Integrated Computer-Aided Engineering -1 (2018) 1–13 1
DOI 10.3233/ICA-180585
IOS Press

Speech gestural interpretation by applying word representations in robotics

Mario Almagro-Cádiz[a,*], Víctor Fresno[a] and Félix de la Paz López[b]
[a]*Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain*
[b]*Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia, Madrid, Spain*

Abstract. Human-Robot Interaction (HRI) is a growing area of interest in Artificial Intelligence that aims to make interaction with robots more natural. In this sense, numerous research studies on verbal and visual interactions with robots have appeared. The present paper will focus on non-verbal communication and, more specifically, gestures related to speech, which is an open question. With the aim of developing this part of Human-Robot Interaction or HRI, a new architecture is proposed for the assignment of gestures to speech based on the analysis of semantic similarities. In this way, gestures will be intelligently selected using Natural Language Processing (NLP) techniques. The conditions for gesture selection will be determined from an assessment of the effectiveness of different language models in a lexical substitution task applied to gesture annotation. On the basis of this analysis, the aim is to compare models based on expert knowledge and statistical models generated from lexical learning.

Keywords: Human-robot interaction, co-verbal gesture, gestural annotation, word representation, robotic speech

1. Introduction

Recent advances in different areas of computing, including machine learning, natural language processing and computer vision have made it possible to extend robotics to sectors more focused on human interaction, such as education and health. The automation of these services has increased demand for new human-robot interfaces that allow people to communicate directly with robots in a simple and fluid way [27]. These interfaces require the inclusion of non-verbal communication aspects to achieve greater naturalness and speed of transmission [47]. To this end, it is important to incorporate gestures in speech, which is one of the main challenges of mentioned process of human-robot communication.

To date there is still no consensus as to what can be considered a gesture or what properties can be used to categorize it into a taxonomy in robotics; in fact, each author usually defines different types of gestures according to the tasks they are going to perform [32]. Among the positions found, some authors such as McNeill consider that gestures consist of spontaneous movements that are part of the communicator's thoughts [28], while others such as Kendon claim that they are communicative actions with intentionality [15]. In spite of these discrepancies, practically all works found in the relevant literature distinguish between those types of gestures focused on interaction with the environment – deictic and manipulation of objects – and those in synchrony with language – also called co-verbal gestures. This paper focuses on a specific type of co-verbal gesture related to the content of speech, also known as iconic gestures.

The most common approaches to synchronizing motions with speech are based on rules [45]. In their simplest form, these approaches make use of trigger words associated with each available gesture, so the system assumes that if one of these words appears in speech, then it must be responsible for executing the associated movement. Since the most commonly used meth-

*Corresponding author: Mario Almagro-Cádiz, Departamento de Lenguajes y Sistemas Informáticos and Inteligencia Artificial, Universidad Nacional de Educación a Distancia (UNED), Calle Juan del Rosal 16, 28040, Madrid, Spain. E-mail: malmagro@lsi.uned.es.

ISSN 1069-2509/18/$35.00

ods are based on exact matching between speech terms and words that represent gestures, they suffer from a lack of flexibility which limits the scope for improvement in human perception. The fact that a gesture is initiated only when some defined word is detected does not seem to simulate natural behaviour.

In a preliminary study [1], a new methodology was proposed to associate co-verbal gestures (those in synchrony with language) with a text representing the speech of a robot. The main idea was to define the meaning of body expressions through relevant terms, giving the robot the ability to execute a motion by finding any word semantically related to those terms. In this way, co-verbal gestures are not only executed by precisely matching the terms of the definitions, but they are also activated after the detection of any word with a high degree of semantic similarity to those terms.

The purpose of this paper is to extend the above study by introducing and evaluating different language models as part of the semantic similarity calculation module within the proposed methodology, and to implement an architecture based on more concrete components along with it. This extension is intended to compare language models generated from lexical learning based on distributed semantics with language models based on semantic schemes prepared by expert linguists. Both types of models represent alternative approaches to the process of language acquisition: while the former configure the acquisition of the meaning of concepts through the different textual contexts in which they appear – in a similar way to how humans acquire the semantics of words within a language – the latter (semantic schemes based on lexical databases) contain meanings derived from a deep and complex manual process of synthesis. The comparative analysis of both approaches aims to infer which of the models best fit the selection of co-verbal gestures in the context of HRI. To this end, the following research questions are raised:

– Is it more effective for a robot to inductively learn its own semantic representations from a large corpus that provides an example of the use of the language in question, or would the use of semantic structures created by expert linguists performing a meticulous and detailed analysis of the concepts work better when trying to establish semantic similarities between terms?
– Is the effort to create and maintain lexical databases or specialized ontologies necessarily restricted to one domain worthwhile in this context, or is it preferable to delve into unsupervised methods based on processing large volumes of textual data to find the meaning of words within a language?

2. Related work

Traditionally, the scientific community has focused its efforts on investigating the recognition of gesticulations, leaving the process of synthesis in the background. This has been reflected in a small number of gesture interfaces in robotics, as well as in the widespread use of the term “gesture” to refer to the manipulation of objects rather than to non-verbal communication [44]. In turn, gestural interfaces developed in robotics tend to focus on collaborative [38,42] or deictic [11] gestures, with the integration of co-verbal gestures being a relatively unexplored field in this area.

The importance of co-verbal gestures lies in their impact on the perception of meanings, since both sound and body expressions are simultaneously assimilated as a single package [35]. In fact, there are many publications that include studies on the impact of these body expressions on human perception [13].

The absence of physical limitations in the development of body expressions has allowed the synthesis of co-verbal gestures to be a more recurrent line of research in the virtual environment with avatars. Some approaches do not contemplate semantic information, but have focused on the use of prosody, simplifying the task to the analysis of metrics extracted from the form of speech [6,23]. The most widespread approaches are those based on rules, which are generally founded on the establishment of mappings between gestures and sets of textual features from a bag of words. Some examples of these approaches are the *GRETA* agent [34], which uses gesture repositories, and the *MAX* agent [18], which is based on speech-gesture pairs. Both Lee and Marsella [22] and Tepper et al. [49] associate lexical, syntactic and semantic information with motions, while Kipp et al. use probabilistic rules [17]. The *BEAT* system [5] manages to group body motions and speech, using a set of heuristic rules according to different types of gestures.

Data-driven approaches have also become popular. Neff et al. [31] use manually annotated semantic tags to train probabilistic models to perform body expressions from new texts. In turn, Endrass et al. apply a model based on a manually generated gesture corpus [8]. The *REA* architecture uses lexical data associated with movements to manage body expressions through natural language generating models. Bergmann and Kopp [3] propose a mixed system based on rules and probabilistic models.

