Quantifiers in a Multimodal World: Hallucinating Vision with Language and Sound

Testoni, A.; Pezzelle, S.; Bernardi, R.

DOI
10.18653/v1/W19-2912

Publication date
2019

Document Version
Final published version

Published in
Cognitive Modeling and Computational Linguistics

License
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Quantifiers in a Multimodal World: Hallucinating Vision with Language and Sound

Alberto Testoni
CI MeC - University of Trento
alberto.testoni@studenti.unitn.it

Sandro Pezzelle
ILLC - University of Amsterdam
s.pezzelle@uva.nl

Raffaella Bernardi
CI MeC, DISI - University of Trento
raffaella.bernardi@unitn.it

Abstract

Inspired by the literature on multisensory integration, we develop a computational model to ground quantifiers in perception. The model learns to pick out of nine quantifiers (‘few’, ‘many’, ‘all’, etc.) the one that is more likely to describe the percent of animals in a visual-auditory input containing both animals and artifacts. We show that relying on concurrent sensory inputs increases model performance on the quantification task. Moreover, we evaluate the model in a situation in which only the auditory modality is given, while the visual one is ‘hallucinanted’ either from the auditory input itself or from a linguistic caption describing the quantity of entities in the auditory input. This way, the model exploits prior associations between modalities. We show that the model profits from the prior knowledge and outperforms the auditory-only setting.

1 Introduction

Quantifiers (words like ‘some’, ‘most’, ‘all’) have long been the holy grail of formal semanticists (see Peters et al. (2006) for an overview). More recently, they have caught the attention of cognitive scientists, who showed that these expressions are handled by children quite early in life (Halberda et al., 2008), even before developing the ability to count (Hurewitz et al., 2006). Though some effort has been paid to model these high-frequency expressions from their use in big corpora of texts (Baroni et al., 2012; Herbelot and Vecchi, 2015), relatively little work has focused on the models’ ability to quantify using these words.

In computer vision, some focus to the task of extracting quantities from images has been expressed through visual question answering, whose benchmark dataset (Antol et al., 2015) contains ‘count questions’ (e.g., ‘How many Xs have the property Y?’) that repeatedly turned out to be rather challenging (Malinowski et al., 2015; Fukui et al., 2016). While this work paid little attention to quantifiers, a few recent studies specifically investigated their computational learning from visual inputs (Sororoc et al., 2016; Pezzelle et al., 2017). These works built on the evidence that (part of) the meaning of quantifiers is grounded in perception. However, they only experimented with the visual modality, though the numerical representations humans derive from sensory inputs have been shown to be shared across modalities, e.g., vision and sound (Feigenson et al., 2004).

In the literature on multisensory integration it is well established that redundant information conveyed through different sensory inputs leads to a better performance on semantic tasks (McGurk and MacDonald, 1976). These findings have brought researchers to propose the ‘Hub and Spoke’ model (hence, H&S): concepts are learned by mutual interaction of the representation produced by sensory specific processors, the ‘spokes’, with a transmodal ‘hub’ (Patterson et al., 2007;
Ralph et al., 2017). The role of the cross-modal hub is to take each of the spokes' output and to reproduce the correct information across the others by back-propagation (Ralph et al., 2017). There is evidence that memory recall is affected by the multisensory context in which the concept was learned. In particular, it has been shown that a congruent pair of audiovisual inputs may facilitate subsequent recall. In other words, we learn to process a sound (e.g., ‘meow’ or ‘woof’) and to associate it to the visual representation of the entity we see making it, and this facilitates the recall of the corresponding concept (i.e., ‘cat’ or ‘dog’).

In this work, we apply the H&S model to the conceptual learning of quantifiers and study how the hub learns to integrate the visual and auditory spoke representations (as illustrated in Figure 1) to perform the quantification task. That is, the model has to learn to say that ‘none’, ‘few’, ‘most’, etc. of the objects in the visual and auditory inputs belong to a given category, that of animals. We focus on 9 common quantifiers and experiment with visual and auditory inputs strongly aligned (viz., aligned at the entity level). We show that

- Using congruent audio visual inputs increases the performance of the model in learning quantifiers within single-sensory models;
- The H&S model can generalize to unseen data quite well. In particular, it generalizes better when trained on small combinations and tested on large ones than *vice versa*.

