

Grounding Words in Visual Perceptions: Experiments in Spoken Language Acquisition

Fabio De Ponte, Sarah Rauchas

Department of Computing, Goldsmiths, University of London, Lewisham Way, New Cross, London SE14 6NW, UK

Abstract

In recent years, Natural Language Processing models have shown compelling progress in generating and translating text. Yet, the symbols that are manipulated by these models are not produced within the models themselves. On the contrary, they are externally given in the form of tokens. The models only measure the probability that a specific token comes after another (or a group of others) and allow the generation of a list of tokens, each of which has a certain probability of following the previous one. There is no connection to sensory perceptions, and the semantic interpretation of the outputs – as well as of the inputs – of these models is completely invisible to the system that produces them. Therefore, language cannot be used by the system to manipulate information about the perceived world. This is commonly referred to as the Symbol Grounding Problem and, as of today, there is not a generally accepted procedure to solve it. This paper explores a possible solution: a sequence-to-sequence model trained over videos characterised by visual elements which reliably predict the presence of acoustic co-occurring elements.

A dataset was created ad-hoc, with videos that include 5 types of objects and 5 actions. Two research questions were considered: whether such a model could map video features onto audio features, in fact producing a categorization without labels, where the categories would emerge from the parallel, simultaneous generalization of both input and target; and whether the model would be able to combine learned information about objects and movements to correctly describe a new combination, shown in a video it was not exposed to during training, a process that is referred to as compositional semantics.

The experiment showed that the model was able to generalize simultaneously over videos and over the utterances that were paired with them. Further, it produced sentences that were in some cases more accurate than the original ones, precisely because of the process of generalization.

However, the results suggest also that the model did not develop the ability to combine information taken from different samples. In other words, while symbol grounding seems to have been achieved, compositional semantics does not. The experiment shows that sensory perceptions can be mapped onto one another with a sequence-to-sequence model trained over a dataset where elements coming from different sensory domains are paired. However, it is not sufficient to develop compositional semantics.

Keywords

Symbol grounding, Compositional semantics, NLP

1. Introduction

Recently, Natural Language Processing models like GPT3 and BERT have shown compelling progress in interpreting and predicting text. Yet, while they help us understand the way

NL4AI 2022: *Sixth Workshop on Natural Language for Artificial Intelligence, November 30, 2022, Udine, Italy* [52]


✉ fabio.deponte@gmail.com (F. De Ponte); s.rauchas@gold.ac.uk (S. Rauchas)

🌐 <https://github.com/fabiodeponte/> (F. De Ponte)

🆔 0000-0001-7139-3057 (F. De Ponte)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

languages work — offering at the same time valuable tools for different applications — they do not address the question of the relationship between words and sensory perceptions and of how language emerged in the first place. These questions date back centuries but have lately become known within the field of artificial intelligence as “the symbol grounding problem,” since the issue was clearly defined by Harnad [13, Abstract], who stated it in these terms: «How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?»

Nowadays, even a model as simple as an n-gram conditional frequency model, trained over a sufficiently large corpus of text, can produce apparently meaningful text. However, the symbols that are manipulated by the system are not inherent to the system itself. Applying the model, the machine only calculates the probability that a certain token comes after another (or a group of others) and then produces a list of tokens, each of which has a certain probability of following the previous one. In principle, this could be done by a human being on an unknown language. As John R. Searle [34] famously argued in his paper “Minds, brains, and programs,” a person could apply a similar method to a Chinese corpus and produce a Chinese text, without knowing the meaning of a single Chinese word.

Yet, recent advancements led a senior engineer at Google, Blake Lemoine, to claim that a language model, LaMDA, short for Language Model for Dialogue Applications, a system for building chatbots, was «sentient» [46]. The company quickly dismissed the claim and, in a statement, a Google spokesperson said: «There was no evidence that LaMDA was sentient (and lots of evidence against it).»[46]. In order to support his claim, Lemoine published a few astonishing fragments of chats between himself and the system. However, as surprising as they are, they do not seem to suggest the existence of a sentient mind. In fact, a simpler explanation is possible. Language models capture language structures and, with them, a fraction of the logic behind language itself. Mikolov, Yih and Zweig [20] were able to design a model such that some seemingly logical operations between word vectors were possible, like “King - Man + Woman”, resulting in a vector very close to “Queen”, corresponding to the analogy “a man is to a king as a woman is to a queen.” Operations like this between embedding vectors do not always work, even in more recent models. However, they show that sometimes it is possible to convert logical operations into geometrical operations. More generally, they suggest that human reasoning reflects in language and that a sufficiently complex language model can capture part of that logic. This could explain the surprising performances of the most advanced language models but leaves us with the unsolved problem of how a system can develop a meaningful semantic interpretation of the symbols that logic is applied to. Therefore, the challenge is to design a technique that would let the system define its own meaningful symbols. According to Harnad, «there is really only one viable route from sense to symbols: from the ground up.» [13, Conclusions]. In other words, symbols have to be grounded to perceptions. There have been many attempts to do that, we will see some of them below.

