

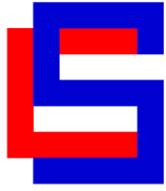
(Dis)embodiment 2022

Proceedings of the 2022 CLASP Conference on (Dis)embodiment

Simon Dobnik, Julian Grove and Asad Sayeed (eds.)



**Gothenburg and online
15–16 September 2022**



CLASP centre for
linguistic theory
and studies in probability

CLASP Papers in Computational Linguistics, Volume 4
©2022 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-67-4
ISSN 2002-9764

Introduction

This volume contains the papers presented at the CLASP Conference on (Dis)embodiment at the Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg, held on September 15–16, 2022.

(Dis)embodiment brings together researchers from several areas examining the role of grounding and embodiment in modelling human language and behaviour – or limits thereof. The conference covers areas such as machine learning, computational linguistics, theoretical linguistics and philosophy, cognitive science and psycholinguistics, as well as artificial intelligence ethics and policy.

Papers were invited on topics from these and closely related areas, including (but not limited to) large-scale neural language modelling, both text-only and multimodal; training corpus and test task development; visual, dialogue and multi-modal inference systems; neurolinguistic and psycholinguistic experimental approaches to human language processing; philosophical discussions of linguistic groundedness and embodiment (or limits thereof) as it pertains to computational modelling; semantics and pragmatics in neural models; dialogue modelling and linguistic interaction; formal and theoretical approaches to language production and comprehension; statistical, machine learning and information theoretic approaches that either avoid or embrace groundedness and/or embodiment; methodologies and practices for annotating dialogue and multi-modal datasets; visual, dialogue and multi-modal generation; text generation in both the dialogue and monologue settings; multimodal and grounded approaches to computing meaning; semantics-pragmatics interface; social and ethical implications of the development and application of neural language models, as well as relevant policy implications and debates.

This conference aims to initiate a genuine discussion between these related topics and to examine different approaches and how they can inform each other. It features 3 invited talks by leading researchers, 9 peer-reviewed archival papers and 7 non-archival presentations.

We would like to thank all our contributors and programme committee members, with special thanks to CLASP for organising the hybrid conference and our sponsors SIGSEM <http://sigsem.org>, the ACL special interest group on semantics, and the Swedish Research Council for funding CLASP.

Simon Dobnik, Julian Grove and Asad Sayeed

Gothenburg

September 2022

Programme Committee:

Kathrein Abu Kwaik	University of Gothenburg
Afra Alishahi	Tillburg University
Alexander Berman	University of Gothenburg
Raffaella Bernardi	University of Trento
Jean-Philippe Bernardy	University of Gothenburg
Yonatan Bisk	Carnegie Mellon University
Ellen Breitholtz	University of Gothenburg
Harry Bunt	Tillburg University
Stergios Chatzikyriakidis	University of Crete
Alexander Clark	University of Gothenburg
Robin Cooper	University of Gothenburg
Ryan Cotterell	Eidgenössische Technische Hochschule Zürich
Simon Dobnik	University of Gothenburg
Markus Egg	Humboldt University
Adam Ek	University of Gothenburg
Katrin Erk	University of Texas at Austin
Chris Fox	University of Gothenburg
Jonathan Ginzburg	Université Paris-Diderot
Dimitra Gkatzia	Edinburgh Napier University
Eleni Gregoromichelaki	University of Gothenburg
Julian Grove	University of Gothenburg
Xudong Hong	Saarland University
Christine Howes	University of Gothenburg
Nikolai Ilinykh	University of Gothenburg
Cassandra Jacobs	State University of New York at Buffalo
Elisabetta Jezek	University of Pavia
Richard Johansson	Chalmers Technical University
Aram Karimi	University of Gothenburg
Ruth Kempson	King's College London
Nikhil Krishnaswamy	Colorado State University
Shalom Lappin	University of Gothenburg
Staffan Larsson	University of Gothenburg
Tal Linzen	New York University
Sharid Loáiciga	University of Gothenburg
Vladislav Maraev	University of Gothenburg
Yuval Marton	University of Washington
Elin McCreedy	Aoyama Gakuin University
Paul McKevitt	Ulster University
Louise McNally	Universitat Pompeu Fabra
Gregory Mills	University of Groningen
Joakim Nivre	Uppsala University
Bill Noble	University of Gothenburg
Manfred Pinkal	Saarland University
Violaine Prince	Université de Montpellier 2
James Pustejovsky	Brandeis University
Christian Retoré	Université de Montpellier
German Rigau	University of the Basque Country

Hannah Rohde	University of Edinburgh
David Schlangen	University of Potsdam
William Schuler	The Ohio State University
Sabine Schulte im Walde	University of Stuttgart
Gabriel Skantze	KTH Royal Institute of Technology
Vidya Somashekarappa	University of Gothenburg
Tim Van de Cruys	KU Leuven
Marten van Schijndel	Cornell University
Eva Maria Vecchi	University of Stuttgart
Carl Vogel	Trinity College Dublin
Alessandra Zarcone	Augsburg University of Applied Sciences
Sina Zarrieß	University of Bielefeld

Invited Speakers:

Afra Alishahi, Tilburg University
Felix Hill, DeepMind
Magnus Sahlgren, AI Sweden

Invited talk 1: Afra Alishahi

Getting closer to reality: Grounding and interaction in models of human language acquisition

Humans learn to understand speech from weak and noisy supervision: they manage to extract structure and meaning from speech by simply being exposed to utterances situated and grounded in their daily sensory experience. Emulating this remarkable skill has been the goal of numerous studies; however researchers have often used severely simplified settings where either the language input or the extralinguistic sensory input, or both, are small-scale and symbolically represented. I present a series of studies on modelling visually grounded language understanding.

Invited talk 2: Felix Hill

Three studies that show that artificial models of general intelligence learn better with language

Having and using language makes humans as a species better learners and better able to solve hard problems. I'll present three studies that demonstrate how this is also the case for artificial models of general intelligence. In the first, I show that agents with access to visual and linguistic semantic knowledge explore their environment more effectively than non-linguistic agents, enabling them to learn more about the world around them. In the second, I demonstrate how an agent embodied in a simulated 3D world can be enhanced by learning from explanations – answers to the question “why?” expressed in language. Agents that learn from explanations solve harder cognitive challenges than those trained from reinforcement learning alone, and can also better learn to make interventions in order to uncover the causal structure of their world. Finally, I'll present evidence that the skewed and bursty distribution of natural language may explain how large language models can be prompted to rapidly acquire new skills or behaviours. Together with other recent literature, this suggests that modelling language may make a neural network better able to acquire new cognitive capacities quickly, even when those capacities are not necessarily explicitly linguistic.

Invited talk 3: Magnus Sahlgren

The Singleton Fallacy: why current critiques of language models miss the point

There is currently a lively debate about the semantic (in)capabilities of current language models: do language models really understand language or are they simply stochastic parrots? Are we wasting our time in the pursuit of bigger and bigger models, and should we instead be climbing some other hill in the NLP landscape? This talk provides an overview over the different positions in the debate, and attempts to disentangle the debate by pointing out an argumentation error that is referred to as the singleton fallacy.

Table of Contents

<i>A Small but Informed and Diverse Model: The Case of the Multimodal GuessWhat!? Guessing Game</i> Claudio Greco, Alberto Testoni, Raffaella Bernardi and Stella Frank	1
<i>A Cross-lingual Comparison of Human and Model Relative Word Importance</i> Felix Morger, Stephanie Brandl, Lisa Beinborn and Nora Hollenstein	11
<i>Dispatcher: A Message-Passing Approach to Language Modelling</i> Alberto Cetoli	24
<i>In Search of Meaning and Its Representations for Computational Linguistics</i> Simon Dobnik, Robin Cooper, Adam Ek, Bill Noble, Staffan Larsson, Nikolai Ilinykh, Vladislav Maraev and Vidya Somashekarappa	30
<i>Can We Use Small Models to Investigate Multimodal Fusion Methods?</i> Lovisa Hagström, Tobias Norlund and Richard Johansson	45
<i>Embodied Interaction in Mental Health Consultations: Some Observations on Grounding and Repair</i> Jing Hui Law, Patrick Healey and Rosella Galindo Esparza	51
<i>Norm Participation Grounds Language</i> David Schlangen	62
<i>Where Am I and Where Should I Go? Grounding Positional and Directional Labels in a Disoriented Human Balancing Task</i> Sheikh Mannan and Nikhil Krishnaswamy	70
<i>From Speed to Car and Back: An Exploratory Study about Associations between Abstract Nouns and Images</i> Ludovica Cerini, Eliana Di Palma and Alessandro Lenci	80

A small but informed and diverse model: The case of the multimodal GuessWhat?! guessing game

Claudio Greco

CIMeC - University of Trento
claudio.greco@unitn.it

Alberto Testoni

DISI - University of Trento
alberto.testoni@unitn.it

Raffaella Bernardi

CIMeC and DISI - University of Trento
raffaella.bernardi@unitn.it

Stella Frank*

Pioneer Centre for AI -
University of Copenhagen
stfr@di.ku.dk

Abstract

Pre-trained Vision and Language Transformers achieve high performance on downstream tasks due to their ability to transfer representational knowledge accumulated during pre-training on substantial amounts of data. In this paper, we ask whether it is possible to compete with such models using features based on transferred (pre-trained, frozen) representations combined with a lightweight architecture. We take a multimodal guessing task as our testbed, GuessWhat?!. An ensemble of our lightweight model matches the performance of the fine-tuned pre-trained transformer (LXMERT). An uncertainty analysis of our ensemble shows that the lightweight transferred representations close the data uncertainty gap with LXMERT, while retaining model diversity leading to ensemble boost. We further demonstrate that LXMERT’s performance gain is due solely to its extra V&L pretraining rather than because of architectural improvements. These results argue for flexible integration of multiple features and lightweight models as a viable alternative to large, cumbersome, pre-trained models.

1 Introduction

Current multimodal models often make use of a large pre-trained Transformer architecture component, which is then fine-tuned for the final task. This setup can lead to high performance, due to the immense amount of data embodied in the pre-trained component, together with its large capacity in terms of parameters. However, these models are also extremely costly to train; even fine-tuning requires non-negligible resources. Here we wonder whether the need for pre-training data, and for large and computationally costly neural networks could be mitigated by feeding the models with richer candidate representations, and by using an ensemble of lightweight models.

* Work done while at CIMeC - University of Trento.

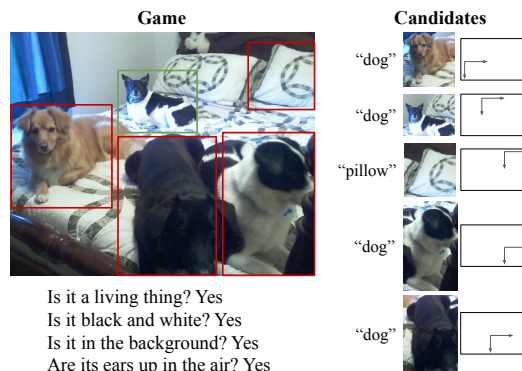


Figure 1: An example of a game from GuessWhat?!. The Guesser receives the image and dialogue as input, and has to pick the correct target (in green) from the list of candidates. We consider different ways of representing the candidates, e.g. using category information, visual features, and/or spatial position.

The motivation for this paper is to disentangle the contributions to good model performance on grounded multimodal tasks: is it due to better architecture, e.g. Transformers for text, vs LSTMs? Or to exposure to large amounts of multimodal data during pretraining? Or learning good representations for the task during fine-tuning?

As the task for this case study we take a multimodal referential game, Guess What?! (GW) (de Vries et al., 2017), specifically the final guessing task. Here the aim is to guess which object in the image is the correct target, based on the dialogue history (the series of questions) and the image. The model receives as input an image and a sequence of question-answer pairs, which are together passed to the multimodal encoder. The final hidden layer of the encoder is then the Guesser’s input representation. The Guesser classifier then generates a representation for each of the candidate objects in the image, and selects the target based on the similarity between the candidate representations and the input (image

and dialogue) encoding. The creation of the Guesser representation is the focus of this paper. We examine both the effect of input encoding (LXMERT vs V-LSTM) and the effect of using different features to represent the candidate targets. Our results show that these two factors interact: a lightweight input encoding can be compensated for by richer candidate targets; conversely, a heavy encoder can work with impoverished features.

When introducing the GW task and the baseline models, [de Vries et al. \(2017\)](#) run a comparative evaluation of the Guesser model; the visual embedding was shown to be not informative in selecting the target object and better results were obtained by using just the embedding learned from the category label and spatial coordinates. All the following work on the GW task retained the baseline candidate representations. We put the attention on this evaluation again situating it in the new context of the pre-trained large encoders we are now familiar with. Our motivation behind re-evaluating the features used by the Guesser is the observation that in GW, guessing the right target could require different sorts of information about the candidates. For instance, in the game illustrated in [Figure 1](#), the Guesser requires candidate representations that encode the ontological information that dogs are living things whereas pillows are not, in order to make sense of the dialogue. The candidate representations also have to distinguish the target dog from the other dogs: here, visual features encoding the colour could differentiate the black and white dogs from the other two dogs. Finally, spatial information is essential to locate the target in the background. Note that if the choice set contained other distractors, other features might have been needed to identify the target. We hypothesize (and our results confirm) that the combination of visual features and spatial location information on the level of individual candidates, plus some kind of semantic information about candidate categories (e.g. task-specific embeddings or general pre-trained embeddings) will lead to the best Guesser.

Along with improving the candidate representations, we also investigate the effect of ensembling multiple models. As well as potentially bringing a boost in performance, ensembling also allows us to inspect the uncertainty of the models under a Bayesian interpretation of deep ensembles as Bayesian model averaging ([Lakshminarayanan](#)

[et al., 2017](#); [Wilson and Izmailov, 2020](#); [Hüllermeier and Waegeman, 2021](#)). Hence, after comparing models based on their task-success, we use data and model uncertainty as a post-hoc analysis to get an in-depth comparison of models' behaviour. This allows us to better understand the effect of richer features: for V-LSTM Guessers, they provide key information, while for pre-trained LXMERT they seem to be redundant with the input encoding. However, for LXMERT without pre-training, the Guesser is not as able to integrate the information from the input encoding and candidate feature representations.

To recap, we investigate whether providing informative candidate representations (which are, themselves, gleaned from pre-trained models) to the Guesser model can make the task more feasible when using lightweight input (dialogue + image) encoders, i.e. V-LSTMs vs LXMERT. LXMERT has the advantage of significant pretraining on a large corpus of V&L data, as well as orders of magnitude more parameters. Hence, we compare V-LSTM ensembles against LXMERT, both alone and ensembled. We also compared the LXMERT architecture trained from scratch on the GW task, to understand the relative contributions of pretraining vs architecture. We compare the models both in terms of task-accuracy as well as doing an uncertainty analysis. Our results show that

- an ensemble of lightweight models with good candidate representations can match the performance of a single LXMERT model;
- while with poor candidate representations V-LSTM models are highly uncertain, richer candidate representations let these models behave similarly to the pre-trained LXMERT in terms of data/model uncertainty;
- better candidate representations lead to V-LSTM ensembles with Guessers that usefully disagree: ensembles can combine these disparate predictions into more accurate overall predictions.

2 Related Work

2.1 GuessWhat?! Guesser

GuessWhat?! ([de Vries et al., 2017](#)) is a dataset of human dialogues collected via Amazon Mechanical Turk in which two players play a guessing game. One player (the oracle) is assigned an object in an

image and the other player (the questioner) has to ask Yes/No questions in order to discover the target object. In the first GW model proposed in [de Vries et al. \(2017\)](#), the questioner player is implemented by two different models: the Question Generator and the Guesser. The Guesser is trained to predict the target object from a set of candidates, using supervised learning. Candidate objects are represented by a learned object category embedding and spatial coordinates.

This simple baseline Guesser has been used in most of the subsequent work on GW. [Shekhar et al. \(2019\)](#) proposed an alternative questioner model (GDSE) in which the Question Generator (QGen) and the Guesser are jointly trained, but the latter still receives the simple candidate representations used by [de Vries et al. \(2017\)](#).

The little work that has focused on the Guesser has retained the baseline candidate representations. [Pang and Wang \(2020\)](#) investigate the dynamics of the Guesser over the course of the dialogue, while [Suglia et al. \(2020\)](#) add an imagination module to improve grounded conceptual learning within the dialogue encoder.

[Greco et al. \(2021\)](#) evaluate the role of the encoder in the Guesser by comparing the blind LSTM encoder, found to work best in [de Vries et al. \(2017\)](#), with a multimodal LSTM (V-LSTM) and a multimodal universal encoder (LXMERT). None of this work has studied the effect of the candidate representation choices within the standard model.

Most recently, [Matsumori et al. \(2021\)](#) propose a new transformer-based architecture for GW, while [Tu et al. \(2021\)](#) evaluate the impact of ViBERT as encoder and design a state-estimator for the Guesser that let it accumulate belief state incrementally. While these models perform well, they are significantly larger and more complex, and do not permit the targeted study done in this paper.

2.2 Deep Ensembles and Uncertainties

Initial work on uncertainty estimation in deep neural networks was within the area of Bayesian Neural Networks ([Gal, 2016](#); [Kendall and Gal, 2017](#); [Depeweg et al., 2018](#)). [Lakshminarayanan et al. \(2017\)](#) showed that deep (non-Bayesian) ensembles can also be used for uncertainty estimation; in fact, in many empirical settings they work better, due to better exploration of the parameter space ([Ashukha et al., 2020](#); [Fort et al., 2019](#)). Deep ensembles are equivalent to Bayesian model averaging, where

averaging over component predictions is analogous to calculating the expected predictive posterior while marginalising over parameters ([Wilson and Izmailov, 2020](#)).

Uncertainty estimation has not received much attention in the multimodal NLP or grounded dialogue setting, with the exception of [Xiao and Wang \(2021\)](#), who use uncertainty decomposition to understand the hallucination behaviour of question generators. [Abbasnejad et al. \(2018\)](#) present a reinforcement learner for grounded dialogue which takes uncertainty into account when learning which questions to ask, and also for deciding when to stop asking questions. This is an orthogonal approach to ours, which uses uncertainty as a post-hoc analysis method, rather than integrating it into the model.

3 Guesser Model

In this section we describe the Guesser model. We use the same Guesser architecture introduced in [de Vries et al. \(2017\)](#) which has been employed in virtually all follow-up work on GW. The Guesser receives as input a 512D vector, encoding the grounded dialogue, and a vector representation of each candidate. This vector representation is the result of feeding the concatenated features for each candidate through a two layer MLP with ReLU activations, resulting in a 512D vector. The Guesser then computes a dot product between the vector representing the grounded dialogue and each candidate representation (processed by the MLP described above). The resulting scores are combined into a softmax layer, resulting in a probability distribution over the candidates. Note that the MLPs share parameters between candidates.

In our experiments, we use as encoder LXMERT or V-LSTM, and study the impact of using an ensemble of encoders together with the enrichment of candidate embeddings.

3.1 Grounded Dialogue Encoder

The encoder generates a grounded dialogue representation from the image and the set of questions and answers. In our experiments, we use two different multimodal encoders following ([Greco et al., 2021](#)):

LXMERT is a transformer-based multimodal encoder ([Tan and Bansal, 2019](#)). It represents an image by the set of position-aware object embeddings for the 36 most salient regions detected by a Faster R-CNN ([Ren et al., 2016](#)) and the text

by position-aware word embeddings. LXMERT is pre-trained on five vision-and-language tasks whose images come from MS-COCO and Visual Genome (Krishna et al., 2017). In our experiments, we fine-tune the pre-trained LXMERT model on GW. We generate our 512D vector representing the grounded dialogue by taking the 768D vector from the [CLS] initial token of LXMERT and by giving it to a feedforward layer with Tanh activation.

We also experimented with LXMERT trained from scratch. In this case, the encoding of the image is still provided by Faster R-CNN (pre-trained on Visual Genome) but the language encoding, as well as the combined representations of L&V, are learned from the GW dataset only.

V-LSTM is a relatively lightweight encoder that represents the dialogue history as the 1024D last hidden state from a LSTM receiving the dialogue, concatenates that vector with a 2048D representation of the image extracted from the penultimate layer of a ResNet-152 pre-trained on ImageNet (He et al., 2016), and gives the concatenation to a feedforward layer with Tanh activation to generate a 512D vector representing the grounded dialogue.

We consider the V-LSTM lightweight because it has $\sim 18\times$ fewer parameters and thus requires much less training (data and time) than LXMERT.

3.2 Candidate representation

In de Vries et al. (2017) and following papers (e.g. Pang and Wang (2020); Greco et al. (2021)), each candidate is represented by a spatial embedding, encoding its bounding box location, and a category embedding learned during training, based on the candidate’s MS-COCO (Lin et al., 2014) label. We question this representation, which could be lacking important information about the candidate with respect to the dialogue and that cause the need of a powerful universal multimodal encoder. According to Shekhar et al. (2019), questions about category and location make up about 65% of the human questions: the baseline model might be sufficient for these cases. However, 15.5% of questions include colour, which the baseline Guesser cannot see. Nearly as many (14.5%) mention an object’s super category (‘animal’, ‘utensil’), which also is information not necessarily included in the embeddings learned from the training games. Hence, we build richer candidate representations starting from the following components:

<https://github.com/airsplay/lxmert>

Spatial information `spatial` is represented by a 8D vector that encodes the location of the candidate’s bounding box. Since the Guesser does not have direct access to the image but only sees it via the encoded grounded dialogue embedding, the spatial coordinates locate the object in the image. Hence, they are very informative for the selection task, especially when multiple candidates look the same at a type/category level and share the most salient visual attributes (like the two black and white dogs in Figure 1.) Moreover, dialogues often refer to objects using their location (e.g. “the dog on the right”) that the Guesser can exploit better by having access to the spatial coordinates.