As for the integration of co-verbal gestures in robotics, the proposed systems have thus far focused on the gestural part rather than the verbal part. Therefore, although more advanced techniques are presented for the execution of body motions – such as the generation of dynamic trajectories – rule-based approaches are the most widespread when it comes to synchronizing gestures with speech. The same as in the virtual environment, interfaces focusing on form of speech or prosody have been proposed; an example of this is the interface created by Salem et al. [44], which allows one to generate movements based on grammatical structure.

Among the approaches that apply iconic gestures, systems based on gesture repositories [20] and lexicons [19] stand out once again. Although other systems pursue greater flexibility and abstraction in movements through behavioral representations, the linguistic aspect is still based on lexicons [43]. Tay et al. propose a new interface for synchronizing language and movements generated in real time from behavior templates and sentiment analysis techniques for intensifying movements [48]. On the other hand, Kim et al. use lexical structure to detect possible words with relevant meanings, which are then used in a database that associates motions with bags of words [16]. Finally, Ng-Thow-Hing et al. propose a new system that filters words using Part-Of-Speech or *POS* tagging and relates them to a type of gesture and a grammatical model based on lexicons [33].

The main objective of this paper is to extend the study of semantic similarity carried out in [1], as well as to use the proposed methodology to implement an architecture that relates phrases and gestures with which to complement verbal communication in robotics through related body expressions. As in [33], the proposed methodology performs a word filter using a POS tagger, as well as assuming that body expressions are usually associated with certain words, and these keywords may be assigned to more than one gesture in different contexts. Therefore, if a gesture is considered to be closely related to a series of words, that relationship could be extended to other similar words, making this process a problem of lexical substitution. In this way, a robot would be able to select the most semantically appropriate co-verbal gesture for a new input text.

3. Architecture

As mentioned above, proposals to synthesize co-verbal gestures into robotic interfaces are scarce [45]. So far, the general trend has been the use of rule-based methods along with other data-based approaches and supervised learning. Both approaches rely on manual annotations, either to define the corresponding rules or to provide the annotated data needed to train the models. The need for annotations reduces the flexibility of the systems in establishing the correspondences between motions and language, which translates into inferior coverage; that is, the associations between gestures and phrases are presented in a very limited number of cases when compared to what a robot could find in a new text, in addition to being limited to a specific semantic context.

The main difficulty in improving communication through gesticulation lies in the immense number of possibilities and meanings. For this reason, this paper proposes an architecture which is adaptive to language. This is intended to reduce manual annotation to the characterization of concepts, increasing the coverage of the system through the application of semantic similarity. In this way, the system could make use of a semantic model and a subsequent application of similarity estimation functions to, given a phrase, find the most relevant gesture among all the defined ones.

As we have found in the relevant literature, the simplest way to characterize those concepts that are attributed to the set of gestures available to the robot is through a set of related terms. Although at first glance it seems that this set of terms would share the same function as the trigger words used in the most basic approximations, it is not simply a matter of locating the same words, but rather of being able to launch a body motion from semantically related words not contained in the set of related terms associated with the motion. In that sense, any word would be a possible candidate for a particular gesture in the absence of any other that is more closely linked to its meaning. For example, the meaning of a concept associated with mountain could be represented by the terms “mountain”, “summit” or “peak”, so that the interface would respond with the corresponding gesture to words such as “hill”, “slope” or “rock”; in this case, the last word could no longer be linked to that gesture if another one related to “stone” were defined, closer to its meaning. Therefore, the greater the enrichment of gestures and the catalogue of available body expressions, the better the gestures will adapt to the message being conveyed by speech.

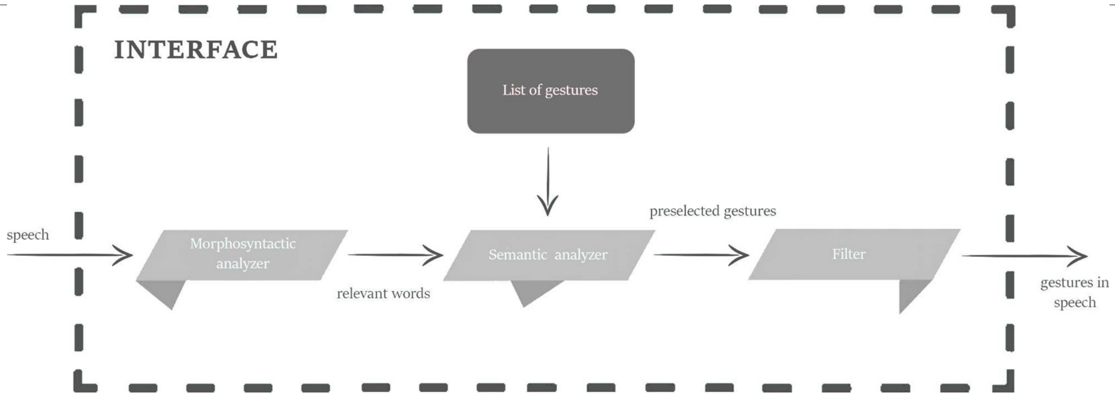

Fig. 1. Proposed achitecture for integrating iconic gestures into robotic interfaces.

Figure 1 shows the outline of the proposed architecture. This requires two entries: the text to be processed by the interface and the list of gestures with their definitions. The output it generates is the text, automatically annotated with the motions it must execute on each line. The first and second layers are a particularization of the methodology proposed in [1], while the third layer has been proposed to adapt the results to the robot. The layers that make up the architecture are detailed below:

- The first layer consists of a morphosyntactic analyzer. It begins by dividing the text into sentences, to which the semantic analysis will be applied independently in the subsequent layer. For each sentence, a tokenization and *POS* tagging process is performed by applying the FreeLing [36] tool to identify words and their grammatical categories. As the objective is to select iconic gestures, it has been decided to discard all those words with a smaller contribution to semantics, considering only nouns, verbs, adjectives and adverbs. This grammatical information will be maintained during the semantic analysis.
- The second layer consists of a semantic analyzer that compares each relevant word in a sentence with each of the terms that define the meaning of gestures. This is the main component of the architecture. The current paper presents an experiment to extend the study of semantic similarity already proposed in [1] through different measures and language models.
- Finally, a third layer outside the methodology is proposed to adapt the set of gestures to the real conditions to which the robot is subject. This last layer acts as a filter, discarding the different gestures that have been pre-selected by the semantic analyzer. In addition, it adapts the output, making it interpretable by the target robotic system (in this case a NAO robot). The time it takes the robot to pronounce the block limits the total execution time. For this reason, it makes no sense to execute too many body expressions in the same sentence when interacting, as this negatively affects fluency of speech. The affinity between word and gesture, execution times or repetitions are some of the factors that are taken into account to rule out gestures.

This paper has focused on optimizing the configuration of the semantic analyzer. To this end, an experiment has been undertaken to study models based on expert knowledge as opposed to models based on learning the lexicon from its use in language, while considering a possible combination of both. The estimation will be carried out by the different models and families of measures that are detailed in the following section.