2. Background Literature

Many attempts have been made at developing a method for grounding symbols to perceptions. One of the earliest was proposed by Roy [29, section 1], who designed «a computational model

which learns from untranscribed multisensory input» where «acquired words are represented in terms associations between acoustic and visual sensory experience.» The model was designed to learn the same way it is claimed children do, by discovering «words by searching for segments of speech which reliably predict the presence of visually co-occurring shapes.» The author recorded a number of sessions of adults speaking to babies in a room. The adults were playing with the babies with toys, one toy at a time. The system included a speech processor that «converted spoken utterances into sequences of phoneme probabilities»; and a visual processor that «extracted statistical representations of shapes and color from images of objects.» Phoneme probabilities and statistical representations of co-occurring images were stored in a short-term memory so that the model was able to predict the most probable next word, given an image. While the experiment was a major step forward, it had an important drawback: not all utterances contained the name of the toy. Sometimes, adults said only, for example, «Here it comes!» referring to the toy car they were playing with. Therefore, the results were inconclusive.

Another interesting experiment was proposed by Tuci *et al.* [48]. The authors set up a virtual environment where subsequent generations of evolving robots were trained «to access linguistic instructions and to execute them by indicating, touching or moving specific target objects.» [48, Abstract]. In the course of the process, they were able to learn that the commands were composed of different parts. For example, “touch the pen” is made of two parts, “touch” and “the pen”, while “move the spoon” is composed of “move” and “the spoon.” Once they had learned this, they were able to generalize to new compositions, e.g. “touch the spoon,” and execute them, even though they had never seen that specific command before. This, according to the authors, demonstrated «how the emergence of compositional semantics is affected by the presence of behavioural regularities in the execution of different actions.» A drawback was that, while this experiment effectively demonstrated the principle, it did not provide a method to enable a system to acquire language, other than within the very limited scope of the experiment itself.

A similar approach was proposed by Sugita and Tani [43]. They focused on the creation of a geometrical n-dimensional space, where the geometric arrangements represented «the underlying combinatoriality» among symbols. They defined 26 actions and recorded 120 corresponding algorithmically generated sensor-motor time series for each. They were then able to create a model where «the composition of symbols is realized by summing up their corresponding vectors,» a compositional semantics potentially much more powerful than the one of Tuci *et al.* [48].

More recently, other approaches have been proposed. One came from Gonzalez-Billandon *et al.*, who converted «the audio signal to an embedding space where embeddings for the same words are closer than different words, regardless of speaker.» [11, Section II]. In order to achieve this, they used a Vector Quantized-Variational Autoencoder network. Then, the embeddings were associated with images. The project is still ongoing and the results have not been released yet.

Another method came from Wang *et al.* [51], who proposed MAXSAT, a SATNet layer that can be integrated into neural network architectures that could successfully be used to learn logical structures. Their technique was subsequently used by Topan, Rolnick and Si [47], to map visual inputs to symbolic variables without explicit supervision, through a self-supervised pre-training pipeline.

A different approach was tried by Tan and Bansal [45]. They proposed LXMERT (Learning

Cross-Modality Encoder Representations from Transformers), a large-scale transformer model. In essence, it is a framework designed to learn direct vision-to-language connections and it represents a step forward in the task of describing the content of images in words. It outperforms previous models on visual question-answering datasets. However, it does not include sound or any perception other than vision. Therefore, it does not offer a direct contribution to tackle the symbol grounding problem and it does not address the question of compositional semantics.

A proposal in this direction was instead put forward by Liu, Li and Cheng [17] who replaced text with audio. They proposed a new general-purpose neural sound synthesis network, based on generative adversarial networks, that was able to generate sound directly from visual inputs. In their work, the task was formulated as a regression problem to map a sequence of video frames to a sequence of raw audio waveform. This model paves the way for the design of a general method to associate visual to acoustic features and vice versa. It represents a significant step in the direction of the solution of the symbol grounding problem. However, it was designed to reproduce any kind of sound, and the authors had in mind mostly noise produced by specific objects (e.g. cars, motorcycles, scrolling water, etc.). It was not designed for spoken language and, again, it does not address the question of compositional semantics.

3. Methodology

3.1. Objectives of the research

Our work focused on the following research objectives:

1. Verify whether it is possible for a system to learn words from an unlabeled dataset of spoken utterances and visual representations, through a sequence-to-sequence neural network.
2. Verify if, once such words are learned, compositional semantic would emerge, that is whether the system would be able to combine them to compose sentences that were not present in the training dataset.

3.2. Dataset

The dataset is composed of 1,000 videos, each showing an object – either staying still or moving – while a voice reads a sentence, for example “this is a pen,” that refers to what we see. There are twenty voices in total: ten voices are natural – i.e. they have been recorded by real people – while the other ten are artificial (produced through the services of the website www.murf.ai). Each video is exactly 3 seconds long. It is 180x180 pixel sized and the audio is sampled to 16 KHZ. The dataset is available online ¹.

Each voice recorded 50 sentences: the first 25 (Table 1) describe still objects, the next 25 (Table 2) refer to the moving of objects. A video was recorded for each sentence. Audio and video files were then paired and saved into a single MP4 file. It is a balanced dataset: each object is represented 5 times not moving and 5 times moving for each group of 50 videos.

¹It can be downloaded from <https://www.kaggle.com/datasets/fabiodeponte/symbolgrounding> and it is released under licence Creative Commons Attribution-ShareAlike 4.0 Generic (CC BY-SA 4.0).