Category information `cat` is given by a 256D category embedding, representing the candidate’s category according to the MS-COCO label. This learned embedding encodes the conceptual representation of the object emerging from its co-occurrences with dialogue and image features within the GW training data.

GloVe embeddings `glove` representations are the 300D pre-trained word embeddings (GloVe (Pennington et al., 2014)) of the word corresponding to the category label, scaled down to 256D using a feedforward layer with ReLU activation. (When the label is a multi-world label, e.g. “*dining table*” we take the mean over the words in the expression). `glove` embeddings, despite some limitations, are shown to be effective at object-property tasks (Lucy and Gauthier, 2017; Forbes et al., 2019) and at capturing taxonomic relations (Da and Kasai, 2019).

Visual information `visual` representations are obtained from a ResNet-152, pre-trained on ImageNet, which receives as input the crop of the object. This visual vector is input to a feed-forward layer with ReLU activation, in order to obtain a 256D vector. This embedding should provide the visual attributes of the entity it represents, which are expected to play a crucial role in games in which there are distractors of the same category of the target objects. For instance, the dialogue identifies the target as a “black and white dog” in Figure 1, but without visual features the dogs are indistinguishable – in contrast with the results reported in the original GW paper (de Vries et al., 2017) about the lower performance obtained by the Guesser when given the visual embedding of the candidates bounding boxes.

We experiment with different combinations of these basic components. We evaluate models with all the 2-input combinations, apart from `glove+cat`, which cannot be sufficiently discriminative in games that contain distractors of the same category of the target object. The spatial and visual embeddings provide token specific complementary information, whereas the `cat` and `glove` embeddings are both meant to encode concept representations. Hence, we experiment only with the following 3-input representations: `cat+visual+spatial` and `glove+visual+spatial`. Finally, to check the degree to which `cat` and `glove` provide redundant information, we try the 4-input embedding containing all the basic components above, `cat+glove+visual+spatial`.

3.3 Training procedure

We minimize the cross-entropy error with respect to the ground-truth annotation during training, using the Adam optimizer (Kingma and Ba, 2014) for V-LSTM and Adam with a linear-decayed learning-rate schedule for LXMERT (Devlin et al., 2018). We perform early stopping with ten epochs of patience.

3.4 Guesser ensembles

We follow the standard deep ensemble setup (Lakshminarayanan et al., 2017) of training independent Guessers by training them with different random seeds. All our Guesser ensembles consist of five Guessers of the same type (i.e., having the same encoder and set of candidate representation input information). Different random seeds mean the Guessers differ in their random initialisations (except for the weights of the pre-trained LXMERT encoder) and the order in which they see the data. An ensemble of Guessers generates predictions using the average of the Guesser prediction distributions.

4 Measuring Uncertainties

The uncertainty of a model, parameterised as θ , is commonly measured by the entropy of the predictive distribution $p_\theta(y|x)$, averaged over a test set. For each example x , a confident model will put most probability mass on a single choice y , leading to low entropy, while an uncertain model will spread its bets, leading to higher entropy.

Within an ensemble, the ensemble *total uncertainty* is the entropy of its predictive distribution, which combines the distributions of the N ensemble components:

$$H[p(y|x)] = H[1/N \sum_{n=1}^N p_{\theta_n}(y|x)]. \quad (1)$$

(This is the sample-based approximation to marginalising over θ .) Note that an ensemble can have high uncertainty (high entropy) either because of noisy or ambiguous data leading to an inability to make a confident decision, or because its components disagree (Depeweg et al., 2018; Hüllermeier and Waegeman, 2021).

We can also measure the average uncertainty of each ensemble component on its own: $1/N \sum_{n=1}^N H[p_{\theta_n}(y|x)]$. This factor is known as *data uncertainty*: it measures whether the datapoint is sufficiently informative for each model to make a confident decision. If x is inherently ambiguous, then all models should have high uncertainty. Total ensemble uncertainty will also be high, due to the combination of uncertain predictions.

The difference between total uncertainty and data uncertainty is *model uncertainty*, which measures the extent to which the models disagree (i.e., the extent to which the ensemble’s predictive distribution does not match the average ensemble component). Model uncertainty is always non-negative.

In this paper we compare different models, differing in their choice representations, as ensembles. Within the GW Guesser, the choice representation should be considered part of the input x . Inadequate choice representation will thus lead to high data uncertainty, since the representation is not sufficient to make confident decisions. Since the humans playing the original GW game, generating the test and training data, guessed correctly, overly high “data uncertainty” values point to problems with data representations, rather than inherently ambiguous data.

Formally, within a Bayesian framework, it is the mutual information between y and the ensemble parameters θ estimated from data D , derived from the difference between entropy and crossentropy: $H[p(y, |x, D)] = E_{\theta|D} H[p(y|x, \theta)] - MI[y, \theta|x, D]$, where the left hand term is total uncertainty (marginalising over θ) and the first term on the right is data uncertainty.

We note here that, while ‘data uncertainty’ has been identified with ‘aleatoric uncertainty’ (Kendall and Gal, 2017; Depeweg et al., 2018; Malinin and Gales, 2018), namely the true uncertainty of the example in the world (Der Kiureghian and Ditlevsen, 2007), this doesn’t hold inasmuch as the *representation* of the data is a modelling decision (see also Hüllermeier and Waegeman (2021), Sec 2.3). Comparing different data representations doesn’t change the true aleatoric uncertainty, which is an lower bound on data uncertainty.

Whether model uncertainty should also be minimised is a different question. In theory, if all ensemble components have found the global optimum, model uncertainty will be zero. In practice, not being able to find the global optimum, we use ensembles to approximate a distribution over good local optima. Ensemble ‘boost’ (the improvement in performance over the component average) also requires model diversity. It is thus more useful to have a collection of strong but different opinions (low data, high model uncertainty) than homogeneous equivocal opinions (high data, low model uncertainty).

5 Experiments

Prior work has shown that LXMERT outperforms V-LSTM when using the standard candidate representation, which uses only spatial and category embeddings: LXMERT accuracy is 69.2% while V-LSTM reaches just 64.5% (Greco et al., 2021). Below we ask whether V-LSTM accuracy can reach LXMERT’s if provided with different candidate representations. Improved candidate representations should make it easier for the fine-tuned model to learn to match the correct target with the image and dialogue encoding, by facilitating the match between features of the candidates and the features discussed in the dialogue.

Secondly we experiment with ensembling our models. We find that ensembling the V-LSTM models leads to a larger boost in accuracy, while ensembling the LXMERT models helps less. This is a very convenient result, since training ensembles of lightweight V-LSTMs is fast and computationally cheap, compared to fine-tuning even a single LXMERT. Analysing ensemble uncertainty confirms that V-LSTM encodings allow the Guesser to make better use of improved candidate representations, while they have less of an effect on LXMERT. Furthermore, we see that an ensemble of V-LSTM Guessers contains sufficient diversity, in terms of model uncertainty, to make ensembling worth it.

5.1 Task success

Candidate representations In this experiment, we evaluate the effect of different candidate representations on Guessers based on V-LSTM encoders. We combine the candidate representations described in Section 3.2: `cat`, `glove`, `visual`, and `spatial`, in various configurations.

The results in Table 1 show that the rep-

Candidate rep.	Guessers	Ens.
<code>cat+sp</code>	64.49±0.12	66.40
<code>cat+vis</code>	59.17±0.23	61.07
<code>glove+sp</code>	64.84±0.18	67.21
<code>glove+vis</code>	58.08±0.52	61.03
<code>vis+sp</code>	55.19±0.55	60.58
<code>cat+vis+sp</code>	66.45±0.25	69.61
<code>gl+vis+sp</code>	66.72±0.19	70.12
<code>cat+gl+vis+sp</code>	66.58±0.26	69.58

Table 1: Test set accuracies for Guessers with different candidate representations, individually and in an ensemble. `cat`, `gl`, `vis`, and `sp` stand for *category*, *glove*, *spatial*, and *visual*. LXMERT with `cat+sp` obtains 69.2% (Greco et al., 2021).

resentation of the candidates has a large effect on Guesser performance. The worst combination, `visual+spatial`, is ten percentage points worse than the best combination, `glove+visual+spatial` (55.19% vs. 66.72%). Models with three or four types of candidate representations outperform models with only two types; however there is not a benefit of combining all four types over only three. Category/type information is crucial for success: the model with only token-level information, `visual+spatial`, clearly underperforms all the others. The category representations, namely `cat` and `glove`, lead to similar results when combined with other representations, and do not benefit from being combined together (unlike the token representations). `glove` representations do seem to be slightly more beneficial than `cat` representations, indicating that the additional world knowledge that they contain can be useful. (See Figure 2 for an example where `glove` representations allow the model to guess correctly.)

Ensembling The results of ensembling five versions of each V-LSTM Guesser are reported in Table 1. We compare them against the accuracy reached by the guesser based on LXMERT, viz. 69.2% (Greco et al., 2021). Surprisingly, the accuracy reached by the simple but well informed ensemble model, V-LSTM with the spatial, visual and `glove` embeddings, is as high as the task-accuracy reached by the guesser based on LXMERT. Indeed, ensembling V-LSTM brings in a boost of 3.4 points from 66.72% to 70.12% accuracy.

We believe this to be a remarkable result, given

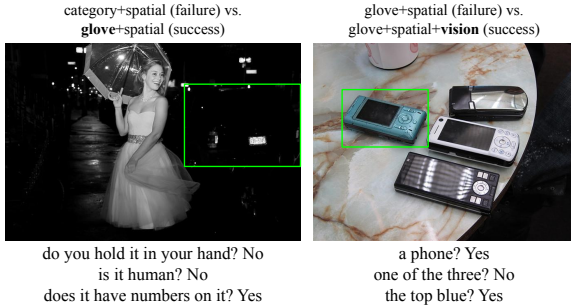


Figure 2: Representative examples of the contribution of different features. On the left: contribution of Glove embeddings on common sense reasoning (cars have numbers on them – plates). On the right: contribution of visual features (colors).

the difference in number of parameters (V-LSTM ensemble: 11M vs. LXMERT: 209M) between the two models and that training a single LXMERT takes significantly longer than training an ensemble of V-LSTM Guessers (V-LSTM Ensemble: $45m \times 5 = 3h54m$; one LXMERT: 21hrs).

We interpret these results as showing that indeed LXMERT has better visual representations of the input image and better lexical grounded information gained through the pretraining on multimodal corpora, but that such gain acquired through its heavy pretraining phase, can be easily reached by simply enriching the candidate representations.

Next, we check whether LXMERT could also benefit from ensembling and from the richer candidate representations. We evaluate both pre-trained LXMERT and LXMERT-scratch trained only on GW. When ensembling pre-trained LXMERT, the only source of randomness is in the order of the training data, while for LXMERT-scratch, the random initializations also differ between ensemble components. For both LXMERT and LXMERT scratch, training an ensemble for GW requires $21hrs \times 5 = 4.3$ days, 27 times more compute than the V-LSTM Ensemble.

As shown in Table 2, ensembling the pre-trained LXMERT brings only a minimal increase in accuracy, while it does provide a boost for LXMERT-scratch. Both the individual models and the ensemble LXMERT-scratch perform only on par with V-LSTM. Interestingly, ensembling LXMERT-scratch shows a similar pattern to V-LSTM, where using improved `glove+visual+spatial` representations leads to a larger ensemble boost than the `cat+spatial` representations.

Together these results indicate that LXMERT’s

improved performance hinges on being able to use the information seen during pretraining, rather than architectural improvements. Furthermore, for LXMERT, adding candidate representations (like `glove` and `visual`) that are extracted from pre-trained models, and thus incorporate similar kinds of pretraining knowledge to the LXMERT encoder, does not help over the weaker candidate representations (`cat+spatial`). We presume this is due to the pre-trained LXMERT model already having learned the relevant ontological information, as well as the ability to localise visual information, and thus not needing it to be provided explicitly. This hypothesis is strengthened by the weaker performance of LXMERT-scratch, which has not been able to learn the relevant features from additional pretraining data.

5.2 Uncertainty analysis

As described in Section 4, we can distinguish between model and data uncertainty in the ensemble. *Data uncertainty* measures the average uncertainty of each ensemble component on an example. In our setting, we expect improved candidate representations to lead to lower data uncertainty, since models should be better informed. As we can see from Figure 3, V-LSTM models show this pattern, with `glove+visual+spatial` candidates leading to lower data uncertainty compared to the `cat+spatial` baseline. However, LXMERT models do not: regardless of candidate representations, they show the same level of data uncertainty. LXMERT-scratch shows a reduction of data uncertainty with `glove+visual+spatial` but to a lesser degree than V-LSTM. In this case the transformer architecture is actually preventing the best use of the information from the candidate features.

The *model uncertainty* measures the extent to which the models disagree (i.e., the extent to which the ensemble’s predictive distribution does not match the average ensemble component). Here again, the pre-trained LXMERTs show no effect of candidate representations. However, for V-LSTM, and again to a lesser extent LXMERT-scratch, the improved representations *increase* the model uncertainty, indicating an increase in model diversity (and subsequently leading to a larger ensemble boost). This again demonstrates the importance of good candidate representations for this model: with `cat+spatial`, all ensemble components were uncertain in the same way, while

Model	Candidate Rep.	Guessers	Ensemble	Boost
V-LSTM	cat+sp	64.49±0.12	66.40	1.9
V-LSTM	gl+vis+sp	66.72±0.19	70.12	3.4
LXMERT-scratch	cat+sp	64.70±0.41	66.50	1.8
LXMERT-scratch	gl+vis+sp	66.3 ±0.39	68.90	2.6
LXMERT	cat+sp	69.73±0.46	71.55	1.8
LXMERT	gl+vis+sp	69.56±0.27	71.57	2.0

Table 2: Average guesser performance vs Ensemble performance: Boost is improvement in performance due to ensembling. Ensembling benefits V-LSTM with good candidate representations most.

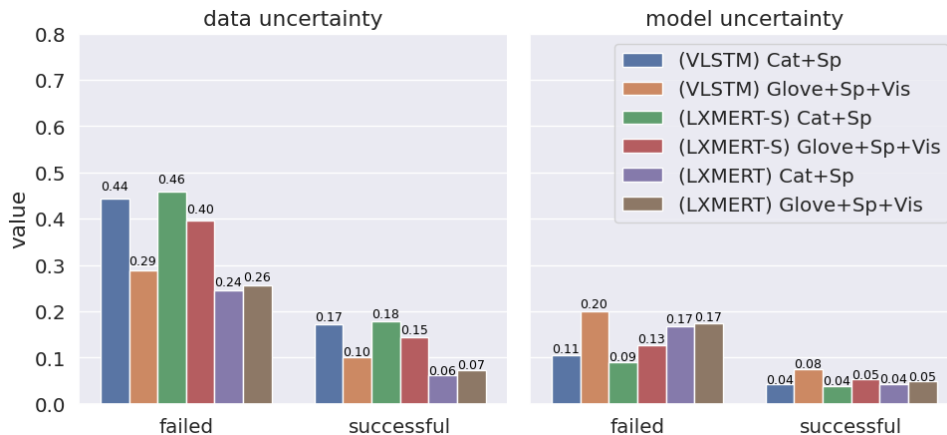


Figure 3: Total, data, and model uncertainty for different ensembles considering games with 5 candidate objects. Improved candidate representations decreases data uncertainty for V-LSTM, but not for LXMERT or LXMERT-scratch.

glove+visual+spatial representations lead to useful diversity.

6 Conclusion

In this paper, we re-evaluated the need for a deep pre-trained multimodal encoder on a testbed multimodal guessing task (GW). We demonstrated that a lightweight V-LSTM model was able to achieve matching performance, given useful features and the reduction in uncertainty enabled by ensembling.

We show that for GuessWhat?!, the candidate representations that lead the V-LSTM ensemble to reach higher accuracy are those encoding ontological (glove), visual and spatial information. The pre-trained model does not profit from either of the richer representation, or the ensemble. The uncertainty analysis of the ensemble models shows that while with poor candidate representations V-LSTM models are highly uncertain, richer candidate representations let these models behave more similarly to the pre-trained LXMERT in terms of both data and model uncertainty.

The richer candidate representations effectively

transfer information from other corpora (glove) or visual recognition models (visual). These features do not match exactly what LXMERT sees during pretraining, but given that LXMERT does not benefit from them, they do not seem to add crucial information for LXMERT. Conversely, V-LSTM benefits from these ‘cheap’ features to the extent of matching deep contextual model performance, indicating a continuing role for these types of representations in grounded language tasks.

Hence, we conclude that the good performance obtained by the Guesser when based on the pre-trained multimodal Transformer is not due to its architecture or to the representation learned during fine-tuning, but rather to the exposure to a large amount of multi-modal data during pre-training. Our results may help mitigate environmental issues given by the training of large models. It remains to be seen whether these results hold for other tasks and other unimodal and multimodal models.

References

- Hsan Abbasnejad, Qi Wu, Javen Shi, and Anton van den Hengel. 2018. What’s to know? Uncertainty as a guide to asking goal-oriented questions.
- Arsenii Ashukha, Alexander Lyzhov, Dmitri Molchanov, and Dmitry Vetrov. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *ICLR*.
- Jeff Da and Jungo Kasai. 2019. [Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations](#). *CoRR*, abs/1910.01157.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.
- Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udfluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *ICML*.
- Armen Der Kiureghian and Ove Ditlevsen. 2007. Aleatory or epistemic? Does it matter? In *Special Workshop on Risk Acceptance and Risk Communication*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. [Do neural language representations learn physical commonsense?](#) In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 1753–1759. cognitivesciencesociety.org.
- Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2019. [Deep ensembles: A loss landscape perspective](#).
- Yarin Gal. 2016. *Uncertainty in deep learning*. Ph.D. thesis, University of Cambridge.
- Claudio Greco, Alberto Testoni, and Raffaella Bernardi. 2021. Grounding dialogue history: Strengths and weaknesses of pre-trained transformers. In *Advances in Artificial Intelligence AIXIA 2020*, volume 12414 of *Lecture Notes in Computer Science*, pages 263–279. Springer Nature Switzerland AG.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Eyke Hüllermeier and Willem Waegeman. 2021. [Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods](#). *Machine Learning*, 110(3):457–506.
- Alex Kendall and Yarin Gal. 2017. [What uncertainties do we need in Bayesian deep learning for computer vision?](#) In *Advances in Neural Information Processing Systems*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). In *NeurIPS*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Li Lucy and Jon Gauthier. 2017. [Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning](#). In *Proceedings of the First Workshop on Language Grounding for Robotics, RoboNLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 76–85. Association for Computational Linguistics.
- Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *NeurIPS*.
- Shoya Matsumori, Kosuke Shingyouchi, Yuki Abe, Yosuke Fukuchi, Komei Sugiura, and Michita Imai. 2021. Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1898–1907.
- Wei Pang and Xiaojie Wang. 2020. Guessing state tracking for visual dialogue. In *ECCV*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. [Beyond task success: A closer look at jointly learning to see, ask, and Guess-What](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alessandro Suglia, Antonio Vergari, Ioannis Konstas, Yonatan Bisk, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2020. [Imagining grounded conceptual representations from perceptual information in situated guessing games](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1090–1102, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.
- Tao Tu, Qing Ping, Govindarajan Thattai, Gokhan Tur, and Prem Natarajan. 2021. Learning better visual dialog agents with pretrained visual-linguistic representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5622–5631.
- Andrew Gordon Wilson and Pavel Izmailov. 2020. [Bayesian deep learning and a probabilistic perspective of generalization](#).
- Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *ACL*.

A Cross-lingual Comparison of Human and Model Relative Word Importance

Felix Morger*
Gothenburg University

Stephanie Brandl
University of Copenhagen

Lisa Beinborn
Vrije Universiteit Amsterdam

Nora Hollenstein
University of Copenhagen

Abstract

Relative word importance is a key metric for natural language processing. In this work, we compare human and model relative word importance to investigate if pretrained neural language models focus on the same words as humans cross-lingually. We perform an extensive study using several importance metrics (gradient-based saliency and attention-based) in monolingual and multilingual models, including eye-tracking corpora from four languages (German, Dutch, English, and Russian). We find that gradient-based saliency, first-layer attention, and attention flow correlate strongly with human eye-tracking data across all four languages. We further analyze the role of word length and word frequency in determining relative importance and find that it strongly correlates with length and frequency, however, the mechanisms behind these non-linear relations remain elusive. We obtain a cross-lingual approximation of the similarity between human and computational language processing and insights into the usability of several importance metrics.

1 Introduction

Large pretrained neural language models, such as BERT (Devlin et al., 2019), have in recent years demonstrated performance equal to that of humans in a range of natural language understanding tasks (Wang et al., 2019). This begs the question of whether the processing and encoding of these models reflect language properties as described by language experts, such as in grammar, semantics, pragmatics and logic and, furthermore, whether the models process language similarly to humans. While extensive research is being done to answer this question, such as inquiries into what linguistic knowledge is encoded into contextual word representations (Clark et al., 2019; Vulić et al., 2020), how linguistic information is processed (Tenney

et al., 2019) and the effects of architectural choices (Rogers et al., 2020), more recent research inspired by psycholinguistics has emerged, which directly compares cognitive signals of language processing to pretrained language models. By using tools such as eye-tracking features and brain activity data (Abdou, 2022; Goldstein et al., 2022; Hollenstein et al., 2020a), this line of research skips the step of having to collect human judgments from speech or text data by directly comparing them to sources of cognitive data. As such, this approach is a direct means of testing whether models process language similarly to humans or, in other words, of evaluating the *cognitive plausibility* of computational language processing.