4. Semantic approaches

The models used in this research present different approaches to the language acquisition process. With some, the contexts of the words are managed from examples, while others start from the exact definitions to compare the meanings of the words. If we look closely at human learning, at the first stage we begin to acquire information about the concepts of a sentence without getting to know its structure [40]. At school, a metalinguistic awareness is acquired that makes it possible to separate meaning from form. Finally, at a more advanced stage of language acquisition, the multiple meanings of words and the ambiguity that this entails, acquire the notion of context. These processes can be approximated in robotic interfaces by establishing the semantic information of words through word representations.

It seems that models based on lexical learning have more properties in common with this first process of semantic learning of language by humans related to linguistic immersion, which is not based on any previous knowledge. They take advantage of a massive amount of textual information by extracting their own relationships – less accurately but with more realistic levels of coverage. In this way, they manage similarity as well as proximity between the different contexts of two words. In contrast, models based on expert knowledge are generated through previous training in an academic environment. The way these models manage information is similar to the process a linguist would use to compare the meaning of words. They are the product of in-depth language analysis and further elaboration, so in principle they are expected to offer higher precision values in decreasing coverage – bearing in mind the manual limitation of design – and efficiency.

4.1. Expert knowledge-based models

Traditionally, the most widespread semantic representation has been addressed through the development of lexical databases for the organization of concepts. Expert knowledge-based models manage words as precise entities with various interpretations and well-defined relationships. Their architecture requires very expensive elaboration, so it does not facilitate the inclusion of new terms. Because of this rigid structure, quantifying relationships is a complex process with a high computational cost [4].

Since Collins and Quillian [7] proposed the use of semantic networks as knowledge stores in the 1970s, a large number of linguistic ontologies have emerged. One of the most popular and complete is *WordNet* [9]. Fellbaum – its creator – describes *WordNet* as a semantic dictionary structured in the form of a network (Fig. 2). Concepts are organized into sets of synonyms or *synsets* associated with each other through a hierarchical structure, the depth of which is linked to specificity. Some of these relationships are synonymies, hyperonymy or homonyms.

Different measures have been designed to estimate similarity between two concepts in lexical databases. Meng et al. [29] review the most popular ones, grouping them into 3 different families according to the principles on which they are based:

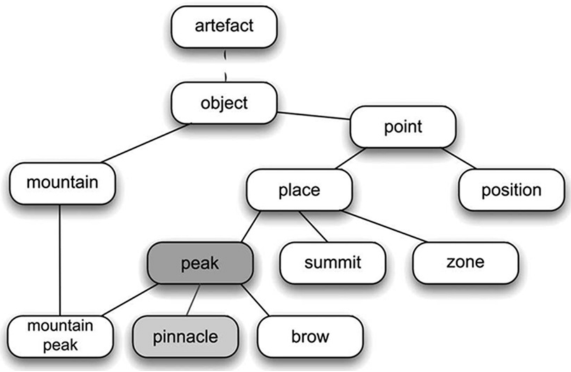

Fig. 2. General architecture for models based on expert knowledge.

- Path. They quantify similarity by the minimum number of separation nodes. In this paper we are going to use the measure proposed by Leacock and Chodorow [21] (*LCH*) and Wu and Palmer [51] (*WUP*), in addition to *Path length* [29].
- Information Content or *IC*. This is independent of the number of nodes that separate the terms. The measure proposed by Resnik [41] (*RES*), Jiang and Conrath [14] (*JCR*) and Lin et al. [26] (*LIN*) will be used.
- Features. They measure the overlapping between the terms of the glosses of two concepts. The measure proposed by Banerjee and Pedersen [2] (*Adapted Lesk*), Patwardhan [37] (*Gloss Vector* and *Gloss Vector Pairwise*), and Hirst and St-Onge [12] (*HSO*) will be applied to the experimentation.

4.2. Models based on lexical learning

In the 1960s, Harris presented the distributional hypothesis [10], positing that words that appear in similar contexts tend to represent similar meanings. This hypothesis, together with the idea that complex semantic entities can be composed from simpler constituents, has motivated the appearance of models that take advantage of the distribution of information in extensive corpora to generate vectors representing words or short phrases. For instance, topic segmentation is addressed through the similarity between vectored phrases in [50]. To generate the semantic space that these vectors form incurs a high computational cost; however, the impulse of deep learning stemming from new computational capabilities has led to the expansion of these models, thereby reaching unprecedented levels of efficiency.

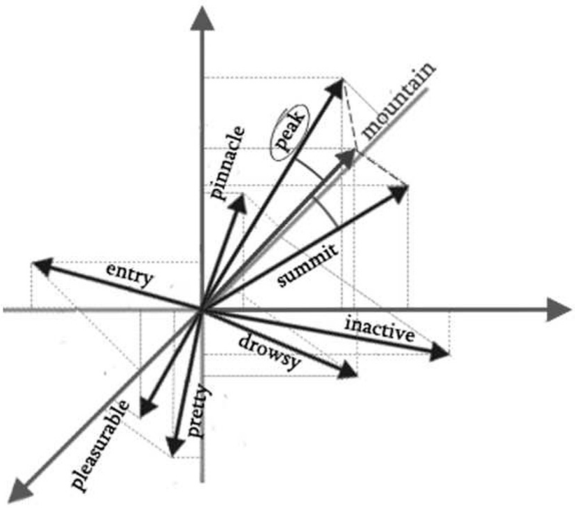

Fig. 3. General architecture for models based on lexical learning.

Most research has focused on word co-occurrence models, known as word embeddings. Pennington et al. consider that there are two families: those based on global matrix factorization methods such as *LSI*, *LDA*, *pLSI* or *sLDA*, and models based on local context window methods such as *skip-gram* or *CBOW*. Among the most popular are Mikolov's *Word2Vec* [30], or Pennington's global log-bilinear regression model called *GloVe* [39]. Levy and Goldberg [25] propose a model based on positive pointwise mutual information or *PPMI* matrices (*PPMIM*). The same authors try to generalize the *skip-gram* model by introducing negative examples (*Dep-Based model*) [24]. Finally, Salle et al. [46] presents two models also enriched with negative examples: one trained with Common Crawl[1] (*LexVec1*), and the other trained with Wikipedia[2] (*LexVec2*).

All these models transform words into vector representations and their relationships into mathematical operations; thus, cosine similarity quantifies the degree of similarity between all contexts that share two words. Figure 3 is a three-dimensional representation of one of these vector spaces.

5. Experimentation

The aim of the experimentation is to determine the best way to group gestures and words based on similarity values. To this end, a set of conditions and restrictions has been evaluated directly on the processing of the semantic analyzer's data of the second layer, at the same time as the different semantic models already mentioned have been compared.

