some of the most representative formulations of that "set of assumptions about the nature of language": "words which are similar in meaning occur in similar contexts"; "words with similar meanings will occur with similar neighbors if enough text material is available"; "a representation that captures much of how words are used in natural context will capture much of what we mean by meaning"; "words that occur in the same contexts tend to have similar meanings". Lenci (2008, p. 3), in turn, proposes the following statement for the same explanatory principle: "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear."

All those different formulations of the distributional hypothesis point to one and the same place: it is the constitutive relationship between language and contexts that explains the capacity of MMs to grasp linguistic meaning. But *what exactly is a context?* And what is its precise role in the production of linguistic meaning? A common principle of explanation is given by what Lenci calls a "'usage-based' perspective on meaning" (Lenci 2008, p. 1), that is, the idea that linguistic meaning is determined by the way language is used. This perspective is not restricted to linguistic research but belongs more generally to the philosophy of language, where, under the name of a "use theory of meaning," it is often attributed to Wittgenstein, and usually summarized as "the meaning of a word is defined by the circumstances of its use" (Manning and Schütze 1999, p. 17). In this customary version, those "circumstances" of use count as the contexts with respect to which the distributional properties of linguistic units (e.g., words) are determined. The image of language that emerges from this common version suggests that natural language speakers use words in different empirical or concrete situations and have a tendency to use similar words—or more precisely words with similar meanings—in similar situations. The result is a correlation between co-occurrence of words and similarity of meaning which grounds the claim that, if a language model uses co-occurrence in linguistic contexts as proxies for those circumstances, such a model has good chances of succeeding in capturing their meaning. This image is also frequently connected with a cognitive perspective anchoring the use of language to associative faculties.[42]

Yet if we subscribe to such an image, a tension arises with respect to the mechanisms underlying the effectiveness of word embedding models. Because, from that point of view, what mediates between observable terms and contexts in a linguistic corpus on one side and the linguistic contents a model can capture on the other, is a formal latent global space defined by discriminating dimensions rather than individual cognitive agents interacting with the environment. And it is not easy to see how those two concurrent candidates to fill that gap between distribution and meaning could be articulated into one and the same explanatory principle. Moreover, while

[42] See, for instance, Spence and Owens (1990), for a classical study on the correlation between co-occurrence and association strength.

in the case of latent spaces the notion of linguistic context admits a rigorous sense, the image of language stemming from that ordinary association between the distributional hypothesis and a use theory of meaning stands on a doubtful parallel between contexts as concrete situations in which a cognitive agent makes use of language, and linguistic contexts as the set of words accompanying the utterance of another word.

Lenci's account of the possible association between distributionalism and usage-based theory of meaning (Lenci 2008) proposes a more subtle image than that customary conception, leading to a refined cognitive interpretation. Through a comprehensive review of the different attitudes toward distributionalism in the literature, the author distinguishes two possible versions of the distributional hypothesis regarding its explanatory power with respect to linguistic meaning, which differ precisely in the status granted to contexts as specifically *linguistic* contexts. The "weak" version considers the distribution of words in linguistic contexts as only *correlated* with the meaning of those words. From this perspective, word meanings, supposed to be established by extra-linguistic means, determine, as an external source, the "combinatorial behavior of words in context" (Lenci 2008, p. 14). This version is then compatible with the ordinary understanding of use as an embedded cognitive practice,[43] although the relation between linguistic and extra-linguistic contexts is not necessarily addressed as such. On the other hand, the "strong" version assigns a *causal role* to contextual distributions with respect to linguistic content. More precisely, the idea is that "[r]epeated encounters with words in different linguistic contexts eventually lead to the formation of a contextual representation as an abstract characterization of the most significant contexts with which the word is used." The nature of the significant contexts of use intended here is not specified, although Miller and Charles—to whom Lenci attributes this strong version—mention pragmatic conditions among the dimensions contextual representations are assumed to include (Miller and Charles 1991, p. 5), suggesting that at least some extra-linguistic features are involved therein. Yet contextual representations are, in this view, extra-linguistic in a stronger sense, since such representations are conceived as abstractions resulting from linguist contexts, finding their source in the cognitive faculties of an individual agent. In the words of Miller and Charles (1991, p. 5): "a word's contextual representation is not itself a linguistic context, but is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts." As the authors point out, the cognitive mechanisms assumed to perform this extra-linguistic abstraction are hardly ever specified, but semantic similarity of individual cognitive agents remains the privileged measure to assess the contextual representations supposed to result from it.

As different as this more sophisticated variant may be in its attempt to attribute explanatory power to the distributional hypothesis, it shares with the other versions a congruent conception of what contexts are, namely the domain or scope within which entities of the same nature can be presented together, or in other terms, co-occur, in such a way that they can be associated by a cognitive agent. Whether it is words within a specific linguistic scope, objects or facts in a circumscribed situation

[43] A solid defense of this position can be found in Glenberg and Mehta (2008).

or concepts within an restricted inferential framework, contexts are seen both here and there as this bounded region against the backdrop of which individual agents perform associative operations. If linguistic contents are connected with the distributional properties of linguistic units, then in one way or another, all those versions provide an image in which the latter need to be somehow correlated with this associative faculty of individual agents, and through it, with the restricted conditions of its exercise that one may then call "contexts".

### 4.2 From Word Co-Occurrence to Term-Context (bi-)Duality

Yet, the elementary mechanisms of MMs we have examined in the previous section have the potential to challenge key aspects of that common conception, providing at the same time an alternative image of language, closer to the results and perspectives the new NN models have been able to reveal in distributional models. A first step toward the reconstruction of such an image is suggested by the last example of the previous section, showing that, in principle, the mechanisms of MMs can determine the similarity between terms that do not share any explicit linguistic context. This implies that, as far as MMs are concerned, *similarity between words is not the direct result of their co-occurrence*.

This point has been strongly raised by Landauer in the framework of a reflection on the theory of meaning conveyed by LSA. As Landauer says: "Whereas LSA starts with a kind of co-occurrence, that of words with passages, the analysis produces a result in which the fact that two words appear in the same passage is not what makes them similar" (Landauer et al. 2007, p. 16).[44] Landauer refers to previous studies (Landauer 2002) that show that the similarity measure produced by LSA correlates only slightly better with the number of times words co-occur in the same context as with the number of times they occur independently in other contexts. In Dennis et al. (2003), the authors report that over 99% of the word-pairs for which LSA can establish a high similarity never appear together in the same context. As an example of this fact, Landauer mentions the expression "a circle's diameter", to which LSA is capable of attributing a cosine similarity of 0.55 with respect to "radius of the sphere", while an expression similar to the latter in terms of word co-occurrence such as "music of the spheres" only measures 0.03 from the point of view of LSA.

As surprising as it may appear at first sight, the most unsophisticated consideration of our use of language provides an intuitive interpretation of this fact. If one takes, for instance, two similar words, such as "my" and "your", or "house" and "bungalow", it is easy to see that there is no intrinsic reason for them to appear in the same linguistic

[44] Incidentally, this entails a critique of a conception of language as a "bag of words," which is common in the field of NLP (including word2vec, the name of one of the models of which—CBOW—makes explicit reference to it): "Some authors have also characterized LSA as a "bag-of-words" technique. This is true in the narrow sense that the data it uses does not include word order within passages. However, what the words are and what the model does with the words is critically different from the keyword or "vector space models" of current search engines with which the sobriquet of "bag-of-words method" is usually associated" (Landauer et al. 2007, p. 21).

context. Rather the opposite is true: if we consider the linguistic context "The day she came to your house in Paris", not only the word "my" is unlikely to appear in that particular context in which "your" already appears, but even more, its occurrence has been actually precluded by that of "your". It is precisely to mark the difference with "my" (but also with "her", "his", "their", etc.) that the word "your" has been used in this particular context. And the same is true of "house" and "bungalow" (or "apartment" or "bedroom" or any other "similar" word). It is an inherent principle *of the very meaning* of a word, as well as *a rule guiding its use*, that its occurrence in a context excludes the occurrence of all other structurally similar words *at that place*.

On the other hand, words such as "my" and "house", which do indeed frequently co-occur,[45] are far from being similar (cosine similarity = 0.04). It follows that, rather than guaranteeing similarity, co-occurrence acts as a formal differentiating or discriminating mechanism with respect to the contents of words. As Landauer correctly points out, "two words that appeared in the same sentence would not be very good evidence that they had the same meaning because there would often be more expressive value in using two different words" (Landauer et al. 2007, p. 16). The same point can be pushed further by observing that, whenever two similar words appear side by side, the resulting effect is that of a magnification of their *differences* and not of their similarities (like in the sentence "She bought a house and a bungalow"). Consider the extreme case of maximal similarity given by the relation of a word to itself, which corresponds to a minimal probability of finding a word in its own context. The rare cases in which we do find it, like in the expressions "to play a play", "such and such," "blah blah blah", or the French "vous vous trompez", imply a repetition of the word utterly unsettling for its identity as a fundamental linguistic unit, since it induces a syntactic or semantic distinction that forces us to think that behind identical occurrences lie two different meanings, if not two different words altogether.[46]

It follows that the explanatory power of the distributional hypothesis with respect to MMs, and through them possibly to NN word embedding models, cannot rely on the correlation between similarity and direct co-occurrence. In the regular use of language, similar words might tend to have similar meanings when they occur in similar contexts, but *not in the same context at the same time*. This basic observation concerning the elementary mechanisms of language should suffice to rule out any customary version of the distributional hypothesis willing to correlate in a simple way linguistic contexts with pragmatic or extra-linguistic ones. What is more, this forces us to attribute a certain degree of autonomy to specifically linguistic contexts in their contribution to linguistic meaning.

However, if similarity is not about simple or direct co-occurrence, the above examples suggest that similarity is nevertheless related to co-occurence in an indirect fashion, namely through *shared* co-occurrence. Indeed, if "my", "your", "her", "their", etc. are in a way similar, the reason, from the point of view of distributional models,

[45] The word "my" appears in 8665 ($\sim$ 0.027%) of the 316,928 different $\pm$2-word contexts surrounding the word "house" within the Corpus of Contemporary American English, compared, for instance, to only 3 of those contexts ($\sim$ 0.0000095%) in which the word "bungalow" appears.

[46] Some aspects of the phenomenon of linguistic reduplication have found an interesting treatment in "The salad-salad paper" (Ghomeshi et al. 2004).

is that they co-occur (or at least they can co-occur) with the same words ("house" or "bungalow", in our example). The notion of "second-order co-occurrence" is sometimes used to characterize this mechanism behind word similarity, a terminology first introduced by Schütze (1998) to underline the contrast with simple or "first-order" co-occurrence. This refined conception of word co-occurrence seems to be appropriate to support the claims of distributionalism to ground semantic similarity through a cognitive interpretation, provided that associative operations are superseded by more complex processes, such as intricate cognitive abstractions like those referred by Miller and Charles in the strong version of the distributional hypothesis.

Yet, even in this elaborated version, the notion of co-occurence fails to capture two essential properties of the elementary mechanism of MMs, namely: the *fundamental disparity between terms and contexts* and the *non-trivial organization of contexts themselves*.

As for the first of these points, the notion of co-occurrence suggests that, within a given context, the entities that co-occur are of the same kind or nature—in this case, words. Thus, in "my house", "my" and "house" are thought to be words co-occurring in the same context, which, together with the co-occurence of "your" and "house" will determine the similarity of "my" and "your". Yet, while the homogeneity of the terms whose similarity is established ("my" and "your") is somewhat required by the model, that of the terms and the contexts through which that similarity is established ("my" and "house", for instance) is not. Indeed, MMs assign distinct principles of representation to terms and contexts, namely row and column vectors in a matrix, and mobilize that distinction in their capacity of extracting similarity from distributional properties. The linguistic relevance of that dissymmetry between terms and contexts is clearly shown by models based on word-by-document matrices (such as LSA). By letting columns represent documents as contexts for words represented in turn by rows, those models are able to establish important semantic properties of both words and documents. Now, words and documents are not, in principle, entities of the same kind: documents are supposed to be constituted of words, but not the other way around. As such, they cannot enter into a relation of co-occurrence, for words do not co-occur with but *within* documents. In spite of this dissymmetry, and maybe even because of to it, MMs have no difficulty in representing their relationship and putting it to work for the reconstruction of meaningful aspects of language.

Of course, columns can also be made to represent words, as shown by models based on word-by-word matrices. It was Sahlgren (2006, 2008) who most convincingly argued for the difference in kind between these models and those relying on word-by-document matrices.[47] Whereas in the latter, contexts are both topical units and unities, which are directly represented by the columns of the matrix, the contextual units word-by-word matrices directly represent are instead the words themselves that can be encountered in linguistic contexts. The notion of context has in this last case a different sense: it refers to an implicit sliding window over a corpus rather than an explicit text region endowed with topical unit-unity.

[47] Understanding by "documents" regions of text of any possible length—from extended regions down to sentences and even below.