One method leveraging cognitive data has been to extract relative word importance, a key metric for natural language processing, from eye-tracking data in order to compare these to relative word importance extracted from state-of-the-art pretrained language models. This has been studied for normal English reading (Hollenstein and Beinborn, 2021; Bensemann et al., 2022), in task-specific reading (Eberle et al., 2022), and in question answering settings (Sood et al., 2020). In this work, we continue in this line of research but apply it across several languages to measure the extent to which pretrained language models focus on the same words as humans cross-lingually. We obtain *human relative word importance* from eye-tracking data (total reading time) and *model relative word importance* from pretrained language models using saliency and attention-based methods. These methods have in recent years been developed for the purpose of explainability, however, which methods serve best for this purpose has been a point of contention (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019).

The goal of this study is two-fold: On the one hand, we aim to obtain a rough estimate of the similarity between human and computational natural

E-mail: felix.morger@gu.se

language processing and, on the other hand, from a usability point of view, see which importance methods best approximate human relative word importance. We compare four methods for calculating relative word importance in pretrained language models, namely first-layer attention, last-layer attention, attention flow, and gradient-based saliency. To investigate whether the same trends hold across multiple languages and are not particular artifacts of one language, we use eye-tracking corpora of four different languages (English, Dutch, German, and Russian) and compare both monolingual and multilingual language models. More precisely, we make this comparison by looking at how human and model relative word importance statistically correlate across different languages.

Lexical properties such as word frequency and length are known to have a large effect on eye movements of any language (Just and Carpenter, 1980; Levy, 2008). Therefore, in an additional investigation, we analyze their impact on the fit between human and model relative importance.

To sum up, this work examines the following research questions:

Q1: Do human and model relative word importance correlate across languages?

Q2: Is there a difference between language-specific and multilingual language models?

Q3: Is there a difference between gradient-based saliency and the attention-based methods first-layer attention, last-layer attention, and attention flow?

Q4: To what extent is human and model relative word importance relying on word length and word frequency?

Contributions We show that human and model word importance correlate strongly in varying degrees across languages (English, Dutch, German and Russian), although the observed differences appear to be more corpus-specific than language-specific (Q1). We observe a slightly stronger correlation of monolingual models over multilingual models, in particular for first-layer attention and attention flow (Q2). We see that other attention-based methods than last-layer attention, i.e., first-layer and attention flow correlate strongly to human eye-tracking data with attention flow being on par with saliency for monolingual models (Q3). We see a strong correlation with the baselines (positive correlation to word length and negative correlation to word frequency). When using linear regression analysis to measure the ability of word length, word

frequency and model relative word importance to predict human relative word importance, we see that word frequency and word length increase the predictive power over model relative word importance alone, indicating that these baselines are not sufficiently accounted for by the models (Q4). The code for our experiments is available online.¹

2 Related Work

This work lies at the intersection of psycholinguistics, interpretability of neural networks, and natural language processing. More specifically, there are two current streams of research that this study directly draws from, namely relative importance metrics and cognitive analysis of natural language processing. Below, we outline the related works in these two subfields.

2.1 Relative Importance Metrics

Approaches for extracting relative word importance of Transformer-based models can be grouped into gradient-based, propagation-based, occlusion-based, and attention-based methods (Bastings and Filippova, 2020, Section 3). In this work, we focus on attention-based and gradient-based methods (see Section 4).

Attention is a key component of Transformer models and multiple studies have analyzed how attention weights are distributed across tokens. It has, for example, been shown that attention at different layers in Transformer models targets different linguistic aspects. For instance, Vig and Belinkov (2019) find that attention in a GPT-2 model targets different parts of speech and depths of dependency relations at different layers within the model and Li et al. (2021) show that different layers of transformer language models perform best when detecting different types of linguistic anomalies. Also, the findings by Tenney et al. (2019) indicate that in BERT earlier layers encode more word-level information than later layers when comparing performance across different language-level tasks from part-of-speech tagging to anaphora resolution.

The methodological merit of attention weights as a measure of relative importance has, however, been questioned. For one, the calculated attention attends to input representations, not the input itself, and these representations can mix in information from other inputs, thus diluting the relative impor-

¹https://github.com/felixhultin/relative_importance.

tance strength of the original input token (Bastings and Filippova, 2020). Moreover, different attention distributions can lead to the same predictions, making the relative importance of attention weights ambiguous (Jain and Wallace, 2019). To address the unreliability of attention weights, Abnar and Zuidema (2020) propose *attention flow*, a mechanism which computes maximum flow values, from hidden embeddings to input tokens.

2.2 Cognitive Analysis of Natural Language Processing

Using cognitive data to evaluate NLP has emerged as a novel method for interpreting NLP systems (Toneva and Wehbe, 2019; Ettinger, 2020; Hollenstein et al., 2019). The motivation behind this research is to assess whether models encode, process, or output language similarly to humans and, thus, provide measurements of their cognitive plausibility (Keller, 2010).

In recent years, more eye-tracking corpora from natural reading have become available in multiple languages (see section 3.1). Although cognitive data, including eye-tracking corpora, have been available as digitized formats for a long time, only recently have they been methodically deployed for the cognitive analysis of NLP systems. For example, the CMCL shared evaluation task uses ZuCo for the modeling of eye-tracking features. In this task, language models, such as BERT, are used to predict eye-tracking features (number of fixations, first fixation duration, total reading time, etc.) (Hollenstein et al., 2021). This work is similar to ours, but instead of fine-tuning the model to predict eye-tracking features, we see if the relative word importance as extracted by different methods correlates to mean total reading time.

Further work using other sources than eye-tracking corpora is for example Ettinger (2020), who proposed a psycholinguistic test suite to diagnose language models’ predictions in context using electroencephalogram (EEG). Moreover, Abnar et al. (2019) use functional magnetic resonance imaging (fMRI) and representational similarity analysis (RSA) to compare representations of the brain and pretrained language models.

In terms of using relative importance metrics, previous studies have shown that attention weights do not correspond to human relative word importance. For example, Sood et al. (2020) compare attention weights from the last layer of Transformer

models to human gaze data. They show that a higher correlation between model attention and human attention does not necessarily yield better performance in downstream NLP tasks. Hollenstein and Beinborn (2021) also find that attention has a weak correlation to human gaze data. Recently, Eberle et al. (2022) have found that attention flow from transformer models correlates strongly with human fixation times in task-specific English reading.

Finally, gradient-based methods have been proposed as a better method than attention weights at approximating the relative importance of input words in neural networks (Bastings and Filippova, 2020). Hollenstein and Beinborn (2021) additionally show that gradient-based saliency might be a cognitively more plausible interpretability metric than attention weights.

We follow these results and provide a large cross-lingual comparison of human eye-tracking data to a range of relative importance metrics, including gradient-based saliency, first and last-layer attention, and attention flow.

3 Data

3.1 Eye-tracking Corpora

We use eye-tracking data collected from native readers of the following corpora to extract the human relative importance metrics based on the mean total reading time of each word (see Table 1). For English, we use the GECO corpus, which contains eye tracking data from English monolinguals reading an entire novel (Cop et al., 2017), and the ZuCo corpus (Hollenstein et al., 2018, 2020b), which includes eye-tracking data of full sentences from movie reviews and Wikipedia articles.² For Dutch, we also use the GECO corpus, which additionally contains eye tracking data from Dutch readers that were presented with the same novel in their native language (Cop et al., 2017). For German, we leverage the Potsdam Textbook Corpus, which contains 12 short passages from college-level biology and physics textbooks, which are read by expert and laymen German native speakers (Jäger et al., 2021). We also use the Russian Sentence Corpus which includes naturally occurring sentences extracted from the Russian National Corpus (Laurinavichyute et al., 2019).³ We exclude a small set

²We use Tasks 1 and 2 from ZuCo 1.0 and Task 1 from ZuCo 2.0.

³<https://ruscorpora.ru>

Language	Corpus	Subjs.	Sents.	Sent. length	Tokens	Types	Word length
English	GECO	14	4,559	10.5 (1–69)	56,410	5,916	4.6 (1–33)
	ZuCo	30	853	19.5 (1–68)	20,545	5,560	5.0 (1–29)
Dutch	GECO	19	4,863	11.6 (1–60)	59,716	5,575	4.5 (1–22)
German	PoTeC	75	89	19.5 (5–51)	1,895	847	6.5 (2–33)
Russian	RSC	103	143	9.4 (5–13)	1,357	993	5.7 (1–18)

Table 1: Descriptive statistics of all eye-tracking datasets. Sentence length and word length are expressed as the mean with the min-max range in parentheses. **Sents.** is the number of the subset of sentences we process for this work, while sentence length and word length are calculated from all sentences in the corpora.

of sentences from the original corpora because of token alignment issues.

3.2 Language Models

All monolingual models and the multilingual model are based on the BERT architecture (Devlin et al., 2019). We use the pretrained checkpoints from the HuggingFace repository of the multilingual base model and language-specific monolingual base models. See Table 3 in the Appendix for the complete list of models and references.

4 Method

For each sentence in the eye-tracking corpora, we calculate human and model relative word importance values. The same sentences are, however, tokenized differently by the built-in tokenizers of the pretrained language models, resulting in longer sequences of relative word importance values than those obtained from humans. To remedy this, we align human and model importance values by discarding the importance values of special tokens (e.g. [SEP] and [CLS]) and merging subtokens and adding their values. Once aligned, we calculate Spearman’s correlation coefficient ρ between human and model relative word importance for each sentence. Finally, we calculate an average Spearman’s ρ across all sentences for each eye-tracking corpus (human relative word importance) and model, corpus, and importance metric tuple (model relative word importance). Additionally, using the same procedure, we explore the correlation of human and model relative word importance to word length and frequency baselines.

We analyze the following importance metrics:

Human relative importance In this work, we use the *total reading time* per word in the eye tracking corpora as the source for defining human relative word importance. It refers to the sum of all fixation durations for each word including regres-

sions (i.e. when a subject goes back to the same word after the first pass). We use the average total reading time across all subjects and normalize the resulting values such that each word is assigned an importance value between 0 and 1, and all values within a sentence sum up to 1. These values are calculated sentence by sentence.

Gradient-based saliency As described in Holtenstein and Beinborn (2021), we define a saliency vector for a masked token to indicate the importance of each of the tokens in the context of correctly predicting the masked token (Madsen, 2019). The saliency s_{ij} for input token \mathbf{x}_j for the prediction of the correct token \mathbf{t}_i is calculated as the Euclidean norm of the gradient of the logit for x_i :

$$s_{ij} = \|\nabla_{\mathbf{x}_j} f_{t_i}(\mathbf{X}_i)\|_2 \quad (1)$$

Last-layer attention We approximate relative importance using the attention values from the last layer and calculate the mean of all heads of each Transformer model as Sood et al. (2020).

First-layer attention Previous work has indicated that earlier layers encode information closer to word-level than later layers (Tenney et al., 2019). Therefore, we also include the first-layer attention weights by averaging over all heads to approximate relative word importance.

Attention flow Finally, we compute attention flow (Abnar and Zuidema, 2020), which has been shown to correlate stronger with human gaze than raw attention weights (Eberle et al., 2022). Attention flow considers the attention graph as a flow network and computes maximum flow values from later attention layers to the input embedding layer. Unlike raw attention weights, which consider token importance at layers in isolation, attention flow computes importance scores that account for mixing of information across layers and, thus, identifies

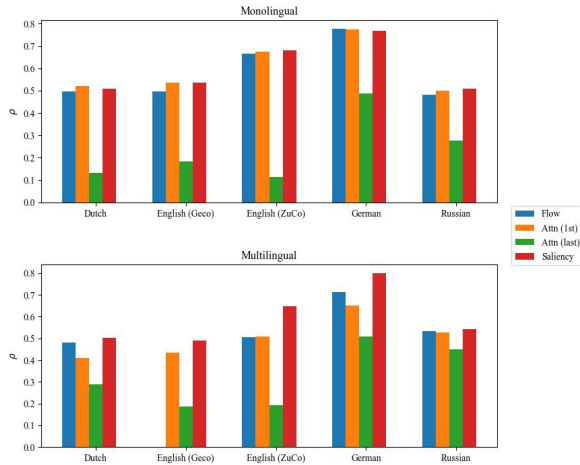


Figure 1: Upper: Spearman’s correlation ρ between human (total reading time) and model relative word importance (BERT monolingual models). Lower: Spearman’s correlation ρ between human (total reading time) and model relative word importance (BERT multilingual model).

the important tokens for the model prediction. Because of the high computational resources needed to calculate attention flow, we do not calculate attention flow of the BERT multilingual model for English (GECO).

Baselines We compare human and model relative word importance to two word-level baselines: word length (the number of characters in a word) and word frequency (the proportion of times a word occurs in a corpus). To obtain the word frequencies, we use the `wordfreq` Python package (Speer et al., 2018) (version 2.3.2), which calculates token frequencies based on corpora from different Internet text resources, such as Wikipedia, Google Books, and Reddit.

Regression Analysis In addition, we use a mixed linear regression analysis (ordinary least squares) to measure the extent to which, model relative word importance, word frequency, and word length can predict human relative word importance. We let human relative word importance be the dependent variable and fit multiple linear regression models with different combinations of model word importance, word frequency and word length as independent variables. This is done to measure each and every variable’s effect in isolation. We analyze the resulting coefficient of determination R^2 .

In the Spearman correlation analysis outlined above, the correlations were calculated per sen-

tence and then averaged. In contrast, we now fit the model to tokens which means that all relative word importance values and all word lengths and frequencies are fitted into the same model.⁴

Since all independent variables (word frequency, word length and model relative word importance) and the dependent variable (human relative word importance) are intrinsically skewed, we log-transform all data. Furthermore, we use an extended version of linear regression (mixed linear regression (Gafcecki and Burzykowski, 2013)) to deal with dependency between samples (i.e., one word appearing more than twice) which otherwise would break the assumption of linear regression models that each observation is independent of each other.

5 Results

5.1 Human vs. Model Word Importance

Figure 1 shows the Spearman correlation between human relative word importance and model relative word importance of importance methods for each eye-tracking corpus. The results show a strong correlation ($\rho > .5$) between human and model word importance across all languages. There are, however, considerable differences between languages. For example, German reaches a Spearman’s ρ of .8, while Russian, English (GECO) and Dutch (GECO) only reach .5 (Q1). When comparing the multilingual BERT model to language-specific BERT models, we observe for some importance metrics, attention flow and first-layer attention, a slightly stronger correlation to monolingual models. In the German and English (ZuCo) monolingual models, in particular, first-layer attention and flow are equally strong as saliency while in the multilingual model they all have more than .1 weaker correlation. Attention first-layer seems even more strongly correlated to monolingual models, where we see +.11 and +.17 for the language-specific BERT model of Dutch (GECO), English (GECO) and ZuCo (English), respectively. Russian, however, is a slight outlier in that it has a +.3 difference in favor of the multilingual BERT model (Q2).

When comparing importance methods, we see similar results to previous findings on English data. Saliency shows a strong correlation to human relative word importance, while last-layer attention shows a weaker correlation. Furthermore, attention flow and, surprisingly first-layer attention in most

⁴Most sentences are too short for the number of independent variables we use to fit a linear regression model.

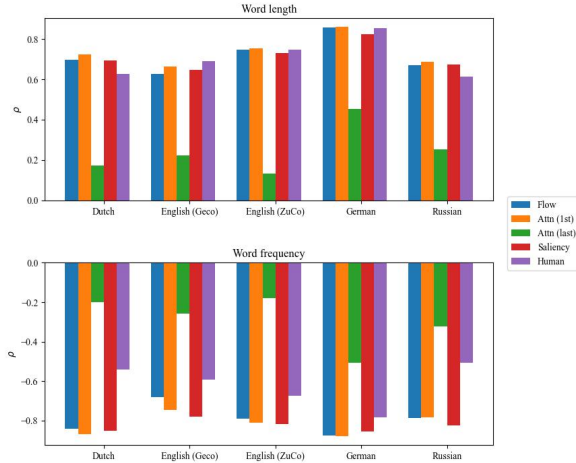


Figure 2: Upper: Spearman’s correlation ρ between word length and human (total reading time) and model (BERT monolingual) relative word importance. Lower: Spearman’s correlation ρ between word length and human (total reading time) and model (BERT monolingual) relative word importance.

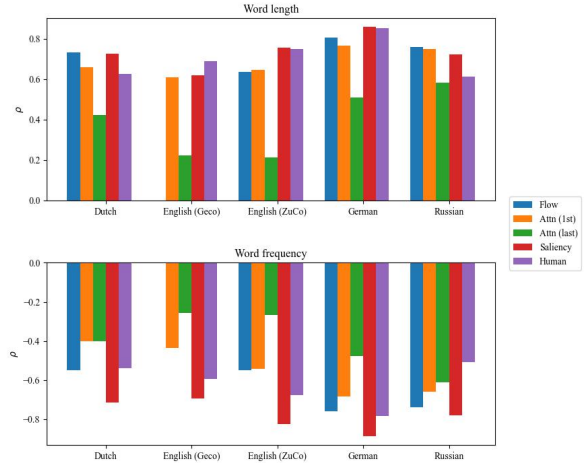


Figure 3: Upper: Spearman’s correlation ρ between word length and human (total reading time) and model (BERT multilingual) relative word importance. Lower: Spearman’s correlation ρ between word length and human (total reading time) and model (BERT multilingual) relative word importance.

cases, show similar strength than saliency, albeit slightly weaker for multilingual models (Q3).

Finally, though not specifically defined in the goals and research questions of this study, we make the separate, but important, observation that the most impactful variable for correlation strength seems to be the size and text domain of the eye-tracking corpora. This becomes apparent when comparing Dutch (GECO), English (GECO) and English (ZuCo). Even though English (GECO) and English (ZuCo) are of the same language, the performance on English (GECO) and Dutch (GECO) are quite similar, while not very similar to English (ZuCo). While the language-specific impact on the results is difficult to grasp due to the differences between the eye-tracking corpora, we, nonetheless, see the same trends hold for all four languages.

5.2 Corpus Statistical Baselines

Figures 2 and 3 show the correlation of word frequency and length baselines to human and model relative word importance. We see a strong correlation between models of all languages and the two baselines word length and word frequency: A strong *positive* correlation to word length and a strong *negative* correlation to word frequency, as also observed by Hollenstein and Beinborn (2021).

For the baselines, however, we see considerable differences between languages. German shows the strongest correlation for word length with 0.85 for

humans and 0.82 and 0.85 for mono- and multilingual BERT, respectively, and also the strongest negative correlation for word frequency with -0.78 for humans and -0.85 as well as -0.87 for mono- and multilingual BERT, respectively. Russian shows the weakest human correlation to the baselines, 0.61 for word length and -0.51 for word frequency, while having a relatively strong saliency baseline correlation of 0.72 and 0.67 for word length as well as -0.78 and -0.82 for word frequency.

Looking at the relative word importance metrics, which had the strongest correlation to human importance, namely saliency and attention flow, we see that their correlation strength with respect to the baselines correlate equally strong or stronger than their human counterparts. This indicates that the more similar their baselines are to the human baselines, the stronger they correlate in terms of relative word importance (see previous section). This is especially the case when looking at (1) attention last-layer, where there weaker correlation to human relative word importance is also reflected in its weaker correlation to word frequency and word length, which are much lower than its human counterpart and (2) word frequency, where the model relative word importance of Dutch (GECO), English (GECO) and Russian have a weaker correlation to human relative word importance but a considerable stronger negative correlation to word frequency than human relative word importance

Table 2: Linear regression (R^2) fitted to predict human relative word importance from word frequency and/or word length.

	freq	length	freq+length
Dutch	0.08	0.15	0.16
English (GECO)	0.07	0.12	0.14
German	0.31	0.38	0.41
Russian	0.22	0.49	0.49
English (ZuCo)	0.13	0.26	0.30

has to word frequency. This effect does not, however, seem to be as pronounced with word length.

These results show that word length and word frequency are powerful indicators of word importance and support the presumption that they play an important role in determining the correlation strength between human and model relative word importance (Q4). In the next subsection, we will try and quantify how much these baselines account for this relation, by measuring the explanatory power word length, word frequency and model relative word importance have in predicting human relative word importance.

5.3 Word Length & Frequency Regression Analysis

We fitted linear mixed models to predict human word importance using either word frequency, word length, model relative word importance, or combinations of the features. Table 2 shows the R^2 results for using word frequency and word length as independent variables, and Figure 4 shows the results for using model relative word importance as an independent variable in combination with word frequency and word length. See Figure 5 and Table 4 in the Appendix for full results.

In Table 2 we see a weak R^2 score or in other words a weak *linear* relationship between human relative word importance and word frequency and word length. Word frequency, however, appears to have a weaker R^2 than word length and differences are large between corpora. Russian, for example, has four times stronger R^2 for word length than English (GECO). Using both word frequency and word length (freq+length) appears only to be as strong as the strongest word length value (length), such that a combination of word frequency and word length (freq+length) does not make the relationship stronger.

Comparing Figure 4 to Table 2 we see a much stronger linear relationship (R^2) when model rel-

ative word importance is used as an independent variable. When combined with word frequency (model+freq) we see a considerable increase of R^2 , but combined with word length (model+length and model+freq+length) we see an even stronger linear relationship. Similarly to Table 2, combining word frequency and word length (freq+length) only gives as much benefit as adding length to model word importance (model+length and model+freq+length).

Comparing the R^2 of model importance, we see different scores than that of the results in section 5.1 and section 5.2. Here, we see saliency achieving the lowest R^2 across all models and corpora, meanwhile the attention-based metrics (attention first/last layer and flow) show a much larger R^2 . Although comparing the R^2 and Spearman’s ρ is not equal due to the methodological differences outlined in Section 4, this difference nevertheless suggest that relation between saliency and human relative word importance is less linear in nature than attention-based ones.