Table 1
Objects

Number	Spoken utterance	Video content	Times
1-5	“This is a pen”	A pen on a table	5
6-10	“This is a phone”	A phone on a table	5
11-15	“This is a spoon”	A spoon on a table	5
16-20	“This is a knife”	A knife on a table	5
21-25	“This is a fork”	A fork on a table	5

The videos were augmented by five operations: flipping image, increasing brightness, decreasing brightness, increasing saturation and zooming in. Audios were augmented by five operations as well: increasing and decreasing speed, increasing and decreasing pitch, and increasing volume. Lastly, data was split into training, validation and test sets.

After completing the data augmentation, each of the five augmented audio sets of 1,000 samples, plus the original one, was paired with each of the five augmented video sets of 1,000 samples, plus the original one. In total, 36,000 samples were generated.

3.3. Model architecture

Once the dataset was ready, a system was designed to map visual elements to spoken utterances. Video files were processed by two pre-trained neural networks, namely Wav2vec (a model introduced by Baevski *et al.* [2] at Facebook for self-supervised learning of representations from raw audio, trained on LibriSpeech corpus) and CLIP (developed by Radford *et al.* [26] at OpenAI, trained on a dataset of 400 million image-text pairs collected from a variety of publicly available sources on the Internet), that extracted respectively acoustic and visual features. Then a sequence-to-sequence neural network mapped the extracted features of the visual part onto the extracted features of the acoustic part ².

Wav2vec returns features vectors of 150 integer values, which in turn can be in turn decoded to a written text by the same library. As we had a dataset composed of 36,000 videos, we got an audio features matrix sized 36,000 x 150. CLIP returns vectors of 512 float values. However, it does not provide a tool to convert the features back to a video or to any other directly readable form. For this reason, while Wav2vec makes it relatively simple to evaluate the predictions of a video-to-audio neural network, CLIP does not do the same for an audio-to-video network. Therefore, we focused on the transformation of video features into audio features and not the reverse, developing a sequence-to-sequence video-to-audio network.

As a preliminary check, audio features were decoded to written text, through one of the libraries made available by Facebook with Wav2vec. From this verification, it emerged that not all features produced intelligible text. In particular, the sentences recorded by real people contained many errors. The reason is probably that, as they are non-native speakers, their pronunciation was not good enough for the library. Moreover, the artificial voices that had their pitch multiplied by a factor of ten for data augmentation purposes resulted in unintelligible

²The code of audio and video extractor and of sequence-to-sequence model is available at https://github.com/fabiodeponte/symbol_grounding.

Table 2
Actions

Number	Spoken utterance	Video content	Times
26	“Move the pen to the left”	A pen moving to the left	Once
27	“Move the pen to the right”	A pen moving to the right	Once
28	“Move the pen up”	A pen moving up	Once
29	“Move the pen down”	A pen moving down	Once
30	“Rotate the pen”	A pen rotating	Once
31	“Move the phone to the left”	A phone moving to the left	Once
32	“Move the phone to the right”	A phone moving to the right	Once
33	“Move the phone up”	A phone moving up	Once
34	“Move the phone down”	A phone moving down	Once
35	“Rotate the phone”	A phone rotating	Once
36	“Move the spoon to the left”	A spoon moving to the left	Once
37	“Move the spoon to the right”	A spoon moving to the right	Once
38	“Move the spoon up”	A spoon moving up	Once
39	“Move the spoon down”	A spoon moving down	Once
40	“Rotate the spoon”	A spoon rotating	Once
41	“Move the knife to the left”	A knife moving to the left	Once
42	“Move the knife to the right”	A knife moving to the right	Once
43	“Move the knife up”	A knife moving up	Once
44	“Move the knife down”	A knife moving down	Once
45	“Rotate the knife”	A knife rotating	Once
46	“Move the fork to the left”	A fork moving to the left	Once
47	“Move the fork to the right”	A fork moving to the right	Once
48	“Move the fork up”	A fork moving up	Once
49	“Move the fork down”	A fork moving down	Once
50	“Rotate the fork”	A fork rotating	Once

decoded text as well. Therefore, all real voices and some of the artificial ones were discarded, along with the corresponding videos. The dataset was thus reduced in size and as a result was composed of only 14,500 samples. Each object was represented 2,900 times: 1,450 times laying still and 1,450 moving.

It should be noted that this does not necessarily mean that the discarded samples are useless. They could be used in a model that does not make use of Wav2vec. In fact, whether the features are intelligible or not for a speech-to-text converter is not relevant for symbol grounding, as long as the same utterances are pronounced consistently in similar ways.

The sequence-to-sequence models are composed of two parts: an encoder and a decoder. Each of them includes one or more Long Short-Term Memory (LSTM) layers. During training, the encoder basically compresses the sequence into a vector, while the decoder learns a series of conditional frequency distributions: for each value and each vector coming from the encoder, it calculates the most likely value to come next. In terms of the translation process, this means that the decoder calculates the most probable next word given the last predicted word and the vector coming from the encoder, which summarizes information about the whole sentence to be translated.