Since the ultimate goal is to improve human perception during robot interaction, making fewer animations that are actually related to the content of speech is preferable to increasing the number of unrelated body expressions. Therefore, all the results have been evaluated in terms of F-measure, with a greater weighting of Precision instead of Recall. Specifically, β with a value of 0.3 has been set.

5.1. *Input data*

The data needed for experimentation could have been generated by manual annotation of gestures in different texts; however, two semi-automatically generated datasets have been used to simulate each semantic analyzer input in order to avoid context-specific dependencies and to simplify the data acquisition process. On the one hand, the most frequent words in language for each grammatical category have been identified from the Corpus of Contemporary American English (*COCA*),[3] and have been used as if they were gestural concepts, to construct a list of sixty gestures. On the other hand, several lexicons of synonyms and related terms such as Thesaurus.[4] have been used, to select twenty words related to each of the gestures under manual supervision. This generates the set of relevant words that should be detected in an input text. Since some of the words used have different meanings, several of the gestures used relate to the same word. This means that some gestures must be associated with more than 20 words out of a total of 1200.

Both datasets allow the simplified simulation of the two inputs of the semantic analyzer, and thus compare the set of measures and models already mentioned to determine the best optimization criteria of this component.

5.2. *Semantic analyzer scenarios*

In total, three different, consecutively proposed scenarios have been considered. In each scenario, the ten measures of similarity mentioned above have been studied on the basis of the lexical data *WordNet*, and the cosine similarity on the six word embedding models.

[1] http://commoncrawl.org/.
[2] http://wikipedia.org/.
[3] http://corpus.byu.edu/coca/.
[4] http://thesaurus.com.

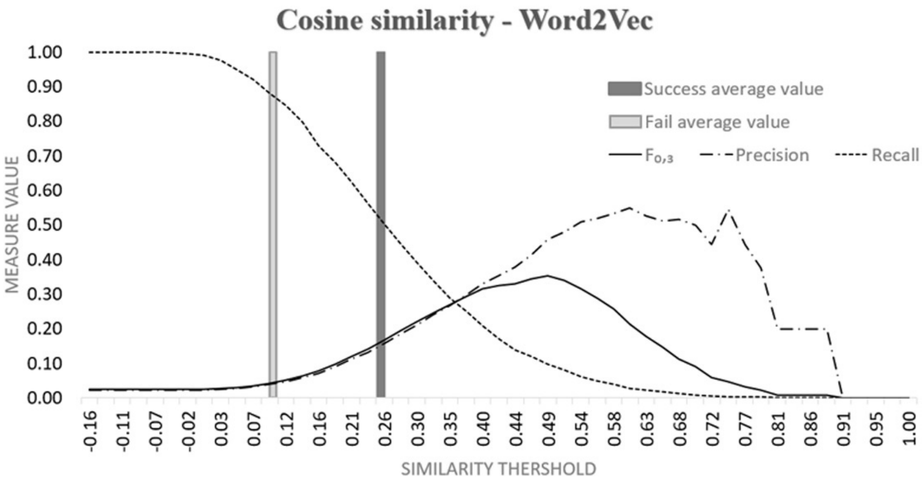

Fig. 4. Multiple assignment with *Cosine Similarity*. Variation of the $F_{0.3}$, precision and Recall as a function of the threshold.

Table 1
Results for the single assignment method without considering grammatical categories

Model	Family	Measure	Global Precision	Recall	$F_{0.3}$
Word2Vec	Geometric	Cos	0.42	0.36	0.41
LexVec2	Geometric	Cos	0.35	0.31	0.35
Dep-based	Geometric	Cos	0.33	0.29	0.33
PPMIM	Geometric	Cos	0.33	0.29	0.33
LexVec1	Geometric	Cos	0.33	0.28	0.32
GloVe	Geometric	Cos	0.29	0.25	0.29
WordNet	Feature	Adapted lesk	0.30	0.26	0.30
WordNet	Feature	HSO	0.30	0.26	0.29
WordNet	Feature	Gloss vector	0.27	0.23	0.26
WordNet	Feature	Gloss vector (pw)	0.16	0.14	0.16
WordNet	Path	WUP	0.22	0.19	0.21
WordNet	Path	LCH	0.20	0.17	0.20
WordNet	Path	Path length	0.20	0.17	0.20
WordNet	IC	JCR	0.19	0.17	0.19
WordNet	IC	LIN	0.19	0.17	0.19
WordNet	IC	RES	0.18	0.16	0.18

5.2.1. *First scenario*

In the first scenario, semantic evaluation of all the relevant words with respect to each of the gestures is proposed. When determining which gestures are associated with each word, the option of using a multiple assignment is considered first, since the existence of different contexts actually makes it a multi-label classification problem. Therefore, the possibility of using a threshold to determine which similarity values should constitute the association of a gesture is considered.

Precision indicates the percentage of correct gesture associations among all associations performed, while Recall represents the percentage of correct gesture associations among the more than 1200 possible associations. In order to examine the effectiveness of the models in selecting these associations by multiple assignment, Precision and Recall are assessed against different overall similarity thresholds. Figure 4 shows one of these graphs, specifically the performance of the *Word2Vec* model, which includes information on the mean of the similarity values of the correct and incorrect associations. If a threshold is set at high similarity values, high Precision and almost no Recall are observed, which means that, perforce, few words will be associated, despite establishing a good correspondence with the gestures. On the other hand, with a low value threshold, there will be a greater number of associations, many of which are unrelated. In any case, in view of the results, it does not seem advisable to set any threshold for multiple association, since the maximum value of $F_{0.3}$ that the model is capable of reaching is 0.35.

Based on this limitation, a single assignment method is proposed with the selection criteria of the one with the highest similarity value. This condition seems to be better suited to the problem, as shown in Table 1, which gives a value of 0.41 for $F_{0.3}$ at best. In general, the highest values are presented by using both cosine similarity on word embedding models and feature-based measures on *WordNet*.