In the light of this difference, one can be tempted to admit that the disparity between terms and contexts as two sides of the matrix is not intrinsic to MMs, but is only an effect of the type of contexts one chooses to represent. However, Sahlgren's clear-cut distinction needs to be relativized. Indeed, Sahlgren associates that distinction with a more fundamental one, borrowed from the Saussurean conceptual framework, namely that of syntagmatic and paradigmatic relations between linguistic units. Syntagmatic relations concern the sequential combination of units in given linguistic contexts (such as the relation between "my" and "home" in our example), while paradigmatic relations hold between units that can be substituted in such contexts (like "my" and "your"). With this distinction in mind, Sahlgren identifies word-document models with models establishing syntagmatic relations between terms and word-word models with those producing paradigmatic ones. Now, as we have seen, word-document models are capable of producing similarities between terms that do not share any context, and hence do not stand in syntagmatic relations. On the other hand, unless the size of the sliding context window is reduced to one, word-word models are always susceptible of collecting syntagmatic information which will then contribute to establishing the similarity of terms. Think about collocations, i.e., words that co-occur very frequently, such as "San Francisco". For a 5+5-sized window in a word-word model, the words "San" and "Francisco" will share 8 out of 10 context words for each one of their co-occurrences, which means the model will establish a high similarity between them, even if collocations such as this are the "prime example" (Sahlgren 2008, p. 40) of syntagmatic relations between words. One could of course decide by alternative means that "San Francisco" is one word and not two.[48] But other cases, such as "of the", "to be", or "it was" cannot afford that simple solution. Of course, in those cases, the isolated occurrences in other contexts of each of the co-occurring words will contribute to discriminating them, compensating and even overturning the similarity effect. Moreover, as Sahlgren appropriately reminds, most terms never co-occur. Yet the point here is that neither word-document models prevent non-syntagmatic similarity nor word-word models exclude syntagmatic ones.[49]

We will have the opportunity to come back to the connection between MMs and the Saussurean distinction between syntagmatic and paradigmatic relations, which we believe needs to be placed at a different level than the one suggested by Sahlgren. What is important for now is that rather than supporting the idea of a difference in kind between document and word models, Sahlgren's use of the syntagmatic-paradigmatic distinction seems to advocate for a difference in degree between those models: word-document models *tend to* capture syntagmatic relations while word-word models *tend to* capture paradigmatic ones. The difference in degree can, in this last case, be easily grasped: in general, for a $n+m$-sized window, co-occurrent words will share $n+m-2$ context words, which implies that the larger the context window,

[48] Which already raises the issue of the identity of words as fundamental units of language.

[49] One could also add that hybrid models are conceivable, in which some columns represent words while others documents, which would result in an increase of precision in the analysis of similarity, without any loss of information. This would be unlikely if both models were different in kind.

the more syntagmatically motivated can the resulting similarity be. Indeed, Sahlgren admits that "a narrow context window is preferable to use for acquiring paradigmatic information" (Sahlgren 2008, p. 46). Interestingly, the narrower the context window, the more the represented words behave like documents, that is, as units and (aggregated) unities (since explicit context words tend to coincide with the implicit sliding context window).

Even if the syntagmatic-paradigmatic distinction does not justify an intrinsic clear-cut separation between word-document and word-word models, it could still be claimed that in word-by-word matrices the terms and contexts respectively represented by the rows and columns of the matrix are entities of the same kind. After all, if context windows are symmetric, so will the resulting frequency matrix be, and its rows and columns could be treated as one and the same thing. Yet symmetric matrices are only a particular form of matrix, induced by the choice of a specific form of context window. Nothing in the mechanisms of MMs (nor in those of language itself) forces us to make that choice. For any other form of sliding window, the resulting matrix will yield different row and column vector representations corresponding to what are supposed to be the same word, splitting, as it were, the representation of the latter into two non reducible dimensions. Such matrices will not fail to capture meaningful features of language nonetheless. On the contrary, important aspects of linguistic content, owing for instance to syntactic properties, will thereby became accessible which were otherwise out of reach.[50] In the case of directional matrices, for example, i.e., matrices collecting either right or left contexts of terms, those two disparate representations will correspond to different combinatorial capacities of linguistic units, with which semantic features can be associated, like being the subject or the object of an action, as in "the man hit the dog" or "the dog hit the man". One could still argue that in both sentences, the meaning of the word "man" is the *same*. But this is less clear as soon as we force both representations to actually relate to each other, as in "the man hits the man".

By producing different representations for terms and contexts, the effectiveness of the mechanisms of MMs forces us to consider that words are complex units determined at the crossroads of two irreducible dimensions: words as actual terms and words as contexts. Those two dimensions do not necessarily coincide, except under very specific conditions which are foreign to the model itself, and which have the effect of restricting the capacity of the model to reveal multiple aspects of linguistic meaning. An analogy with the roots of a quadratic equation might help to understand the coincidence between those two dimensions (i.e., the equality of the corresponding rows and columns in symmetric matrices): just because under very specific conditions both roots can have the same value does not mean that the equation has only one root. This analogy has nevertheless its limitations, since roots can indeed be thought as being of the same kind. However, there is no reason to resort to any analogy here, since the very form upon which MMs rely provide everything we need. Indeed, the

[50] It might be relevant to remind here that both models applying SVD as well as word2vec and other related NN models produce different representations for terms and contexts.

representation of terms and contexts in a matrix literally induces between them a relation as the one between row and column vectors. Far from being trivial, all the benefit of the application of the power of linear algebra to linguistic analysis in MMs hinges upon this circumstance. Now, this suggests that the disparity that at the same time distinguishes and connects rows and columns in a matrix is being projected upon linguistic units. As it is known, in spite of both being represented as vectors, rows and columns are related as the elements of a vector space to those of its *dual* space. This is why the latter are usually referred to as "covectors" instead. This is not just a terminological issue, vectors and covectors differ in kind: while the first ones constitute elements of some space, the second ones represent (linear) functions over those elements.

We can then rely in those formal properties of MMs to characterize the ontology such models induce on linguistic units. In particular, the *argument-function* distinction seems to accurately characterize the disparity between terms and contexts: linguistic contexts can be understood as functional expressions ranging over linguistic terms. In an expression such as "my house", "my" as a term is seen as a possible argument for the functional or unsaturated[51] expression "( ) house". And certainly, matrices can be read from both sides, so that, for instance "my house" can be decomposed into "my ( )" and "house" as well. The disparity between (words as) terms and (words as) contexts is not erased because of that.

All the contexts we have encountered can be formally represented by such functional expressions, even if that representation requires the recourse to complex structures.[52] This leads us to the second essential property of the elementary mechanisms of MMs that the notion of co-occurrence tends to disregard: the non-trivial organization of contexts themselves. By focusing on terms and on the establishment of their similarity, the idea of (shared) co-occurrence usually neglects the difficulty involved in determining when two context can be considered as being the same.

The trivial answer pointing to documents and words hides the complexity of their functional structure, and the non-triviality of the problem of the equivalence of functional expressions. Yet the question of contexts in the framework of MMs is non-trivial also in a more apparent way. If we recall the fact that MMs can establish the similarity of terms that do not co-occur in any explicit context, then we can understand that, even if the identity of explicit contexts is settled, the structure of contexts which is relevant to the similarity of terms is not given because of that, since it requires an analysis of the global distribution of contexts themselves. This problem lies behind the idea that semantic similarity is not about occurring in the same contexts, but in *similar* ones. An immediate consequence of this fact is then that a procedure equivalent to the one applied on terms has to be performed at the level of contexts for the former to take place. Interestingly, this is not possible without having

[51] To use a classical Fregean way of referring to functional expressions. Cf. Frege (1984).

[52] Landauer's conception of a passage as a linear equation (Landauer et al. 2007, p. 13) and Sahlgren's idea of null-weights for positions in context windows (Sahlgren 2008, p. 45) provide a hint on how such representations could be constructed in the case of document and word contexts.

first determined the similarity of terms that was expected as the result of the whole process.[53]

Of course, the relation between terms and context appears from this perspective to be circular. Yet, this is not a defect of the model but a property of language itself, revealed by the ontology the former projects on the latter. Indeed, from a strictly linguistic point of view, one could say that an adjective is in part defined by its relation to a noun, which is in turn defined in part by its capacity of being affected by adjectives. By neglecting the difference in kind between terms and contexts, and trivializing the organization of contextual units, the notion of co-occurrence evades this difficulty, at the expense of disregarding important dimensions of linguistic content. However, if distributionalism is to be understood as the correspondence between similarity of terms and similarity of contexts, no version of co-occurrence seems to be the right way to understand the ontology that underlies it. Indeed, to occur in similar contexts can imply not to co-occur at all.

Through the lens of MMs, distributionalism is then less about co-occurrence of words in context, than about *simultaneous and articulated discrimination* between terms and contexts: "my" and "house" are not words co-occurring in the same contexts, but "house" is the context of the actual term "my", while "my" is the context of the actual term "house". A subtle originality of MMs, and a fundamental source of their strength, consists precisely in their capacity of assigning different yet interrelated representations for those two heterogeneous dimensions of the same linguistic units, and putting it at work for the derivation of linguistic content.

The example by which we illustrated the similarity between non-co-occurring words can provide here an alternative figure, closer to the essence of distributionalism than that of co-occurrence. That example suggested that the similarity between "she" and "we" could be established by "she" appearing in the context of "says" and "we" in that of "say" (p. 22). Similarity relies then on the *equivalence of relation* between pairs of linguistic units, or in other terms, on *analogy*: if "she" can be said to be similar to "we", it is because "she" is to "says" as "we" is to "say". Shared co-occurrence could be seen as a restricted version of analogy ("my" is to "house" as "your" is to "house"). While the latter can contribute to semantic similarity, only the general form of analogy is capable of grasping the structural contents of that similarity, like the ones word2vec and alike NN models were able to exhibit.

We will see that the same notion of analogy was central in Saussure's structuralism, and in his view of the relation the speakers of a language maintain with that language. However, the analysis of the mechanisms of MMs provide the means of characterizing the articulated opposition between terms and contexts in a more formal way. We can draw on the formal relation between the spaces defined respectively by row and column vectors to understand such a relation as a genuine *duality*: contexts are the duals of terms as much as terms are the duals of contexts. Borrowing from a

[53]Terms do not have any priority over contexts in this back and forth, since, as we have seen, if terms are arguments for functional contexts, the inverse can also be true, by simple transposition of the matrix. Terms and contexts differ only formally, the only important thing is that one is seen as saturated and the other as unsaturated.

suggestive use of a related notion of duality in contemporary logic (cf. Krivine 2001; Girard 2001), we can then say that "house" is a (right) dual of "my" as well as "my" is a (left) dual of "house", as well as a document is a dual of all the terms it contains and vice versa. Thus understood, duality is neither reducible to co-occurrence nor is it the direct source of similarity (dual words, like "my" and "house", are not similar). And yet, it is through duality that word similarity is attained in the end: if "my" and "your" are similar, it is because, among others, they are both *duals* of "house" no less than of "bungalow" (whose similarity is in turn determined in part by the fact of being duals of both "my" and "your"). It follows that content similarity between terms is not about co-occurrence, but about *bi-duality*, i.e., about the relation of duality a term maintains with the dual contexts of another term (a term is similar to another if it is a dual of the latter's duals). Such a mechanism can be pushed further by considering the duality between a term and the type of its dual contexts established by bi-duality. Thus, if "says" and "say" are shown to be bi-dual (for instance, with respect to "this" and "that" as (right) dual terms), then the duality of "she" with respect to this bi-dual type of its dual context "says" permits to establish a bi-dual relation between "she" and "we".[54]

### 4.3 From Use to Strategy: Contexts as Formal Dimensions of Language

The combined action of the duality between actual terms and their corresponding contexts, and the bi-duality by which the similarity and dissimilarity of terms can be determined, and vice versa—to which we will henceforth refer as (bi-)duality—constitutes a fundamental component of the image of language engraved on the intimate mechanisms of MMs.[55] However, unlike cognitive attempts to ground the effectiveness of distributionalism upon linguistic meaning, the notion of context stemming from that image is a strictly *formal* one, referring to an internal dimension of the organization of language itself, rather than to cognitive capacities of individuals. Contexts are just the result of a formal distinction at the level of linguistic units, which is the condition for determining their content. Indeed, the procedures that lead to the analysis of word meanings in terms of vector representations require that we consider the contexts of actual terms; but such contexts are made of nothing more than other terms, whose meaning is determined by inspecting their own contexts in

[54] This understanding of articulated opposition and structural similarity in terms of duality and bi-duality, as well as its possible connection to the conditions for typing logical terms through bi-orthogonality relations owes everything to a joint work in progress with Luc Pellissier.

[55] If attention is paid to the technical details of word2vec models, it appears that such duality is pervasive: not only the duality between central words and context words motivates the introduction of two different models (CBOW and Skip-gram, the former producing vector representations of context words out of central words, and vice versa), but for each of those models, two different sets of vector representations are produced corresponding to words taken as central or as context words, and the combination of those two representations for the same word has not received a satisfactory conceptual solution (in the original word2vec model, representations of output vectors are simply discarded; in other models, both vectors are added). For a careful assessment of the technical implications attached to the possible ways of associating both sets of vectors, see Levy et al. (2015, §3.3).

turn, which might happen to be composed of the initial terms. Hence, through the lens of our models, each word, insofar as it has a meaning, is both an actual term and a component of contexts for other terms. Its identity is split, as it were, by those two incompatible aspects (since the same word cannot be a term and a context *at the same time*) at the intersection of which its meaning can be established.

Landauer provides the means to understand the purely formal character of the relation between terms and contexts when he attaches the latter to a system of simultaneous equations such as $A + 2B = 8$ and $A + B = 5$, in which "neither equation alone tells the value of either $A$ or $B$, but the two together tells both." (Landauer et al. 2007, p. 13). The meanings of contexts are then like equations between terms, as much as the meanings of terms are equations between contexts, neither of which can be determined other than by simultaneously considering their respective entire systems. We must be careful not to take Landauer's image as a mere metaphor, since it finds its strict realization in the system of equations defined by the term-context matrix, where the duality between terms and contexts corresponds to the duality between the spaces defined by row and column vectors. What is more important to us here is that such an image of formal contexts arising from MMs deeply contrasts with that of contexts as concrete cognitive situations suggested by the customary associations of the distributional hypothesis with a use theory of meaning. Indeed, nothing of a concrete extra-linguistic situation nor of empirical cognitive operations seems to find its clear correspondence in the formal organization of contexts within the language, other than indirectly through word distributions in the corpus. To affirm that linguistic contexts are a faithful proxy for pragmatic contexts seems to constitutes therefore an ungrounded claim, which does not do justice to the image of language that arises from the computational models under consideration.