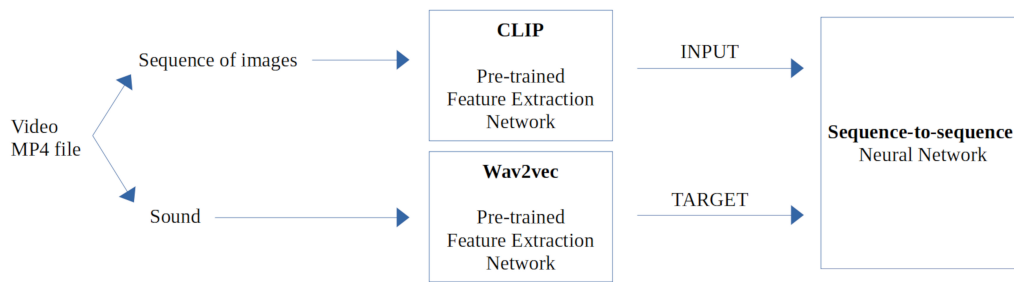


Figure 1: The structure of the model.

In our case, we can consider each word of the input sentence one value of the video features vector and each word of the target translated sentence a value of the audio features target vector. Then, mapping video features onto audio features may be interpreted as a translation task: on the input side, we have a sequence that represents the contents of a video; on the output side, we expect a sequence that represents the same contents expressed in audio format.

3.4. The compositional semantics experiment

As we mentioned, while the first aim was to design a model that would be able to map videos onto audio recordings, the second was to verify whether such a model could develop compositional semantics. In order to do that, an experiment was set up. The videos containing a specific moving object were removed from the dataset, while videos showing the same object laying still were kept, along with all other objects, both moving and still. The model was trained on the reduced dataset, and it was tested against the removed videos. The audio features predicted by the model were converted into text and so it was possible to compare them with the intended ones.

For example, at one point all videos of moving pens were removed, so that the model was trained only on videos showing pens laying still (paired with the utterance “this is a pen” read by different voices) and other objects (spoons, forks, knives and phones) both staying still (paired with utterances like “this is a spoon”) and moving (paired with utterances like “move the spoon to the right”). No videos showing a pen moving to the right and the utterance “move the pen to the right” were included in the training set. The experiment consisted of verifying whether the model was able to compose the utterance “move the the pen to the right,” once such a video was given to the network. In order to do that, the model should have been able to combine the information coming from the shape of the pen and the information coming from the movement of other objects. We repeated the experiment five times, with one object at a time.

Table 3
Original and predicted utterances.

N.	ORIGINAL UTTERANCE	PREDICTED UTTERANCE
0	MOVE THE FORK TO THE RIGHT	MOVE THE FORK TO THE RIGHT
1	ROTATE THE KNIFE	WROTATE THE KNIFE
2	THIS IS A SPOON	THIS IS A SPOON
3	MOVE THE FORK TO THE RIGHT	MOVE THE FORK TO THE RIGHT
4	MOVE THE SPOON DOWN	MOVE THE SPOON DOWN
5	HIS IS A SPURN	THIS IS A SPERN
6	MOVED THE FORK TO THE LEFT	MOVE THE FORK TO THE LEFT
7	ROTATE THE SPOON	ROTATE THE SPOON
8	LUSA PHONAP	MOVES THE PHONE UP
9	THIS IS A SPOON	THIS IS A SPOON
10	MOVE THE PHONE TO THE RIGHT	MOVE THE PHONE TO THE RIGHT

4. Results and discussion

4.1. Symbol grounding

Twenty different configurations of the sequence-to-sequence model were tried³, modifying the size of the layers, the learning rate, the optimizer and the directionality. The best performing one had an LSTM layer of size 1,024 for the encoder and an LSTM layer of the same size for the decoder. It was not bidirectional, as bidirectionality made the training slower without offering any gain in terms of performance. Categorical cross-entropy loss function and Adam optimizer were adopted, with learning rate 0.001, and a batch size of 64. The video features values were normalized to the range 0-1, multiplied by 100 and rounded to integer. Therefore, each value was one-hot encoded in a vector sized 101. This model showed a test loss of 0.0336 and a test accuracy of 0.9890. Along with loss and accuracy, another performance metric was adopted: the cosine distance between the predicted features and the expected ones. This measure was calculated comparing predicted and target vectors on 100 samples extracted from the train set and 100 from the test set. The showed a test cosine distance of 50.12.

Another important method of evaluation was the visual inspection. Due to space constraints, we show here only the results of the best performing model⁴. We can see in Table 3 a sample of 11 randomly picked sentences, as predicted by the model. The left column shows the sentence originally paired with the video (as converted by Wav2vec from the original audio into text), while the right column shows the predicted sentence. As we can see, the model was able to produce sentences that in some cases were even more understandable than the original. In the batch of 11 sentences that we randomly extracted from the test dataset:

- The model correctly predicted “move the fork to the right” (sample number 0), “this is a spoon” (n. 2), “move the fork to the right” (n. 3), “move the spoon down” (n. 4), “rotate the spoon” (n. 7), “this is a spoon” (n. 9) and “move the phone to the right” (n.10).

³The detailed description of this process can be found on https://github.com/fabiodeponte/symbol_grounding.

⁴The results of the other models are described on the Github repository.

- It predicted “wrotate the knife” (n. 1), a substantially correct output with the minor error of adding a “W” to the word “rotate”.
- It predicted “This is a spern” (n. 5) for the video of a spoon staying still that was originally paired with a scarcely comprehensible utterance, namely “his is a spurn”, instead of “this is a spoon”. In fact, generalizing over similar videos and over the utterances that were paired with them, the model showed a result that turned out to be better than the original. This was again the case with the utterance (n. 6). The model returned features corresponding to “move the fork to the left”, whereas the original video had been paired with “moved the fork to the left”, with a “d” at the end of “move” that made the utterance slightly incorrect. And again (n. 8) the model returned “moves the phone up”, which was a strong improvement over the utterance originally paired with the video, that was interpreted by the Wav2vec library as “lusa phonap”.