5.2.2. *Second scenario*

In the second scenario, an analysis by categories is proposed, in such a way that a word is only evaluated against those terms that belong to the same grammatical category. This is a method of avoiding associations between words and terms from different fields. In addi-

Table 2
Separation by categories. *P* and *R* symbols represent Precision and Recall metrics, respectively

			Global			Nouns			Verbs			Adjectives			Adverbs		
Model	Family	Measure	P	R	$F_{0.3}$	P	R	$F_{0.3}$	P	R	$F_{0.3}$	P	R	$F_{0.3}$	P	R	$F_{0.3}$
Word2Vec	Geometric	Cos	0.50	0.44	0.50	0.70	0.62	0.69	0.45	0.39	0.45	0.5	0.42	0.50	0.36	0.31	0.36
GloVe	Geometric	Cos	0.44	0.38	0.44	0.64	0.57	0.63	0.38	0.33	0.37	0.49	0.40	0.48	0.27	0.23	0.27
LexVec2	Geometric	Cos	0.42	0.36	0.42	0.60	0.53	0.59	0.35	0.31	0.35	0.47	0.38	0.46	0.26	0.22	0.25
LexVec1	Geometric	Cos	0.39	0.34	0.39	0.58	0.52	0.58	0.31	0.27	0.30	0.4	0.33	0.39	0.26	0.22	0.26
Dep-based	Geometric	Cos	0.38	0.33	0.37	0.56	0.50	0.55	0.33	0.29	0.33	0.41	0.33	0.40	0.20	0.17	0.20
PPMIM	Geometric	Cos	0.38	0.33	0.38	0.54	0.48	0.54	0.33	0.28	0.32	0.43	0.35	0.42	0.23	0.19	0.22
WordNet	Feature	Gloss vector	0.38	0.33	0.37	0.31	0.28	0.31	0.40	0.34	0.39	0.51	0.42	0.51	0.31	0.27	0.31
WordNet	Feature	Adapted lesk	0.34	0.29	0.33	0.30	0.27	0.30	0.34	0.30	0.34	0.49	0.40	0.49	0.23	0.20	0.23
WordNet	Feature	HSO	0.32	0.28	0.31	0.35	0.31	0.34	0.34	0.29	0.33	0.48	0.39	0.47	0.14	0.12	0.13
WordNet	Feature	Gloss vector (pw)	0.25	0.22	0.25	0.26	0.23	0.26	0.26	0.23	0.26	0.22	0.18	0.22	0.27	0.23	0.26
WordNet	Path	WUP	0.23	0.20	0.23	0.40	0.36	0.40	0.38	0.33	0.38	0.07	0.06	0.07	0.07	0.06	0.07
WordNet	Path	LCH	0.22	0.19	0.22	0.39	0.35	0.39	0.33	0.29	0.33	0.07	0.06	0.07	0.07	0.06	0.07
WordNet	Path	Path length	0.22	0.19	0.22	0.39	0.35	0.39	0.33	0.29	0.33	0.07	0.06	0.07	0.07	0.06	0.07
WordNet	IC	LIN	0.22	0.19	0.21	0.36	0.33	0.36	0.35	0.31	0.35	0.07	0.06	0.07	0.07	0.06	0.07
WordNet	IC	JCR	0.21	0.18	0.21	0.36	0.32	0.36	0.34	0.29	0.33	0.07	0.06	0.07	0.07	0.06	0.07
WordNet	IC	RES	0.21	0.18	0.20	0.31	0.28	0.31	0.36	0.32	0.36	0.07	0.06	0.07	0.07	0.06	0.07

tion, this separation enables an individual assessment of the measures on each category. Precision, Recall and $F_{0.3}$ values can be seen in Table 2, which shows higher overall values than in the previous scenario.

It is interesting to observe the behavior of the different measures used in the estimation of similarity. In general, measures based on *IC* and *Path* reach similar values and do not appear to perform well on adjectives and adverbs. In contrast, feature-based measures behave more robustly, maintaining higher values in all categories and resulting in higher overall values. In particular, they are very efficient at calculating similarities between adjectives, reaching a $F_{0.3}$ value of 0.51. Word embeddings also have a more homogeneous function and better characterize the semantics between nouns, since the 0.69 of $F_{0.3}$ is practically double the average value of the other measures. Specifically, cosine similarity and the *Word2Vec* model outperform all other measures and models in all categories except adjectives.

The huge difference between the values of $F_{0.3}$ achieved in nouns and adverbs, which rose from 0.69 to 0.36 in the best cases, could be explained by the fact that concepts and their semantics are better reflected by nouns, while adverbs represent the circumstantial scope to a greater extent. On the other hand, there is also a slight increase in the $F_{0.3}$ values of adjectives with respect to verbs, perhaps due to greater semantic specificity of adjectives, facilitated by their inherent polarity, as opposed to the greater ambiguity of verbs.

Because of the lower occurrence of adverbs in language, as well as the low number of existing synonyms, one might think that the poorer results obtained with adverbs are partly due to the distribution of data; that is, an over-representation of adverbs has led to the definition of associations in the goldstandard with non-existent semantic similarities. For this reason, a third scenario is proposed by readjusting the evaluation collection for each grammatical category with a decrease in the number of adverbs. A brief glance at the corpus *COCA* allows us to estimate the frequency of adjectives and adverbs in general-purpose texts at 6%, while nouns and verbs account for 21% of the corpus.

5.2.3. Third scenario

The third and final scenario contemplates a redistribution of data according to the different frequencies of grammatical categories in language, as well as the combination of different measures. The results in Table 3 show a slight increase in $F_{0.3}$ in all categories. In short, there is a significant but smaller than expected increase in the $F_{0.3}$ value of adverbs, which would validate both arguments: over-representation and low contribution of adverbs to semantics.

Since the *Gloss Vector* measure and cosine similarity are based on similar principles and handle the same range of values, a number of combinations have been evaluated. Observing that the percentage of overlap between the results of the different measures and models is approximately 70%, a direct combination is now proposed by choosing the measure with the highest similarity value in each comparison between term and word. The combination that gives the best results (*Word2Vec* + *Comb2* + *Comb3* in Table 3) reaches 0.59 $F_{0.3}$. This combination consists of using only *Word2Vec* with nouns, cosine similarity of *Word2Vec*

Table 3
Redistribution of data according to each grammatical category. Symbols *P* and *R* represent Precision and Recall metrics, respectively