Such a formal image of contexts is the key to address the question of the place of an individual cognitive agent as a necessary intermediary between distribution and content. Landauer, for instance, advances the idea that meaning is the effect of "autonomous manipulations of strings of words that convey abstract combinations of ideas such as imaginary numbers" (Landauer et al. 2007, p. 7). Of course, Landauer recognizes that, for language to be useful, a certain "mapping" has to exist between some of the those words and perceptual experience. "However," he adds, "once the mappings have been obtained through the cultural evolution of a language, there is no necessity that most of the knowledge of meaning cannot be learned from exposure to language itself" (Landauer et al. 2007, p. 7). The capacity of vector representations to capture multiple aspects of linguistic meaning based only on unsupervised analysis of corpora provides, for him, good evidence of that fact.

Yet if contexts are only a formal dimension of language rather than extra-linguistic concrete situations, and important aspects of meaning are still deducible without the need of any other source, then meaning must be to a large extent independent from the empirical situations of individual cognitive agents. The only empirical source of meaning is language itself, as a series of "autonomous manipulations of strings of words." Rather than to the instrumental practice of a cognitive subject, linguistic meaning should be attributed to a large extent to *the structure of language itself*. This does not mean that subjective dimensions are absent from meaningful linguistic phenomena, but that subjectivity has more to do with the *effects* of those mechanisms

than with their causes. In Landauer's terms: "It is [...] almost entirely the relations that are represented and activated by words and collections of words that create verbal meaning. And it is primarily these abstract relations that make thinking, reasoning, and interpersonal communication possible"; and shortly after: "Memory and language are not physical objects, they are properties of an information-processing system" (Landauer et al. 2007, p. 8). Yet nowhere is he as explicit in this respect as in the following passage:

> LSA axes are not derived from human verbal descriptions; they are underlying points in a coordinate system, in LSA's case, one that relates meanings to each other. LSA's theory of meaning is that the underlying map is the primitive substrate that gives words meaning, not vice versa. (Landauer et al. 2007, p. 8)

Although Landauer is referring here only to LSA models, our previous analyses suggest that these remarks can be extended to all MMs. This could answer why new NN models are capable of reconstructing the underlying organization of language based on an almost raw linguistic corpus without the need of any other human intervention. Such a perspective suggests that linguistic meaning is not immediately about cognitive use in pragmatic context, but about duality relations between terms and contexts as a formal distinction within language itself. Use is, however, not entirely absent from that image. If contexts are the formal duals of terms, the ordinary use of language consists in submitting oneself to the formal constraints of matching the right terms with the right contexts. Each time a word is "used" as an actual term, a fairly precise series of dual contexts are selected within the structured space of language, virtually determining the meaning of such term. To speak (or to write) is to choose among those highly restricted and organized possible contexts the next actual term, which will in turn act backwards on the previous term now turned into a context of the present one, and opening forward a new structured series of possible contexts that will keep the ball moving. If, as formal dimensions of words, term and context are like the two faces of the same coin, meaningful use of words is like flipping coins on their side and back, one at the time.

Such a conception of use is not entirely foreign to Wittgenstein's own profound views, which are not reducible to a strictly cognitive interpretation. Landauer does not miss out on referring to his work in this respect. However, rather than Wittgenstein's notion of "use," this unusual image of language is best captured by his conception of language as a "game." Language is less like a hammer, that could be used by different subjects in different situations in different ways, than like chess or Go (or flipping coins, for that matter), where strict rules, and not subjective decisions, govern the conditions under which certain practices are to be considered meaningful practices of that game. Unlike tools, games do not give the subject the freedom of use, but the organized space of a *strategy*.[56]

[56] Of course, linguistic constraints are not as minimal, rigid and explicit as those of chess or Go, and leave the room for strategies to be at the origin of new rules. This is how a deeper notion of use can find its place in the new landscape. Such creations are, however, rare at a large scale and constantly guided by existing regularities at each state of the language.

All this shows us a picture of language radically different from the one ordinarily associated with the distributional hypothesis as it is commonly adopted in the recent literature around word embeddings. Rather than words similarly used by natural language speakers in similar concrete situations, we are confronted with an image of language in which both the identity and the meaning of words are the result of a fundamental duality between two internal dimensions of an abstract and highly autonomous system of simultaneous co-determination, defining a space of formal structured constraints conditioning subjective linguistic practices. Form rather than substance, game rather than tool, strategy rather than use.

### 4.4 The Limits of Semantics

Drawing this image of language as close as possible to the capacities of the intrinsic mechanisms of MMs suggested by new NN word embeddings is not just a speculative exercise. Among others, that image can suggest new ways in which the use of classic MMs methods could be broaden in order to extend their capacity of accounting for linguistic phenomena.[57] For, as we have suggested, the large spectrum of possibilities those mechanisms offer, in particular concerning the complex organization of contexts, but also of their implicit relation with terms, can contribute in unexpected ways to revealing significant aspects of language. It is not implausible that NN models are taking advantage of that full spectrum to obtain their surprising results. What can then be the reason for the frequent restriction of MMs to the principles dictated by a small number of models, and by those based upon symmetric word-by-word matrices in particular?

Part of the answer for this limitation must be sought in the *overwhelming semantic orientation of distributional approaches of the last decades*. Indeed, as already mentioned, MMs emerged and evolved in the context of mainly semantic tasks, such as information retrieval and word similarity.[58] This almost exclusive semantic understanding of distributionalism keeps it tied to an image in which the source of meaning is to be sought either in objective "states of affairs" or in subjective or intentional cognitive operations. Indeed, be it empirically or theoretically, word similarity by individual agents remains the judge in last resource of the pertinence of contextual representations, whether it is conceived to find its source in pragmatic or extralinguistic situations, as in the customary or the weak version of the distributional hypothesis, or in the subjective faculty of abstraction as in Miller's view.

Such a circumstance entails a series of epistemological obstacles preventing from exploiting the manifold possibilities offered by MMs, starting with the *almost*

[57] This is, for instance, what Levy et al. (2015) carry out in a way, although they only focus on the technical details, without assessing the conceptual aspects concerning language itself.

[58] This circumstance motivates the frequent conflation between MMs and DSMs (distributional *semantic* models). See, for instance, Baroni and Lenci (2010).

*complete exclusion of syntax* from the original interests of MMs.[59] Moreover, owing to that semantic orientation, distributional models have extensively granted a *privilege to words as fundamental units of linguistic analysis*. Finally, and connected to those two aspects, the meaning extracted from the distributional properties of language has been above all conceived as organized in terms of mere similarities, thus neglecting the possible *underlying structures* and the mechanisms by which they can contribute to the emergence and reconstruction of linguistic meaning.

If we concentrate on the first of those obstacles, it is easy to see that this a point in which the task-independent NN models of word embedding show all their conceptual originality. As we have seen, in this case, the underlying organization of the embedding space is not only semantic (capital cities, company CEOs, traditional dishes, etc.) but also syntactic (adverbial forms of adjectives, comparatives, superlatives, gerunds, verb tenses, plurals, etc.). What is more, both semantic and syntactic structures emerge *from one and the same analytic procedure*. Although barely acknowledged as such, this is a remarkable fact if we recall that, since its introduction, the distinction between syntax and semantics as two independent dimensions has extensively oriented the study of language both within and outside the field of linguistics.

To better grasp the significance of this fact, the notion of syntax requires some clarification. While semantics generally refers to the meaning of words as given by (the features of) the objects and ideas those words refer to, the notion of syntax is more ambiguous. In the framework of the evaluation of NN models of word embeddings we have seen in previous sections (and of analogies in particular), syntax is implicitly associated with grammatical principles which can govern the combination of words into supra-lexical constructions. Yet, in this restricted sense, which follows the standard use of the term in linguistics, syntax is generally considered in contrast, not to semantics, but to morphology, that is, the form of words as given by the arrangement of sub-lexical units. When opposed to semantics, the notion of syntax tends however to assume a broader scope, namely the manipulation of linguistic units of any level (from sub-lexical to supra-lexical) independently of the meaning of both units and manipulations. In this wider sense, which encompasses the restricted one, the scope of syntax tends to cover also those of phonology and morphology insofar as all those

[59] In the particular case of LSA, for instance, although implied in the idea of manipulations of strings of words and abstract mutual constraints between terms and contexts, syntactic properties are entirely disregarded as dimensions of language worth capturing by the model. In fact, syntax is discarded twice in the original formulation of LSA: first, because the model is not supposed to capture syntactic features, but only semantic content; second, because syntactic contribution to semantic content is neglected by design, namely by neglecting word order. Certainly, those two exclusions are not oversights and find valid grounds and explicit explanations: the first one in the fact that LSA was conceived above all as a system of information retrieval (and not as a general language model), the second in the fact that, from the specific perspective of information retrieval, syntactic structure (as given by word order) conveys negligible information. Landauer and his team provide fairly convincing arguments that estimate the contribution of word order to the meaning LSA is supposed to capture at around 10–15% (Landauer et al. 2007, pp. 25–29).

fields are considered in complete independence from the possible meaning of their objects. This notion of syntax as opposed to semantics, which is close to the general idea of a formal grammar, can certainly be traced back to the works of Carnap (2001) and Morris (1938), becoming since one of the cornerstones of the philosophy of language.[60]

In this broad sense, the distinction between syntax and semantics is considered an important evolution in our understanding of linguistic phenomena, and confusions between both dimensions are identified as possible sources of helpless paradoxes.[61] Now, against this firmly established view, the syntax-semantic distinction does not play any role whatsoever in NN word embedding models such as word2vec: the model produces word vector representations bearing both syntactic and semantic content without ever orienting the training procedure in one sense or another. It is true that the window size and the symmetry or asymmetry of the contexts have been shown to correlate with higher performances on syntactic or semantic tasks (small and asymmetric context windows performing better on syntactic tasks, while large and symmetric improving semantic ones (Pennington et al. 2014a, §4.4)). However, rather than proving a clear-cut distinction, such behavior argues in favor of a continuity between syntax and semantics. The fact remains that recent NN model of word embeddings are able to reach high performance in both tasks following one and the same training procedure, and the vector representations resulting from those models inherit that lack of distinction. Indeed, it is the same vector representing the word "house" that bears with "houses" the same relation that "foot" maintains with "feet", with "housed" the same relation as "catch" with "caught", with "farmhouse" the same as "city" with "countryside". Only the directions of the offset in the embedding space distinguish those relations, for which there is no a priori classification. If syntactic and semantic tasks are distinguished, such distinction is only external and entirely arbitrary, in principle, from the point of view of the vector representations themselves. It appears that against one of the best-established features of the classical image of language, word embeddings install a continuum between syntax and semantics that blurs any categorical attempt to make of that distinction a precondition

[60] Carnap's broadening of the notion of syntax was originally intended to provide a syntactic and linguistic conception of logic: "logic will become a part of syntax, provided that the latter is conceived in a sufficiently wide sense and formulated with exactitude" (Carnap 2001, pp. 2). He thus introduced the notion of "logical syntax" of a language to refer to "the formal theory of the linguistic forms of that language—the systematic statement of the formal rules which govern it together with the development of the consequences which follow from these rules." And he defined "formal" in the following terms: "A theory, a rule, a definition, or the like is to be called formal when no reference is made in it either to the meaning of the symbols (for example, the words) or to the sense of the expressions (e.g. the sentences), but simply and solely to the kinds and order of the symbols from which the expressions are constructed." (Carnap 2001, pp. 1). Morris, in turn, defined "syntactics" as the "the study of syntactical relations of signs to one another in abstraction from the relations of signs to objects or to interpreters" (Morris 1938, p. 13), while semantics "deals with the designation of signs to their designata" (Morris 1938, p. 21).

[61] Certainly, from the viewpoint of the study of natural languages, multiple relations between both dimensions have been proposed, and what is known as "syntax-semantics interface" remains an active domain in the field (see Rappaport Hovav and Levin 2015 for an overview). Yet rather than eroding the frontier between both dimensions of language, those approaches presuppose it and intend to specify it.

of linguistic phenomena and of their scientific study and raises serious objections to the exclusion of syntactic properties for the establishment of linguistic meaning.

However, in spite of the better conceptual insight offered by word embeddings in this respect, that outcome is not restricted to NN models. Its underlying mechanisms bear practically no difference with those of MMs. This leads us to the second of the epistemological obstacles mentioned before. Since, after all, the exclusive semantic orientation of word-document and word-word MMs, as well as their usual inadequacy to grasp syntactic features, are the result of an arbitrary choice concerning the restriction of terms and contexts to the sole form of words (and even in most cases to non-syntactic words, such as nouns) and of passages or even documents as combinations of words, that is, strictly semantic units. Yet nothing in the principles of MMs and (bi)-duality relations prevents from defining terms and contexts otherwise, along the line of the syntactic-semantic continuum, so as to capture more syntactic content. It is then not difficult to see that, for instance, contexts such as "those" and "all the" would be the duals of "houses" and "feet", but not of "house" and "foot", and the reconstruction of the corresponding underlying structure should in principle be able to capture that and other syntactic analogies exhibited by word embeddings.