4.2. Compositional semantics test

Once video features were mapped onto audio features, the experiment described in 3.4 was performed. In order to do that, five subsequent tests were performed, one for each object: pen, phone, spoon, knife and fork. For each of them, a reduced dataset was prepared, removing the videos that showed the object moving to the left, to the right, up, down and rotating. The model was trained on the reduced dataset, resulting in 13,050 samples, and then tested against the 1,450 videos that had been removed. The results are shown in Table 4.

Table 4

Model trained on reduced dataset and tested on removed moving object.

OBJECT	TRAIN AND TEST ON REDUCED DATASET				TEST ON MOVING OBJECT		
	Train loss	Train acc.	Test loss	Test acc.	Loss	Acc.	Cosine dist.
PEN	0.02	0.99	0.04	0.99	0.73	0.88	59.81
FORK	0.01	0.99	0.03	0.99	0.8	0.88	66.28
PHONE	0.03	0.99	0.04	0.99	0.73	0.89	59.59
KNIFE	0.03	0.99	0.04	0.98	0.86	0.87	64.59
SPOON	0.02	0.99	0.04	0.99	0.97	0.87	61.47

As we can see, loss increased dramatically when the model was tested over videos of moving objects that it had not been exposed to during training. We can clearly see that it is not a generalization problem, because the Test loss column shows that, when tested against videos previously unseen but belonging to known groups, the model showed a loss ranging between 0.030 and 0.043. Yet, when exposed to new kinds of videos, its loss increased dramatically, ranging between 0.73 and 0.97. The model was able to generalize when similar videos were present during training. However, it was not able to combine information from the videos that showed the objects staying still and information about the movement applied to other objects, to form a sentence composed by “move” and the name of the object.

In Table 5, ten predictions (out of 1,450) are shown. As we can see, the model did not predict a sentence containing either “move the pen” or “rotate the pen” when exposed to a moving pen,

Table 5

Model tested on PEN videos removed from dataset during training.

Number	ORIGINAL UTTERANCE	PREDICTED UTTERANCE
1	MOVED THE PAN TO THE LEFT	THIS IS A PEN
2	MOVE THE PEN TO THE RIGHT	THIS IS A PEN
3	MOVE THE PEN UP	THIS IS A PEN
4	MOVE THE PEN DOWN	THIS IS A PEN
5	ROTATE THE PEN	THIS IS A PEN
6	MOVE THE PEN TO THE LEFT	MOVE THE KNIFE TO THE LEFT
7	MOVE THE PEN TO THE RIGHT	THIS IS A PEN
8	MOVE THE PEN DOWN	MOVE THE KNIFE DOWN
9	MOVE THE PEN DOWN	THIS IS A PEN
10	ROTATE THE PEN	THIS IS A PEN

Table 6

Model tested on PHONE videos removed from dataset during training.

Number	ORIGINAL UTTERANCE	PREDICTED UTTERANCE
1	MOVED THE PHONE TO THE LEFT	THIS IS A PHONE
2	MOVED THE PHONE TO THE RIGHT	MOVE THE SPOON TO THE LEFT
3	MOVE THE PHONE UP	THIS IS A PHONE
4	MOVE THE PHONE DOWN	THIS IS A PHONE
5	ROTATE THE PHONE	THIS IS A PHONE
6	MOVE THE PHONE TO THE LEFT	THIS IS A PHONE
7	MOVE THE PHONE TO THE RIGHT	THIS IS A PHONE
8	MOVE THE PHONE	THIS IS A PHONE
9	MOVE THE PHONE DOWN	THIS IS A PHONE
10	ROTATE THE PHONE	THIS IS A PEN

nor was it able (Table 6) to predict “move the phone” or “rotate the phone” when exposed to the moving phone videos. Similarly, it could not predict any of the correct sentences for the other three objects. In the dataset of 14,500 videos, each object was represented 2,900 times: 1,450 times laying still and 1,450 moving. The model was trained five times, each time leaving aside the videos of a particular object depicted while moving. Each time it was tested against those videos. In total, it returned 7,250 audio features, converted then to text. Again, it cannot be shown here due to space constraints, but a through inspection showed that the combination never occurred⁵.

The model mostly favoured the shape of the object, predicting a sentence in the form “this is...,” thus ignoring the movement. In a minority of cases, it recognised the movement, but it failed to combine the information with the shape of the object and simply predicted the utterance “move” accompanied by a different object, often the most similar. For example, “move the pen” was frequently mistaken for “move the knife.”

⁵See COMPOSITIONAL SEMANTICS RESULTS - FOR EACH OBJECT COMPARE ORIGINAL AUDIO FEATURES AND PREDICTED AUDIO FEATURES.ipynb on Github repository.

5. Conclusions

5.1. Symbol grounding and compositional semantics

This project had two aims. The first was to design a system, composed of three neural networks, that had to be able to directly map, without labels, visual elements onto spoken utterances, to achieve symbol grounding. The second aim was to verify whether, once the mapping was done, such a system would develop the ability to compose sentences that were not present in the training dataset, combining information gathered from different samples, in order to achieve compositional semantics.