6 Discussion

6.1 Findings on Human vs. Model Relative Word Importance

This cross-lingual study shows that model relative word importance has a strong correlation to human relative word importance. We confirm the findings of other English-based studies that saliency (Hollenstein and Beinborn, 2021), first-layer attention (Bensemann et al., 2022) and attention flow (Eberle et al., 2022) show a strong correlation to human relative word importance as well as last-layer attention showing a weak correlation (Q1 & Q3). This research, thus, from a usability perspective supports the critique against using attention weights for explanation, while providing supporting evidence for the use of attention flow. Comparing monolingual and multilingual models, we see slightly stronger results for monolingual models, in particular for attention first-layer and attention flow, indicating that some importance-bearing information are more readily available in the attention weights of monolingual models. However, given that these results are not vastly different, there is still a strong argument for training multilingual models over monolingual models because of their resource-efficiency and saving of computational resources (Q2).

The secondary finding of this study that corpus-specific differences have a big impact on correlation strength, indicates the need to control the size

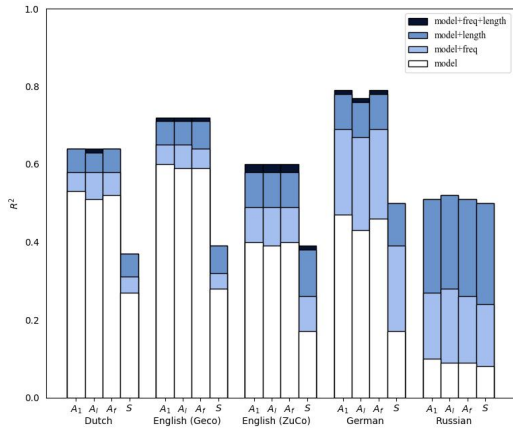


Figure 4: Linear regression (R^2) fitted to predict human relative word importance from model word importance (model), word frequency (freq), word length (length) or combinations thereof (+) (BERT monolingual). A_1 , A_l , A_r , and S are short for first-layer attention, last-layer attention, attention flow, and saliency, respectively.

and text domain for cross-lingual comparison. A promising avenue for future work could be to apply our analyses to the recent multilingual eye-tracking corpus MECO (Siegelman et al., 2022).

6.2 The Role of Word Length And Word Frequency

Measuring the effect of word length and word frequency using Spearman’s ρ and linear regression we find that they play an important role in determining relative word importance. First, we see that similarity in correlation strength to the word length and especially word frequency baseline mirrors a stronger correlation between model and human relative word importance (Q4). Secondly, we see that when using linear regression to quantify the linear relationships between the two lexical baselines, model relative word importance, and human relative word importance, stronger predictions can be made when model relative word importance is combined with word frequency or word length, especially the latter (see Figure 4). This suggests that word frequency and especially word length might not be sufficiently accounted for by the language models. Furthermore, the discrepancy between the lower impact of word frequency and the higher impact of word length has several potential explanations. Firstly, word frequency might be better approximated by the model than word length (which is supported by the fact that word length is not

explicitly processed in Transformer-based architectures). Secondly, the relationship between word frequency and relative word importance is probably less linear than that of word length and relative word importance (as suggested by the results in Table 2) and could, therefore, not be as adequately fitted by linear regression. The nature of the relationships between word length, word frequency, and human relative word importance, thus, remains elusive. To gain clarity on this, future work could control for word length and frequency more explicitly, by, for example, grouping and comparing relative word importance by length and frequency in isolation as well as using probing tasks to test the extent to which contextual word representations themselves can predict word length and frequency.

7 Conclusion

In this work, we show that the strong correlation between relative word importance of neural language models and humans holds across several languages, namely, English, German, Dutch, and Russian. This is the case for both monolingual as well as multilingual pretrained Transformer models, which yield similar performance in our correlation analyses.

We also find that several relative importance metrics for pretrained language models, both first-layer attention and attention flow as well as saliency, perform similarly well and that these importance values, as their human counterparts, strongly correlate to word length and word frequency. However, as expected, we have found that last-layer attention correlates more weakly.

Comparing the correlations of relative word importance is a simple, easily interpretable metric for evaluating the similarity of human and computational language processing. Using this metric, we can evaluate the extent to which the model’s attention compares to approximate human language processing and, thus, get a gauge of their cognitive plausibility. In addition to the BERT-based architectures we studied, looking at more recent cross-lingual models such as GPT, T5 and XLNet as well as multimodal language models would give further insights into the role of pre-training tasks as well as non-textual modalities.

Acknowledgements

We thank Alexander Koplein for his assistance in assembling the results for the regression analysis.

References

- Mostafa Abdou. 2022. Connecting neural response measurements & computational models of language: a non-comprehensive guide. *arXiv preprint arXiv:2203.05300*.
- Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. [Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.
- Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.
- Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.
- Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. [Eye gaze and self-attention: How humans and transformers attend words in sentences](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.
- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. [Bertje: A dutch BERT model](#). *CoRR*, abs/1912.09582.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. [Do transformer models show similar attention patterns to task-specific human gaze?](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.
- Allyson Ettinger. 2020. [What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models](#). *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Andrzej Gałeccki and Tomasz Burzykowski. 2013. *Linear Mixed-Effects Model*, pages 245–273. Springer New York, New York, NY.
- Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nasta, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.
- Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020a. [Towards best practices for leveraging human language processing signals for natural language processing](#). In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.
- Nora Hollenstein and Lisa Beinborn. 2021. [Relative importance in sentence processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.
- Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. Cmc1 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78.
- Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*.
- Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.
- Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020b. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 138–146.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings*

- of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online. Association for Computational Linguistics.
- Lena Jäger, Thomas Kern, and Patrick Haller. 2021. Potsdam Textbook Corpus: Potsdam textbook corpus (potec): Eye tracking data from experts and non-experts reading scientific texts. available on OSF, DOI 10.17605/OSF.IO/DN5HP.
- Sarthak Jain and Byron C. Wallace. 2019. **Attention is not Explanation**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329.
- Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 60–67.
- Evan Kidd, Seamus Donnelly, and Morten H Christiansen. 2018. Individual differences in language acquisition and processing. *Trends in cognitive sciences*, 22(2):154–169.
- Yuri Kuratov and Mikhail Arkipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- AK Laurinavichyute, Irina A Sekerina, SV Alexeeva, and KA Bagdasaryan. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Cyrillic. *Behavior research methods*, 51(3):1161–1178.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is bert surprised? layer-wise detection of linguistic anomalies. *arXiv preprint arXiv:2105.07452*.
- Andreas Madsen. 2019. **Visualizing memorization in rnns**. *Distill*. <https://distill.pub/2019/memorization-in-rnns>.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. **A primer in BERTology: What we know about how BERT works**. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, pages 1–21.
- Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. **Interpreting attention models with human visual attention in machine reading comprehension**. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.
- Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinight/wordfreq: v2. 2. *Zenodo [Computer Software]*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. **BERT rediscovers the classical NLP pipeline**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.
- Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.
- Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. **Probing pretrained language models for lexical semantics**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.
- Sarah Wiegrefe and Yuval Pinter. 2019. **Attention is not not explanation**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Limitations

The results and conclusions of this paper should be read with the following limitations in mind: (1) Differences between human subjects can be quite significant, thus, the results will also reflect this uncertainty (Kidd et al., 2018). (2) Comparing model relative word importance of saliency-based

and attention-based methods to that of human relative word importance only reflects these methods' ability to mimic human behavior, but does not say anything about their ability to accurately represent the inner workings, i.e., the *faithfulness* of pretrained language models (Jacovi and Goldberg, 2020).

Table 3: List of models used for each corpora. Hugging Face path refers to the model path used to identify the model in Hugging Face repository. For models without explicit paper reference, we refer to the Hugging Face website.

Corpus	Model name	Hugging Face path	Reference
GECO (en)	BERT	bert-base-uncased	Devlin et al. (2019)
GECO (en)	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
GECO (nl)	BERT	GroNLP/bert-base-dutch-cased	de Vries et al. (2019)
GECO (nl)	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
ZuCo	BERT	bert-base-uncased	Devlin et al. (2019)
ZuCo	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
Potsdam	BERT	dbmdz/bert-base-german-uncased	https://huggingface.co/dbmdz/bert-base-german-uncased (accessed 2022-03-15)
Potsdam	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
RussSent	BERT	DeepPavlov/rubert-base-cased	Kuratov and Arkhipov (2019)
RussSent	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)

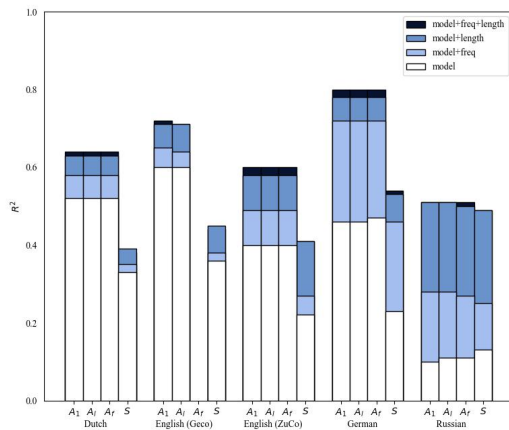


Figure 5: Linear regression (R^2) fitted to predict human relative word importance from model word importance (model), word frequency (freq), word length (length) or combinations thereof (+) (BERT multilingual). A_1 , A_l , A_f , and S are short for attention (first-layer), attention (last-layer), attention flow, and saliency, respectively.

Table 4: Linear regression models R^2 measuring impact of word length, word frequency and word importance.

Language	Model	Importance	model	model+freq	model+length	model+freq+length
Dutch	BERT	Attn (1st)	0.53	0.58	0.64	0.64
		Attn (last)	0.51	0.58	0.63	0.64
		Flow	0.52	0.58	0.64	0.64
		Saliency	0.27	0.31	0.37	0.37
	mBERT	Attn (1st)	0.52	0.58	0.63	0.64
		Attn (last)	0.52	0.58	0.63	0.64
		Flow	0.52	0.58	0.63	0.64
		Saliency	0.33	0.35	0.39	0.38
English (Geco)	BERT	Attn (1st)	0.6	0.65	0.71	0.72
		Attn (last)	0.59	0.65	0.71	0.72
		Flow	0.59	0.64	0.71	0.72
		Saliency	0.28	0.32	0.39	0.39
	mBERT	Attn (1st)	0.6	0.65	0.71	0.72
		Attn (last)	0.6	0.64	0.71	0.71
		Saliency	0.36	0.38	0.45	0.44
		English (ZuCo)	BERT	Attn (1st)	0.4	0.49
Attn (last)	0.39			0.49	0.58	0.6
Flow	0.4			0.49	0.58	0.6
Saliency	0.17			0.26	0.38	0.39
mBERT	Attn (1st)		0.4	0.49	0.58	0.6
	Attn (last)		0.4	0.49	0.58	0.6
	Flow		0.4	0.49	0.58	0.6
	Saliency		0.22	0.27	0.41	0.41
German	BERT	Attn (1st)	0.47	0.69	0.78	0.79
		Attn (last)	0.43	0.67	0.76	0.77
		Flow	0.46	0.69	0.78	0.79
		Saliency	0.17	0.39	0.5	0.5
	mBERT	Attn (1st)	0.46	0.72	0.78	0.8
		Attn (last)	0.46	0.72	0.78	0.8
		Flow	0.47	0.72	0.78	0.8
		Saliency	0.23	0.46	0.53	0.54
Russian	BERT	Attn (1st)	0.1	0.27	0.51	0.51
		Attn (last)	0.09	0.28	0.52	0.52
		Flow	0.09	0.26	0.51	0.51
		Saliency	0.08	0.24	0.5	0.5
	mBERT	Attn (1st)	0.1	0.28	0.51	0.51
		Attn (last)	0.11	0.28	0.51	0.51
		Flow	0.11	0.27	0.5	0.51
		Saliency	0.13	0.25	0.49	0.49

Dispatcher: A Message-Passing Approach To Language Modelling.

Alberto Cetoli †

alberto.cetoli@fractalego.io

Abstract

The transformer architecture has achieved state-of-the-art performance on language modelling. Nonetheless, a more efficient algorithm would allow for a larger number of tokens, thus a wider context and better grounding of the model’s predictions. In this spirit, we introduce a more efficient layer type that aims to substitute self-attention for unidirectional sequence generation tasks. The system is shown to be competitive with existing methods: Given N tokens, the computational complexity is $\mathcal{O}(N \log N)$ and the memory complexity is $\mathcal{O}(N)$ under reasonable assumptions. The Dispatcher layer is seen to achieve comparable perplexity to self-attention while being more efficient¹.

1 Introduction

The introduction of self-attention (Vaswani et al., 2017) has produced a considerable surge of language models (Devlin et al., 2018; Liu et al., 2019; Yang et al., 2019; Lan et al., 2020). Originally, self-attention had been envisioned as a three elements algorithm (key, query, and value) to be applied onto an encoder-decoder framework for machine translation. It soon became evident that the Transformer architecture can successfully master the most relevant NLP tasks (Devlin et al., 2018) with some marginal modifications. One key application of self-attention has been language generation (Radford et al., 2019), where typically the model attempts to predict the next token given a limited window of prior elements. A full text can thus be generated word by word. Self-attention - bidirectional in nature - needs to be masked in order to avoid backward propagation of information.

Until a few years ago recurrent models (Sutskever et al., 2011; Graves, 2014; Merity et al.,

2017; Melis et al., 2018) outperformed every other method for language modelling and generation. This has changed with the introduction of masked self attention (MSA) models, which have achieved the state-of-the-art in language generation, culminating in some unexpected results for multi-task zero-shot learning (Radford et al., 2019) as well as intriguing few-shot abilities (Brown et al., 2020).

The main argument of this work is to show that language modelling can efficiently rely on a message passing approach to perform, proposing a method that *does not leverage upon self-attention*. Instead, the system builds a tree-like structure of forward message passing weighed by *dispatching coefficients*. In the end, the Dispatcher architecture can generate texts as well as the original Transformer model, more efficiently. The main contributions of the paper are to introduce the novel algorithm as well as compare perplexity to the "standard" self-attention on the task of language modelling.

2 Model

The original Transformer architecture is composed of a number of self-attention, skip connection, and feed-forward layers. Given N tokens, the self-attention block has a computational and memory complexity of $\mathcal{O}(N^2)$ and is therefore problematic for long sequences. Here we propose to substitute each self-attention layer with a different algorithm.

Within the Dispatcher layer, information is pushed forward onto the next tokens in a recursive fashion. The algorithm is given a list of embeddings as *input*, with the aim to create *output* embeddings that contain a mixture of the tokens that precede them, without any leakage from the tokens that follow. The system achieves this goal by summing the tokens with themselves shifted by a power of two, iteratively. Each of these steps is labelled *shift and sum* in Fig. 1.

In the pseudo-code shown in Alg. 1 the *dis-*

¹The code is available at <https://github.com/fractalego/dispatcher>

† Work done while at QBE Europe.

Algorithm 1: The Dispatcher Layer Algorithm

```
c ← Sigmoid(Linear1(input));
c ← c ⊙ mask;
V ← Linear2(input);
for row = 0 → log2N − 1 do
  | V ← V + c[row] ⊙ RollRight(V, 2row);
end
output ← Linear3(V);
```

patching coefficients are written as $c \in \mathbb{R}^{N \times \log_2 N}$, whereas $V \in \mathbb{R}^{N \times d}$ is the tensor containing the hidden states used as a working memory in the main loop, with embedding dimension d . The Linear functions are dense layers, while RollRight shifts the tokens to the right.

The message coming from the prior tokens follows a binary tree structure, as depicted in Fig. 1. The sum is weighed by the *dispatching coefficients*, which effectively decide whether information coming from the left of the tree should propagate further, and by what amount. These weights are computed through a dense layer applied to the original tokens. A constant mask is applied to the tensor c after it has been computed to avoid leakage after the RollRight operation.

The algorithm presented above describes a single-head unit. As with self-attention, this layer can be split into a set of Dispatcher heads to improve performance. The number of heads then becomes another hyper-parameter to tune during training. Finally, if the number of input embeddings is not a power of two, the loop stops when the shift value is greater than the input length.

2.1 Dispatcher Dropout

A quick modification of Alg. 1 can introduce an effective dropout by randomly skipping a *shift and sum* step in training with a probability given by a dropout value between 0 and 1. Notice that in this procedure dropout makes the algorithm quicker, albeit with the same computational complexity.

2.2 Computational complexity

The creation of the *dispatching coefficients* is linear in time, as a dense layer is applied to every input token. The main algorithm repeats $\log_2 N$ times a weighted sum. If d is the dimension of the embeddings, the computational complexity is $\mathcal{O}(dN \log N)$.

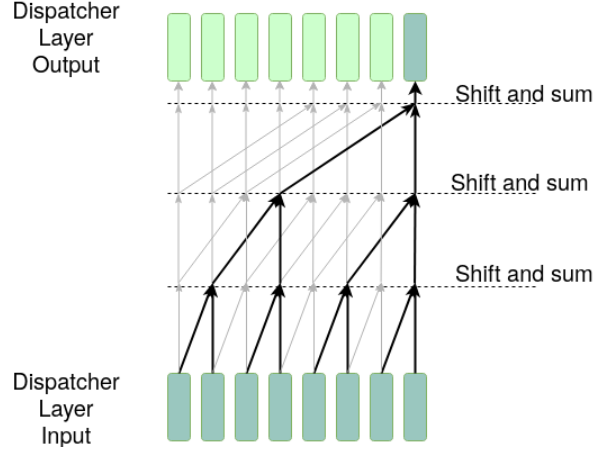


Figure 1: A representation of how information is passed from the input tokens to the output within the Dispatcher Layer. At every vertex of this directed graph the embeddings are summed together, each sum weighed by the *dispatching coefficients*. These weights determine how much of the message from the left needs to be passed onto the right. For clarity the paths to the last output item are painted in a darker color.

2.3 Memory complexity

The system computes the dispatching coefficients in every layer with a space complexity of $\mathcal{O}(N \log N)$. In addition, the algorithm uses at every step a set of embeddings V with complexity $\mathcal{O}(N \times d)$. In a typical scenario $d \gg \log_2 N$, yielding an effective asymptotic linear memory consumption $\mathcal{O}(N \times d)$.

3 Evaluation

3.1 Datasets

The algorithm is evaluated on the following datasets: PTB (Mikolov and Zweig, 2012), WikiText2 and WikiText103 (Merity et al., 2017), and One Billion Word (Chelba et al., 2013). A simple pre-processing step uses the special token $\langle \text{EOS} \rangle$ to indicate the end of each sentence.

Among the sets, PTB and WikiText2 are the smallest, with only 4.9MB and 11MB of text data for training respectively. This is to be compared to the 515MB training set of WikiText103 and 3.9GB of 1BW. While the larger dataset, 1BW only models short-term dependency because the sentences have been shuffled. An additional corpus called OpenWebText (Cohen and Gokaslan, 2020) is used to train the larger Dispatcher model. This set was created as an open alternative to the one used when training GPT-2 (Radford et al., 2019) and consists of about 40 GB of text data.

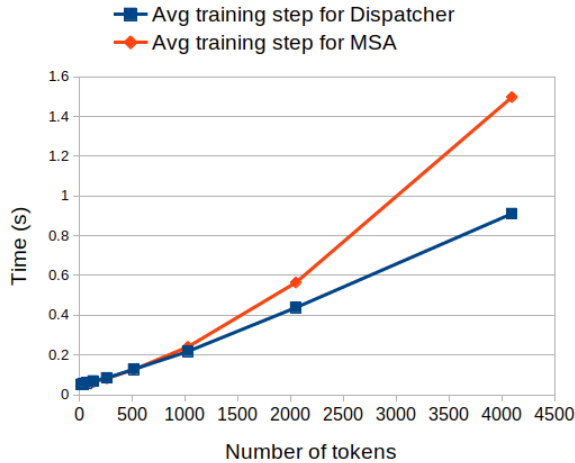


Figure 2: Average time in seconds for a single training step on WikiText2 as a function of the number of input tokens. Both models are trained on the same single-GPU instance. The asymptotic behavior appears remarkably different.

3.2 Training

At first, the Dispatcher algorithm is compared against a masked self-attention model. Rather than trying to optimize for the MSA and the Dispatcher separately, we choose an identical set of hyperparameters (embedding size, number of heads, layers) and compare their perplexity. While this approach might not give the best results for each model, it helps to show that the two algorithms perform similarly under similar conditions.

Secondly, a slightly bigger Dispatcher model (Sec. 3.4) is trained for *one epoch only* on the OpenWebText corpus. The goal is not to achieve state-of-the-art results, rather to prove that the proposed architecture can reach comparable perplexity to Transformer-based models. In the spirit of simplicity, we use a single head for all the models, which are trained on the same single-GPU machine. All the models are implemented in *PyTorch* (Paszke et al., 2019).

3.3 Masked self-attention (MSA) and Plain Dispatcher

These models have embedding and inner dimension size 512, 6 layers, and only 1 head. The training batch size is 20 and dropout is set to 0.2, using 512 tokens. The only difference between the two models is the self-attention/Dispatcher layer. Tokenization is done using a pre-trained WordPiece tokenizer made available by HuggingFace (Wolf et al., 2020). On training and evaluation the vo-

cabulary is further restricted on each dataset to improve performance, as a consequence the number of parameters changes depending on the relevant dataset’s vocabulary size.

3.4 Dispatcher after OpenWebText

This model has a single head, 480 embedding and inner dimension size, and 12 layers with a mini-batch size of 5 and no dropout. A BPE tokenizer - pre-trained on OpenWebText - is used. The number of tokens used for this model is 1024. First pre-trained on OpenWebText, the model is then fine-tuned onto the relevant sets.