It follows that modifications of the boundaries defining which units are to be taken as terms and which as contexts give us the possibility of navigating indistinctly between semantic and syntactic features within the same linguistic space. As a consequence, it appears that under a general conception of syntax the lexical level does not hold any a priori privilege. Significantly, the principle of (bi)-duality between terms and contexts can hold at all the levels of articulation of language. Take for instance the sub-lexical units "er" and "est" as the (right) duals for both terms "big" and "larg"; or the units "ing" and "ed" as duals of "play" and "lov". An analysis of the implicit space of dependencies in this case will capture the system of the corresponding constraints in the exact same way as for words in documents or window contexts. Only in this case we would be extracting latent morphological dimensions of the same space, which, incidentally, are intimately related to other syntactic features at a lexical (and supra-lexical) level, no less than to semantic aspects of the resulting expressions (the first group of dualities expressing the comparative, the second the verb tense, although there is as yet no guarantee that a global analysis can extract those exact features as independent dimensions,[62] which is a matter of empirical research).

Yet, one could wonder why to stop there and not continue down to the character level? We will then easily see that "a", "e", "i", "o", and "u" appear in contexts such as "s", "t", "r", "n", "d", "l", "c", and "m", more frequently than in any other, and vice versa. Certainly, the terms involved at this level can hardly be considered as words: their semantics is hopelessly absent. Yet the content of those corresponding groups of bi-dual terms is no less clear: they correspond to the phonological categories of vowels and consonants, as a meaningful phonological distinction of the English language. A meaningful distinction whose underlying mechanisms, as far as

[62]Consider for instance the fact that "er" will probably be an observed right context for "play" and "lov" as well, although for different reasons, involving here also the relations between syntactic and semantic features.

our computational models are concerned, bear no substantial difference with the distinction between words such as "house" and "bungalow". Incidentally, the fact that each of the units of this level is devoid of semantic reference does not prevent those units from bearing decisive semantic *effects*: think about the possible substitution, within higher level units, of character units belonging to the same category, as in "bleed" and "plead" (we will return to this question in Section 5.2).

Instead of invalidating the possibility of modifying at will the limits of terms and contexts, the absence of lexical units below a certain level of analysis argues *against the privilege traditionally granted to words as fundamental units of analysis*. Such privilege is strongly rooted in an old conception of language and associated, more recently, with the semantic orientation that prevails in the field of NLP.[63] However, as we have shown, it is inscribed neither in the mechanisms governing the models we have been studying, nor in their results. If all the kinds of contexts we referred to could be taken into account at once, the result could in principle embrace the series of all the relations we encountered at different levels of analysis—syntactic (in the narrow sense), morphological, phonological—as part of a unified structured space in which determination of those possible levels would intertwine. Not that those levels would be indistinguishable, but relevant distinctions would respond to internal criteria of the embedding space itself, not necessarily matching the ones we are used to introduce in the study of language from without. The remarkable results exhibited in recent years by a new generation of NN NLP models, such as FastText (Bojanowski et al. 2016), BPE (Sennrich et al. 2016), ELMo (Peters et al. 2018), BERT (Devlin et al. 2018), GPT (Radford 2018), or XLNet (Yang et al. 2019), in which standard NN word embeddings are complemented with or replaced by character-based and contextual embeddings—i.e., with possible units at the sub-lexical and supra-lexical levels—in multiple ways, seem to provide an indirect confirmation of this point of view.[64]

It goes without saying that, in spite of their promising results, NN word embedding models are far from attaining such degree of syntactic or grammatical reconstruction. In particular, when we consider units of higher levels, syntactic regularities tend to depend increasingly on implicit classes or categories of terms rather than on explicit individual terms and it is not entirely clear how existing models could overcome this difficulty in their present state. And yet, this is what the results they exhibit seem to accomplish up to a certain extent, determined by the limited syntactic resources they operate on (such as the lexical lower bound for their minimal units). Provided that syntactic categories can be distributionally defined, a progressive and stratified derivation of them could in principle handle important aspects of this problem, although this is of course a mainly empirical question.

[63] For a critical perspective avoiding the privilege of the lexical level in the philosophy of language and the field of NLP, see (Rastier et al. 2001; Rastier 2001).

[64] The study of these more recent models, which have redefined the current state of the art, could therefore significantly contribute to the questions raised in these pages. Unfortunately, such an inquiry falls outside the scope of the present paper.

We thus arrive to the last of the obstacles identified: the neglected possibility of reconstructing linguistic *structures* and establishing their connection to the mechanisms of linguistic meaning. We find no better way to address this problem than exploring the structuralist image of language that lies behind the one we have drawn from our treatment of MMs and word embeddings.

## 5 The Structuralist Soil

### 5.1 Harris's Original Distributionalism

As unsettling as it may seem, the image of language that arises from word embeddings and best matches its simple mechanisms and its surprising capabilities corresponds quite faithfully to the stakes of Harris's original distributionalism. As we have already mentioned, practically no scholar in the recent state of the field of NLP has failed to quote at some point Harris's 1954 article "Distributional Structure" (Harris 1970) to justify the conceptual basis of their endeavors. And yet, given the strong semantic orientation of the field, one can be surprised that almost none of those references has dwelled upon an elementary aspect of his theory, mentioned since the very first lines of those pages, namely that in Harris's view *the distributional character of language is perfectly independent from meaning*.[65] Certainly, in section 2 of the paper, entitled "Distribution and meaning," Harris explores several ways in which distributional structure of language might indirectly relate to meaning as a "general characteristic of human activity" (p. 780), going as far as to maintain that "[i]n certain important cases it will even prove possible to state certain aspects of meaning as functions of measurable distributional relations." (p. 785). However, if attention is paid to the general intention of all those pages as well as to the overall purpose of his entire work, it is easy to see that those are only concessions with respect to the main idea that language is an autonomous system. To such an extent that to the question "Is there a Parallel 'Meaning Structure'?" the answer is clear: "the structure of language does not necessarily conform to the structure of subjective experience, of the subjective world of meanings" (p. 780).

The original distributionalism set out by Harris confirms in this way that distributionalism cannot be interpreted in terms of a use theory of meaning, if by "use" one understands the act of ordinary subjects freely using words as a vehicle of meaning in different situations. Harris could not be more explicit on this point:

> The perennial man in the street believes that when he speaks he freely puts together whatever elements have the meanings he intends; but he does so only by choosing members of those classes that regularly occur together, and in the order in which these classes occur. [...] the restricted distribution of classes

[65] In the first paragraph, Harris affirms: "Here we will discuss how each language can be described in terms of a distributional structure, i.e. in terms of the occurrence of parts (ultimately sounds) relative to other parts, and how this description is complete without intrusion of other features such as history or meaning" (Harris 1970, p. 775).

> persists for all their occurrences; the restrictions are not disregarded arbitrarily, e.g. for semantic needs. (Harris 1970, pp. 775-776)

Like Landauer's reconstruction, Harris's distributionalism conveys an image of language as an autonomous system of regularities and constraints, not governed by any external source of meaning. However, while the former was the result of an attempt to draw the image that best fitted the inner mechanisms of MMs, Harris's conception is rooted in a different ground. In line with Bloomfield's seminal work in linguistics, Harris's aim has above all an epistemological motivation, namely to *constitute language as an independent object of scientific inquiry*. To that end, assuming that meaning cannot be altogether detached from linguistic phenomena, its explanatory role needs at least to be put on hold, or else linguistics will end up being subjected to extraneous fields of knowledge, such as psychology, history, sociology or logic. If the study of language ought to acquire a scientific status, it is of vital importance that the properties of language can be accounted for without recourse to anything other than language itself.

Not unlike other sciences, in the case of linguistics, those properties take the form of *regularities*.[66] If one can positively show that there are substantial regularities ascribable to language, and to language alone—and not, for instance, to psychological behaviors, sociological practices, historical events, or logical principles—then the existence of language as an independent phenomenon is verified and the possibility of a science of language is guaranteed. This is what an account of linguistic distribution achieves. Harris presents all the elements of such endeavor in the following terms, in which we can find another elementary definition of what he means by distribution:[67]

> Descriptive linguistics [...] is a particular field of inquiry which deals [...] with the regularities in certain features of speech. These regularities are in the distributional relations among the features of speech in question, i.e. the occurrence of these features relatively to each other within utterances. (Harris 1960, p. 5)

Put in those general terms, Harris's distributionalism can be taken as an accurate characterization of the general approach of the computational models we have been examining, and in more than one sense, the fact that computational analysis of regularities alone is capable of capturing large regions of linguistic properties is a confirmation of Harris's views: language exists as an autonomous object of inquiry. Autonomous, that is, above all, independent from considerations of meaning, even when it is meaning we are interested in. However, unlike the models we have examined so far, in the case of Harris, that independence entails a complete *absence of privilege of the lexical level*. One should see here not only the effect of the traditional association already mentioned between semantics and lexical level, but also the fact

[66] Harris follows Bloomfield's program on this point: "As Leonard Bloomfield pointed out, it frequently happens that when we do not rest with the explanation that something is due to meaning, we discover that it has a formal regularity or 'explanation'. It may still be 'due to meaning' in one sense, but it accords with a distributional regularity." (Harris 1970, 785).

[67] Shortly after, he also characterizes distribution as "the freedom of occurrence of portions of an utterance relatively to each other" (Harris 1960, p. 5).

that, as we have already suggested, distributional regularities are stronger at sub-lexical levels. Hence, Harris's distributionalism concerns essentially phonological and morphological elements.

Besides the epistemological motivation and the difference in level of analysis, Harris's formulations bear remarkable convergences with recent computational models. For instance, despite the multiple differences between phonological and morphological levels, the formal procedures of distributional analysis apply indistinctly to both of them. Moreover, distributional analysis stands on the same principles as the models we have seen: mutual dependencies among terms or "elements"—as Harris generically calls them—are determined by inspecting the entire series of their contexts or "environments." Here, no less than in those models, the relation between elements and environments is conceived, not as simple "co-occurrence,"[68] but as a real duality relation in which the categorization of the multiple occurrences of elements is achieved by bi-duality with respect to the corresponding environments within a corpus.

Presenting the details of such distributional analysis at the level of phonemes or morphemes would take us too far.[69] Nevertheless, it is worth noticing that not only the conceptual framework set up by Harris corresponds intimately to the mechanisms underlying word embeddings, but the formal objects governing the analytical procedure are also virtually the same. Of course, it is impossible to find in Harris's formulations any attempt to make use of advanced algebraic methods such as SVD or tensor calculus, let alone neural network models. Yet term-context matrices are no less explicitly used, as we can see in Fig. 8. Significantly, the main motivation for their introduction is to establish the structural relation between terms that we have referred to as "analogy" (as opposed to shared co-occurrence). Indeed, following his analytical procedure, two elements (rows) can be merged into one formal unit if the set of their environments (columns) are complementary; in other words, if they *do not share any explicit context*. The corresponding elements are then established as contextual variants of the same formal unit. Of course, there can be more than one way of performing such reduction, which are all equivalent from the formal viewpoint of the description, but those reductions are privileged which entail the most optimal description of the whole system.[70]

Unlike the latent dimensions of the models we have examined, the formal units stemming from the reduced dimensions are not only interpretable in this case, but they constitute the elementary units of language resulting from linguistic description: phonemes and morphemes. In the case presented in Fig. 8, elements represented by different typographical variants of the same letter should be reduced to the same formal unit. For instance, the distributional complementarity of the elements K, k, and ĸ, respectively indicating the sounds for "c̲ar", "doc̲tor", and "k̲ey", leads to consider them as three different variants (back, central, and front) of the same phonological

[68] Harris despises as much as Landauer a conception of language as a "bag of words." See Harris (1970, p. 785).

[69] The interested reader could simply refer to Harris's (1960).

[70] Note that Harris's procedure can be understood as an elementary technique of dimensionality reduction.

| Segments | Environments | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #— r̥ | #—r | #—l | e i —C | æ—C | a o u —C | s— e i | s—æ | s— a o u | … | t̥— | C³— |
| t̥ | √ | | | | | | | | | | | |
| t | | √ | | √ | √ | √ | √ | √ | √ | | | |
| K | | | | | | √ | | | √ | | | |
| k | | √ | √ | | √ | | | √ | | | | |
| ĸ | | | | √ | | | √ | | | | | |
| G | | | | | | √ | | | | | | |
| g | | √ | √ | | √ | | | | | | | |
| ɢ | | | | √ | | | | | | | | |
| r | | | | √ | √ | √ | | | | | | √ |
| r̥ | | | | | | | | | | | √ | |
| | | | | | | | | | | | | |

**Fig. 8** Term-context matrix showing the distribution of the different sounds corresponding to the phonemes /t/, /k/, /g/, and /r/ in Harris's (1960, p. 74)

unit /k/. Yet the identity of the latter is purely formal. Indeed, there is no substantial (i.e., acoustic or articulatory) reason to associate k and ĸ with K rather than with G ("gods"). It also follows in this case that the elementary building blocks of language as they can be scientifically established by a distributional description relies on nothing more than the latent discriminating space derived from the overall system of constraints given by the distributional structure presented by the matrix.