The results shown above suggest that the ability to directly produce a correct sentence when exposed to a video has been achieved. The model was able to generalize on both sides of the dataset and map videos onto audio recordings without labels. However, the results also suggest that the model did not develop the ability to combine information taken from different samples.

With this experiment, we tried to capture the power of categorization carried by words into an artificial system. In fact, the co-occurring audio features associated with each video, once generalized, play the role of categories. However, they are categories that the model itself extracts from the sensory perceptions through a process of generalization, rather than from externally given labels. The fact that words are intimately linked to categories is not surprising, as with language comes indeed «the ability to generalize,» as Oliver Sacks [33, p. 42] pointed out in “Seeing voices”, a book devoted to the relationships between sensory perceptions and language. As Lupyan [19] argues, «merely perceiving an object does not require categorizing it. In contrast, naming an object (whether to communicate to another individual or for your own benefit) does require placing it into a category.» However, classification into categories is not enough. Once things have names, the necessity of a grammar arises, in order to allow the acquired categories to be manipulated. And, as Corballis [7, p. 37] illustrates, the origin of grammar could very well have been compositional semantics: “The simplest events consist of objects and actions, such as *baby screams*, *snake approaches*, or *apple falls*. Suppose, for example, that an animal’s experience includes five meaningful objects and five meaningful actions. If each object is associated with a single action, so that only babies scream or only apples fall, then there are only five events to be signaled, and five event symbols will do the trick; the objects do not need to be distinguished from the actions associated with them. But if all possible combinations of objects and actions can occur, then it would be more economical to learn five symbols for the objects and five for the actions, making ten in all, than to learn twenty-five symbols to cover all their possible combinations. This might be the source of protolanguage, leading eventually to grammar.”

5.2. Further work

This experiment shows that it is possible to achieve symbol grounding through a sequence-to-sequence model trained over a dataset with audio and video co-occurring elements. The work could be expanded in several directions:

1. Apply an attention layer as introduced by Bahdanau *et al.* [3] to the sequence-to-sequence model, something that might allow the model to achieve compositional semantics.

2. Give up pre-trained features extraction networks. For real symbol grounding to be fully achieved, the system should not include networks that were trained over labels. They could be replaced by autoencoders, a solution that could offer the possibility of reconstructing the original audio and video files.
3. Expand the dataset. The dataset developed for this project was relatively small, with one thousand original videos and only five objects represented. It could be expanded, with the addition of more objects, either staying still or moving, and new voices.
4. Apply the model to the “something something” video database, that offers images of thousands of actions on objects, paired with a linguistic label, composed by a description of the action in the form “do <something> on <something>” (hence the name). That dataset seems perfectly suitable for purposes of compositional semantics. However, it lacks audio. In order to overcome the problem, captions – at least a fraction of them – could be converted to spoken utterances through artificial voices.

References

- [1] Anderson, P., Wu, Q., Teney D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3674–3683. (2018). Available at: https://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html. Last Accessed 16 September 2022.
- [2] Baevski, A., Zhou, H., Mohamed, A.R., Auli, M.: Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada (2020). Available at: <https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf>. Last Accessed 14 September 2022.
- [3] Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: 3rd International Conference on Learning Representations (2015). Available at: <https://arxiv.org/pdf/1409.0473.pdf>. Last Accessed 6 September 2022.
- [4] Cangelosi, A., Greco, A., Harnad, S.: From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, Vol.12, No. 2, 143-162 (2000). Available at: <https://doi.org/10.1080/09540090050129763>. Last Accessed 8 May 2022.
- [5] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724-1734 (2014). Available at: <https://aclanthology.org/D14-1179>. Last Accessed 19 September 2022.
- [6] Chollet, F.: Character-level recurrent sequence-to-sequence model, Keras, 29 September (2017). Available at: https://keras.io/examples/nlp/lstm_seq2seq. Last Accessed 5 September 2022.