3.5 Discussion

The MSA model and the Plain Dispatcher are evaluated against four different datasets, as shown in the first two rows of Table 1. The results are quite similar, with the Dispatcher architecture seen performing better on the smaller sets PTB and WikiText2. This is arguably due to the model having fewer parameters and being less prone to overfitting. Conversely, the larger MSA model wins on WikiText103. The Dispatcher overtaking MSA on 1BW is more challenging to explain in terms of model size and seems to suggest its enhanced ability to model short-term dependencies, at least in this one-headed configuration.

A striking difference between the MSA and the Dispatcher is however shown in Fig. 2, which plots the average time for a single training step as a function of the number of input tokens. While the recorded times are configuration-specific, the asymptotic behavior looks radically different, suggesting the Dispatcher architecture as a better candidate for longer sequences.

A single epoch of training onto the OpenWebText dataset boosts the Dispatcher performance into competitive results for a model of this size, after fine-tuning on the relevant corpus. This is shown in the third row of Table 1, presenting our top results.

The rest of Table 1 is a showcase of the most recent self-attention based models. Notably, our results on PTB and WikiText2 are among the best in the literature, surpassing the results in (Wang et al., 2019) which are obtained by fine-tuning a pre-trained BERT model. This is most likely due to the OpenWebText corpus being a better set for language generation than BookCorpus (used by BERT), but it bodes well for the algorithm presented here that it can compete against models with one order of magnitude more parameters. The last

Model Type	PTB	WikiText2	WikiText103	1BW
Masked Self-Attention (18M / 30M / 39M / 41M)	40.58	55.05	22.56	66.52
Plain Dispatcher (17M / 27M / 36M / 38M)	35.40	50.23	24.39	53.32
Dispatcher after OpenWebText (59M)	18.95	22.74	20.38	36.76
(Fan et al., 2020) (44M)	-	-	22.4	-
(Wang et al., 2019) (395M)	31.34	34.11	20.42	-
(Tay et al., 2021) (100M)	-	-	-	21.5
(Dai et al., 2019) (257M / 0.8B)	-	-	18.3	21.8
(Radford et al., 2019) (1.5B)	35.7	18.34	17.48	42.16
(Shoeybi et al., 2020) (355M)	-	-	19.31	-
(Shoeybi et al., 2020) (8.3B)	-	-	10.81	-

Table 1: Top: The Dispatcher architecture is evaluated concurrently with a masked self-attention model yielding similar results. Bottom: The Dispatcher pre-trained on OpenWebText compared to some recent results achieved using a variant of the Transformer architecture. All the results refer to the test perplexity.

three rows relate to zero-shot results. Omninet’s impressive result (Tay et al., 2021) is achieved by extending self-attention to all tokens in all the layers, while here the dispatcher layer is only aware of the embeddings within a single layer.

4 Related works

The way information is funneled to higher layers in Fig. 1 is reminiscent of convolutional neural networks (CNN) (Liu et al., 2020; Gehring et al., 2017). It is especially evocative of *dilated convolutions* as presented in (van den Oord et al., 2018). While similar, the method presented here is not technically a convolution, which by definition requires the same operator being translated over the input elements. In this paper the dispatching coefficients are *local* to the tokens.

Another way to visualize the Dispatcher algorithm is as a set of overlapping Recursive NN acting on binary trees (Goller and Kuchler, 1996; Socher et al., 2011) which share parameters where the trees overlap. It is however important to keep in mind that the *shift and sum* iteration only performs a weighted sum of the input embeddings, achieving competing performance only when repeated within a multi-layer structure.

The computational cost of large models has become a source of concern in terms of scalability as well as energy consumption (Strubell et al., 2019). For this reason, a growing number of approximations (Wang et al., 2020; Kitaev et al., 2020; Zahoor et al., 2021; Choromanski et al., 2020; Zhai et al., 2021) have appeared in the literature, suggesting modifications to the main self-attention layer.

These approximations tend to leverage linear algebra properties to speed up calculations, capturing the essence of the Transformer architecture into more efficient algorithms. In many cases the approximation makes the model irreducibly bidirectional, thus hindering language generation tasks.

More recently, inductive biases alternative to self-attention have been shown to achieve comparable results on language tasks using the Fourier Transform in place of the MSA layer (Lee-Thorp et al., 2021) and on vision tasks by means of spatial MLPs (Yu et al., 2021).

Finally, the concept of message passing is understood to describe Graph Convolutional Networks (Kipf and Welling, 2017; Geerts et al., 2020) and by extension the self-attention mechanism in the Transformer architecture. The Dispatcher algorithm makes message passing explicit by keeping the routing topology constant while relying on the coefficients to distribute the message within a set of binary trees.

5 Conclusions

A novel architecture dedicated to language modelling is introduced and shown to achieve comparable perplexity with self-attention based models, requiring less computational and memory resources. A larger number of tokens can allow for a wider context window and more detailed grounding of the model’s predictions.

Finally, low perplexity in the task of language modelling is often predictive of high-quality text generation. This intriguing possibility will be pursued in a future work.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. [One billion word benchmark for measuring progress in statistical language modeling](#). *CoRR*, abs/1312.3005.
- Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarrlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. [Rethinking attention with performers](#).
- Vanya Cohen and Aaron Gokaslan. 2020. [Opengpt-2: Open language models and implications of generated text](#). *XRDS*, 27(1):26–30.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. [Transformer-XL: Attentive language models beyond a fixed-length context](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Angela Fan, Thibaut Lavril, Edouard Grave, Armand Joulin, and Sainbayar Sukhbaatar. 2020. [Accessing higher-level representations in sequential transformers with feedback memory](#).
- Floris Geerts, Filip Mazowiecki, and Guillermo A. Perez. 2020. [Let’s agree to degree: Comparing graph convolutional networks in the message-passing framework](#). *CoRR*, abs/2004.02593.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. [Convolutional sequence to sequence learning](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.
- C. Goller and A. Kuchler. 1996. [Learning task-dependent distributed representations by backpropagation through structure](#). In *Proceedings of International Conference on Neural Networks (ICNN’96)*, volume 1, pages 347–352 vol.1.
- Alex Graves. 2014. [Generating sequences with recurrent neural networks](#).
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). *ICLR 2017*.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. [Reformer: The efficient transformer](#).
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [Albert: A lite bert for self-supervised learning of language representations](#).
- James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. [Fnet: Mixing tokens with fourier transforms](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Chunguo Li, and Luxi Yang. 2020. [Convtransformer: A convolutional transformer network for video frame synthesis](#). *CoRR*, abs/2011.10185.
- Gabor Melis, Chris Dyer, and Phil Blunsom. 2018. [On the state of the art of evaluation in neural language models](#). In *International Conference on Learning Representations*.
- Stephen Merity, Caiming Xiong, James Bradbury, and R. Socher. 2017. [Pointer sentinel mixture models](#). *ArXiv*, abs/1609.07843.
- Tomas Mikolov and Geoffrey Zweig. 2012. [Context dependent recurrent neural network language model](#). In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239.
- Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. 2018. [Parallel WaveNet: Fast high-fidelity speech synthesis](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926. PMLR.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning](#)

- library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. [Megatron-lm: Training multi-billion parameter language models using model parallelism](#).
- Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, page 129–136, Madison, WI, USA. Omnipress.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. [Energy and policy considerations for deep learning in NLP](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, page 1017–1024, Madison, WI, USA. Omnipress.
- Yi Tay, Mostafa Dehghani, Vamsi Aribandi, Jai Prakash Gupta, Philip Pham, Zhen Qin, Dara Bahri, Da-Cheng Juan, and Don Metzler. 2021. [Omninet: Omnidirectional representations from transformers](#). In *ICML 2021*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Chenguang Wang, Mu Li, and Alexander J. Smola. 2019. Language models with transformers. *ArXiv*, abs/1904.09408.
- Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. [Linformer: Self-attention with linear complexity](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Huggingface's transformers: State-of-the-art natural language processing](#).
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.
- Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. 2021. [Metaformer is actually what you need for vision](#).
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. [Big bird: Transformers for longer sequences](#).
- Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. 2021. [An attention free transformer](#).

In search of meaning and its representations for computational linguistics

Simon Dobnik^{*}, Robin Cooper[◇], Adam Ek^{*}, Bill Noble^{*}, Staffan Larsson[◇],
Nikolai Ilinykh^{*}, Vladislav Maraev^{*} and Vidya Somashekarappa^{*†}

Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden

^{*} name.surname@gu.se [◇] name.surname@ling.gu.se

Abstract

In this paper we examine different meaning representations that are commonly used in different natural language applications today and discuss their limits, both in terms of the aspects of the natural language meaning they are modelling and in terms of the aspects of the application for which they are used.

1 Introduction

A crucial component to produce a “successful” NLP system is sufficiently expressive representations of meaning. We consider a sufficiently expressive meaning representation to be one that allows a system’s output to be considered acceptable to native speakers given the task. In this paper we present several features of meaning and discuss how different methods of deriving meaning representations capture these features. This list is by no means exhaustive. It might be viewed as a first attempt to discuss ways of establishing a general methodology for evaluating meaning representations and characterising what kinds of applications they might be useful for.

2 Formal meaning representations

The rigour of the work on semantics by Richard Montague (Montague, 1973; Partee, 1976) inspired early work on computational semantics (perhaps the earliest was Friedman and Warren, 1978; Friedman et al., 1978). Two high-points of the literature on computational semantics based on Montague are Blackburn and Bos (2005), using logic programming, and van Eijck and Unger (2010), using functional programming. Montague’s semantic techniques have also played an important role in semantic treatments using Combinatory Categorical Grammar (CCG, Bos et al., 2004).

One problem with Montague’s treatment of semantics was that it was limited to the level of the sentence. It could not, for example, deal with cross-sentence anaphora such as *A dog_i barked. It_i was upset by the intruder.* This, among several other things, led to the development of Discourse Representation Theory (DRT, Kamp and Reyle, 1993; Kamp et al., 2011) and other variants of *dynamic* semantics such as Heim (1982) and Groenendijk and Stokhof (1991). Here “dynamic” is meant in the sense of treating semantic content as context change potential in order, among other things, to be able to pass referents from one sentence to a subsequent sentence in the discourse. This is a much less radical notion of dynamic interpretation than we discuss in Section 4, where the meaning associated with a word or phrase may change as a dialogue progresses. DRT has played an important role in computational semantics from early work on the Verbmobil project (Bos et al., 1996) to work by Johan Bos and others on the Groningen Meaning Bank¹ and the Parallel Meaning Bank².

One of the cornerstones of Montague’s approach is **compositionality**, the ability to compute the meaning of phrases on the basis of the meanings of their immediate sub-constituents. Another central feature in the Montague tradition is the ability to derive conclusions based on **logical inference**, including logical inferences based on the semantics of logical constants such as *and*, *not* and logical quantifiers, and the ability to characterise additional axioms or “meaning postulates”. Defeasible reasoning has been added to this kind of framework (e.g., Asher and Lascarides, 2003) and systems have been connected to theorem provers and model builders (Blackburn and Bos, 2005). The variants of dynamic semantics discussed above gave us the

^{*} All authors contributed equally.

¹<https://gmb.let.rug.nl/>

²<https://pmb.let.rug.nl/>

ability to treat **discourse phenomena**. That is, phenomena occurring in texts or utterances of more than a single sentence, including cases of discourse anaphora. **Underspecified meaning representations** are single representations which cover several meanings in cases where there is systematic ambiguity. While there is some work on underspecification of meaning in the theoretical literature (Reyle, 1993), the most interest has been devoted to it in computational work based on formal semantics (such as Alshawi, 1992; Bos, 1996; Copestake et al., 2005). **Model theory** deals with representing the relationship between language and the world, in computational terms to database queries (Blackburn and Bos, 2005; van Eijck and Unger, 2010).

What we have sketched above might be called the classical canon of formal semantics as it relates to computational semantics. One of the features lacking in the classical canon includes **dialogue**. The notion that language is actually used in interaction between agents engaging in communication (and therefore going beyond the notion of discourse in texts discussed earlier) came quite late to formal semantics though there is now a significant body of theoretical work. Notions of dialogue semantics covering plan-based approaches to dialogue (Allen, 1988), questions under discussion (Ginzburg, 1994, 2012) and communicative grounding (Traum, 1994) became central in the literature on formal approaches to dialogue. This gave rise to the Information State Update approach to dialogue (Larsson and Traum, 2000; Larsson, 2002). TTR (a theory of types with records, Cooper, 2005, *forthc*) has played an important role in this. **Similarity of meaning** is another feature. In addition to meaning relations such as entailment there is a notion of words, phrases and sentences having similar meanings in various respects. In a formal meaning representation this can be represented, for example, by the use of record types in TTR. Yet another feature is **robust non-logical inference** which is represented, for example, in work on textual entailment now commonly referred to as Natural Language Inference (NLI). This is hard to square with the logic-based inference discussed above. Rather than representing something that follows logically, it corresponds to what conclusions people might draw from a given utterance or text. It is often reliant on background knowledge and is to a large extent defeasible. The work on topoi by Breitholtz (2020) and probabilistic TTR (Cooper et al., 2015)

is suggestive of a computational approach to this. Finally, while model theory purports to relate language and the world it tells us little about how we relate our perception of the world and action in the world to the meaning of words and phrases which is known as **perceptual grounding**. Such issues become important, for example, if we want to put natural language on board a robot. This has become central to theories such as TTR (Cooper, *forthc*; Larsson, 2013; Dobnik et al., 2013) and Dynamic Syntax (Kempson et al., 2016). There is important formal work on **multimodal nature of communication**, for example (Lücking, 2016; Pustejovsky and Krishnaswamy, 2020).

Above we have mentioned examples of formal approaches which attempt to incorporate features which are not present in the classical canon. An alternative strategy is to try to incorporate features from the classical canon in non-formal approaches (e.g. Coecke et al., 2010) or to combine aspects of non-formal and formal approaches in a single framework (e.g. Erk and Herbelot, 2020).

3 Distributional meaning representations

Meaning as a function of its usage can be traced back to Wittgenstein (1953), but were popularised by Firth (1957). The idea at its core is that the meaning of a word is given by its context. Wittgenstein (1953) primarily speaks about meaning in relation to the world and real world activities while Firth (1957) speaks about language in relation to language. The second notion of meaning is the basis for distributional semantics. The notion that meaning in language can be found based on language context is related to the observation that if two words occur in the same context, their meaning is likely related.

The two predominant approaches to constructing distributional meaning representations today is to use machine learning to construct distributed and contextualised word representations (Sahlgren, 2006; Mikolov et al., 2013a; Peters et al., 2018). In these approaches, the meaning of a word is encoded as a dense vector of real valued numbers. The values of the vector are obtained by training a neural network to perform some task, using a (possibly annotated) corpus. The task then helps guide the neural network to produce meaningful representations. Distributed word representations focus on building static representations of words given a corpus. Popular techniques for obtaining

these representations are BoW (Bag-of-Words) or SGNS (Skip Gram with Negative Sampling), popularised by (Mikolov et al., 2013a). The main trick to BoW and SGNS is to construct a training schema such that given a random meaning representation for the word x , the representation is transformed so it can be used to identify the word in a context³, or can be used to identify the context of the word. The BoW or SGNS meaning representations can then be used as a component in another system. Contextualised representations on the other hand build dynamic word representations, that is, a single word will have different vector representations in different contexts. These representations are typically informed by the output from a language model. Thus, to really exploit contextual representations effectively a sentence is needed when extracting the meaning representation. This is in contrast to the BoW and SGNS representations which are fixed after being constructed.

With distributed representations we may also **reason analogically** about words and combinations of concepts, e.g. "Russia" + "River" = "Volga" (Mikolov et al., 2013b). That is, we may construct complex meaning by combining simpler parts. By combining the representation for "Russia" and "river" we obtain some vector z which contains information about the contexts of both "Russia" and "river". By querying the vector space for words with a *similar* representation to that of z we find other words with similar context. The success of distributed meaning representations, both static and contextualised, can in part be attributed to the ability of a model to **predict similarity** between units of language. Because meaning is defined as the context in which words occurs, two vector representations can be compared and their similarity measured. (Conneau and Kiela, 2018; Vulic et al., 2020). This similarity can be explored in terms of words (Hill et al., 2015; Artetxe et al., 2016) and in term of sentences (Cer et al., 2017).

The ability to model similarity allows models to **discover relationships** between units of language. It allows models to transfer knowledge between languages. For example, unsupervised word translation can be done by aligning monolingual vector spaces. (Lample et al., 2018; Artetxe et al., 2018). Transformer models (Vaswani et al., 2017) have also enabled zero-shot and transfer learning, e.g. by learning English word representations and evaluat-

³Context here is typically a n -gram containing x .

ing on a task in another language (Pires et al., 2019). The simplicity of static and contextualised meaning representations allows us to construct them for *any* unit of language, be it words, sentences (Conneau et al., 2018), documents (Lau and Baldwin, 2016) or languages (Östling and Tiedemann, 2017).

However, a word or a sentence may mean different things depending also on a larger context. For example a sentence in different domains will express different meanings even if the words are exactly the same. This presents a problem for distributed representations, as our observation of a word or sentence in the real world includes additional context from what we have recorded in the data. However, the effects of different domains may be counteracted by **domain adaptation** techniques (Jiang and Zhai, 2007).

Distributed representations enjoy success across a wide variety of NLP tasks. However, a consequence of automatically learning from a corpus results in some inherent shortcomings. A corpus is a snapshot of a subset of language and only captures language as it was used then and there. This means that the resulting meaning representations do not inherently capture language change (though they can be used to study it, see Section 4). Additionally, the meaning representations are generally created from observing language in a corpora, not from language use in the world. A consequence of this is that distributional meaning representations don't capture the state-of-affairs in the world, i.e. the context in which the language was used. In practical terms this means that for tasks that depend on the state-of-affairs in the world, such as robot control, dialogue or image captioning, a system must incorporate this information somehow which is further explored in the remaining sections.

4 Dynamic meaning representations

To see how meaning is context dependent in (at least) two different ways we can make the distinction between *meaning potential* and *situated meaning* (Norén and Linell, 2007). The situated meaning of a word is its disambiguated and contextually enriched interpretation in a particular context of use. Meaning potential (or *lexical meaning*) is the system of affordances (Gibson, 1966; Gregoromichelaki et al., 2020) that a word offers for sense-making in a given context. In this conception, situated meaning is context dependent by construction, but we also claim that the meaning

potential of a word depends on context of a certain kind. In particular, it depends on what is *common ground* (Stalnaker, 2002) between a speaker and their audience. At a linguistics conference, a speaker might use words like *token* or *modality*—words that would mean something completely different (or nothing at all) at a family dinner. The conference speaker expects to be understood *based on* their and their audience’s joint membership in the computational linguistics community, where they (rightly or wrongly) consider certain specialised meanings to be common ground. The communities that serve as a basis for semantic common ground can be as broad as *speakers of Spanish* (grounding the “standard” Spanish lexicon), or as small as a nuclear family or close group of friends (grounding specialised meanings particular to that group of people) (Clark, 1996).

Recent work in NLP has demonstrated the value of modelling context, including sentential (Section 3) and multimodal context (Section 5) for representing situated meanings. The dynamic representations given by language models like BERT depend on the *local context* in which the word appears, but don’t the context of the community or individual speakers involved. Little work has been done in NLP to explicitly incorporate *social context*, which provides the basis for semantic common ground. Recent work has shown that neural language models can be used to detect and analyse variation and change in post-hoc way (Del Tredici et al., 2019a; Giulianelli et al., 2020). This suggests that explicitly modelling social context may be a fruitful way forward.

In the following, we identify three kinds of social context that might be accounted for with dynamic meaning representations

Variation As demonstrated in the conference example, lexical meaning is community dependent. This doesn’t necessarily mean that every NLP application needs to mimic the human ability to tailor our semantic representations to the different communities we belong to, but some applications may serve a broader set of users by doing so (Hovy, 2015). Consider, for example, an application that serves both the general public and experts in some domain. Even where variation is not explicitly modelled, it is an important factor to consider on a meta level. In practice, NLP models typically target the most prestigious, hegemonic dialect of a given language, due in part to biases in what train-

ing data is easily available on the internet (Bender et al., 2021). This results in applications that favour users who are more comfortable with the dominant language variety. Furthermore, many applications *assume* a single variety of a given language, when in fact the training data of the models they rely on is rather specific. The standard English BERT model, for example, is trained on a corpus of unpublished romance novels and encyclopedia articles, but is applied as if it represents English written large.

Alignment Semantic common ground is not only based on joint community membership—it can also be built up between particular agents through interaction. Experiments have shown that pairs of speakers develop shorter lexicalised referring expressions when they need to repeatedly identify a referent (Mills and Healey, 2008). Additions or modifications to existing common ground can take place implicitly (through *semantic accommodation*) or *meaning accommodation* (Larsson, 2010) or explicitly, as in *word meaning negotiation* (Myrendal, 2015).

There is some hope for developing models that dynamically update their meaning representations based on interaction with other agents. Larsson and Myrendal (2017) suggest an inventory of semantic update functions that could be applied to formal meaning representations based on the results of an explicit word meaning negotiation. Dynamic Interpretation Theory (Bunt, 2000) offers a way of representing meaning as change to the conversational context, including social context, and has been incorporated in the implementation of several dialogue managers (Keizer et al., 2011; Malchanau, 2019). On the distributional side, one- or few-shot learning may eventually allow models to generalise from a small number of novel uses by drawing on existing structure in the lexicon (Lake et al., 2019). One question that remains unexplored in both these cases is which updates to local (dialogue or partner-specific) semantic ground should be propagated to the agent’s representation of the communal common ground (and to which community). This naturally brings up the issue of community-level semantic change.