Furthermore, a second part of our work is based on an ongoing debate in multisensory integration, namely whether the processing of sensory inputs is passive or rather influenced by previous experience that creates cross-sensory associations. Within this debate, one of the most influential frameworks is the Predictive Coding Model (hence, PCM), according to which prior knowledge affects the representation of perceptual inputs (Friston, 2010). There is a general agreement on the *predictive* effects between visual and auditory inputs, whereas the role of language in priming visual perception is still under debate (see Simanova et al. (2016) for an overview).

Inspired by this work, we compare a single auditory sensory model with a model in which the processing of the auditory stimuli is facilitated by prior expectation elicited by either the visual spoke (implemented as a mapping from the experienced auditory input to its corresponding visual representation) or the language input (again implemented as a mapping from language to visual representations). In Figure 1, the ‘prior’ arrow illustrates this predictive factor. Simplifying somewhat, we simulate a setting where a model, trained to quantify from co-occurring synchronous audio visual inputs, is tested on a situation where (a) it *hears* but does not *see* the entities (audio-vision association prior) or (b) it *reads* a description of the entities and *hears* their sounds but does not *see* them (language-vision association prior). We show that

- Using priors hallucinating the visual representation improves the performance of the model compared to when it receives only auditory inputs;
- Language prior is slightly more effective than sound prior to hallucinate concurring vision.

2 Related Work

2.1 Multimodal Models

Fueled by the explosion of deep learning, much effort has been paid in recent years to develop models that exploit information from various modalities. Attention has been mostly on language and vision, for which various tasks have been proposed, i.e. image captioning (Hodosh et al., 2013), visual question answering (Antol et al., 2015; Goyal et al., 2017), visual reasoning (Andreas et al., 2016; Johnson et al., 2017; Suhr et al., 2017), visual storytelling (Huang et al., 2016; Gonzalez-Rico and Fuentes-Pineda, 2018), and visual dialogue (De Vries et al., 2017). While all this work combines images with *written* text, some other studies employed *spoken* language to perform various tasks, such as image-audio retrieval (Chrupała et al., 2017; Harwath et al., 2018). Overall, these works repeatedly showed that combining information from language and vision leads to representations that are beneficial in virtually any task.

A relatively recent strand of research focused on the integration of visual and *sound* information, where the latter is, e.g., the ‘roar’ of a fast car (Owens et al., 2016, 2018; Zhao et al., 2018). More akin to our work is Aytar et al. (2017), who jointly investigated language, vision, and sound. By training a deep convolutional network
for aligned representation learning across the three modalities, they showed that the emerging alignment improved both retrieval and classification performance. Interestingly, their results also suggested that, even though the network was never exposed to pairs of sounds and text inputs during training, an alignment between these two modalities was learned, possibly due to the use of images as an internal ‘bridge’. We explore the same three modalities studied by Aytar et al. (2017). However, we use different models and evaluation settings (to mimic the PCM) and tackle a different task, namely quantification.

2.2 Computational Models of Quantification

The task of quantification (in the broad sense of providing some quantitative information), has been largely explored in computer vision (Seguí et al., 2015; Zhang et al., 2015a; Arteta et al., 2016). In these works, the focus is to provide the exact number of objects in a scene, and only rarely it is inspired by cognitive abilities (Zhang et al., 2015b; Chattopadhyay et al., 2017). Similarly, in the visual question answering community, the so called ‘number’ questions are almost exclusively about cardinals, with some exceptions including generalized quantifiers like every or more than half (Suhr et al., 2017; Kuhnle et al., 2018).

Inspired by the cognitive skill of Approximate Number Sense (ANS) is instead Stoianov and Zorzi (2012), which tested hierarchical generative networks and showed that they learn ANS as a statistical property of images. Practically speaking, the model was able to compare one approximate ‘numerosity’ against another and to perform a more/less task. Similar high-level cognitive abilities are required to humans to use vague quantifiers such as few, many, or most, whose meaning is heavily dependent on contextual factors. Using visual scenes as context, a recent strand of work has focused on the computational learning of quantifiers with neural networks. One approach tackled the task in a visual question answering fashion (Sorodoc et al., 2018), while another aimed at learning to apply the correct quantifier to a given scene (Sorodoc et al., 2016; Pezzelle et al., 2017).