- [7] Corballis, M.C.: From hand to mouth. The origins of language. Princeton University Press, Princeton and Oxford (2002).
- [8] Fanello S.R., Ciliberto C., Natale L., Metta G.: Weakly supervised strategies for natural object recognition in robotics in IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, May 6-10 (2013). Available at: <https://ieeexplore.ieee.org/document/6631174>. Last Accessed 23 May 2022.
- [9] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional Sequence to Sequence Learning. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70 - 1243–1252 (2017). Available at: <https://arxiv.org/abs/1705.03122>. Last Accessed 6 September 2022.
- [10] Gentner, D.: Some interesting differences between verbs and nouns. In: Cognition and Brain Theory, 1981, vol. 4, 161-178 (1981). Available at: <https://groups.psych.northwestern.edu/gentner/papers/Gentner81c.pdf>. Last Accessed 22 May 2022.
- [11] Gonzalez-Billandon, J., Grasse, L., Sciutti, A., Tata, M., Rea, F.: Cognitive Architecture for Joint Attentional Learning of word-object mapping with a Humanoid Robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2019). Available at: https://www.researchgate.net/publication/344432053_Cognitive_Architecture_for_Joint_Attentional_Learning_of_word-object_mapping_with_a_Humanoid_Robot. Last Accessed 12 June 2022.
- [12] Hannagan, T., Agrawal, A., Cohen, L., Dehaene S.: Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. In: Pnas, Vol. 118 - No. 46 (2021). Available at: <https://www.pnas.org/doi/10.1073/pnas.2104779118>. Last Accessed 6 September 2022.
- [13] Harnad, S.: The Symbol Grounding Problem. *Physica D* 42: 335-346 (1999). Available at: <https://arxiv.org/abs/cs/9906002v1>. Last Accessed 7 May 2022.
- [14] Harnad, S.: Symbol grounding and the origin of language, University of Southampton Institutional Research Repository (2002). Available at: <https://eprints.soton.ac.uk/256471>. Last Accessed 16 September 2022.
- [15] Huang, H., Wong, R.K.: Attention-based Seq2seq Regularisation for Relation Extraction. In: International Joint Conference on Neural Networks (IJCNN) (2021). Available at: <https://ieeexplore.ieee.org/abstract/document/9533807>. Last Accessed 6 September 2022.
- [16] CLIP, Video Features Documentation, https://iashin.ai/video_features/models/clip. Last Accessed 5 September 2022.
- [17] Liu, S., Li, S., Cheng, H.: Towards an End-to-End Visual-to-Raw-Audio Generation with GAN. In: IEEE Transaction on Circuits and Systems for Video Technology, Vol. 32, Issue 3 (2021). Available at: <https://ieeexplore.ieee.org/document/9430540>. Last Accessed 6 September 2022.
- [18] Luong, M.T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015). Available at: <https://arxiv.org/abs/1508.04025>. Last Accessed 6 September 2022.
- [19] Lupyan, G.: Carving nature at its joints and carving joints into nature: how labels augment category representations, *Progress in Neural Processing*, Vol. 16, 2005, 87-96 (2005). Available at: https://doi.org/10.1142/9789812701886_0008. Last Accessed 8 May 2022.

- [20] Mikolov, T., Yih, W., Zweig, G.: 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. Association for Computational Linguistics, 746–751 (2013). Available at: <https://aclanthology.org/N13-1090>. Last Accessed 16 September 2022.
- [21] Nolfi S.: “Emergence of communication and language in evolving robots” in Lefebvre, C., Comrie, B. and Cohen H., *New Perspectives on the Origins of Language*, 533–554 (2013). Available at: <https://doi.org/10.1075/slcs.144.20nol>. Last Accessed 8 May 2022.
- [22] Pasquale, G., Ciliberto, C., Odone, F., Rosasco, L., Natale, L.: Are we done with object recognition? The iCub robot’s perspective. In: *Robotics and Autonomous Systems*, Volume 112, February 2019, 260–281 (2019). Available at: <https://www.sciencedirect.com/science/article/pii/S0921889018300332>. Last Accessed 12 June 2022.
- [23] Pinker, S.: *Language Learnability and Language Development* (1984/1996). Cambridge, MA: Harvard University Press.
- [24] Pinker, S.: How could a child use verb syntax to learn verb semantics?. In: *Lingua* Vol. 92, April 1994, 377–410 (1994). Available at: <https://www.sciencedirect.com/science/article/pii/0024384194903476>. Last Accessed 22 May 2022.
- [25] Prickett, B., Traylor, A., Pater, J.: Seq2Seq Models with Dropout can Learn Generalizable Reduplication. In: *Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 93–100, Brussels, Belgium, October 31 (2018). Available at: <https://aclanthology.org/W18-5810.pdf>. Last Accessed 5 September 2022.
- [26] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, A., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: *Proceedings of the 38 th International Conference on Machine Learning*, PMLR 139 (2021). Available at: <http://proceedings.mlr.press/v139/radford21a/radford21a.pdf>. Last Accessed 14 September 2022.
- [27] Ramanathan, V., Tang, K., Mori, G., Fei-Fei, L.: Learning Temporal Embeddings for Complex Video Analysis in *Proceedings of International Conference on Computer Vision (ICCV)* (2015). Available at: <https://arxiv.org/abs/1505.00315>. Last Accessed 26 June 2022.
- [28] Roger, E., Banjac, S., Thiebaut de Schotten, M., Baciua, M.: Missing links: The functional unification of language and memory in *Neuroscience & Biobehavioral Reviews*, Volume 133, February (2022). Available at: <https://www.sciencedirect.com/science/article/pii/S0149763421005601>. Last Accessed 26 June 2022.
- [29] Roy, D.: Grounded speech communication. In: *Proceedings of the 6th International Conference on Spoken Language (ICSLP)*, vol. 4, 69–72 (2000). Available at: https://www.isca-speech.org/archive_v0/archive_papers/icslp_2000/i00_4069.pdf. Last Accessed 22 May 2022.
- [30] Roy, D.: Learning Visually Grounded Words and Syntax of Natural Spoken Language, *Computer Speech & Language*, Volume 16, Issues 3–4, July–October 2002, 353–385 (2002). Available at: https://www.media.mit.edu/cogmac/publications/evol_comm_2002.pdf. Last Accessed 12 May 2022.
- [31] Roy, D.: Grounded Spoken Language acquisition: Experiments in Word Learning. In: *IEEE Transactions on Multimedia*, Vol. 5, No. 2 (2003). Available at: https://www.media.mit.edu/cogmac/publications/ieee_multimedia_2003.pdf. Last Accessed 12 May 2022.