			Global			Nouns			Verbs			Adjectives			Adverbs		
Model	Family	Measure	P	R	$F_{0.3}$	P	R	$F_{0.3}$	P	R	$F_{0.3}$	P	R	$F_{0.3}$	P	R	$F_{0.3}$
Word2Vec	Geometric	Cos	0.53	0.43	0.52	0.69	0.65	0.69	0.46	0.42	0.46	0.50	0.42	0.50	0.40	0.34	0.39
LexVec2	Geometric	Cos	0.51	0.42	0.50	0.63	0.59	0.63	0.46	0.41	0.46	0.50	0.45	0.50	0.39	0.33	0.38
Glove	Geometric	Cos	0.48	0.39	0.47	0.62	0.59	0.63	0.39	0.35	0.37	0.50	0.42	0.50	0.32	0.29	0.31
LexVec1	Geometric	Cos	0.47	0.39	0.46	0.62	0.59	0.62	0.38	0.34	0.37	0.44	0.39	0.44	0.38	0.33	0.37
Dep-based	Geometric	Cos	0.46	0.38	0.45	0.58	0.55	0.58	0.44	0.40	0.44	0.45	0.40	0.44	0.30	0.25	0.29
PPMIM	Geometric	Cos	0.47	0.38	0.46	0.57	0.54	0.57	0.42	0.37	0.41	0.47	0.41	0.46	0.35	0.30	0.35
WordNet	Feature	Gloss vector	0.40	0.33	0.39	0.29	0.27	0.28	0.45	0.41	0.45	0.51	0.43	0.51	0.36	0.31	0.36
WordNet	Feature	Adapted lesk	0.38	0.31	0.37	0.31	0.30	0.31	0.38	0.34	0.39	0.50	0.43	0.51	0.30	0.26	0.29
WordNet	Feature	HSO	0.35	0.29	0.35	0.31	0.30	0.33	0.40	0.37	0.39	0.47	0.40	0.48	0.20	0.18	0.17
WordNet	Feature	Gloss vector (pw)	0.27	0.22	0.26	0.25	0.24	0.26	0.29	0.26	0.26	0.22	0.19	0.22	0.33	0.28	0.33
WordNet	Path	WUP	0.25	0.20	0.24	0.35	0.33	0.37	0.43	0.39	0.43	0.08	0.07	0.07	0.08	0.07	0.08
WordNet	Path	Path length	0.23	0.19	0.23	0.34	0.32	0.36	0.38	0.34	0.37	0.08	0.07	0.07	0.08	0.07	0.08
WordNet	Path	LCH	0.23	0.19	0.23	0.34	0.32	0.36	0.38	0.34	0.37	0.08	0.07	0.07	0.08	0.07	0.08
WordNet	IC	LIN	0.23	0.19	0.23	0.33	0.31	0.35	0.40	0.36	0.4	0.08	0.07	0.07	0.08	0.07	0.08
WordNet	IC	RES	0.23	0.19	0.23	0.31	0.29	0.33	0.41	0.37	0.41	0.08	0.07	0.07	0.08	0.07	0.08
WordNet	IC	JCR	0.22	0.18	0.22	0.32	0.30	0.34	0.37	0.33	0.37	0.08	0.07	0.07	0.08	0.07	0.08
Comb1 – Cos (Word2Vec) ‖ Cos (Lexvec2)			0.55	0.45	0.54	0.69	0.65	0.69	0.49	0.44	0.49	0.54	0.48	0.54	0.42	0.36	0.41
Comb2 – Cos (Word2Vec) ‖ Gloss Vector			0.54	0.44	0.53	0.53	0.50	0.52	0.55	0.49	0.54	0.61	0.54	0.60	0.47	0.41	0.47
Comb3 – Cos (Lexvec2) ‖ Gloss Vector			0.53	0.43	0.52	0.52	0.49	0.52	0.52	0.47	0.51	0.57	0.51	0.60	0.51	0.44	0.50
Word2Vec + Comb2 + Comb3			0.60	0.49	0.59	0.69	0.65	0.69	0.54	0.49	0.54	0.61	0.54	0.60	0.51	0.44	0.50

Table 4
Dialog of the story annotated with gestures

Sentence	Word	Term defining the gesture	Cosine similarity value
Teo was a little fearful.	Fearful	Frightened	0.70
He was afraid of witches…	Witches	Magic	0.36
…aliens and clowns.	Clowns	Circus	0.47
If he plays with a ball…	Ball	Kick	0.59
…he feared it could hit in the eyes.	Eyes	Head	0.38
His dog scared him, so his mother caress it for him.	Dog	Cat	0.68
He was afraid of stars and even birds.	Birds	Fly	0.39
At breakfast he believed that heating milk into the microwave…	–	–	–
…may occur an explosion	Explosion	Bomb	0.66
He was feared certain types of music…	Music	Guitar	0.53
…and lightning	Lightning	Flash	0.37
But one day…	Day	Night	0.64
Teo went into a mysterious shop and…	–	–	–
…bought a terrible mask.	Mask	Sword	0.36
For a time, he no longer feared monsters…	Monsters	Monster	0.63
…and noise while wearing it.	Noise	Light	0.31
Until the day…	Day	Night	0.64
…that Teo was frightened when he saw his reflection…	Reflection	Contemplation	0.56
…in a mirror.	Mirror	Camera	0.42
Teo then cut the mask…	Mask	Sword	0.36
…into one hundred and thirty little pieces.	–	–	–
Teo, what are you doing? -his mother exclaimed.	–	–	–
There is nothing to fear	Fear	Frighten	0.35
-said Teo- I'm now Batman!	–	–	–
The bravest one!	–	–	–

versus *Gloss Vector* for verbs and adjectives, and cosine similarity of *Lexvec2* versus *Gloss Vector* for adverbs.

Finally, it is proposed to use a minimum similarity threshold to avoid associations with low correlation values. Figure 5 shows the overall variation of $F_{0.3}$ for each grammatical category as a function of the threshold for the best combination already mentioned. By selecting thresholds 0.2, 0.3, 0.35 and 0.5 the $F_{0.3}$ values 0.68, 0.54, 0.64 and 0.67 are reached for nouns, verbs, adjectives and adverbs respectively, achieving an overall $F_{0.3}$ value of *0.63*.

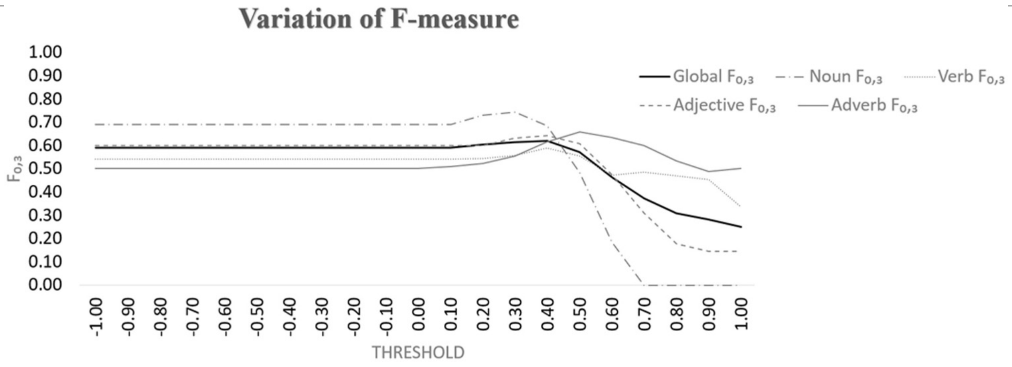

Fig. 5. Minimum threshold for association. Variation of measure $F_{0.3}$ for each grammatical category.

6. Results

The experiment concludes that the optimal configuration for the semantic analyzer would be to evaluate the similarity between terms and words under the following conditions:

- Single assignment. The proposed architecture manages the assignment of a single gesture per word, selecting the one that presents the highest values of semantic similarity.
- Restriction by grammatical categories. As experimentation has shown, it is advisable to restrict comparisons so that only the similarity between a word and terms corresponding to a gesture that are of the same grammatical category is evaluated.
- Combination of measures. The best combination is to use cosine similarity in the *Word2Vec* model to compare nouns, cosine similarity in the *Word2Vec* model versus the *Gloss Vector* measure to evaluate verbs and adjectives, and cosine similarity in the *Lexvec2* model versus the *Gloss Vector* measure to estimate the correspondence between adverbs. The combination of two different measures is resolved by selecting the maximum value.
- Minimum threshold. A threshold is applied for each grammatical category to discard all those similarities that do not reach that value, thus avoiding the assignment of less related gestures.