More related to our work is Pezzelle et al. (2018b), which tested a model in the task of predicting the probability of each quantifier to be used in a given scene. The network was trained with probabilities from human participants by Pezzelle et al. (2018a). We use the same human annotation but make two steps further: First, we also experiment with auditory inputs; second, we experiment with different settings inspired by the literature on multisensory integration.

3 Task and Datasets

3.1 Task

Given an input (a scene) consisting of entities that are either animals (targets) or artifacts (distractors), the model has to quantify the former. For instance, given the image in Figure 2 on the left, it should assign a high probability to ‘most’, whereas for the image on the right it should assign a high probability to ‘few’. The inputs are either unimodal (sound, vision) or multimodal (sound+real vision, sound+hallucinated vision). We inherit and adapt to our multimodal datasets the gold standard annotation collected by Pezzelle et al. (2018a): Human participants were asked to select, out of nine quantifiers (‘none’, ‘almost none’, ‘few’, ‘the smaller part’, ‘some’, ‘many’, ‘most’, ‘almost all’, ‘all’), the one that best referred to the set of animals depicted in a briefly-presented visual scene (these scenes were similar, but not identical to those in Figure 2). Each quantifier turned out to be used to refer to various proportions of animals. For instance, ‘most’ could apply when animals corresponded to 57%, 60%, 67%, 75% and 80% of the objects. At the same time, various proportions had different probabilities to be referred by a given quantifier. With a proportion of 60% animals, for example, the probability to choose ‘most’, ‘many’ and ‘some’ is 0.52, 0.20 and 0.18, respectively. The models have to learn the probability distribution associated with each proportion. Intuitively, ‘none’ and ‘all’ are almost exclusively used with, respectively, 0% and 100% animals.

3.2 Datasets

Following Pezzelle et al. (2018a), our datasets consist of scenes containing animals and artifacts with a minimum of 3 and a maximum of 20 entities in total. There are in total 17 proportions, out of which 8 contain more animals than artifacts, 8 contain more artifacts than animals, and 1 contains an equal number of them. For each proportion

---

1 The proportions obtained by having min. 3 max 20 objects are: 0%, 10%, 17%, 20%, 25%, 33%, 40%, 43%, 45%, 50%, 57%, 60%, 67%, 75%, 80%, 83%, 90%, 100%.
we generated scenes containing all possible combinations of cardinalities: For the proportion 0%, for example, 17 combinations were built, ranging from 0:3 (0 animals, 3 artifacts) to 0:20.

We built visual and auditory datasets aligned at the entity level: For each image, we created the corresponding auditory datapoint containing the sound of each entity in the image. By so doing, using the terminology of (Aytar et al., 2018), we obtained strongly aligned visual and auditory datasets. In total, we used 55 unique animals and 55 unique artifacts. We only used those entities for which we could have whole-depicting images (not just parts) and for which we had a corresponding sound. Furthermore, for each audio-visual input we created a corresponding linguistic caption describing the quantities of the entities in it. Details on the three datasets are provided below.

**Visual Dataset** Similarly to Pezzelle et al. (2018b), we built a large dataset of synthetic visual scenes depicting a variable number of animals and artifacts on top of a neutral grey background (see Figure 2). The scenes were automatically generated using the following pipeline: (a) Natural images depicting target objects (e.g., a dog) or distractors (e.g., a car) were randomly picked up from the 110 entities pre-selected from the dataset by Kiani et al. (2007). As opposed to the synthetic dataset of Pezzelle et al. (2018a), where multiple copies of the same animal/artifact were reproduced in the scene, we have different target/distractor instances in each scenario (e.g., different instances of ‘car’ as in Figure 2 (right)). However, we do not vary the size and orientation of entities; (b) The proportion of targets in the scene was chosen by selecting only those matching the 17 pre-defined proportions mentioned above. We generated 17K scenes balanced per proportion (1K scenes/proportion), and split them into train (70%), validation (10%), and test (20%) sets. The distribution of proportions per total number of objects in the training set is illustrated in Figure 3.