The few elements we have considered show that Harris's original distributionalism provides an adequate basis for the intelligibility of MMs and word embeddings, not despite his refusal to associate his theory of language with the *structure* of meaning, but precisely *because of it*. It is by holding to the idea that, irrespective of the multiple aspects of meaning, linguistic phenomena are endowed with their own gravity, characterized by their own properties and defined by their own rules that Harris develops a formal account of language whose mechanisms correspond closely to those of the computational models we examined. Indeed, as we have seen for the latter, the image of language stemming from Harris's distributionalism is that of an autonomous system of dependencies relying on a co-determination between elements and environments related as dual terms in a term-context matrix. The essence of language can thus be attributed to an implicit or latent space of that matrix distribution and not the cognitive faculties of individual subjects. Moreover, like those models, the same mechanism of dependency works at different levels of languages (phonological and morphological).

After examining Harris's own formulations, we understand above all that language is about regularities in utterances and that, as long as computational models are capable of extracting regularities from linguistic corpora, they will be capturing relevant linguistic features. However, the originality Harris's theory has to offer to our image of language resides in the idea that, since those regularities are purely formal (i.e., independent from meaning), *they are not organized as mere similarities*. Far from relying in similarities, distributional analysis establishes those regularities as a complex *structure* holding all the pieces of language together. Indeed, the regularities extracted from the term-context matrix provide the elementary building blocks of language (phonemes and morphemes) as formal units whose identity depends on the global constraints of that space of distributions. That very definition of the elementary units deploys a system of internal dependencies that makes that each term maintains with all the others coordinated relations of mutual opposition (e.g., by construction, the phonemes /t/, /k/, /g/, and /r/ of Fig. 8 are mutually exclusive). Unlike the spirit of semantically oriented MMs, Harris uses the same principles giving access to a latent space (i.e., a duality relation between terms and contexts) to derive rigorous structural relations rather than mere proximities and similarities. Only now can we grasp the significance of what our analysis of the mechanisms of MMs revealed, namely that similarities were a side effect of a system of discriminating axes: it is not discriminating axes that are an efficient way of capturing similarities of linguistic units, it is rather similarity of linguistic units—and even their very identity, as in the case of phonemes or morphemes—that is an effect of spaces structured by discriminating axes.

In this way, Harris provides us with the means to understand the notion of structure that we have been needing to characterize underlying regularities of analogical relations in the embedding space. "Structure" here has a technical sense: it refers to a system of dependencies of formal units of different levels which are the elementary building blocks of language. Such an idea of structure, which is also conveyed by the notion of "grammar" in its most general sense, is to be opposed to that of embedding "space," in which terms are not considered for their mutual stratified dependencies, but only for their higher or lower similarity. Distributionalism is not about similarity of use, but about regularity of structures of oppositions built upon discriminating principles. It is precisely that structure that lies behind the regularities in the dual relations between terms and contexts. Incidentally, since that structure has no particular semantic import, the distributional properties concern above all syntactic regularities. It should not be surprising then if NN word embedding models are capable of unearthing analogical structures at the syntactic level when looking for semantic similarities. Relaxing the semantic orientation of previous models, the original sense of distributionalism emerged on the surface: the regularities of language are necessarily structured at all levels, starting with syntax. Syntactic structure is what language is *made of*.

### 5.2 The Structuralist Hypothesis

However, a last question remains open: if distributional structure is independent from meaning, how is it that, after all, by a pure distributional analysis, word embeddings

and other related models are capable of grasping so much of the meaning language conveys? In other terms, if distribution is above all a matter of syntax, how is it that not only syntactic but also semantic structure appears in those new distributional analysis. Harris gave us the means of understanding why linguistic distribution is necessarily structural. Yet the price to pay was to lose any insight on why linguistic structures can be meaningful nonetheless.

Among the scholars who have addressed the distributional hypothesis within the framework of computational models, Sahlgren is certainly the one who most seriously considered this difficulty in Harris's distributionalism. What is more, he perceived that the answer resided in the fact that distributional approaches "are rooted, and thrive, in structuralist soil" (Sahlgren 2008, p. 34). However, in his view, the issue can be resolved by falling back on the notion of semantic similarity as a perfect equivalent of that of structural differences. In this way, the problem of meaning within Harris's framework is seen as that of the lack of precision of the concept of semantic similarity, "too broad to be useful" (Sahlgren 2008, p. 37). As we have already seen, Sahlgren's solution is then to further characterize similarity by resorting to Saussure's distinction between syntactic and paradigmatic relations, and identifying two models that can respectively grasp those two kinds of relations (word-document and word-word models), thus substantially specifying the otherwise unusable notion of semantic similarity.

We have sufficiently discussed the irreducibility of structural differences to semantic similarity, as well as the difficulties associated with the correlation between documents and word models on one side and syntagmatic and paradigmatic relations on the other. However, there is a last aspect of Sahlgren's account whose rectification gives the opportunity to address the problem of the meaning of structural differences in a different way.

Sahlgren is right in invoking the name of Saussure to asses this difficult question. Indeed, the idea we found in Harris that linguistic phenomena are better understood if conceived under the form of a structural organization goes back to Saussure's structural linguistics (Saussure 1959). Significantly, as in the case of Harris, the structuralist conception of language is, for Saussure, also tied to epistemological concerns. Both here and there, structuralism is presented as a way of assuring an autonomous scientific account of linguistic phenomena. Yet for Saussure, that starting point assumes a dramatic tone, since, as soon as we resolve to derive some basic properties from linguistic phenomena—even if only distributional properties—we realize that we are hampered by a fundamental circumstance: *linguistic units are not immediately given in experience*. Not that we cannot have access to actual utterances for one reason or another. However, faced with an actual utterance, *there is no immediate, simple or natural way to determine which units that utterance is composed of*. Ultimately, there is even no essential way of knowing if we are in front of a linguistic utterance at all.

Among the many pieces of evidence he provides to support this claim, Saussure evokes the case in which we are confronted with an unfamiliar language (Saussure 1959, p. 103). In that situation, not only do we not understand what is being said, in the sense that we do not know what the words *mean*, but we are not even capable of identifying what words *are*. More than that, we are unable to determine what are

the *different sounds* that have been successively pronounced, as we can easily verify by trying to reproduce the entire utterance under the supervision of our interlocutor. And it is known that no model exists that can convincingly determine the clear limits of sentences in spoken language. Eventually, if no other principle comes to our aid, we could never tell if our interlocutor is actually speaking or just spouting gibberish to make fun of us or for any other reason—in which case we would not be in front of a linguistic phenomenon at all.

From that basic fact, we are forced to conclude that in the materiality in which it is given in experience, language appears as a continuous and "shapeless mass" (Saussure 1959, p. 111 sq.),[71] for which no physical property can provide a natural analysis into relevant *linguistic* units.[72] It belongs to the essence of linguistic units to be established through an *arbitrary* segmentation of that material continuum, understanding by arbitrary, not that they can be determined at will at any time, but that their established form is not motivated by any substantial reason (Saussure 1959, p. 67 sq.). The multiple names the same object receives depending on the language (e.g. "house", "maison", and "casa") is a piece of evidence of that fact, as is the diversity with which different languages organize the continuum of sounds, which makes that, for instance, Japanese speakers recognize the sounds corresponding to the English /r/ and /l/ as one and the same sound, while Spanish are not able to distinguish between /b/ and /v/.

It follows that, without recourse to some other tool allowing us to make the relevant distinctions, no distributional analysis could ever begin to be performed, since distribution presupposes that we can already identify the same terms or units in different environments. More than establishing distributional properties, the principal task of the linguist is then to perform a complex segmentation procedure at multiple levels, establishing the relevant linguistic units upon which an entire language is built. And her situation is in this no different than that of the ordinary speaker dealing with a new language, including the latter's mother tongue at the moment of its acquisition.

Saussure's famous solution to this problem consists in affirming that relevant linguistic distinctions are made and recognized only through the *intervention of another similarly structured albeit heterogeneous system* that will help polarize, as it were, the continuous materiality of language. As an example, take the case in which, as non-English speakers, we hear the word "plead" in spoken English, and we try to analyze it in relevant units. We might at some point consider the fact that such utterance corresponds to non-comparable situations, for instance a discourse in court and red liquid flowing out of someone's body. Eventually, those incomparable situations will end up helping us distinguish what we initially experienced as continuous and undifferentiated sounds into the corresponding different utterances "plead" and "bleed".[73]

[71] This is no less true of written or any other kind of language than of spoken language.

[72] This fact has been verified since by multiple empirical studies. See for instance (Liberman 1957).

[73] It would not be too difficult to interpret NN supervised training along these lines, the training set being typically composed, as we have seen, of pairs of vector representations belonging to heterogeneous domains of content (e.g. the image of an animal and its written name). For an interpretation of NN training in terms of structuralist procedures, see Maniglier (2016).

The heterogeneous elements that helped us operate that otherwise arbitrary acoustic distinction can legitimately be called the *meanings* of the resulting terms. However, two remarks are immediately necessary here so as not to fall back into a classic referential theory of meaning. On the one hand, such meaning relation does not hold between words and things (or states of affairs), but between the discriminating power of a field of experience and the acoustic continuum going from /p/ to /b/ (and not from "plead" to "bleed"). The word "plead" is at the crossroads of a multiplicity of those sub-lexical discriminating effects (such as the ones resulting from the opposition between "plead" and "played" or between "plead" and "please") and its meaning is not given by the aforementioned situation in court, as a state of affairs, but by the set of all those discriminating directions.

On the other hand, unlike a referential conception of meaning, while an heterogeneous field of experience helps operate a distinction in the acoustic continuum, the opposite is also true. For instance, in the presence of two initially indistinguishable water courses, the English utterances "river" and "stream"—already distinguished through mechanisms related to their sub-lexical units, while not necessarily to their respective meanings—will help establish a relevant polarity in experience which would be otherwise non-existent, as it is indeed the case for other languages. The discriminating action between the two systems does therefore not take place unidirectionally (from meanings to utterances), as in the case of standard referential theories of meaning, but operates simultaneously in both ways. This is why Saussure prefers the words "signifier" and "signified" instead of utterance and meaning to qualify those two systems as internal aspects of the same complex phenomenon (Saussure 1959, p. 65 sq.). In this respect, his disciple Louis Hjelmslev is right to characterize the distinction between the two systems or "planes" as a purely formal one, calling those planes "expression" and "content" arbitrarily, without regard to any substantial consideration (Hjelmslev 1953, §13). Hjelmslev's generalization has the additional advantage of considering under one and the same framework the discriminating action between pairs of systems of any kind, for instance graphical and acoustic, in which the graphic characters "g" and "c" help operate the distinction between /G/ and /K/, and sounds such as /b/ and /d/ permit to dissipate the possible confusion between the symmetric graphics "b" and "d".

By taking into account the discriminating action of another system in our original series of linguist phenomena, relevant linguistic units can finally be identified at different levels of language. Yet the most important thing to us is that, by doing so, *we are also revealing fundamental aspects of the meaning of those units*, since that meaning is nothing but the effect, in another field of experience, of the discriminating principles implied in the very procedure of derivation of those linguistic units.

Due to the singular way in which they are established, units are not standalone identities, but differential or oppositional *values* at the crossroads of a whole series of discriminating axes. We have seen the same idea acting in Harris's formal procedure for the definition of phonemes. Saussure provides a more intuitive illustration of the underlying dimensions that contribute to the delimitation of a lexical unit, namely the French word "*enseignement*" (corresponding to the English noun "teaching") in Fig. 9. It is impossible not to recognize in Saussure's diagram the mechanism at play

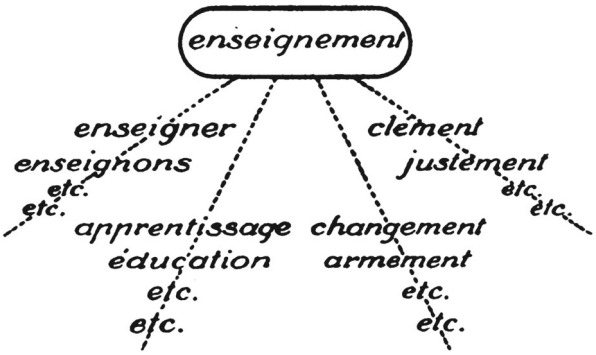


**Fig. 9** Saussure's diagram of the underlying dimensions contributing to the delimitation of the lexical unit "*enseignement*" (Saussure 1995, p. 175)

in embedding spaces as we characterized them earlier (p. 38) through all the analogical directions identifiable at the intersection of the word "house" (i.e., "houses", "housed", "bungalow").

It should not be surprising if, here again, we find duality and bi-duality as the basic relations through which those axes can be defined. Indeed, if we come back to our example of "plead" and "bleed", we see that, as the result of that heterogeneous polarization of the continuous acoustic matter, we now have the means to distinguish the sound /p/ at once from /b/ and from /li:d/. In other terms, we have acquired the capacity of defining the relation of duality /p/ maintains with /li:d/ and the bi-duality that relates /p/ to /b/ through their shared duality with /li:d/. These are precisely what Saussure respectively calls syntagmatic and paradigmatic relations, which define in his view the fundamental mechanism of language, not only at a phonological but at every level (Saussure 1959, II, ch. VI).

We can understand now why Sahlgren's appeal to Saussurean structuralism is conceptually flawed: syntagmatic and paradigmatic relations cannot rely on words, sentences, passages, or documents to specify the organization of meaning because those linguistic units are not given before the syntagmatic and paradigmatic relations into which they will enter. On the contrary, those relations are introduced precisely as means to establish the relevant units at work within a language. This is not to say that MMs bear no connection whatsoever with syntagmatic and paradigmatic relations. Quite the opposite, such relations are almost all those models implement. But they are not what those models produce in result, based on existing units. They are what inform the mechanisms of the model in order to derive the implicit differential dimensions by which the relevant linguistic units can be established in the end. More precisely, the relation between terms and contexts corresponds to what Saussure calls syntagmatic relations.[74] The global organization of those relations will then

[74]Incidentally, MMs help to understand that syntagmatic relations are not symmetrical but connect saturated terms with unsaturated ones—an idea that is absent in Saussure's *Course*. Hjelmslev's functional reformulation of structuralism will make this asymmetry explicit.

contribute to establishing paradigmatic relations both at the level of terms and of contexts, the co-determination of which will result in relevant discriminating dimensions giving shape to the implicit structure that governs the existence of significant units of language.