- [32] Roy, D.: Grounding words in perception and action: computational insights. In: Trends in Cognitive Sciences, Vol.9, No.8 (2005). Available at: https://lsm.media.mit.edu/papers/Roy_TICS_2005.pdf. Last Accessed 22 May 2022.
- [33] Sacks, O.: Seeing voices. 3rd edn. London: Picador books (2012).
- [34] Searle, J.R.: Minds, brains, and programs. In: The behavioral and brain sciences, 417-457 (1980). Available at: <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A>. Last Accessed 12 May 2022.
- [35] Selsam, D., Lamm, M., Bunz, B., Liang, P., de Moura, L., Dill, D. L.: Learning a sat solver from single-bit supervision. In: Proceedings of The International Conference on Learning Representations (ICLR) (2018). Available at: <https://arxiv.org/abs/1802.03685>. Last Accessed 12 June 2022.
- [36] Shao, L., Migimatsu, T., Zhang, Q., Yang, K., Bohg, J.: Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. In: The International Journal of Robotics Research, Vol. 40, Issue 12-14, October (2021). Available at: <https://journals.sagepub.com/doi/abs/10.1177/027836492111046285>. Last Accessed 16 September 2022.
- [37] Siskind, J.M.: Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition. In: AI technical reports 1964-2004 (1993). Available at: <https://dspace.mit.edu/handle/1721.1/6784>. Last Accessed 22 May 2022.
- [38] Siskind, J.M.: Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. In: Journal Of Artificial Intelligence Research, Volume 15, 31-90 (2001). Available at: <https://arxiv.org/abs/1106.0256>. Last Accessed 22 May 2022.
- [39] Speech to Text with Wav2Vec 2.0, KdNuggets. <https://www.kdnuggets.com/2021/03/speech-text-wav2vec.html>. Last Accessed 5 September 2022.
- [40] Steels, L.: The symbol grounding problem has been solved, so what's next?. In: Manuel de Vega, Arthur Glenberg, and Arthur Graesser (eds), Symbols and Embodiment: Debates on meaning and cognition, Oxford, 2008. Online edn, Oxford Academic, 2012. Available at: <https://doi.org/10.1093/acprof:oso/9780199217274.003.0012>. Last Accessed 14 November 2022.
- [41] Steels, L., Van Eecke, P., Beuls, K.: Usage-based learning of grammatical categories. In: Belgian/Netherlands Artificial Intelligence Conference (BNAIC) 2018 Preproceedings, 253-264. Available at: <https://arxiv.org/pdf/2204.10201.pdf>. Last Accessed 14 November 2022.
- [42] Steels, L., Verheyen, L., van Trijp, R.: An experiment in measuring understanding. In: Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence 2022, 241-242. Available at: <https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA220203>. Last Accessed 14 November 2022.
- [43] Sugita, Y., Tani, J.: A sub-symbolic process underlying the usage-based acquisition of a compositional representation: Results of robotic learning experiments of goal-directed actions. In: 2008 7th IEEE International Conference on Development and Learning, ICDL (2008). Available at: <https://ieeexplore.ieee.org/document/4640817>. Last Accessed 16 September 2022.
- [44] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. In: Proceedings of the 27th International Conference on Neural Information Processing

- Systems - Volume 2 (2014). Available at: <https://arxiv.org/abs/1409.3215>. Last Accessed 6 September 2022.
- [45] Tan, H., Bansal, M.: LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 5100–5111 (2019). Available at: <https://aclanthology.org/D19-1514>. Last Accessed 1 July 2022.
- [46] Tiku, N.: 'The Google engineer who thinks the company's AI has come to life', The Washington Post, 11 June (2022). Available at: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine>. Last Accessed 1 July 2022.
- [47] Topan, S., Rolnick, D., Si, X.: Techniques for Symbol Grounding with SATNet. In: 35th Conference on Neural Information Processing Systems (2021). Available at: <https://proceedings.neurips.cc/paper/2021/hash/ad7ed5d47b9baceb12045a929e7e2f66-Abstract.html>. Last Accessed 12 June 2022.
- [48] Tuci, E., Ferrauto, T., Zeschel, A., Massera, G.: An Experiment on Behaviour Generalisation and the Emergence of Linguistic Compositionality in Evolving Robots, IEEE transactions on autonomous mental development, 176–189 (2011). Available at: <https://dx.doi.org/10.1109/TAMD.2011.2114659>. Last Accessed 8 May 2022.
- [49] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (2017). Available at: <https://arxiv.org/abs/1706.03762>. Last Accessed 6 September 2022.
- [50] Wang P., Donti P.L., Wilder B., Kolter Z.: Bridging deep learning and logical reasoning using a differentiable satisfiability solver, Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97 (2019). Available at: <https://arxiv.org/abs/1905.12149>. Last Accessed 12 June 2022.
- [51] Wang, W., Wu, M., Pan, S.J.: Deep Weighted MaxSAT for Aspect-based Opinion Extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 5618–5628, November 16–20 (2020). Available at: <https://aclanthology.org/2020.emnlp-main.453.pdf>. Last Accessed 6 September 2022.
- [52] D. Nozza, L. Passaro, M. Polignano, Preface to the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), November 30, 2022, CEUR-WS.org, 2022.