Therefore, a robotic interface that aims to integrate iconic gestures under this architecture should first have a list of pre-configured gestures along with their definitions in the form of relevant sets of terms. Next, the speech to be used would have to be analyzed with a tokenizer and a POS tagger, redirecting the output to the semantic analyzer specified above. This component would attempt to associate the animations most closely related to speech words among all the gestures defined in the list. Finally, a series of rules defined by the programmer would be followed to rule out gestures that are candidates for the same sentence. For example, one could select the body motions with the highest similarity value per phrase, with a greater weighting of the value of gestures related to nouns and adjectives versus verbs and adverbs. It would also be advisable to penalize gestures that have been performed previously.

7. Discussion

Initially, it was expected that expert knowledge-based models would apply similarity estimates much better, due to their greater precision in handling semantics. Despite this, and contrary to forecasts, experimentation shows that both models have similar effects. Therefore, in response to the first research question raised in this paper, it is necessary to look at efficiency. The cost of calculating similarity on the basis of the models already generated is undoubtedly higher in the methods for lexical databases. The latter require navigation techniques in graphs for this estimation, while it is a simple geometrical operation for representations generated through the corpus. Although it is true that feature-based measures may be independent of the location of concepts, due to limited resources they end up needing the properties of neighboring nodes. In short, from the point of view of efficiency rather than effectiveness, the use of models based on lexical learning seems more feasible.

The attractive qualities of unsupervised methods that define meanings from large volumes of textual data have become apparent. However, the greater complexity during the estimation of similarities of lexical databases could cast doubt on their computational cost-effectiveness. Nevertheless, experimentation shows that a significant percentage of the similarities calculated by these methods differ from models based on lexical learning. In addition, lexicons group words by meaning, unlike unsupervised methods that encode those meanings in one-hot representations, with the ambiguity that comes with it. In our opinion, a combination of both approaches is the best option for comparing linguistic meanings, so it is worthwhile to maintain and use both of them.

As for the proposed architecture, the different components have been developed on a Nao robot for implementation. On the one hand, a set of gestures provided by the manufacturer has been used in the animations library. As already mentioned, FreeLing has been used as a morphosyntactic analyzer, isolating words and categorizing them, and semantic comparisons have been applied using the models described between those words and the set of gestures. Finally, the whole process of writing down a story has been applied. The complete video[5] can be found at the address at the bottom of the page.

It should be noted that the values obtained during experimentation correspond to an evaluation of the study of the estimation of semantic similarity carried out at the level of gestural association. However, as already mentioned, what is really pursued in this paper is the perception of naturalness and fluidity in Human-Robot Interaction. The robot is not expected to perform all the possible gestures associated with a sentence as the speech pronunciation times are a constraint to the execution times. Therefore, this perception would have to be evaluated in the output of the proposed architecture.

Considering the difficulty of carrying out a quantitative evaluation of the complete architecture, the relationships between the gestures and the phrases of the story are shown in the Table 4 so that the reader can directly evaluate the gestures recorded in the story.

As it is a system based on models that are adaptive to language, the gestures associated with a sentence can be more or less related depending on the number of gestures that are established and the quality of related terms that define their meaning, or, in other words, their enrichment. In this sense, the words of speech will be adapted to the available gestures. The greater the number of gestures, the greater the likelihood of finding stronger associations for the words. Similarly, the better the choice of terms that will define the gestures, the more accurate the system will be in finding related words.

In our example, "sword" is one of the terms that defines the animation related to the concept of sword. As can be seen in the video, there is no gesture closer to the meaning of mask, and although there is a more distant semantic relationship, it is strong enough to exceed the established threshold. Another similar example is the association between the term "noise" and the gesture related to the concept of glare.

There are proposals for *WordNet* in multiple languages, such as *MultiWordNet*, as well as numerous word embeddings in other languages. This allows the proposed architecture to be multilingual. A demo in Spanish is available on the website of this project.[6]

8. Conclusions and future work

In a future where robots are expected to play a key role in society, it is critical to facilitate interactions between robots and humans. This motivation has led to the application of semantic similarity techniques in the present article, which we believe have yielded promising results. For this reason, we believe that a greater inclusion of natural language processing in the HRI field is a prerequisite for its future evolution.

As regards the experimentation carried out, two types of word representation models have been studied: those based on expert knowledge that offer a better defined structure despite the maintenance costs involved, and those based on lexical learning, which handle ambiguity but achieve greater efficiency and lexical wealth. Although experimentation concludes that in the proposed gestural framework both models are quantitatively similar in Precision and Recall, their opposite nature leads to entirely different behaviors. A more in-depth examination of results shows that a majority do not overlap, so both types of models can fit together.

The semantic analysis component that is included in the proposed gestural interaction architecture is determined with this combination of models. Implantation in a Nao robot has enabled the video attached to the article to be produced and we consider it a good reflection of the range of possibilities offered by semantic analysis for the integration of co-verbal gestures. In

[5] https://youtu.be/itslGVDCSlU.

[6] http://www.ia.uned.es/delapaz/tfm_NAONLP.html.

spite of this, we are aware that this architecture is a first approximation and there is still much work to be done to improve the calculation of correspondences and the set of heuristic rules to discard pre-selected gestures. Although the focus thus far has been on semantics, it would be interesting to try combining the semantic analyzer with one component of sentiment analysis and another of rhetorical techniques, in the same architecture. In this way, sentiment analysis could, for example, detect different degrees of effusiveness. With the analysis of rhetoric, on the other hand, the relationships between different nuclei of the phrases could be used to associate rhetorical gestures as expressions of causality or enumeration.

Acknowledgments

This work has been supported by the Spanish Ministry of Science and Innovation MAMTRA-MED Project (TIN2016-77820-C3-2-R).

References

[1] Almagro-Cádiz M, Fresno V, de la Paz López F. Smart gesture selection with word embeddings applied to nao robot. In: International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer. 2017; 167-179.

[2] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. 2003; 3: 805-810.

[3] Bergmann K, Kopp S. Increasing the expressiveness of virtual agents: Autonomous generation of speech and gesture for spatial description tasks. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems. International Foundation for Autonomous Agents and Multiagent Systems. 2009; 1: 361-368.

[4] Bollegala D, Alsuhaibani M, Maehara T, Kawarabayashi KI. Joint word representation learning using a corpus and a semantic lexicon. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. 2016; 2690-2696.

[5] Cassell J, Vilhjálmsson HH, Bickmore T. Beat: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). ACM. 2001; 477-486.