Interestingly, for Saussure, structural properties are built upon a specific configuration of syntagmatic and paradigmatic relations that he calls "analogy," covering the exact same phenomena we have examined in the previous section (Saussure 1959, II, ch. IV). Yet in the general framework set up by Saussure's structuralism, analogies are not just an external fact verifiable at the level of language, but the very procedure by which language organizes its meaningful regularities against the irrationality introduced by linguistic use. The following example is prototypical in this sense:

> The nominative form of Latin *honor*, for instance, is analogical. Speakers first said *honōs : honōsem*, then through rhotacization of the *s*, *honōs : honōrem*. After that, the radical had a double form. This duality was eliminated by the new form *honor*, created on the pattern of *ōrātor : ōrātōrem*, etc., through a process which subsequently will be set up as a proportion:
>
> *ōrātōrem : ōrātor = honōrem : x*
>
> *x = honor*
>
> Thus analogy, to offset the diversifying action of a phonetic change (*honōs : honōrem*), again unified the forms and restored regularity (*honor : honōrem*). (Saussure 1959, p. 161)

It appears that, from a structuralist point of view, analogical relations in language are not a surprising feature of a formal account of language as it seemed to be for NN word embedding models, but on the contrary, the general form under which language identifies and organizes its elementary units. Yet analogy is only the local manifestation of such organization, related to the way individual speakers practice a language. The underlying structure is made of multiple interconnected discriminating dimensions from which concrete units at all levels (phonological, morphological, lexical, etc.) borrow their syntactic identity and their semantic effect as the result of a combination of distinctive features.

No one has dedicated more efforts than Hjelmslev to draw the formal conceptuality of such latent structure and to specify the procedures through which it could be practically derived in the scientific study of language. A brief account of his conception can then indicate the way in which the structuralist perspective can contribute to grasping the underlying structure that embedding spaces seem to manifest.

For Hjelmslev, as for Saussure, the aim of linguistics is to derive the regularities underlying linguistic utterances by which linguistic units of all levels could be identified as oppositional entities stemming from a continuous substrate. Yet he will mainly put the accent on the systematic character of such regularities. To this end, one of the key components of his theory is given by a unified formal procedure to describe any linguistic *category*, such as grammatical case or comparison. Categories are composed of *members* related to each other by a complex structure of oppositions. For instance, the category of case in German is composed of the interrelated members: nominative, accusative, dative, and genitive, while the category of comparison in English contains the three members: comparative, superlative, and positive.

Members of categories contribute to determining the form and the content of specific linguistic units (e.g., "better", "best", and "good" in the case of the English comparative), such that no significant linguistic unit can be identified without identifying at the same time properties of the categories that unit manifests.

The structure of any category ultimately depends on a continuous "conceptual zone'—as Hjelmslev sometimes calls them—specific to each category, the particular occupation of which characterizes the latter's members.[75] As we have revealed for MMs, the challenge of linguistic analysis within Hjelmslev's framework is to provide the general means to describe the complex relations of opposition structuring any category in terms of those highly unstructured conceptual zones. And for Hjelmslev as well, the solution will be to conceive a system of discriminating axes as a sort of intermediary layer between conceptual zones and categories. Thus, on the one hand, a small number of dimensions (empirically never more than three) will be drawn to describe the conceptual zone; on the other, each member of the category will be described in terms of the possible values of discriminating axes built upon those dimensions. For instance, if a certain conceptual zone is described by three dimensions, and if $\alpha$ is a possible value of the axis built upon the first of them, $\beta$ a value of the second and $\gamma$ of the third, then a possible member of that category could be characterized by the combination $\alpha\beta\gamma$, acting as a kind of discrete coordinates determined by the three axes.

For a particular axis, possible values can be understood as the result of a partitioning of the axis. Values also reflect a specific way of occupying the corresponding dimension in the conceptual zone. However, the key point of this construction is that *the partitioning of the axis does not correspond to a partitioning of its corresponding dimension*. The whole interest of intermediate axes is to provide a partitioning over which a structured category can be built, without equally partitioning the conceptual zone, so that the members of linguistic categories can be described as a system of overlapping or non-exclusive partial coverings of conceptual zones.

Take the example of the English comparative. As stated above, the members of this category are the comparative, manifested by the suffix "-er" on adjectives[76] (e.g. "better", "lower", "broader" ), the superlative, manifested by the suffix "-est" (e.g. "best", "lowest", "broadest" ) and the positive, corresponding to the absence of suffix (e.g. "good", "low", "broad"). For Hjelmslev, the description of this category requires only one dimension, which he characterizes as "intensity" (i.e. the intensity with which those adjectives are attributed to nouns[77]). Intuitively, we would be

[75] While in his early work Hjelmslev thought of those conceptual zones in semantic terms, his theory evolved into a purely formal conception based on abstract notions of correlation, participation and exclusion. See Hjelmslev's (1975, §Gb3.1) and Herreman's introduction to this difficult text (Herreman 2011).

[76] We presuppose here the existence of adjectives, although such a class should in principle be structurally derived as well, partly based on the fact that the terms that fall under it are subject to a comparative category, following the basic mechanisms of (bi)-duality.

[77] Here again, the class of nouns is presupposed. Its actual inference would in part stand on the possible relation to the class of adjectives. As we already mentioned, this semantic characterization of comparison in terms of intensity will be abandoned in Hjelmslev's later works, in favor of purely formal definitions.

inclined to make each member of the category correspond to a particular degree of intensity (weak, strong and neutral, respectively), and hence to a particular region of the conceptual space described by the intensity dimension. However, it is easy to verify that this is not what the members of the category usually mean in *actual* utterances. For instance, the English superlative can be sometimes used to express a weak intensity, like the comparative, as in "of these two, which is best?"; likewise, the positive, neutral in principle, can also express weak intensity, as when a room is called "the small room" only because it is smaller than others, without actually being small.[78] Phenomena such as these are so pervasive in natural language that a perfect correspondence between members of a category and regions of the corresponding conceptual zone would continuously force us to denounce the ambiguities and abuses of natural language, with the consequent shift from a descriptive to a normative attitude.

Hjelmslev's intermediary axes avoid that pitfall. For Hjelmslev, dimensions of conceptual zones are divided into *fields*: positive (+), negative (÷) and neutral (0) (e.g. strong, weak and neutral, in the case of the intensity dimension of comparison). Yet, fields are not exclusive, therefore a category member can cover more than one field. Since categories are nevertheless structured as oppositions, different forms of oppositions will be defined out of non-exclusive ways of occupying those fields. In its most simplified version, three basic forms of oppositions can be defined: contrariness, contradiction and a form of opposition which is neither contrariness nor contradiction and which is, for Hjelmslev, the simplest and most common in natural language. The latter is given by two values, which Hjelmslev notes $\alpha$ and A, the first of which corresponds to occupying only one field of the corresponding dimension (usually the positive one), while the second can occupy any of the three fields. The English grammatical case is an example of a category structured by this opposition. Such category has only two members: the genitive ($\alpha$), occupying a precise region of the corresponding dimension and marked by the morpheme "-s" (e.g. "John's book") and the non-genitive (A), which is unmarked, and can occupy any position, including that of the genitive (e.g. "The book of John"). The other two forms of oppositions are complex, which means that, for each value, one or two fields are necessarily occupied while the others may be occupied as well, but not necessarily. In the contrary opposition, the values—noted β and B—relate to each other as the necessary occupation of the positive and the negative field respectively (with the possible but not necessary occupation of the remaining fields in each case); the contradictory opposition appears, in turn, as the simultaneous necessary occupation of the positive and the negative field ($\gamma$) against the necessary occupation of the neutral (Γ), with possible occupation of the remaining fields. Figure 10a shows the diagram provided by Hjelmslev to illustrate these relations.

[78] This is not necessarily the case for other languages, in which the comparative category is otherwise structured. See Hjelmslev (2016, §7).

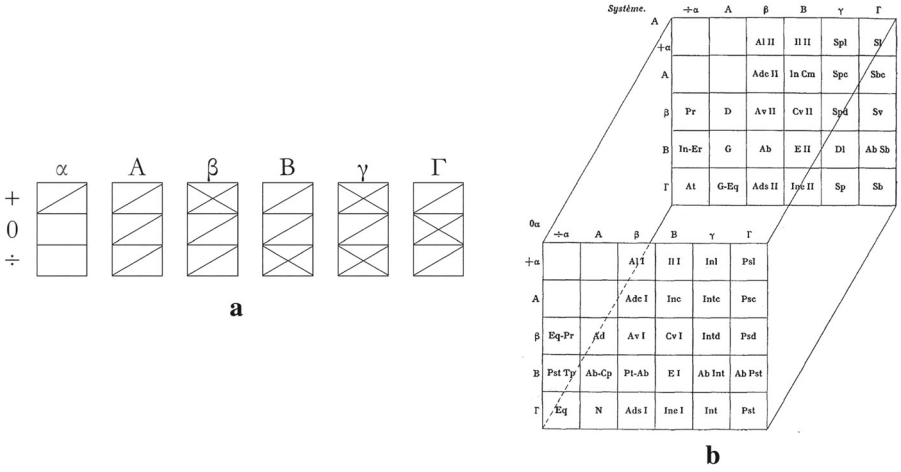


**Fig. 10** **a** Possible values for discriminating axes (neither contrary nor contradictory = $\alpha$−A, contrary = β−B, contradictory = $\gamma$−Γ) in terms of the occupation of the fields of a conceptual zone (+ = positive, 0 = neutral, ÷ = negative). **b** Case category of the Tabassaran language (Hjelmslev 1935, p. 146)

In this way, three pairs of oppositional values are identified as possible discriminating values for axes: $\alpha$−A; β−B and $\gamma$−Γ. However, an axis can be partitioned into more values than those given by only one pair. In the extreme case, it can hold all six of them.[79] As already stated, it is through these values that members of categories will be finally characterized. If we return to the example of the English comparison, the three members of the category (comparative, superlative and positive) can be now respectively characterized through the axis partitioned into the values β, B and $\gamma$, all of which are defined in terms of different possible occupations of the fields +, ÷, 0 of the intensity dimension. Those three values can now adequately grasp the overlapping uses of the linguistic units corresponding to the members of the English comparative, precisely qualifying, instead of denouncing, their apparent ambiguity.

The English comparison is a fairly simple category, with only three members. More complex categories, with a larger number of members can however be described in the same way, increasing the number of dimensions if necessary. Thus, if a category has six members or less, it can in principle be described by one axis (like the aforementioned category of case in German, described by the four values β, B, $\gamma$, Γ) although other descriptions with more axis are also possible. The description of categories containing more than six members require more than one dimension, and as we said, Hjelmslev never found the empirical need to appeal to more than three dimensions. In the latter case, counting six values at most for each dimension, the

[79] Due to internal dependencies between the six values (which tend to respect the original pairs), not every combination is actually possible. Hjelmslev identifies a total of seven, defining axes of two up to six values: $\alpha$, A; β, B, $\gamma$; β, B, Γ; β, B, $\gamma$, Γ; $\alpha$, A, β, B, $\gamma$; $\alpha$, A, β, B, Γ and $\alpha$, A, β, B, $\gamma$, Γ.

maximum number of possible members is $6^3 = 216$. Within configurations of this sort, not all the possible combinations of values need to be realized as members of a category. Figure 10b shows the configuration of the category of case of the Tabassaran language, the largest encountered by Hjelmslev, defined by three axes with five, six and two possible values respectively, only 52 of the 60 possible combinations of which are actually realized in the language.

In this manner, Hjelmslev provides an explicit theory of the way in which discriminating axes structure the latent space that underlies the distribution of linguistic units and conditions their semantic effect. Since that latent structure constitutes the fundamental principle by which units can be identified as such,[80] Hjelmslev's perspective offers a suggestive connection between Saussure's conception of language and the contemporary linguistic models we have studied. Therefore, while the compatibility and fruitfulness of Hjelmslev's approach with respect to present-day models is still an empirical question, from a philosophical perspective his conceptual efforts can nevertheless help understanding the mechanisms that guarantee the success of those models and even suggest new research orientations.

More generally, all the elements belonging to classic structuralist linguistics that we have exposed achieve the image of language that we were retrospectively reconstructing, by allowing us to think the way meaning can emerge as a direct effect of the formal organization of language itself.[81] From the structuralist perspective, that formal organization of language appears as a complex structure of discriminating operators resulting from a mutual segmentation process between two systems. Insofar as our analysis relies on the distributional properties of one of those systems (utterances, signifiers, expressions, sounds, etc.) and extract the multiple discriminating dimensions that determine the categories organizing its units, we will be capturing significant distinctions of the other system (meaning, signified, contents, thought, etc.) which are carved in the former as its deep imprint. If we accept to generalize the reference to "thought" and "sound" to any pair of heterogeneous systems, the resulting image of language could be condensed in Saussure's formula: "The characteristic role of language with respect to thought is not to create a material phonic means for expressing ideas but to serve as a link between thought and sound, under conditions that of necessity bring about the reciprocal delimitations of units." (Saussure 1959, p. 112). That is what we can call the *structuralist hypothesis*: the idea that distributional properties convey meaning only through the action of a latent structure

[80] Although we presented the elementary mechanisms of Hjelmslev's theory through mostly morphological categories, their formal definition could, in theory, be applied to any linguistic level in which a latent structure is to be drawn as the underlying principle of identification and characterization of significant units.