Change How words change in meaning has long been an object of study for historical linguists (e.g., Paul, 1891; Blank, 1999). Historical change may not seem like a particularly important thing for NLP applications to model. After all, we can accommodate for changes over decades or cen-

turies by simply retraining models with more current data, but significant *semantic shift* can also take place over a much shorter timeline, especially in smaller speech communities (Eckert and McConnell-Ginet, 1992). The issue of semantic change also intersects with that of variation, since coinages and shifts in meaning that take place in one community can cause the lexical common ground to diverge from another community. Conversely, variants in one community may come to be adopted by another (possibly broader) community. The recent widespread use of distributional semantics to study semantic change suggests that distributional representations are capable of capturing change.⁴ Diachronic distributional representations have been used to study semantic change on both a historic/language level (e.g., Dubossarsky et al., 2015; Hamilton et al., 2016) and on a short-term/community level (Rosenfeld and Erk, 2018; Del Tredici et al., 2019b; Noble et al., 2021). While social context is not often taken into account in meaning representations, ongoing research on semantic variation and change suggests that such dynamic representations are possible as extensions of the formal and distributional paradigms.

5 Grounded meaning representations

The meaning of words is not merely in our head. It is grounded in our surroundings and tied to our understanding of the world (Regier, 1996; Bender and Koller, 2020), particularly through visual perception (Mooney, 2008). Mapping language and vision to get **multi-modal** meaning representations imposes an important challenge for many real-world NLP applications, e.g. conversational agents. Such agents typically learn by finding statistical relations and often lack causal reasoning about the world (Agarwal et al., 2020) and common-sense knowledge (Hwang et al., 2021). This section describes how different modalities are typically integrated to get a meaning representation for **language-and-vision (L&V)** tasks and what is still missing in the respective **information fusion** techniques.

Historically, modelling of situated language has been influenced by ideas from language technology, computer vision and robotics, where a combination of top-down rule-based language systems was connected with Bayesian models or other kinds of classifiers of action and perception (Kruijff et al.,

2007; Dobnik, 2009; Tellex et al., 2011; Mitchell et al., 2012). In these approaches, most of the focus was on how to ground words or phrases in representations of perception and action through classification. Another reason for this hybrid approach has also been that such models are partially interpretable. Therefore, they have been a preferred choice in critical robotic applications where security is an issue. The compositionality of semantic representations in these systems is ensured by using semantic grammars, while perceptual representations such as SLAM maps (Dissanayake et al., 2001) or detected visual features (Lowe, 1999) provide a model for interpreting linguistic semantic representations. Deep learning, where linguistic and perceptual features are learned in an interdependent manner rather than engineered, has proven to be greatly helpful for the task of image captioning (Vinyals et al., 2015; Anderson et al., 2018a; Bernardi et al., 2016) and referring expression generation (Kazemzadeh et al., 2014).

A more in-depth analysis of how meaning is represented in these models is required. Ghanimifard and Dobnik (2017) show that a neural language model can learn compositionality by grounding an element in the spatial phrase in some perceptual representation. In terms of methodology for understanding what type of meaning is captured by the model, attention (Xu et al., 2015; Lu et al., 2017) has been successfully used. Lu et al. (2016) have shown that co-attending to image and question leads to a better understanding of the regions and words the model is focused on the most. Ilinykh and Dobnik (2020) demonstrate that attention can struggle with fusing multi-modal information into a single meaning representation based on the human evaluation of generated image paragraphs. This is because the nature of visual and linguistic features and the model’s structure significantly impact what representations can be learned when using an attention mechanism. Examining attention shows that attention can correctly attend to objects, but once it is tasked to generate relations (such as prepositional spatial relations and verbs), attention visually disappears as these relations are non-identifiable in the visual features utilised by the model. This leads several researchers to include specifically geometric information in image captioning models (Sadeghi et al., 2015; Ramisa et al., 2015). On the other hand, it has also been shown that a lot of meaning can be extracted solely from word dis-

⁴See (Tahmasebi et al., 2018), (Tang, 2018), and (Kutuzov et al., 2018) for recent surveys.

tributions. Choi (2020) demonstrates how linguistic descriptions encode common-sense knowledge which can be applied to several tasks while Dobnik and Kelleher (2013); Dobnik et al. (2018) demonstrate that word distributions are an important part of the semantics of spatial relations.

Interactive set-ups such as visual question answering (VQA) (Antol et al., 2015; de Vries et al., 2017) or visual dialogue (Das et al., 2017) make first attempts in modelling multi-modal meaning in multi-turn interaction. However, such set-ups are asymmetric in terms of each interlocutor’s roles, which leads to homogeneous question-answer pairs with rigid word meaning. *Conversational games* have been proposed as set-ups in which the meaning of utterances is agreed upon in a collaborative setting (Dobnik and Storckenfeldt, 2018). These settings allow for modelling meaning coordination of grounded perceptual classifiers (Larsson, 2013) and phenomena such as clarification requests. Several corpora of perceptual dialogue exist where conversational partners need to leverage dialogue and visual information to achieve mutual understanding of a scene, for example MeetUp! (Ilinykh et al., 2019), PhotoBook (Haber et al., 2019) and Cups (Dobnik et al., 2020).

Examining L&V models and representations they learn points to several significant and interesting challenges. The first relates to the structure of both datasets and models. Many corpora contain prototypical scenes where the model can primarily optimise on the information from the language model to generate an answer without even looking at the image (Cadene et al., 2019). Secondly, information captured by a language model is more compact and expressive than patterns of visual and geometric features. Thirdly, common-sense and visual information are not enough (Lake et al., 2017; Bisk et al., 2020; Tenenbaum, 2020): we also rely on mental simulation of the scene’s physics to estimate, for example, from the appearance and position of a person’s body that they are making a jump on their skateboard rather than they are falling over a fire hydrant. Such representations are necessary for modelling *embodied agents* (Anderson et al., 2018b; Das et al., 2018; Kottur et al., 2018). Fourthly, adding more modalities and representations puts new requirements on inference procedures and more sophisticated models of attention (Lavie et al., 2004) that weighs to what degree such features are relevant in a particular con-

text. In recent years we have seen work along these lines implemented with a transformer architecture (Lu et al., 2019; Su et al., 2020; Herdade et al., 2019). However, the interpretability of how individual parts (self-attentions) of large-scale models process information from different modalities is still an open question (Ilinykh and Dobnik, 2022).

6 Meaning expressed with our body

Meanings can result in bodily reactions and, conversely, they can be expressed with our bodies, for example non-verbal vocalisations, gaze and gestures.

Emotions Meanings perceived from the environment can change our emotional states and be expressed in bodily reactions: evaluating events as intrinsically unpleasant may result in gaze aversion, pupillary constriction and some of the other components (Scherer, 2009). On the other hand, our emotional states can be expressed and the expressions can be adjusted by emotional components, such as mood (Marsella et al., 2010).

Over the last years *appraisal theories* became the leading theories of emotions (for overview, see Oatley and Johnson-Laird, 2014). These theories posit that emotion arises from a person’s interpretation of their relationship with the environment or *appraisal*. The key idea behind cognitive theories is that emotions not only reflect physical states of the agents but also emotions are judgements, depending on the current state of the affairs (depending on a person, significance/urgency of the event etc.). In our view, linguistic events can as well enter the calculation of appraisal on the level of information-state of the agent which can be modelled by formal theories. For instance, following Oatley and Johnson-Laird (2014) we can distinguish emotions as either free-floating or requiring an object such as a linguistic entity, entity in the environment or a part of agent’s information-state (e.g., obstruction of the agent’s goal can lead to anger or irritation, and, vice versa, agent’s sadness can lead to the search for a new plan). Several attempts implement emotional appraisal in text and speech (e.g., Alm, 2012), and within the agent models (e.g., Marsella et al., 2010).

Non-verbal vocalisations Non-verbal vocalisations, such as laughter, are ubiquitous in our everyday interactions. In the British National Corpus laughter is a quite frequent signal regardless of gender and age—the spoken dialogue part of the

British National Corpus contains approximately one laughter event every 14 utterances. In the Switchboard Dialogue Act corpus non-verbally vocalised dialogue acts (whole utterances marked as non-verbal) constitute 1.7% of all dialogue acts and laughter tokens make up 0.5% of all the tokens that occur in the corpus.

Despite a distinct bodily reaction (laughter causes tensions and relaxations of our bodies), it is believed that we laugh in a very different sense from sneezing or coughing (Prusak, 2006). Many scholars agree that laughter is not involuntary but we laugh for a reason, *about* something and that laughter performs a social function (Mehu, 2011). It is associated with senses of closeness and affiliation, establishing social bonding and smoothing away discomfort. For example, tickling not only requires the presence of the other but also it is more likely if subjects have close relationships (Harris, 1999).

Therefore, the meaning of laughter ought to be represented so that an artificial agent can understand it and react to it accordingly (Maraev et al., 2018; Mazzocconi et al., 2021). Mazzocconi (2019) presents a function-based taxonomy of laughter, distinguishing functions such as indication of pleasant incongruity or smoothing the discomfort in conversation. Ginzburg et al. (2020) propose a formal account of laughter within the information-state of dialogue participants which includes scaling up to non-verbal social signals such as smiling, sighing, eye rolling and frowning.

Gaze Gaze has many functions. It can dictate attention, intentions, and serves to give communicative cues in interaction (Somashekarappa et al., 2021). Gaze following can infer objects that people are looking at. While we scan a visual scene, our brain stores fixation sequences in memory and reactivates them when visualising the scene later in the absence of any perceptual information (Brandt and Stark, 1997). Scan-path theory illustrations indicate that meaning representations on scanned areas depend on the semantics of sentences (Bochynska and Laeng, 2015). Semantic eye fixations supports the view of mental imagery that is flexible and creative. Being grounded in previous experiences, by selecting a past episode we are able to generalise the past information to novel images that share features (Martarelli et al., 2017). Spatial representations associated with different semantic categories launch eye movements during retrieval

(Spivey et al., 2000).

For dialogue participants gaze patterns represent resources to track their stances. Interlocutors engage in mutual gaze while producing agreeing assessments (Haddington, 2006). Gaze shifts follow sequentially a problematic stance and are followed by a divergent stance by the person who produced the gaze shift. Gaze patterns are not meaningful per se but acquire interpretation within their linguistic and interactional contexts.

Eye movement patterns, EEG signals and brain imaging are some of the techniques that have been used to augment traditional qualitative text-based features. Temporal course and flexibility of the speaker's eye gaze can be used to disambiguate referring expressions in spontaneous dialogue. Eye-tracking data from reading experiments provide structured information with fine-grained temporal resolution which closely follows sequential structure of speech and is related to the cognitive workload of speech processing (Barrett and Hollenstein, 2020). Deep convolutional neural networks have been used to classify text to gaze using eye movements. Their performance has improved when human readers were tackling semantic challenges (Mishra and Bhattacharyya, 2018). For multi-modal and multi-party interaction in both social and referential scenarios, (Somashekarappa et al., 2020) calls for categorical representation of gaze patterns.

Gestures Gestures are hand and body movements that help to convey information (Kita and Özyürek, 2003). The observational, experimental, behavioural and neuro-cognitive evidence indicates that language and gestures are linked both during comprehension and production (Wilkins, 2006; Willems et al., 2007). Speech and gestures are semantically and temporally coordinated and therefore involved in co-production of meaning. Gestures convey meaning through iconicity and spacial proximity providing information that is not necessarily expressed in speech.

While shaping of gestures is related to conceptual and semantic aspects of the accompanying speech, gestures cannot be unambiguously interpreted by naïve listeners (Hadar and Pinchas-Zamir, 2004). However, Morett et al. (2020) showed that the semantic relationship between representational gestures vs their lexical affiliates and language is evaluated similarly. Mentions of referents for the first time in a discourse are often accompanied by

gestures. For example, [Debreslioska and Gullberg \(2020\)](#) report that the “entity” gesture accompanies referents expressed by indefinite nominals. As referents are introduced in clauses, inferable referents referred to by definite nominals are identified by the contrasting “action” gestures. Head movements are produced to give feedback ([Petukhova and Bunt, 2009](#)) and it is possible to identify a specific pattern for a specific movement and that movements can be easily measured and their extent can be quantified ([Allwood and Cerrato, 2003](#)).

Fixing gesture functions, integrating different modalities and determining their composite meanings is challenging. For artificial agents multi-modal output planning is crucial and timing must be explicitly represented. [Lücking \(2016\)](#) takes a qualitative formal approach from a type-theoretic perspective, representing iconic gestures in TTR and linking them with linguistic predicates. ([Pustejovsky and Krishnaswamy, 2020](#)) take a hybrid approach linking qualitative representations in VoxML with machine learning classification.

7 Conclusions

We surveyed formal, distributional, interactive, multi-modal and body-related representations of meaning used in computational linguistics. They are able to deal with compositionality, under-specification, similarity of meaning, inference and provide an interpretation of expressions but in very different ways, capturing very different kinds of meaning. These aspects can be broadly categorised into (i) aspects that are related to the construction of linguistic forms and (ii) aspects concerned with the interpretations and understanding the world and human activity in it.

Current mainstream computational linguistics is a practical field which is not working toward a uniform model of human language but focuses on several sub-tasks which, although related, are frequently considered in isolation. For example, natural language understanding and natural language generation frequently use entirely different approaches and representations even when the linguistic context is the same, for example texts or image captions. Solutions are provided given the practical goals and limitations of each task. Secondly, the solutions are also limited by what linguistic information can be feasibly collected for this task and by our understanding of human language and behaviour (or its lack-of) as witnessed

by ongoing work in linguistics and psychology.

If our goal is to translate documents or answer general fact-based questions then a reasonable performance can be achieved even if the system is able to ground representations only indirectly in linguistic contexts over situations rather than situations themselves. However, for a situated robot semantic grounding in texts, although relevant, is not enough as it has to connect language with its environment that it accesses with its sensors and actuators. For example, word embeddings for *left* and *right* will tell us that they are similar relations but also that they have slightly different selectional preferences for objects that they relate.

Humans rely on different aspects of meaning for different kinds of descriptions and contexts and it may be perfectly fine for our task-specific computational models to only use some dimensions and in an indirect way. What is wrong to claim, however, is that any of these models have reached human-like intelligence. Humans can re-evaluate linguistic descriptions against different dimensions of meaning and this is something that our systems are not capable of. Work on the cognitive notion of attention informs us how aspects of meaning representations are selected to disambiguate under-specified linguistic utterances by balancing information from different modalities.

A stronger connection between representations from different tasks is certainly desirable and important progress has been made for example in integration of formal grammars in situated agent systems or integration of vision and language representations learned by neural networks. However, the challenge remains precisely because “linguistic experiences” of our systems are limited by the narrow tasks and domains that they are specialised on. Transferring models across contexts of language use and aligning their representations is by no means straightforward as there may be very little overlap. At the same time we also expect that models learn generalisations that apply across contexts. In line with this we suggest that future work should focus on developing benchmarks that test different representations in different contexts. For this we need datasets of instances requiring some type of linguistic inference where instances are labelled by the context type and the type(s) of modality required for successful inference. We hope that this paper points to some of the aspects of representations that need to be taken into account.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

- Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9690–9698.
- James Allen. 1988. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.
- Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*, pages 7–22. Copenhagen.
- Cecilia Ovesdotter Alm. 2012. The role of affect in the computational modeling of *Natural language*. *Lang. Linguistics Compass*, 6(7):416–430.
- Hiyan Alshawi, editor. 1992. *The Core Language Engine*. MIT Press.
- P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.
- P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2289–2294. The Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 789–798. Association for Computational Linguistics.
- N Asher and A Lascarides. 2003. *Logics of conversation*. Cambridge University Press.
- Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pages 610–623, New York, NY, USA. Association for Computing Machinery.
- Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. *arXiv*, arXiv:2004.10151 [cs.CL].
- Patrick Blackburn and Johan Bos. 2005. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.
- A. Christian Blank. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*. De Gruyter Mouton.
- Agata Bochynska and Bruno Laeng. 2015. Tracking down the path of memory: eye scanpaths facilitate retrieval of visuospatial information. *Cognitive processing*, 16 Suppl 1.
- J. Bos. 1996. Predicate logic unplugged. In *Proceedings of the Tenth Amsterdam Colloquium*, pages 133–143, Amsterdam. ILLC/Department of Philosophy, University of Amsterdam.

- Johan Bos, Stephen Clark, Mark Steedman, James R Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1240–1246.
- Johan Bos, Björn Gambäck, Christian Lieske, Yoshiki Mori, Manfred Pinkal, and Karsten Worm. 1996. Compositional semantics in Verbmobil. *arXiv preprint cmp-lg/9607031*.
- Stephan Brandt and Lawrence Stark. 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9:27–38.
- Ellen Breitholtz. 2020. *Enthymemes in Dialogue*. Brill.
- Harry Bunt. 2000. Dialogue pragmatics and context specification. In Harry Bunt and William Black, editors, *Abduction, belief and context in dialogue: studies in computational pragmatics*, pages 81–150. John Benjamins Publishing, Amsterdam/Philadelphia.
- Remi Cadene, Corentin Dancette, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases for visual question answering. In *NeurIPS*, pages 841–852.
- Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.
- Yejin Choi. 2020. Intuitive reasoning as (un)supervised language generation. Seminar, Paul G. Allen School of Computer Science and Engineering, University of Washington and Allen Institute for Artificial Intelligence, MIT Embodied Intelligence Seminar.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.
- Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.
- Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $\&\#\&$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.
- Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.
- Robin Cooper. forthc. *From Perception to Communication: a Theory of Types for Action and Meaning*. Oxford University Press.
- Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic Type Theory and Natural Language Semantics. *Linguistic Issues in Language Technology*, 10(4):1–45.
- Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.
- Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sandra Debreslioska and Marianne Gullberg. 2020. What’s new? gestures accompany inferable rather than brand-new referents in discourse. *Frontiers in Psychology*, 11.
- Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019a. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019b. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1 (Long and Short Papers), pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.
- M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241.
- Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen’s College, Oxford, United Kingdom.

- Simon Dobnik, Robin Cooper, and Staffan Larsson. 2013. [Modelling language, action, and perception in Type Theory with Records](#). In Denys Duchier and Yannick Parmentier, editors, *Constraint Solving and Language Processing: 7th International Workshop, CSLP 2012, Orléans, France, September 13–14, 2012, Revised Selected Papers*, volume 8114 of *Lecture Notes in Computer Science*, pages 70–91. Springer Berlin Heidelberg.
- Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. [Exploring the functional and geometric bias of spatial relations using neural language models](#). In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.
- Simon Dobnik and John D. Kelleher. 2013. [Towards an automatic identification of functional and geometric spatial prepositions](#). In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.
- Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. [Local alignment of frame of reference assignment in English and Swedish dialogue](#). In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.
- Simon Dobnik and Axel Storckenfeldt. 2018. [Categorisation of conversational games in free dialogue over spatial scenes](#). In *Proceedings of AixDial – Semdial 2018: The 22st Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, Aix-en-Provence, France.
- Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. *NetWords 2015 Word Knowledge and Word Usage*, page 5.
- Penelope Eckert and Sally McConnell-Ginet. 1992. Communities of practice: Where language, gender, and power all live. *Locating Power, Proceedings of the 1992 Berkeley Women and Language Conference*, pages 89–99.
- Jan van Eijck and Christina Unger. 2010. *Computational Semantics with Functional Programming*. Cambridge University Press.
- Katrin Erk and Aurelie Herbelot. 2020. How to marry a star: probabilistic constraints for meaning in context. *arXiv preprint arXiv:2009.07936*.
- J. R. Firth. 1957. *Papers in Linguistics, 1934-1951*. Oxford University Press, London.
- Joyce Friedman, Douglas B. Moran, and David S. Warren. 1978. Two Papers on Semantic Interpretation in Montague Grammar. *American Journal of Computational Linguistics*. Microfiche 74.
- Joyce Friedman and David S Warren. 1978. A parsing method for Montague grammars. *Linguistics and Philosophy*, 2(3):347–372.
- Mehdi Ghanimifard and Simon Dobnik. 2017. [Learning to compose spatial relations with grounded neural language models](#). In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.
- James J Gibson. 1966. *The senses considered as perceptual systems*. Mifflin, New York [u.a.].
- Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the 1st International Workshop on Computational Semantics*, Tilburg University. ITK Tilburg.
- Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.
- Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1).
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. [Analysing lexical semantic change with contextualised word representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- Eleni Gregoromichelaki, Stergios Chatzikyriakidis, Arash Eshghi, Julian Hough, Christine Howes, Ruth Kempson, Jieun Kiaer, Matthew Purver, Mehrnoosh Sadrzadeh, and Graham White. 2020. Affordance Competition in Dialogue: The Case of Syntactic Universals. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*.
- Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, pages 39–100.
- Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. [The PhotoBook dataset: Building common ground through visually-grounded dialogue](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.
- Uri Hadar and Lian Pinchas-Zamir. 2004. [The semantic specificity of gesture](#). *Journal of Language and Social Psychology - J LANG SOC PSYCHOL*, 23:204–214.
- Pentti Haddington. 2006. [The organization of gaze and assessments as resources for stance taking](#). *Text & Talk - TEXT TALK*, 26:281–328.

- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Christine R Harris. 1999. The mystery of ticklish laughter. *American Scientist*, 87(4):344.
- Irene Heim. 1982. *The Semantics of Definite and Indefinite NPs*. Ph.D. thesis, University of Massachusetts at Amherst.
- Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. [Image captioning: Transforming objects into words](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. [Simlex-999: Evaluating semantic models with \(genuine\) similarity estimation](#). *Comput. Linguistics*, 41(4):665–695.
- Dirk Hovy. 2015. [Demographic factors improve classification performance](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.
- Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. [\(comet-\) atomic 2020: On symbolic and neural commonsense knowledge graphs](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.
- Nikolai Ilinykh and Simon Dobnik. 2020. [When an image tells a story: The role of visual and semantic information for generating paragraph descriptions](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.
- Nikolai Ilinykh and Simon Dobnik. 2022. [Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.
- Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. [Meet up! a corpus of joint activity dialogues in a visual environment](#). In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.
- Jing Jiang and ChengXiang Zhai. 2007. [Instance weighting for domain adaptation in NLP](#). In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.
- Hans Kamp, Josef van Genabith, and Uwe Reyle. 2011. [Discourse Representation Theory](#). In Dov Gabbay and Franz Guenther, editors, *Handbook of Philosophical Logic*, volume 15. Springer Science+Business Media B.V. .
- Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.
- Simon Keizer, Harry Bunt, and Volha Petukhova. 2011. [Multidimensional Dialogue Management](#). In Antal van den Bosch and Gosse Bouma, editors, *Interactive Multi-modal Question-Answering, Theory and Applications of Natural Language Processing*, pages 57–86. Springer, Berlin, Heidelberg.
- Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as Mechanisms for Interaction. *Theoretical Linguistics*, 42(3-4):203–276.
- Sotaro Kita and Asli Özyürek. 2003. [What does cross-linguistic variation in semantic co-ordination of speech and gesture reveal?: Evidence of an interface representation of spatial thinking and speaking](#). *Journal of Memory and Language*, 48:16–32.
- Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. [Visual coreference resolution in visual dialog using neural module networks](#).
- Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Brenden M. Lake, Tal Linzen, and Marco Baroni. 2019. [Human few-shot learning of compositional instructions](#). *arXiv:1901.04587 [cs]*.
- Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. [Building machines that learn and think like people](#). *Behavioral and Brain Sciences*, 40:e253.

- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ran-
zato, Ludovic Denoyer, and Hervé Jégou. 2018. [Word translation without parallel data](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, University of Gothenburg.
- Staffan Larsson. 2010. Accommodating innovative meaning in dialogue. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 83–90.
- Staffan Larsson. 2013. [Formal semantics for perceptual classification](#). *Journal of Logic and Computation*, 25(2):335–369.
- Staffan Larsson and Jenny Myrendal. 2017. [Dialogue Acts and Updates for Semantic Coordination](#). In *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 52–59. ISCA.
- Staffan Larsson and David R Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural language engineering*, 6(3-4):323–340.
- Jey Han Lau and Timothy Baldwin. 2016. [An empirical evaluation of doc2vec with practical insights into document embedding generation](#). In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 78–86. Association for Computational Linguistics.
- Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. [Load theory of selective attention and cognitive control](#). *Journal of Experimental Psychology: General*, 133(3):339–354.
- David G Lowe. 1999. [Object recognition from local scale-invariant features](#). In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE.
- J. Lu, C. Xiong, D. Parikh, and R. Socher. 2017. [Knowing when to look: Adaptive attention via a visual sentinel for image captioning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, page 289–297, Red Hook, NY, USA. Curran Associates Inc.
- Andy Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, pages 383–392. IEEE.
- A. Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 383–392.
- Andrei Malchanau. 2019. *Cognitive Architecture of Multimodal Multidimensional Dialogue Management*. Ph.D. thesis, Saarland University, Saarbrücken.
- Vladislav Maraev, Chiara Mazzocconi, Christine Howes, and Jonathan Ginzburg. 2018. [Integrating laughter into spoken dialogue systems: preliminary analysis and suggested programme](#). In *Proceedings of the FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 9–14.
- Stacy Marsella, Jonathan Gratch, Paolo Petta, et al. 2010. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46.
- Corinna Martarelli, Sandra Chiquet, Bruno Laeng, and Fred Mast. 2017. [Using space to represent categories: insights from gaze position](#). *Psychological Research*, 81.
- Chiara Mazzocconi. 2019. *Laughter in interaction: semantics, pragmatics and child development*. Ph.D. thesis, Université de Paris.
- Chiara Mazzocconi, Vladislav Maraev, Vidya Somashekarappa, and Christine Howes. 2021. [Looking for laughs: Gaze interaction with laughter pragmatics and coordination](#). In *Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI ’21*, page 636–644, New York, NY, USA. Association for Computing Machinery.
- Marc Mehu. 2011. Smiling and laughter in naturally occurring dyadic interactions: Relationship to conversation, body contacts, and displacement activities. *Human Ethology Bulletin*, 26(1):10–28.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. [Efficient estimation of word representations in vector space](#). In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

- Gregory Mills and Pat Healey. 2008. Semantic negotiation in dialogue: The mechanisms of alignment. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 46–53.
- Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Scanship Complexity: Modeling Reading/Annotation Effort Using Gaze Information: An Investigation Based on Eye-tracking*, pages 77–98. Springer, Singapore.
- Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.
- Richard Montague. 1973. The Proper Treatment of Quantification in Ordinary English. In Jaakko Hintikka, Julius Moravcsik, and Patrick Suppes, editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 247–270. D. Reidel Publishing Company, Dordrecht.
- Raymond J. Mooney. 2008. Learning to connect language and perception. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3, AAAI’08*, page 1598–1601. AAAI Press.
- Laura Morett, Sarah Hughes Berheim, and Raymond Bulger. 2020. Semantic relationships between representational gestures and their lexical affiliates are evaluated similarly for speech and text. *Frontiers in Psychology*, 11.
- Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. PhD Thesis, University of Gothenburg, University of Gothenburg.
- Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37, Online. Association for Computational Linguistics.
- Kerstin Norén and Per Linell. 2007. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, 17(3):387–416.
- Keith Oatley and P.N. Johnson-Laird. 2014. Cognitive approaches to emotions. *Trends in Cognitive Sciences*, 18(3):134–140.
- Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.
- Barbara H. Partee, editor. 1976. *Montague Grammar*. Academic Press.
- Hermann Paul. 1891. *Principles of the History of Language*. London ; New York : Longmans, Green.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.
- Volha Petukhova and Harry Bunt. 2009. Grounding by nodding. In *Proceedings of GESPIN, Conference on Gestures and Speech in Interaction, Poznań*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Bernard G Prusak. 2006. The science of laughter: Helmut plessner’s laughing and crying revisited. *Continental philosophy review*, 38:41–69.
- James Pustejovsky and Nikhil Krishnaswamy. 2020. Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. *Revue TAL*, 61(3):17.
- Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal. Association for Computational Linguistics.
- Terry Regier. 1996. *The human semantic potential spatial language and constrained connectionism*. Neural network modeling and connectionism. MIT Press, Cambridge.
- Uwe Reyle. 1993. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10(2):123–179.
- Alex Rosenfeld and Katrin Erk. 2018. Deep Neural Models of Semantic Shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.
- Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction

- and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.
- Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Institutionen för lingvistik.
- Klaus R Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351.
- Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 759–765, Marseille, France. European Language Resources Association.
- Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2021. A deep gaze into social and referential interaction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).
- Michael Spivey, Daniel Richardson, Melinda Tyler, and Ezekiel E Young. 2000. Eye movements during comprehension of spoken descriptions. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*.
- Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy*, 25(5-6):701–721.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. *Vi-bert: Pre-training of generic visual-linguistic representations*. In *International Conference on Learning Representations*.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. *Survey of Computational Approaches to Diachronic Conceptual Change*. *arXiv:1811.06278 [cs]*.
- Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.
- Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. *Understanding natural language commands for robotic navigation and mobile manipulation*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1).
- Joshua B. Tenenbaum. 2020. *Cognitive and computational building blocks for morehuman-like language in machines*. Acl 2020 keynote, Center for Brains, Minds and Machines, MIT.
- David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. *Show and tell: A neural image caption generator*.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. *Guesswhat?! visual object discovery through multi-modal dialogue*.
- Ivan Vulic, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. *Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity*. *CoRR*, abs/2003.04866.
- David Wilkins. 2006. *Adam kendon (2004). gesture: Visible action as utterance*. *Gesture*, 6.
- Roel Willems, Asli Özyürek, and Peter Hagoort. 2007. *When language meets action: The neural integration of gesture and speech*. *Cerebral cortex (New York, N.Y. : 1991)*, 17:2322–33.
- Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. *Show, attend and tell: Neural image caption generation with visual attention*. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Can We Use Small Models to Investigate Multimodal Fusion Methods?

Lovisa Hagström¹ Tobias Norlund^{1,2} Richard Johansson^{1,3}

¹Chalmers University of Technology, ²Recorded Future,

³University of Gothenburg

{lovhag, tobiasno, richajo}@chalmers.se

Abstract

Many successful methods for fusing language with information from the visual modality have recently been proposed and the topic of multimodal training is ever evolving. However, it is still largely not known what makes different vision-and-language models successful. Investigations into this are made difficult by the large sizes of the models used, requiring large training datasets and causing long train and compute times. Therefore, we propose the idea of studying multimodal fusion methods in a smaller setting with small models and datasets. In this setting, we can experiment with different approaches for fusing multimodal information with language in a controlled fashion, while allowing for fast experimentation. We illustrate this idea with the math arithmetics sandbox. This is a setting in which we fuse language with information from the math modality and strive to replicate some fusion methods from the vision-and-language domain. We find that some results for fusion methods from the larger domain translate to the math arithmetics sandbox, indicating a promising future avenue for multimodal model prototyping.

1 Introduction

Having models learn language from text alone has been criticised based on several aspects, from fundamental arguments about how language works (Bender and Koller, 2020; Bisk et al., 2020) to findings on certain information lacking in text (Gordon and Van Durme, 2013; Paik et al., 2021). Consequently, there is much interest in creating models that learn from more than text, i.e. “multimodal models”. Many different multimodal models that fuse different types of information with text have been developed, ranging from vision-and-language models (VL models) (Zhang et al., 2020) to language models fused with knowledge graphs (Yu et al., 2022). In this work, we mainly focus on the vision-and-language domain, while our results may generalize to other multimodal domains as well.

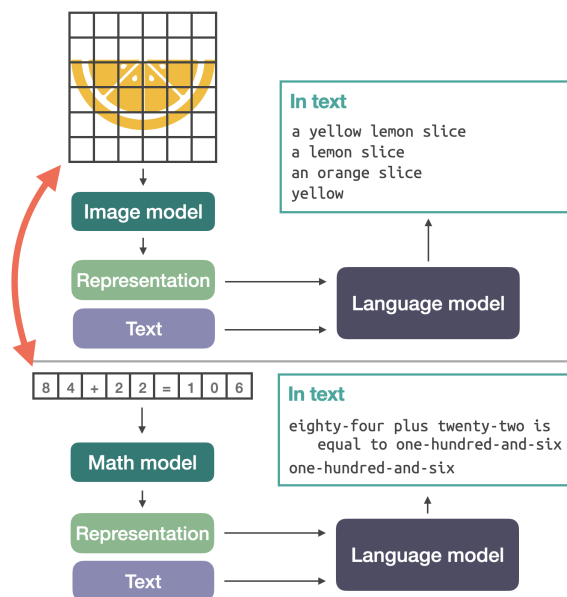


Figure 1: An overview of the vision-and-language fusion method and our proposed math-language fusion method. The image model and the math model are backbones.

A standard approach for fusing information from images with language is firstly to use a *backbone*, a large neural network that has been trained to create good representations of images. One such model is ResNet trained on Visual Genome (Anderson et al., 2018; Krishna et al., 2016). The training data in this standard approach typically consists of image samples paired with linguistic information, such as MS COCO (Lin et al., 2014). Multimodal fusion is then achieved by generating representations of image samples using the backbone and feeding those together with paired linguistic samples to a large language model pre-trained on text. The language model is then expected to learn to leverage the visual information for its linguistic usage if trained on a sufficient amount of image-text pairs. We illustrate this approach in Figure 1.

Significant success on several multimodal vision-and-language tasks has been achieved with this ap-

proach (Chen et al., 2020; Li et al., 2019, 2020; Tan and Bansal, 2019). However, it has seldom been investigated from a methodological perspective, and existing methodological investigations indicate that performance on different vision-and-language tasks may not originate from what we would expect. For example, differences in performance for different VL models have previously been ascribed to different model architectures, while Bugliarello et al. (2021) recently showed that training data and embedding layers matter more than e.g. if the model is dual- or single-stream. Hessel and Lee (2020) also showed that performance improvements for some VL models on image-text classification tasks mainly originate from unimodal signals, and not cross-modal interactions. Additionally, Frank et al. (2021) found that VL models are not necessarily symmetrical in their cross-modal interactions.

Additionally, the standard approach for fusing information from images with language does not necessarily encourage experiments with different methods for multimodal fusion, mainly since the associated compute power is substantial, as a consequence of large models and data sizes. For example, the size of the Faster R-CNN visual features used by Bugliarello et al. (2021) is approximately 1.6 TB, significantly larger than e.g. the 20.3 GB for English Wikipedia¹.

In this work, we hypothesize that methods for multimodal fusion can be developed in a smaller domain for more efficient investigations. We experiment with a *sandboxed* experimental setting for investigating different multimodal fusion methods. It is based on synthetic multimodal data, in a very constrained math-and-language domain consisting of simple arithmetic math statements, as illustrated in Figure 1. Using this setup, we are free to investigate different multimodal fusion methods like the one seen in Figure 1 *in silico*, using models that are easier to work with, with faster training times and full control of the data since we can synthetically generate it. The sandbox is described in Section 2 and we release the corresponding code to enable other researchers to build on it.²

We reason that results in the math-and-language domain could generalize to more complex domains, such as the vision-and-language domain (Section 2) and provide empirical support for this with a few

¹Provided by Huggingface via <https://huggingface.co/datasets/wikipedia>

²Available at <https://github.com/lovhag/small-math-language-multimodal-fusion>

experiments (Section 3).

To summarize, our contributions are two, 1) proposal of idea in using smaller domains for easier investigations of vision-and-language fusion methods, and 2) demonstration of idea in the math arithmetics sandbox set in the math-and-language domain. We couple this demonstration with validating experiments that compare against recent results in the vision-and-language domain.

2 The math arithmetics sandbox

Similarly to how images are described with pixel values over some grid, equations can be described with numerical values over potential positions in an equation. Also similarly to how we use image specific models to generate representations of images for language fusion, we can use math specific models to generate representations of equations for language fusion, illustrated in Figure 1.

In the math arithmetics sandbox we limit ourselves to two-digit numbers and equations describing sums of these numbers. An example of a data sample in the math modality of the sandbox is then $54 + 21 = 75$, with the corresponding string “fifty-four plus twenty-one is equal to seventy-five”. We can work with a total of $10^4 = 10,000$ different such math-language pairs in our sandbox. This number of possible samples is much smaller than e.g. the corresponding number for image data, for which the size of the support is $(V^3)^{32 \times 32}$ for 32×32 -pixel images with V possible values for each RGB channel, with underlying probability distributions of the pixel values.

An example of a potential task in the math arithmetics sandbox is to predict the continuation of the string “fifty-four plus twenty-one is equal to” given the math information $54 + 21 = 75$. This task can be compared to the task of visual question answering in the vision-and-language domain, for which a question could be “What is the color of the fruit?” provided with an image of a yellow lemon. The underlying format of the two tasks is essentially the same, a text prompt is provided together with information from another modality that encodes the information necessary to answer the prompt.

We hypothesize that results from experiments in the math arithmetics sandbox can give intuitions about the vision-and-language domain, since the difference between the math arithmetics sandbox and more complex multimodal domains mainly lies in the size of the support, while the underlying for-

mat for the multimodal tasks are similar. Thus, it is interesting to investigate whether we can acquire valuable knowledge in the math arithmetics sandbox and exploit it in more complex multimodal domains, such as the vision-and-language domain.

3 Experiments

To investigate whether insights gained in the math arithmetics sandbox are comparable to other multimodal domains, we perform a set of experiments.

3.1 Setup

We validate our math arithmetics sandbox by comparing findings from Huang et al. (2021) and Bugliarello et al. (2021) with findings we obtain on the same, but sandboxed, cases. The findings we investigate and aim to reproduce are:

1. Adding information to a model from a modality with a sufficient amount of train samples improves on the performance of a model (Huang et al., 2021).
2. Training with an additional modality with too few training samples weakens the performance of a model compared to not adding the extra modality (Huang et al., 2021).
3. The design of the embedding layers matters for VL models (Bugliarello et al., 2021).
4. Dual- and single-stream VL model architectures perform on par (Bugliarello et al., 2021).

Task The task of the validating experiments is to accurately predict the answer to a text version of a sum equation x_T , given the corresponding complete math equation x_M . An input example for this task is $x_T = \text{“one plus two is equal to”}$, $x_M = 1+2=3$ and with the correct answer $y = \text{“three”}$.

During validation, the multimodal model is to generate the continuation of an incomplete string equation given the corresponding complete math equation. For simplicity, we use a greedy decoding scheme and measure the $k = 1$ accuracy.

Models As illustrated in Figure 1 our problem setup firstly consists of a math model M_M that generates a representation $M_M(x_M)$ of a math input x_M . This representation is then given to a language model M_L together with the incomplete text input x_T to get a prediction $y' = M_L(x_T, M_M(x_M))$.

We model M_M with a small version of GPT2³

³Embedding dimension of 64, 4 Transformer layers, inner dimension of 256 and 8 attention heads.

Name	Embedding	Stream	Backb
GPT2text			
VisualBERT	VisualBERT	Single	99%
UNITER	UNITER	Single	99%
LXMERTs	LXMERT	Single	99%
LXMERTd	LXMERT	Dual	99%
LXMERTb	LXMERT	Single	5%

Table 1: Backb denotes the backbone, which was trained on either 99% or 5% of the math data. GPT2text is a standard unimodal GPT2 model only taking text input, serving as a unimodal baseline.

(Radford et al., 2019) with a vocabulary restricted to only include the tokens in the math equations. The output of $M_M(x_M)$ is then a set of vector representations, one for each token in x_M .

M_L is also modelled with a model similar to GPT2. The difference here is that we need to adapt the model to accept a multimodal input (x_T and $M_M(x_M)$). For this, we design a special embedding layer and choose between a single- or dual-stream architecture for the model.

Since one goal of this article is to investigate the effect of different embedding layer designs and dual- versus single-stream model architectures, we create and evaluate four different model variations, seen in Table 1. The embedding designs essentially determine how the multimodal model input x_T and $M_M(x_M)$ is processed before being passed to the encoder of M_L . The dual- or single-stream architectures essentially determine how early linguistic information and visual information is fused while being processed in M_L . We name the model variations after their embedding design, stream type and amount of training data for the math model (backbone). The embedding designs are named after the VL model that incorporates the same design. Detailed descriptions of how we create the model variations can be found in Appendix A.

Data The data we have available are the 10,000 math-language pairs. We train the math model M_M on 9,900 (99%) of the pure math equations samples such that it attains a validation accuracy of 0.99, to ensure that the model is able to generate information-rich math features. We then train the language model M_L on math-text pairs. We experiment with different sizes of training data and evaluate on the remaining data samples, to investigate the effect of having access to many or few text samples. We train the models until they have

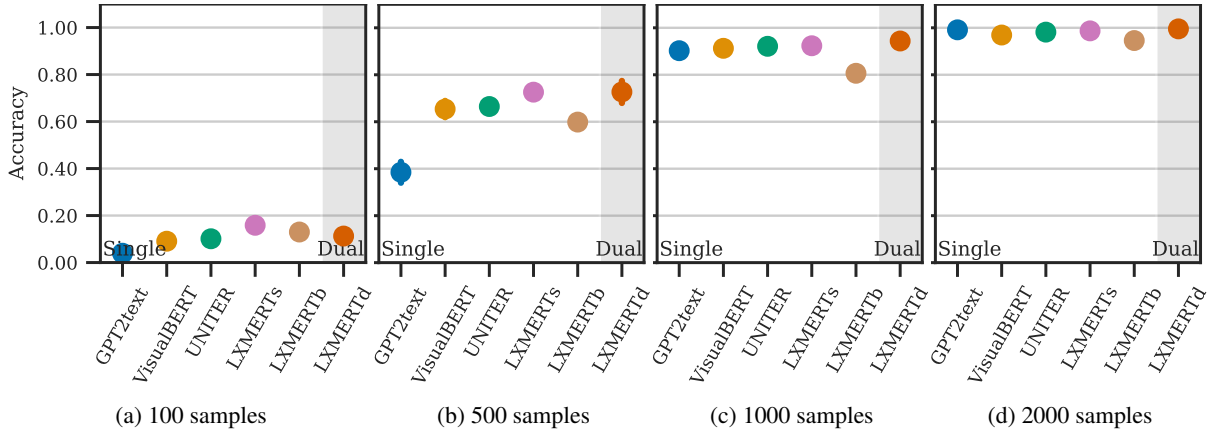


Figure 2: The validation accuracies for the different sandboxed models trained on 100, 500, 1000 and 2000 math-text samples respectively. The validation task was to predict the answer to a string version of a sum equation. Each configuration was trained 5 times.

converged in performance on the validation set and report their accuracy score on the same set.

We also train a math model on only 500 samples (5%) of the pure math equations to a validation accuracy of 0.67. We use this backbone to investigate 2) the effect of training with too few samples for the additional modality and train a single-stream model with LXMERT embeddings with this backbone, denoted ‘LXMERTb’.

3.2 Results

The results from the validating experiments are shown in Figure 2.

Does adding multimodal information improve model representations? Yes, for all cases in which the language data is more scarce (<2000 samples), adding multimodal information leads to a better model performance than only using text.

Does training with insufficient data for the additional modality weaken model performance? Predominantly yes, when there is more text data (500< samples) adding information from the “bad” math backbone has a deteriorating impact on the model performance and the multimodal model even performs worse than the unimodal version. When there is less paired math-text data, additional information from the backbone trained on little data leads to a better performance than in the pure-text case. Potentially, this is due to the pure text case also having too little data, such that additional information, despite bad quality, still is helpful.

Do embeddings matter for our models? No, we do not observe significantly different perfor-

mances for the models with different embedding layers. Potentially, this is due to the features from the math modality being so simple that the embedding does not really matter, compared to the case of e.g. features from an object detector that also encode locations of bounding boxes.