**Auditory Dataset** We followed a similar procedure to build the auditory scenes. We took Audioset (Gemmeke et al., 2017) as our starting point to obtain sounds corresponding to the entities since it contains a huge collection of human-labeled 10-sec sound clips. It is organized as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds. We took sounds belonging to the categories of ‘animals’ and ‘tools’. We built our auditory dataset starting from the visual one described above and obtained the strongly aligned auditory version. Hence, as in the case of the visual datapoint, an auditory datapoint can contain different instances of the same type of animal/artifact. The auditory dataset consists of 17K scenes again balanced per proportion (1K scenes/proportion), with the same split as the visual one and each ‘scene’ containing min 3 max 20 entities out of 110 entities.

**Linguistic Dataset** For each aligned visual and auditory input pair, we built a linguistic caption describing the exact quantities of the entities present in it (for instance, for the image in Figure 2 (left), we obtain ‘There are one butterfly, two automobiles and two mammals’). The procedure, illustrated in Figure 4, is as following: (a) We manually annotated each of the 110 entities used to
Figure 4: Linguistic dataset construction. In red: randomly selected nouns for each entity. Bottom: generated caption.

We build the dataset (55 animals and 55 artifacts) with 3 nouns expressing different levels of an ontological hierarchy (e.g., ‘cat’, ‘feline’, ‘mammal’).\(^2\) (b) For each entity present in the audio-visual scene, we randomly picked one of the three nouns. (c) For each noun, we counted the number of entities present in the audio-visual input, assigned that number to the noun and pluralized it, if necessary. (d) In order to account for more variability, we started the linguistic caption by choosing one of six possible starting phrases.\(^3\) We obtained captions with on average 10.5 nouns (standard deviation: 4.53).

**Sensory Representations** The vector representation of the visual scene is extracted using Inception v3 CNN (Szegedy et al., 2016) pretrained on ImageNet (Deng et al., 2009) from the last average pooling layer which consists of 2048-d visual vectors.

For the auditory dataset, we built the representation of each entity and the scenes containing them as following. We started from the audio features computed with the VGG-inspired auditory model described in Hershey et al. (2017) which has been trained on a preliminary version of YouTube-8M.\(^4\) For each second of a sound clip, the model produces a 128-d vector; hence each 10-sec sound clip of the Audioset dataset (Gemmeke et al., 2017) would be represented by a 1280-d vector. To work with smaller and more representative vectors, we selected the two central seconds of each 10-sec audio clip (the 5th and 6th) and used the resulting 256-d vector as the representation of the corresponding entity. Out of these entity representations we built the representation of the scene by concatenating the entity vectors. Scenes can contain min 3 and max 20 entities, hence we use vectors of 20 ‘cells’. When there are less than 20 entities, there are ‘empty cells’ which are visually represented by the grey background. We represented their auditory counterpart with a ‘silent sound’ computed as following: we recorded a 10-sec sound clip of silence, picked the 5th and 6th seconds and obtained the 256-d auditory vector using the model of Hershey et al. (2017). The 20 total ‘cells’ are then shuffled, resulting in a 5120-d auditory vector.

As for the linguistic scenes, for each caption we extracted the features through the Universal Sentence Encoder (USE) (Cer et al., 2018) producing 512 dimensional vectors for each sentence. Alternatively, we could have used LSTM modules to process from scratch both the linguistic and acoustic inputs exploiting their sequential nature. We rejected this alternative mainly to avoid that, during the training process, the neural network learns task-dependent representations and arbitrary associations. It has been shown (e.g., in Cer et al. (2018)) that USE provides sentence-level embeddings with strong transfer performance on several NLP tasks. We consider this point as a strong motivation for our choice: in this way, we get more consistent representations across different modalities and the overall architecture turns out to be easier, more scalable and less prone to learn task-specific representations.

The semantic spaces containing the entity representations of the three modalities are rather different. It is interesting to note that the auditory dataset is much more dense than either the visual or the linguistic one: The average cosine similarity between entity pairs is 0.73 for sound vs. 0.44 for vision and 0.43 for language. In other words, entities are visually and linguistically much more distinct than auditorily. This could be possibly due to the fact that, as highlighted by Owens et al. (2018), sound undergoes less transformations than vision, which is affected by, for instance, lighting, scene composition, and viewing angle. In other words, sound could be denser than vision since it

---

\(^2\) Note that in the case of animals, this hierarchy is much more easier to build (e.g. Linnaean taxonomy) while for the artifacts the 3 nouns are generally more often synonyms and often do not represent a real hierarchy/taxonomy.