[81] This does not imply that the complex phenomenon of meaning is reducible to the linguistic mechanisms by which language organizes its semantic effects, since a theory of meaning could hardly avoid the difficult question of the contribution of perception to meaning. The structuralist perspective just exposed suggests already a possible treatment of this question when it appeals to an heterogeneous plane as a condition for establishing discriminating criteria for linguistic units, since that other plane could in principle be related to perceptual properties. But the role played by perception in this case is still far from being clear.

determining possible semantic values, and which is inseparable from the principles of identification of the elementary units of language, since meaning is the effect of discriminating operations performed through segmentation procedures of which the units of language keep the trace.

## 6 Conclusions

Before summarizing the principal features of the image of language that best conforms to the deeds of NN word embeddings, it is worth recalling the main steps that guided us through its reconstruction.

After presenting the basic functioning and most promising results of word embeddings through its pioneering model, word2vec, we recognized their conceptual and philosophical significance in three main circumstances. First, the idea that the automatic reconstruction of the underlying organization of language does not require more human intervention than the one implied in the most ordinary use of language as recorded in a practically *raw linguistic corpus*. Second, the evidence that in that reconstruction both *semantic and syntactic contents* of words are determined *at once and as the result of the same procedure*. Third, the perspective that word vector representations are not simply disposed in similarity neighborhoods, but that the vector space itself is also *structured following precise directions* at the crossroads of which syntactic and semantic contents are established. It is the conjunction of those three circumstances, namely that linguistic structure defining semantic and syntactic content is successfully derived through an unsupervised automatic procedure relying exclusively on bare linguistic performances, that suggests that NN word embeddings are not just a successful new technology, but convey an altogether novel image of language.

As a first step to unravel that image, we focused on the underlying mechanisms upon which word embedding models stand. We found then that those mechanisms where essentially attached to the *factorization of a term-context matrix*, that is, a matrix collecting (an information-theoretical transformation of) the frequency in which linguistic terms appear in the context of other terms. We were thus led to examine those matrix models and the effects of factorization, and identified that their fundamental mechanism was capable of drawing a latent organization underlying observable linguistic units, defined by *multiple dimensions of discriminating action* between groups of terms and groups of contexts. We also showed how the action of those implicit or latent dimensions was irreducible to explicit representations in their capacity of mobilizing the *global organization of the space* to determine the relation between terms or between contexts, thus allowing in principle to connect terms that do not share any explicit context.

However, once we identified such a mechanism as the basis of word embeddings, the reasons for it to succeed in grasping so many aspects of linguistic content were still to be addressed. This necessitated turning our attention from the technical aspects of word embedding models to that of the image of language itself. In other words, the question "why can computers understand natural language?" required that we stop interrogating computers and we turn to natural language instead, to raise another

question, namely: "what must natural language be for those mechanisms to succeed in the way they do?".

To that question, the literature around NN word embeddings only provided a frail answer, under the name of "distributional hypothesis", stating that "words that occur in similar contexts tend to have similar meanings". We turned, therefore, to the treatment of this question that accompanied the development of previous (non-NN) MMs, and found that the most salient versions of the distributional hypothesis appealed to a "use theory of meaning" as an attempt to attribute explanatory power to the notion of linguistic context conventionally attached to distributionalism. We suggested then that all those attempts shared a common interpretation of the connection between distributionalism and a usage-based perspective, namely, the idea that the distributional properties of linguistic units are somehow correlated with the cognitive faculties of individual agents, owing to the position of contexts—whether linguistic or extra-linguistic—conceived as nothing more than the restricted conditions of the exercise of those faculties. Within this setting, the notion of co-occurrence played a decisive role as the support of the cognitive operations (of association or abstraction) of subjects over the homogenous units appearing in those restricted contexts.

However, that image conflicted in several points with the elementary mechanisms of MMs we had identified. First, most results of MMs could not be ascribed to direct co-occurrence. Yet the alternative accepted idea of shared co-occurrence failed to capture two essential properties of the mechanisms of MMs, namely: the *fundamental disparity between terms and contexts* and the *non-trivial organization of contexts themselves*. The first one led us to understand the relation between terms and contexts as a functional one, that is, as a relation between saturated and unsaturated expressions. The second one raised the question of the necessary mutual conditioning between the similarity of terms and that of contexts, asking for much more complex explanatory principles that those proposed by the usual images of language attached to distributionalism. Based upon those aspects, we advanced that distributionalism is less about co-occurrence of words in context than about *simultaneous and articulated discrimination* between terms and contexts. The figure of *analogy* corresponding to it provided then a possible alternative to that of co-occurence to qualify that mechanism from the perspective of individual speakers. However, drawing from the formal principles governing the mechanisms of MMs, and in particular, from the dual spaces corresponding to row and column vectors of matrices, we proposed a formal characterization of that articulated discrimination by resorting to the notion of *(bi-)duality*: terms and contexts stand in relations of duality in such a way that the relatedness between two terms is established through bi-duality, that is, through the duality the first maintains with the dual contexts of the second. We also suggested that such a mechanism could be pushed further by considering the duality between a term and the type of its dual contexts established by bi-duality.

As a consequence of this characterization, the notion of context appeared as a strictly *formal* one, less related to the cognitive capacities of individuals than to an internal dimension of the organization of language itself. The conceptual framework for the analysis of language is thus freed from the figure of individual subjects since language is less about subjective use than about relatively autonomous regularities in the global organization of language before which individuals are left with the

possibility of a strategy. This provides a plausible explanation of why word embedding models are capable of deriving significant linguistic content without the need for any more human intervention than the one implied in raw linguist corpora.

To address the other two initial questions of our inquiry, we noted that if the joint derivation of semantic and syntactic properties might seem surprising, that should not be attributed to a new capacity of NN models but to a neglected possibility within MMs due to the mainly semantic orientation governing their development. We identified *three epistemological obstacles associated with that semantic orientation*, namely the *almost complete exclusion of syntax* from the original interests of MMs, the *privilege granted to words as fundamental units of linguistic analysis* and the *disregard for possible underlying structures of language* in favor of an organization of the linguistic space in terms of similarities. We showed then how the first two aspects were intimately connected: accounting for syntactic content in the framework of MMs requires to consider terms and contexts at sub- and supra-lexical levels, and conversely, introducing the latter naturally induces the establishment of syntactic regularities. We suggested then that, if NN word embeddings were able to grasp semantic and syntactic contents as the result of one and the same procedure, the reason might reside in their capacity of taking advantage of this neglected possibility of MMs.

The abandon of the lexical privilege and the corresponding consideration of syntactic content were for us the principal hint to understand why the embedding space can be informed by an emergent structure. To follow that lead, we entered into "structuralist soil" by examining Harris's original distributionalism. In accordance with our interpretation, Harris's formulations show that, given that the very sense of a distributional description is to free the scientific account of language from the external viewpoint of semantics, the link between the analysis of linguistic distribution and the derivation of structural features is necessary. Indeed, outside a semantic approach, lexicality has no privilege for distributional analysis, which takes place essentially at the phonological and morphological levels. Unconcerned with semantics and standing on phonological or morphological features, *distribution is necessarily structural distribution.* Significantly, we found that, despite a natural lack of technical complexity, the key analytic instruments of Harris's distributionalism are highly convergent with the mechanisms of modern MMs, namely a term-context matrix with dimensionality reduction techniques. However, we saw that the latent space derived through those techniques provides a complete series of formal units structurally related to each other rather than a set of dimensions to measure approximate similarity. Harris's theory was not free of difficulties, nonetheless. Indeed, by making such a clear-cut distinction between semantics and distributional features, Harris precluded the possibility of conceiving a direct relation between those structured formal units and any semantic content. This is what led us to conclude our inquiry by assessing the structuralist theoretical background of Harris's distributionalism.

We found the answer to that last open issue in an idea stemming from the singular approach to language of European structuralism (i.e. that of Saussure and Hjelmslev), which we called "the structuralist hypothesis", namely the idea that *meaning is the effect of structure*. Such hypothesis is based on the fact that linguistic structure is not the result of the composition of units but of unmotivated *decomposition*

*or segmentation* of an originally shapeless material continuum. Such segmentation can only be done through the intervention of a second heterogeneous continuum operating a discriminating action on the first one, while receiving from the latter a comparable action in return. Structure is then always double, since it is the result of a process of simultaneous segmentation between two heterogeneous systems or planes, of which one can be thought of as the content or meaning of the other. This is why the analysis of segmentation of one plane—which is what distributional analysis in fact achieves—can capture meaningful features belonging to the other one, carved, as it were, in the form of the former. Moreover, an overview of Hjelmslev's development of Saussure's program suggested the way in which that structuralist conception of language could be related to the basic formal principles of the contemporary models under study.

From word2vec's word contexts and implicit matrix factorization to Saussure's syntagmatic and paradigmatic relations and Hjelmslev's categories, through MMs, SVD and Harris's element-environment charts, we have found the action of one and the same mechanism underlying the capacity of linguistic analysis to grasp significant features of language: the *duality between terms and contexts*, and a principle of *equivalence between terms and between contexts related through bi-duality*. Interestingly, that mechanism tends to characterize bottom-up approaches to language, that is, approaches in which the organization of language is not supposed to be given from without, but derived from within the practice of language itself, as an emergent structure. NN vector representations (whether it is of words, characters or contexts) appear then only as the latest tool capturing the effects of that mechanism of duality and bi-duality in language. Compared to previous perspectives of the same nature, its importance lies in that, with almost minimal presupposed structure (let alone computational costs), it can account for those effects as something more than mere semantic similarity satisfying most needs of MMs; but also, in that it can actually provide a surprising amount of semantic content, the derivation of which classic structuralist theories could, at best, only conjecture.

As we have suggested, that basic mechanism of (bi-)duality is to be attributed to language itself rather than to the linguistic models studying it. This was a common conception among the authors we have studied, from Landauer affirming that the model "is the real-thing" (Landauer et al. 2007, p. 8), to Harris stating that "the position of the speakers is after all similar to that of the linguist" (Harris 1970, p. 779), to Saussure, for whom, the units of language "are not given from the outside, language must draw them from within itself" (quoted in Maniglier 2006, p. 207, our translation). This is why we need to think that embeddings not only provide a successful technique, but also convey a whole image of language, or in other terms a general but elementary conception of what language is, of what it means to speak and write.

Animated by (bi-)duality as a fundamental principle governing the linguistic practice of speakers, language appears more like a *game* than like a tool. (Bi-)duality and the regularities attached to it are in this sense like the rules of chess or of Go, which are independent of the will of individual players, who cease to be players of those games as soon as they stop submitting to those rules. Rather than an instrumental use, subjects develop a *strategy within language*, which far from violating its regularities,

reinforces them in a way models like word embeddings and MMs in general can efficiently identify. The elements of that game are not to be confused with the words as elementary material units whose identity is determined by their referential meanings, since both the identity and the meaning of words are the result of the fundamental duality between terms and contexts, as two internal dimensions of linguistic *formal units*. Yet units are not necessarily lexical, and against a purely semantic approach of meaning, we were confronted with only a difference of degree between the syntactic and the semantic levels. Finally, as units are not given from the outset, linguistic practice and the production of meaning is less the result of a composition of words than the effect of the simultaneous *segmentation* of pairs of material continuums into formal units of different levels. From this perspective, if the linguistic subject computes, it is a strategy of segmentation rather than a compositional rule that it computes.

(Bi)-duality rather than co-occurrence, game rather than tool, strategy rather than use, form rather than substance, segmentation rather than compositionality. This is the image of language that we have been able to draw behind word embeddings by tracing back the meaning of its underlying mechanisms. That image contrasts with many of the principles generally orienting the present development of the very field that begets it. Our hope is that, by bringing that underlying image to the surface and making explicit its connections with the tradition from which it stems, we will contribute to providing novel orientations to the field, in adequacy with the nature of its latest advances. In particular, the image of language we have derived can allow us to relativize the predominantly semantic perspective of the field in favor of the derivation of complex interconnected structures at different linguistic levels. In this sense, one of the clearest hints offered by our inquiry is that a segmentation procedure should be an integral part of the process by which structural features are derived, and more precisely, of the construction of the representation of the elementary units of language. Since, as we have seen, structural features are intimately connected with the segmentation of units at all levels, the consideration of segmentation for the construction of vector representations should help capturing the structural features those vectors will be expected to represent. By presupposing the segmentation of words and discarding other possible levels of segmentation, classical word embedding models rely only on partial information for the reconstruction of linguistic structure. Incidentally, that presupposition also relativizes the unsupervised character of the learning for, as insignificant as it may seem, word segmentation already involves a great amount of linguistic content (that is precisely the content word embeddings are able to grasp). By including segmentation in the analytic procedure, one would be contributing to achieving a truly unsupervised model, while increasing its capabilities at the same time, since the structure of the segmentation bears a direct relation with the linguistic structures the model is intended to capture. The recent orientations of the field, obtaining remarkable results with character and context-based embeddings, tend to confirm this idea.