[6] Chiu CC, Marsella S. How to train your avatar: A data driven approach to gesture generation. In: International Workshop on Intelligent Virtual Agents. Springer. 2011; 127-140.

[7] Collins AM, Quillian MR. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior. 1969; 8(2): 240-247.

[8] Endrass B, Damian I, Huber P, Rehm M, André E. Generating culture-specific gestures for virtual agent dialogs. In: International Conference on Intelligent Virtual Agents. Springer. 2010; 329-335.

[9] Fellbaum C. Wordnet: An Electronic Lexical Database. MIT Press. 1998.

[10] Harris ZS. Distributional structure. Word. 1954; 10: 146-162.

[11] Hato Y, Satake S, Kanda T, Imai M, Hagita N. Pointing to space: modeling of deictic interaction referring to regions. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE. 2010; 301-308.

[12] Hirst G, St-onge D. Lexical chains as representations of context for the detection and correction of malapropisms. In: WordNet: An Electronic Lexical Database. MIT Press. 1998; 305-332.

[13] Huang CM, Mutlu B. Modeling and evaluating narrative gestures for humanlike robots. In: Proceedings of Robotics: Science and Systems. 2013; 57-64.

[14] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th Research on Computational Linguistics International Conference. 1997; 19-33.

[15] Kendon A. Gesture: Visible action as utterance. Cambridge University Press. 2004.

[16] Kim HH, Lee HE, Kim YH, Park KH, Bien ZZ. Automatic generation of conversational robot gestures for human-friendly steward robot. In: Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE. 2007; 1155-1160.

[17] Kipp M, Neff M, Kipp KH, Albrecht I. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In: International Workshop on Intelligent Virtual Agents. Springer. 2007; 15-28.

[18] Kopp S, Wachsmuth I. Synthesizing multimodal utterances for conversational agents. Journal Computer Animation and Virtual Worlds. 2004; 15(1): 39-52.

[19] Le QA, Hanoune S, Pelachaud C. Design and implementation of an expressive gesture model for a humanoid robot. In: 11th IEEE-RAS International Conference on Humanoid Robots. IEEE. 2011; 134-140.

[20] Le QA, Pelachaud C. Generating co-speech gestures for the humanoid robot nao through bml. In: Gesture and Sign Language in Human-Computer Interaction and Embodied Communication. Springer Berlin Heidelberg. 2012; 228-237.

[21] Leacock C, Chodorow M. Combining local context and wordnet similarity for word sense identification. WordNet: An Electronic Lexical Database 1998; 49(2): 265-283.

[22] Lee J, Marsella S. Nonverbal behavior generator for embodied conversational agents. In: International Workshop on Intelligent Virtual Agents. Springer. 2006; 243-255.

[23] Levine S, Theobalt C, Koltun V. Real-time prosody-driven synthesis of body language. In: ACM Transactions on Graphics (TOG). ACM. 2009; 28: 172.

[24] Levy O, Goldberg Y. Dependency-based word embeddings. In: Proceedings of the 52nd; Annual Meeting of the Association for Computational Linguistics. 2014; 2: 302-308.

[25] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. In: Proceedings of the 27th International Conference on Neural Information ProcessingSystems. 2014; 2: 2177-2185.

[26] Lin D, et al. An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1998; 98: 296-304.

[27] Mavridis N. A review of verbal and non-verbal human-robot interactive communication. Robotics and Autonomous Systems. 2015; 63: 22-35.

[28] McNeill D. Gesture and thought. University of Chicago Press.

2005.

[29] Meng L, Huang R, Gu J. A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology. 2013; 6(1): 1-12.

[30] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013; 2: 3111-3119.

[31] Neff M, Kipp M, Albrecht I, Seidel HP. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG). 2008; 27(1): 5-24.

[32] Nehaniv CL, Dautenhahn K, Kubacki J, Haegele M, Parlitz C, Alami R. A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE. 2005; 371-377.

[33] Ng-Thow-Hing V, Luo P, Okita S. Synchronized gesture and speech production for humanoid robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE. 2010; 4617-4624.

[34] Niewiadomski R, Bevacqua E, Mancini M, Pelachaud C. Greta: An interactive expressive eca system. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems. 2009; 2: 1399-1400.

[35] Özyürek A, Willems RM, Kita S, Hagoort P. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. Journal of Cognitive Neuroscience. 2007; 19(4): 605-616.

[36] Padró L, Stanilovsky E. Freeling 30: Towards wider multilinguality. In: LREC2012. 2012.

[37] Patwardhan S. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, University of Minnesota, Duluth. 2003.

[38] Pellegrinelli S, Pedrocchi N. Estimation of robot execution time for close proximity human-robot collaboration. Integrated Computer-Aided Engineering. 2018; 25(1): 81-96.

[39] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP). 2014; 14: 1532-1543.

[40] Räsänen O. Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. Speech Communication. 2012; 54(9): 975-997.

[41] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995; 1: 448-453.

[42] Riek LD, Rabinowitch TC, Bremner P, Pipe AG, Fraser M, Robinson P. Cooperative gestures: Effective signaling for humanoid robots. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE. 2010; 61-68.

[43] Salem M, Kopp S, Wachsmuth I, Joublin F. Towards meaningful robot gesture. In: Human Centered Robot Systems: Cognition, Interaction, Technology. Springer Berlin Heidelberg. 2009; 173-182.

[44] Salem M, Kopp S, Wachsmuth I, Joublin F. Towards an integrated model of speech and gesture production for multimodal robot behavior. In: Proceedings of the 19th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE. 2010; 614-619.

[45] Salem M, Kopp S, Wachsmuth I, Rohlfing K, Joublin F. Generation and evaluation of communicative robot gesture. International Journal of Social Robotics. 2012; 4(2): 201-217.

[46] Salle A, Idiart M, Villavicencio A. Matrix factorization using window sampling and negative sampling for improved word representations. The 54th Annual Meeting of the Association for Computational Linguistics (ACL). 2016; 419-424.

[47] Sauppé A, Mutlu B. Robot deictics: How gesture and context shape referential communication. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction. ACM. 2014; 342-349.

[48] Tay J, Veloso M. Modeling and composing gestures for human-robot interaction. In: Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE. 2012; 107-112.

[49] Tepper P, Kopp S, Cassell J. Content in context: Generating language and iconic gesture without a gestionary. In: Proceedings of the Workshop on Balanced Perception and Action in ECAs at Automous Agents and Multiagent Systems (AAMAS). 2004; 4: 8.

[50] Wu JW, Tseng JC, Tsai WN. A hybrid linear text segmentation algorithm using hierarchical agglomerative clustering and discrete particle swarm optimization. Integrated Computer-Aided Engineering. 2014; 21(1): 35-46.

[51] Wu Z, Palmer M. Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL). 1994; 133-138.