\(^3\) ‘There are . . . ’, ‘It seems to me that there are . . . ’, ‘I’m thinking of . . . ’, ‘I can spot . . . ’, ‘There exists . . . ’, ‘I can spot . . . ’).

\(^4\) https://research.google.com/youtube8m/
‘abstracts’ from all the possible visual transformations that we encounter in the other modality. It follows that integrating these modalities requires some degree of generalization over a variety of transformations, which is intuitively not trivial.

4 Models and Test Settings
Below we describe the ‘Hub and Spoke’ model (H&S) that takes as input strongly aligned auditory and visual inputs, and the ‘Predictive Code Model’ (PCM) which differs from the former only at testing time, when it takes as input the vector processed by the auditory spoke and the visual representation obtained by prior knowledge, viz. through an external mapping. We take as baselines the single-modality (visual, auditory inputs) versions of the model.

Hub and Spoke model (H&S) As illustrated in Figure 5 (up), this model takes the 2048-d and 5120-d visual and auditory vectors, reduces them to vectors of the same dimensions (512-d) and merges them in the Hub through multiplication. The multimodal output is reduced to 128-d via a ReLU hidden layer, then a softmax layer is applied to output a 9-d vector with the probabilities to assign each of the 9 quantifiers.

Unimodal model The three layers of the hub described above are trained to perform the quantification task from either the visual or auditory representations alone.

Predictive Code Model (PCM) We take the hub trained using the representations produced by the visual and auditory spokes (namely the hub of the H&S) and evaluate it on new types of audio-visual inputs: the auditory vectors are produced by the auditory spoke as for the H&S, while the visual vectors are obtained via a linear mapping function that simulates prior knowledge which ‘hallucinates’ the visual perception. The mapping function takes as input either (a) the auditory input itself (auditory prior) or (b) the corresponding linguistic caption (language prior), as illustrated in Figure 5 (bottom, (a) left vs. (b) right). For sake of simplicity, the mapping function is trained outside the model. It is implemented as a linear neural network which is exposed to the aligned data of the training and validation sets used for the H&S. Hence, when used in the PCM setting it is applied to data that was never seen before. The mapping is trained using Mean Squared Error (MSE).

We only experimented with hallucinated visual representations and left for the future the other direction – a visual experience facilitated by the corresponding imagined auditory. Since the semantic space of the auditory input is rather dense, we expect that a non-linear mapping might be necessary to obtain the latter.

Implementation details We used ReLU activation function for all the hidden layers, and Adam optimizer (Kingma and Ba, 2015) with learning rate = 0.0001 and default weight decay. All models were trained for no more than 150 epochs (using early stopping) by minimizing the Kullback-Leibler (KL) divergence loss between the activations by softmax and the probability distribution of human responses for each proportion by Pezzelle et al. (2018a). All models were implemented in PyTorch v0.4.

5 Experiments and Results
Evaluation All models are evaluated by computing the Pearson product-moment correlation coefficient between the Softmax probabilities and the 9-d vectors from Pezzelle et al. (2018a), which encode the probability of each quantifier to be used with respect to a given proportion based on human choices.

5.1 Experiments
Unimodal vs. multimodal models Testing the models on the unimodal and multimodal data might lead to results that are influenced by the different sizes of data seen during training. To rule out this possibility, we use unimodal and multimodal datasets of equal size. We take 11,900 datapoints for each single modality; and in the multimodal model, we use 5950 instances for each modality which sum up to 11,900 datapoints.

Incongruent visual-auditory inputs In order to test the effectiveness of the integration of the two modalities, we take the H&S trained on aligned (congruent) visual-auditory data and we test it with incongruent data, viz. inputs that do not have the same proportion of animals. Given a visual input containing, e.g., 3 animals and 2 artifacts (as in Figure 2 left), we pair it with an auditory input having 3 artifacts and 2 animals. This way, the corresponding probability distributions are different, hence we refer to these pairs as incongruent auditory input. Similarly, we generate incongruent vi-
**5.2 Quantitative Results**

**Unimodal vs. multimodal models** Table 1 reports the Pearson’s $r$ correlation results comparing the unimodal and multimodal models. As we can see, the visual data is slightly more informative than the auditory one for learning the quantification task (0.68 vs. 0.72). The first main result is that the multimodal model outperforms the unimodal ones to a large extent. The H&S obtains 0.18 and 0.14 higher correlation than the auditory and visual model, respectively. This result shows that the multimodal data provide complementary information that the model manages to exploit. Regarding the effect of prior knowledge,
Figure 6: A: Density plot reporting the frequency of human responses for the 9 quantifiers (y-axis) against the proportion of targets in the scene (x-axis). B-F: Average probabilities predicted by models in test set (same axes).