Yet the consequences of reconstructing the alternative image of language implied in the new NN language technologies can transcend the limits of the NLP field, and attain other regions of knowledge and research, such as social sciences, epistemology and even logic or computer science, in which the notion of language occupies a crucial position. With respect to those fields, the emergence of a new image of

language can provide the opportunity of significant conceptual shifts which could be, moreover, accompanied by a fruitful transfer of methodologies, introducing in those fields analytical tools from NLP and enriching the latter with the specific treatment language receives in each of the former. In this respect, we need to recognize a major stake in the possibility, provided by that new image, of understanding language and meaning independently of purely individual practices, and yet not resorting to absolute principles valid without restriction that would make the analysis of actual practices inconsequential. At equal distance from those two positions, the image of language we have reconstructed in the previous pages allows us to envisage language as a collective playground, as a reservoir where the significant distinctions resulting from a collective construction of signs are deposited as the most intimate treasure of a *culture*—to borrow a Saussurean metaphor. It is, after all, the image of those cultures that models like word embeddings give us the means to depict.

**Acknowledgments** The author wishes to thank the reviewers for their insightful remarks and generous suggestions.

**Funding Information** This work was partly funded by the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 839730.

## References

Baroni, M., & Lenci, A. (2010). Distributional memory: a general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721. https://doi.org/10.1162/coli_a_00016.

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 (Volume 1: Long Papers)* (pp. 238–247): Association for Computational Linguistics, https://doi.org/10.3115/v1/P14-1023.

Bengio, Y. (2008). Neural net language models. *Scholarpedia*, *3*(1), 3881. https://doi.org/10.4249/scholarpedia.3881. Revision #140963.

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blitzer, J., Weinberger, K., Saul, L.K., Pereira, F.C.N. (2005). Hierarchical distributed representations for statistical language modeling. In *Advances in neural and information processing systems*, Vol. 17. Cambridge: MIT Press.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2016). Enriching word vectors with subword information. arXiv:abs/1607.04606.

Bullinaria, J.A., & Levy, J.P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods*, *44*(3), 890–907. https://doi.org/10.3758/s13428-011-0183-8.

Bybee, J.L., & Hopper, P.J. (Eds.) (2001). *Frequency and the emergence of linguistic structure. No. 45 in Typological studies in language*. Amsterdam: Benjamins. OCLC: 216471429.

Carnap, R. (2001). *Logical syntax of language*. Routledge. OCLC: 916122384.

Caron, J. (2001). *Computational information retrieval. chap. Experiments with LSA scoring: optimal rank and basis*, (pp. 157–169). Philadelphia: Society for Industrial and Applied Mathematics.

Chater, N., Clark, A., Goldsmith, J.A., Perfors, A. (2015). *Empiricism and language learnability*, 1st edn. Oxford: Oxford University Press. OCLC: ocn907131354.

Chen, J., Tao, Y., Lin, H. (2018). Visual exploration and comparison of word embeddings. *Journal of Visual Languages and Computing*, *48*, 178–186. https://doi.org/10.1016/j.jvlc.2018.08.008.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning, ICML '08* (pp. 160–167). New York: ACM. https://doi.org/10.1145/1390156.1390177.
Dahl, G.E., Yu, D., Deng, L., Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 30–42. https://doi.org/10.1109/TASL.2011.2134090.
Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.
Deleuze, G. (1994). *Difference and repetition*. New York: Columbia University Press.
Dennis, S., Landauer, T., Kintsch, W., Quesada, J. (2003). Introduction to latent semantic analysis. Slides from the tutorial given at the 25th anual meeting of the cognitive science society. Boston.
Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:abs/1810.04805.
Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 213–252.
Elman, J.L. (Ed.) (1996). *Rethinking innateness: a connectionist perspective on development. Neural network modeling and connectionism*. Cambridge: MIT Press.
Firth, J.R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford: Blackwell.
Frege, G. (1984). Function and concept. In McGuinness, B. (Ed.) *Collected papers on mathematics, logic, and philosophy, pp. 137–156. Basil Blackwell*.
Ghomeshi, J., Jackendoff, R., Rosen, N., Russell, K. (2004). Contrastive focus reduplication in english (the salad-salad paper). *Natural Language & Linguistic Theory*, *22*(2), 307–357. https://doi.org/10.1023/B:NALA.0000015789.98638.f9.
Girard, J.Y. (2001). Locus solum: from the rules of logic to the logic of rules. *Mathematical Structures in Computer Science*, *11*(3), 301–506.
Gladkova, A., Drozd, A., Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *SRW@HLT-NAACL* (pp. 8–15): The Association for Computational Linguistics.
Glenberg, A.M., & Mehta, S. (2008). Constraint on covariation: it's not meaning. *Special issue of the Italian Journal of Linguistics*, *1*(20), 241–264.
Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.
Hale, B., & Wright, C. (1997). *A companion to the philosophy of language*. Oxford: Blackwell Publishers.
Hamilton, W.L., Leskovec, J., Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. arXiv:abs/1605.09096.
Harris, Z. (1960). *Structural linguistics*. Chicago: University of Chicago Press.
Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics* (pp. 775–794). Dordrecht: Springer.
Herreman, A. (2011). Analyser l'analyse, décrire la description. *texto! Textes & Cultures*, *16*, 2. http://www.revue-texto.net/index.php?id=2875.
Hewitt, J., & Manning, C.D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4129-4138). Minneapolis: Association for Computational Linguistics, https://doi.org/10.18653/v1/N19-1419.
Hinton, G.E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth annual conference of the cognitive science society* (pp. 1–12). Hillsdale: Erlbaum.
Hjelmslev, L. (1935). *La catégorie des cas*. Munchen: Wilhelm Fink.
Hjelmslev, L. (1953). *Prolegomena to a theory of language*. Baltimore: Wawerly Press.
Hjelmslev, L. (1975). *Résumé of a Theory of Language. No. 16 in Travaux du Cercle linguistique de Copenhague*. Copenhagen: Nordisk Sprog-og Kulturforlag.
Hjelmslev, L. (2016). *Système linguistique et changement linguistique*. Paris: Classiques Garnier.
Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22Nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '99* (pp. 50–57). New York: ACM. https://doi.org/10.1145/312624.312649.
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1–2), 177–196. https://doi.org/10.1023/A:1007617005950.
Howard, J., & Ruder, S. (2018). Fine-tuned language models for text classification. arXiv:abs/1801.06146.
Google inc. (2013). word2vec, https://code.google.com/archive/p/word2vec/.

Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition. In *Proceedings of Interspeech 2012*.

Jansen, S. (2017). Word and phrase translation with word2vec. arXiv:abs/1705.03127.

Jurafsky, D., & Martin, J.H. (2008). *Speech and language processing*, 2nd edn. Upper Saddle River: Prentice Hall.

Krivine, J.L. (2001). Typed lambda-calculus in classical Zermelo-Fraenkel set theory. *Archive for Mathematical Logic*, *40*(3), 189–205.

Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S. (2014). Statistically significant detection of linguistic change. arXiv:abs/1411.3315.

Landauer, T.K. (2002). *On the computational basis of learning and cognition: arguments from LSA*, (pp. 43–84): Academic Press.

Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. New Jersey: Lawrence Erlbaum Associates.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. From context to meaning: distributional models of the lexicon in linguistics and cognitive science. *Italian Journal of Linguistics*, *1*(20), 1–31.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254.

Levy, O., & Goldberg, Y. (2014a). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014* (pp. 171–180).

Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th international conference on neural information processing systems - volume 2, NIPS'14* (pp. 2177–2185). Cambridge: MIT Press.

Levy, O., Goldberg, Y., Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *TACL*, *3*, 211–225.

Li, J., Chen, X., Hovy, E.H., Jurafsky, D. (2016). Visualizing and understanding neural models in NLP. In *HLT-NAACL* (pp. 681–691): The Association for Computational Linguistics.

Liberman, A.M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America*, *29*(1), 117–123. https://doi.org/10.1121/1.1908635.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. arXiv:abs/1606.07736.

Liu, S., Bremer, P., Thiagarajan, J.J., Srikumar, V., Wang, B., Livnat, Y., Pascucci, V. (2018). Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 553–562. https://doi.org/10.1109/TVCG.2017.2745141.

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

Luong, T., Pham, H., Manning, C.D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA* (pp. 151–159).

MacWhinney, B. (Ed.) (1999). *The emergence of language. Carnegie Mellon symposia on cognition*. Mahwah: Lawrence Erlbaum Associates.

Maniglier, P. (2006). *La vie énigmatique des signes*. Paris: Léo Scheer.

Maniglier, P. (2016). Milieux de culture. In Beividas, W., Lopes, I.C., Badir, S. (Eds.) *Cem Anos com Saussure, Textos de Congresso Internacional*, Vol. 2. Sao Paulo: Annablume editora.

Manning, C.D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

McEnery, A.M., & Wilson, A. (2001). *Corpus linguistics: an introduction*. Edinburgh: Edinburgh University Press.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013a). Efficient estimation of word representations in vector space. arXvi:abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., Strohmann, T. (2013b). Learning representations of text using neural networks. NIPS Deep learning workshop 2013 slides. NIPS Deep Learning Workshop 2013.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. arXiv:abs/1310.4546.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., Strohmann, T. (2013d). Learning representations of text using neural networks (nips deep learning workshop 2013 slides).

Mikolov, T., Yih, W.t., Zweig, G. (2013e). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 746–751): Association for Computational Linguistics.

Miller, G.A., & Charles, W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28. https://doi.org/10.1080/01690969108406936.

Mnih, A., & Hinton, G. (2007). Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on machine learning, ICML '07* (pp. 641–648). New York: ACM. https://doi.org/10.1145/1273496.1273577.

Morris, C.W. (1938). *Foundations of the theory of signs*. Chicago: The University of Chicago Press.

Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. arXiv:abs/1705.08039.

Nielsen, M.A. (2015). *Neural networks and deep learning*. Determination Press. http://neuralnetworksanddeeplearning.com/.

Österlund, A., Ödling, D., Sahlgren, M. (2015). Factorization of latent variables in distributional semantic models. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 227–231). Lisbon: Association for Computational Linguistics, https://doi.org/10.18653/v1/D15-1024.

Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., Elmqvist, N. (2018). Conceptvector: text visual analytics via interactive lexicon building using word embedding. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 361–370. https://doi.org/10.1109/TVCG.2017.2744478.

Pennington, J., Socher, R., Manning, C.D. (2014a). Glove: global vectors for word representation. In *EMNLP*, (Vol. 14 pp. 1532–1543).

Pennington, J., Socher, R., Manning, C.D. (2014b). Glove: global vectors for word representation. https://nlp.stanford.edu/projects/glove/, https://nlp.stanford.edu/projects/glove/.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv:abs/1802.05365.

Qiu, Y., & Frei, H.P. (1993). Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '93* (pp. 160–169). New York: ACM. https://doi.org/10.1145/160688.160713.

Radford, A. (2018). Improving language understanding by generative pre-training.

Rappaport Hovav, M., & Levin, B. (2015). The syntax-semantics interface. In Lappin, S., & Fox, C. (Eds.) *The handbook of contemporary semantic theory, chap. 19* (pp. 593–624): Wiley-Blackwell. https://doi.org/10.1002/9781118882139.ch19.

Rastier, F. (2001). *Arts et sciences du texte*. Paris: Presses Universitaires de France.

Rastier, F., Cavazza, M., Abeillé, A. (2001). *Semantics for descriptions: from linguistics to computer science*. Chicago: The University of Chicago Press.

Rumelhart, D.E., & McClelland, J.L. (1986). *Parallel distributed processing*. Cambridge: MIT Press.

Sahlgren, M. (2006). *The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. thesis*. Stockholm: Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Special issue of the Italian Journal of Linguistics*, *1*(20), 33–53.

Salton, G. (1971). *The SMART retrieval system: experiments in automatic document processing*. Upper Saddle River: Prentice-Hall.

Salton, G., Wong, A., Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Saussure, F. de. (1959). *Course in general linguistics*. New York: McGraw-Hill. Translated by Wade Baskin.

Saussure, F. de. (1995). *Cours de linguistique générale. Grande Bibliothéque Payot*. Paris: Editions Payot & Rivages.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Schnabel, T., Labutov, I., Mimno, D.M., Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pp. 298–307*.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE conference on supercomputing, supercomputing '92* (pp. 787–796). Los Alamitos: IEEE Computer Society Press.

Schütze, H. (1993). Word space. In *Advances in neural information processing systems 5, [NIPS Conference]* (pp. 895–902). San Francisco: Morgan Kaufmann Publishers Inc.

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

Schwenk, H., & luc Gauvain, J. (2002). Connectionist language modeling for large vocabulary continuous speech recognition. In *International conference on acoustics, speech and signal processing* (pp. 765–768).

Senel, L.K., Utlu, I., Yücesoy, V., Koç, A., Çukur, T. (2017). Semantic structure and interpretability of word embeddings. arXiv:abs/1711.00331.

Sennrich, R., Haddow, B., Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1715–1725). Berlin: Association for Computational Linguistics, https://doi.org/10.18653/v1/P16-1162.

Smilkov, D., Thorat, N., Nicholson, C., Reif, E., Viégas, F.B., Wattenberg, M. (2016). Embedding projector: interactive visualization and interpretation of embeddings. arXiv:abs/1611.05469.

Spence, D.P., & Owens, K.C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*(5), 317–330. https://doi.org/10.1007/BF01074363.

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

Turney, P.D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd international conference on computational linguistics - volume 1, COLING '08* (pp. 905–912). Stroudsburg: Association for Computational Linguistics.

Turney, P.D., & Pantel, P. (2010). From frequency to meaning: vector space models of semantics. arXiv:abs/1003.1141.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V. (2019). Xlnet: generalized autoregressive pretraining for language understanding.