Table 2: Unimodal vs. multimodal models tested on unseen combinations which have smaller or larger number of entities than the seen data.

<table>
<thead>
<tr>
<th></th>
<th>Pearson’s $r$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>large $\rightarrow$ small</td>
<td>small $\rightarrow$ large</td>
</tr>
<tr>
<td>Sound</td>
<td>0.55</td>
<td>0.73</td>
</tr>
<tr>
<td>Vision</td>
<td>0.64</td>
<td>0.76</td>
</tr>
<tr>
<td>H&amp;S</td>
<td>0.74</td>
<td>0.85</td>
</tr>
</tbody>
</table>

we see that hallucinating the visual representations improves over processing only the auditory input. Using the latter to hallucinate the visual scene leads to an increase of 0.10 in correlation, and an even higher increase (+0.13) is obtained when the hallucination is induced by a linguistic description of the scene. It is worth noticing, however, that the correlation values obtained by the PCMs are slightly lower than the one obtained by the H&S. This is intuitive since the latter can capitalize on first-hand information from both modalities.

To better understand the behavior of the multimodal model, we scrutinize its results by investigating whether the absolute difference between the animals and artifacts sets has an impact on the performance of the model. Figure 7 reports Pearson’s $r$ obtained by the H&S model for the smallest and highest combination of each proportion (we do not plot proportion 0.5 since the distance is 0 for all its combinations). For instance, for proportion 67%, the smallest combination is 2/3 (2 targets, 1 non-targets), the largest combination is 12/18 (12, 6), and their absolute difference is equal to 1 and 6, respectively. As can be seen from the plot, smaller absolute differences are always harder than higher ones.

Incongruent sensory pairs As the results in Table 1 show, the model is strongly sensitive to these incongruent data, suggesting that cross-modal integration is actually part of the models.
Unseen combinations Table 2 shows that models are able to generalize to unseen combinations quite well. In particular, they turn out to be always better in generalization when they learn from small combinations and are tested on large ones. This pattern of results reflects the findings illustrated in Figure 7, assuming that a model trained on hard cases and tested on easier ones would lead to higher results compared to the opposite ‘direction’.

5.3 Qualitative Results

Figure 6 compares the probability distributions learned by the tested models (panels B-F) against the distribution of responses by humans (panel A) from Pezzelle et al. (2018a). As can be clearly seen, both unimodal models (B-C) show a much lower correlation with human data compared to either H&S (D) or PCMs (E-F). In particular, the unimodal models tend to produce very similar curves for all quantifiers, thus predicting them with a similar probability at any proportion (i.e., there are no clear ‘peaks’). Both the H&S and the PCMs, in contrast, output a distribution that is very similar to that by humans (mirrored in the results of Table 1). While plots D-F are almost identical, it can be noted that the H&S is slightly better than both PCMs at the ‘extreme’ proportions, particularly 0% and 100%. We conjecture this ability is responsible of the slightly higher correlation obtained by this model compared to the PCMs.

6 Conclusion

In this paper, we show that concurrent multisensory information bootstraps models performance in a semantic task, namely grounding quantifiers, in line with the results on human perception. Also, we provide computational evidence that the predicting code hypothesis advocated in the cognitive literature is an interesting and useful source of inspiration for computational models. We plan to further investigate how predictions from prior knowledge can be compared with those obtained through sensory experience to further improve the performance on semantic tasks.

Acknowledgements

We kindly acknowledge the support of NVIDIA Corporation with the donation to the University of Trento of the GPUs used in our research. We thank Aurélie Herbelot, Manuela Piazza, and Marco Marelli for their valuable comments. The second author is funded by the Netherlands Organisation for Scientific Research (NWO) under VIDI grant no. 276-89-008, Asymmetry in Conversation.

References


Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet

Stanley Peters, and Dag Westerståhl. 2006. *Quantifiers in language and logic*. Oxford University Press.