Do dual- and single-stream architectures perform on par? Yes, despite the fact that the sandboxed dual-stream model is larger than the single-stream models, the model performs on par with the single-stream models. This is in agreement with results found on the vision-and-language domain.

4 Conclusions

We propose the idea of studying vision-and-language fusion methods from a smaller domain. We reason that it would be easier and less resource demanding to develop performant multimodal fusion methods if we could investigate them using small models and domains. We exemplify this with the math arithmetics sandbox and get promising results. However, more experiments on other small domains are necessary to evaluate the potential of our proposed idea. It would also be beneficial to investigate small domains that are more similar to e.g. the vision-and-language domain.

Acknowledgements

We would like to thank the anonymous reviewers of this paper for their valuable feedback. Additionally, this work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. [Experience grounds language](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.
- Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. [Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs](#). *Transactions of the Association for Computational Linguistics*, 9:978–994.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.
- Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. [Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jonathan Gordon and Benjamin Van Durme. 2013. [Reporting bias and knowledge acquisition](#). In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.
- Jack Hessel and Lillian Lee. 2020. [Does my multimodal model learn cross-modal interactions? it’s harder to tell than you might think!](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.
- Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#).
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. [Oscar: Object-semantics aligned pre-training for vision-language tasks](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. [The World of an Octopus: How Reporting Bias Influences a Language Model’s Perception of Color](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Survey (CSUR)*.
- Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493.

A Language Model variations

To investigate the effect of different vision-and-language fusion methods, we implement versions

of these in the model M_L in the math-and-language domain.

A.1 Design of embedding layers

We wish to investigate the effect of different embedding layer designs for text x_T and math features $M_M(x_M)$. The embedding designs essentially determine how the multimodal model input x_T and $M_M(x_M)$ is processed before being passed to the encoder of M_L . Similarly to the work by [Bugliarello et al. \(2021\)](#), we switch between different embedding designs, for which we test designs similar to those used in VisualBERT, UNITER and LXMERT ([Li et al., 2019](#); [Chen et al., 2020](#); [Tan and Bansal, 2019](#)). The embedding designs we use mainly differ in how layer normalization and dropout are applied to the x_T and $M_M(x_M)$ inputs.

In the vision-and-language domain the embedding designs are more differentiated since they process positional information differently. As a consequence of not having any extra positional information in the math-and-language domain we cannot imitate this positional embedding design difference in the math arithmetics sandbox.

A.2 Dual- and single-stream models

We also wish to investigate dual- and single-stream multimodal models. The single-stream model is already provided by the standard GPT2 architecture in the sense that we give it the concatenation of the math and linguistic features of an math-text pair as input, to allow for information fusion from the start. This design is similar to that of the single-stream VisualBERT model.

The dual-stream architecture we use is based on the dual-stream LXMERT architecture, in which we switch BERT specific layers for the corresponding GPT2 layers, such that the math and linguistic features are first processed by two independent stacks of Transformer layers before they are fed into cross-modal Transformer layers. Dual-stream models are typically larger, where e.g. LXMERT with 228M trainable parameters is two times larger than the 110M parameters of single-stream VisualBERT. For our sandboxed dual-stream model, we go by the same principle and set the number of linguistic Transformer layers to three, the number of relational layers to two and the number of cross-attention layers to two, such that we get a sandboxed dual-stream model that is approximately two times larger than the sandboxed single-stream models, corresponding to the size difference between

VisualBERT and LXMERT.

The number of trainable parameters for the sandboxed single-stream models is approximately 211K. For the sandboxed dual-stream model the number of trainable parameters is 495K.

Embodied Interaction During Mental Health Consultations: Some Observations on Grounding and Repair

Jing Hui Law, Patrick G.T. Healey and Rosella P. Galindo Esparza

Queen Mary, University of London
Wolfson Institute of Population Health and
School of Electronic Engineering and Computer Science
London E1 4NS
jing.law@qmul.ac.uk

Abstract

Shared physical space is an important resource for face-to-face interaction. People use the position and orientation of their bodies—relative to each other and relative to the physical environment—to determine who is part of a conversation, to manage conversational roles (e.g. speaker, addressee, side-participant) and to help co-ordinate turn-taking. These embodied uses of shared space also extend to more fine-grained aspects of interaction, such as gestures and body movements, to support topic management, orchestration of turns and grounding. This paper explores the role of embodied resources in (mis)communication in a corpus of mental health consultations. We illustrate some of the specific ways in which clinicians and patients can exploit embodiment and the position of objects in shared space to diagnose and manage moments of misunderstanding.

1 Background

Non-verbal signals are integral to natural human interaction. The best known are the facial expressions of emotion (e.g. anger, fear, sadness, surprise, happiness) (Ekman, 1979; Chovil, 1991b). However, there are also a range of non-verbal signals that are specific to conversation (Bavelas et al., 1995; Chovil, 1991a; Kaulard et al., 2012). These include large-scale configurations of body position and orientation that can tell us e.g., who is participating in a conversation, what their role is (e.g. speaker, addressee, listener or bystander) and their relative levels of interest and engagement (Schefflen, 1973; Kendon, 2010; Bull, 2016). There are also a range of small-scale conversational gestures. For example, the use of hand gestures to hold or yield a turn, to enlist help with finding a word—or expression and the use of facial gestures such as raised eyebrows to emphasise particular words or to display “thinking” (Bavelas et al., 1995; Ekman, 1979; Chovil, 1991a).

Embodied signals can be produced in parallel with verbal contributions by both the speaker and by other participants (Bavelas, 2007; Bavelas et al., 1995; Deppermann et al., 2021). This facilitates real-time, incremental checking and feedback. Speakers can produce gestures that complement or augment their speech and listeners can simultaneously display their reactions through concurrent backchannel signals (Chovil, 1991b,a). These concurrent signals shape a speaker’s turn in real-time and—if problems are apparent—can cause a speaker to rephrase, change direction or cut-off their turn (Goodwin, 1979).

Some non-verbal signals are associated with potential problems with shared understanding. A frown and briefly averted gaze before a turn can suggest that a speaker is about to say something potentially problematic (Kaukomaa et al., 2014) and gaze aversion by an addressee following a turn can prompt the speaker to rephrase what they said (Kendrick and Holler, 2017). Some facial gestures, such as raised eyebrows and widening of the eyes, can act as stand-alone clarification requests (Kendrick, 2015; Seo and Koshik, 2010). Similarly, temporary suspension of hand movements described as non-verbal ‘holds’ or ‘freezes’ can provide signals of ongoing repairs (Seo and Koshik, 2010; Floyd et al., 2016; Bavelas et al., 1995). Quantitative data from motion captured conversations shows that the velocity and height of head and hand movements changes during both self-repairs/disfluencies and other-repairs—and that these changes are different for speakers and listeners (Healey et al., 2013, 2015; Özkan et al., 2021).

2 Communication in Healthcare

Communication in healthcare settings is critical to the quality of patient-clinician relationships, affecting outcomes such as patient satisfaction and treat-

ment adherence. It is arguably even more important in mental healthcare settings, since talk is the primary means of diagnosis and treatment (McCabe and Healey, 2018; Mahmoodi et al., 2020; McCabe et al., 2013; McVittie et al., 2020; Wu, 2020). However, difficulties in balancing good communication practices with work pressures have often been reported in the NHS (NHS Improvement, 2018).

This has prompted an interest in developing protocols and tools to help healthcare professionals organise their thoughts and structure consultations. However, overly rigid protocols can have a problematic impact on the naturalness of healthcare professional’s interaction, limiting their ability to adapt to individual patient’s needs, concerns or understanding (NHS Improvement, 2018). At its worst, the artificial structure imposed by the protocol disrupts the flow of conversation and can become counterproductive for clinical interactions (NHS Improvement, 2018).

Here we examine these issues in the context of a tablet application (DIALOG+) designed to promote communication in face-to-face mental health consultations. We focus on a detailed qualitative analysis of the moments where misunderstandings arise—and the combination of verbal and non-verbal resources that are used to address such problems. Our analysis shows how the position of the physical device and people’s orientation toward it plays a role in both causing and mitigating misunderstandings.

2.1 DIALOG+

The DIALOG+ protocol is designed to support and structure conversations in routine community mental healthcare consultations for patients suffering from psychosis (Priebe et al., 2015, 2017). This intervention applies principles of solution-focused therapy to promote assessment of all relevant aspects of patients’ lives. It also uses this information to help patients initiate change and improve their situations. The overall aim is to improve the therapeutic benefits of the consultation process (Priebe et al., 2015, 2017).

The DIALOG+ application is a tablet-based system, built around a central screen with a sequence of eleven quality of life questions (items) that cover various aspects of patients’ mental and physical health, job situation, relationships, medication and practical help received. Clients and clinicians work through the list together—and as the users select each item, a slider appears, so that clients can pro-



Figure 1: Screenshot of the DIALOG+ Application User Interface

vide their current satisfaction rating for each item. After performing the ratings, a structured, iterative 4-step process is followed of (1) choosing items that the patients would like to discuss, (2) understanding what determines the ratings, (3) considering options for what can be done to improve their satisfaction with these items and (4) agreeing on some next steps and action plans that can be adopted by the clients, or with assistance from the clinician, family members and/or support workers, to get to potential solutions.

Clinical trials suggest the intervention is effective, with patients reporting fewer general psychopathological symptoms, fewer unaddressed needs, higher levels of treatment satisfaction and better objectively measured social outcomes—such as in terms of housing situations and employment status—at one-year follow ups (Priebe et al., 2015, 2017). The DIALOG+ app has been translated into over 15 languages and has been implemented and tested in various studies in more than 10 countries. Nevertheless, little is understood about how the intervention is actually incorporated into interactional practices within consultations. This is important for refining clinician communication skills and also because new versions of the application are being developed, including for remote use.

Since the COVID-19 pandemic, many consultation services in healthcare have moved online, at least temporarily (Liberati et al., 2021; Khan et al., 2021). In mental healthcare settings, this has led to reported difficulties in the establishment or maintenance of meaningful trust and rapport between clients and clinicians (Liberati et al., 2021; Olwill et al., 2021; Khan et al., 2021). The use of standardised tools in mediated communication

is a potential concern, given the complexities of balancing flexibility with system-driven structures of service delivery (Drew et al., 2021). In this context, it is important to understand the interactional features of DIALOG+ and how this may impact on its delivery online. To gain a more granular perspective, we explore the details of the DIALOG+ consultations to observe what exactly happens in these interactions.

3 The current study

The data for this study comes from video recordings of mental health consultations. Extracts from these interactions are described below using conversation analysis (CA) techniques. CA involves detailed qualitative analysis of conversations in naturally occurring circumstances—and the material for such analyses is recorded and/or transcribed talk (Sacks et al., 1978; Sacks, 1992; Silverman, 1998). The focus of the analysis is on the organisation of fine-grained features of interaction such as timing, pauses, repetitions, restarts and the details of concurrent non-verbal signals such as gaze and gesture.

The examples considered below are selected to illustrate the types and trajectories of (mis)communication observed in this corpus, how they are affected by the presence of DIALOG+ application, and how this affects the management of shared understanding. The two overarching research questions in this study are:

1. Where and how does shared understanding break down in the DIALOG+ interactions?
2. What role does embodiment play in detecting and dealing with these breakdowns?

4 Methods

Design. The DIALOG+ trials are described elsewhere (Priebe et al., 2017). Our dataset consists of 40 video recordings of 32 clinical consultations; 16 are from a control group receiving treatment-as-usual and 17 from an intervention group using DIALOG+. The average length of the recordings is 30.12 minutes in the control group and 39.68 minutes in the intervention group.

Participants and Ethics. Participants who consented to join the original DIALOG+ trials provided informed consent to be video-taped or audio-taped on at least one session for use in future studies on potential improvements to the DIALOG+

procedure and technology. The study protocol, including data collection and storage procedures, was audited by the National Research Ethics Service (NRES) London, Stanmore (12/LO/1145). All personal data has been removed from the extracts and faces have been blurred.

Procedure and data analysis. The authors familiarised themselves with the data by watching and listening to the recordings repeatedly, making brief notes about the general form of the conversations and any features that appeared striking. A second pass then focused on selecting particular episodes with overt evidence of misunderstanding. The clearest of these were transcribed in detail using Jefferson's orthography which includes features like pauses, overlap and intonation (Sacks et al., 1978; Ekberg, 2021). Embodied conduct that is relevant to the interaction—such as nods, gestures and postures shifts—were also included in the transcriptions.

5 Results

5.1 Overview

The recorded conversations were mostly follow-up sessions to previous consultations. Although DIALOG+ is designed to encourage input by both parties to the interaction in practice, the clinicians typically took the initiative and controlled all input to the device. Patients only rarely touched the tablet or laptop—and the devices were often positioned in ways that obscured the patients' view of the screen (although see Figure 2).

The full DIALOG+ protocol is not strictly followed in the consultations. All participating clinicians were trained in the use of the protocol—and the application in front of them is also structured according to the protocol and contains prompts. However, steps are sometimes skipped, merged or discussed in varying orders. These deviations typically arise as adaptations to the immediate conversational context, particularly where clinicians' and patients' expectations or interpretations are misaligned. In general, this merging and skipping was designed to prioritise the flow of the conversation over the protocol.

A specific structural problem arises where patients mention a concern early in a consultation and then this same concern is reintroduced later in the conversation by the DIALOG+ protocol. This creates a sequential problem in the conversation—and the pragmatic effect of these repetitions is to

imply that either the patient's original description was somehow inadequate, or that the clinician had not been listening (see also below).

Another recurring issue is understanding the relevance of patients' responses to questions about particular items. There are often natural interconnections between, for example, people's accommodation situation and their family situation (see e.g., Excerpt 1 and Excerpt 2). The protocol questions imply a level of conceptual independence between different quality of life domains, but this may be misaligned with the concerns and practical circumstances of individual patients.

A third common source of trouble is clinicians' concurrent typing or note-taking, which was interspersed throughout discussions. These activities disrupt the flow of conversation and sometimes cause a misalignment on what should come next in the conversation.

In the control group, clinicians usually make notes throughout the consultations, but spend less time on this than clinicians in the intervention group (partly because typing on the tablet keyboard is a slow process). Similar issues in the patients' lives are explored—but more rapidly and lightly in conversation. This sometimes led to sessions with no action plans agreed upon. In other sessions, time is spent actually taking action on and resolving a particular issue the patient is facing (e.g. making calls; filling out forms). There are cases of clinicians finding out that pressing issues in the patient's life have been missed from previous consultations, simply because a particular topic never came up. Additionally, there are fewer opportunities to review or make conclusions about what has been discussed during the sessions, contributing to consultations that appear even less structured.

5.2 Understanding and Misunderstanding

As noted, the standardised format of the protocol questions can be a source of trouble. One problem arises from differences in the interpretation of apparently simple phrases. For example, the interpretation of "practical help" was a source of difficulty in more than one consultation. Each patient's personal circumstances are different and can involve problems that are partly or wholly outside the practical, ethical and institutional competence of the clinicians. Patients typically recognise that these limits exist but they do not have the context to be able to decide what "practical help" they can rea-

sonably seek. In addition, different clinicians place different boundaries around what they consider to be practically possible. These multiple sources of indeterminacy lead to clinicians to paraphrase and extend these key phrases (see e.g. below) or list possible examples of what they consider might constitute "practical help".

In CA terms, many of the problems created by the protocol relate to sequential appropriateness. For example, a clinician and patient are discussing the latter's physical health, specifically the pain in his leg that had been troubling him for a while. The patient mentions that he had been taking ibuprofen to help relieve the pain, but the medication has not been very effective. He also mentions that he is going to see his General Practitioner. The clinician then asks the patient what he thinks the "best case scenario" would be. The patient has already talked about the steps he is taking to improve the situation and finds it difficult to interpret the question in context. He pauses and eventually explicitly says he does not understand the question. In response, the clinician tries to reformulate "best case scenario" as something that the patient would "look forward to" or see as being "more satisfying" for his physical health. In response, the patient tuts, sits back and repeats that the pain in his leg has been "holding [him] back" from his daily activities; he wants to get rid of it. The timing of the responses, the repetition and the posture changes convey his sense of frustration at the apparently irrelevant question.

5.3 Typing, Distraction and Repetition

Points of potential miscommunication are often associated with clinicians dividing their attention between typing on the tablet and engaging with the patients' conversations. The notes are often only partially visible, in the sense that the patient can see that the clinician is typing, but often not *what* they are typing. Although the patients pay attention to the concurrent activity and usually suspend speaking, they do still sometimes make follow-up comments—but this may or may not tie into what the clinician is typing. Clinicians sometimes try to compensate, e.g. using concurrent *outlouds* of what is being written (Heath and Luff, 1992) to mitigate this. The concurrent activities place significant cognitive demands on the clinician and make it more difficult for them to track what is said. In the example immediately above and in Extracts 1 and 2 below, the clinician hears the patient but has

trouble integrating what they are saying with the current step in the protocol, one that is often not visible to the patient.

Repetition of questions can be especially problematic. In one case, a patient had complained, several times, that her medication is causing her a lot of physical symptoms—including vomiting—and that she needs to change her medication. However, the clinician is following the protocol and only noting responses relevant to the current item. When the clinician then asks about the vomiting problem and what she could do about it, the patient becomes visibly frustrated; she tosses her hands up in the air, says “I don’t know” and sighs.

The preceding examples illustrate some common sources of trouble in the consultations and also the integration of verbal and non-verbal resources used in response to them. An important feature of the way misunderstandings are addressed in these interactions is the way they make use of the shared space.

5.4 The Tablet as a Resource for Coordination

The layout of chairs and tables in the consultation rooms has a direct influence on how the participants orient to each other (Kendon, 2010). Direct face-to-face positioning is rare. The typical arrangement is an l-shape with two chairs arranged at a 45-60 degree angle and a table in-between (see Figure 2). The tablet is typically placed on the table and angled towards the clinician. The importance of its position in-between the participants and the influence of this on the management of the conversations is significant in all the intervention videos.

Changes in posture, reflecting shifts in orientation between the tablet and the other person present, help to mark important changes in participants’ focus and level of engagement (Bull, 2016). One example is the way clinicians start a new question by simultaneously displaying a shift of attention from the tablet to the patient by sitting back and turning their head toward the patient. Similarly, gestures to the tablet and gestures placed between the tablet and the other participant are used to propose or reintroduce items on the tablet screen as relevant to the ongoing discussion, e.g. as a prompt to shift from complaints to the possible solutions that need to be entered in a dialogue box (see Figure 2 and 3).

The significance of gestures is also illustrated by

examples of how misunderstandings occur when the tablet is *not* used effectively for reference coordination. At discussion stages, some clinicians review the action plans agreed on with their patients for each item, before moving on to discuss the next item. They usually wrap things up by asking patients if there is anything else they want to add, while gesturing towards that item on the tablet. In the absence of such gestures, patients can interpret the question more widely. For example, a clinician and patient were discussing action plans for the latter’s job situation when the clinician asks, while typing, if the patient thought there was anything she could do for him. The patient starts describing how she could possibly get in touch with the housing association in charge of his case; something that the patient had mentioned earlier during the consultation because he wanted to speed up his relocation (with his wife and child) to permanent housing. At this point, however, the clinician points to the item on the tablet that is specific to job situation to help reinterpret their original question—in CA terms a form of third-position repair.

In another example, the wrap-up discussion is on physical health, when the clinician asks—again while typing—if there is anything else the patient wants to tell her. The clinician glances up briefly after asking this question and notices the patient looking down at the tablet. She then adds a more explicit statement about the item she was referring to. This coincides with the patient’s request for clarification. As such, these examples illustrate how the tablet display not only has a role in managing ambiguity, but also what people are engaged with, as a means of diagnosing potential trouble sources.

5.5 Embodied Resources for Managing Misunderstanding

The complex configuration of bodies and artefacts in shared space during misunderstandings can be illustrated by considering two extracts from one consultation.

In Excerpt 1, C reorients from the screen to P and asks about P’s family situation, to which she responds with a comment about accommodation. C fails to understand the relevance of the response, partly because they have just talked about a different aspect of accommodation and he initiates a repair in line 3 by quickly pointing to the screen whilst still looking at his notes (2) on the first “its” before verbally repeating the topic. P tries again



Figure 2: C Points to Screen with Left Hand

in overlap on line 4 and when she gets the floor in line 5—and reformulates her point about accommodation preceded by a small gesture for emphasis. C nods twice towards P to acknowledge the importance of P’s reformulation which contains a significant complaint about being denied access to her possessions but then queries who “they” are, possibly thinking this could be a reference to family, which P answers in overlap (“housing”). At this point C is unable to integrate P’s turns into his understanding of what is going on. He switches to a generic clarification question at line 8. In line 9, P repeats that she was locked out and after the truncated question in line 10 C clearly orients to P. P’s gestures then become more emphatic, peaking at “things” and then reducing as C turns back to his notes.

Excerpt 1: Discussion of Relationships and Family (Key: C = Clinician, P = Patient, Underlining = emphasis, ↑ = rising intonation, :: = lengthened sound, // = onset of overlapping speech, [] = non-verbal action, (.) = short pause).

1 C: [selects item] looking at [sits back, looks at P] what makes you dissatisfied with your family situation [moves forward to pick up pen]
 2 P: its it was originally to do with my accommodation (pause)
 3 C: well no it’s to [points at screen] // it’s to do with your family yeah ↑
 4 P: // but when when it
 5 P: when they um:: [raises hands slightly] when they didn’t let me get my things and the power of attorney was lost [C

nods twice] (.)
 6 C: when your say they they wouldn’t let you get // your things
 7 P: // the housing
 8 C: [head down] explain to me what you mean
 9 P: // was locked out of my hostel
 10 C: // is this to do with your quest-
 11 P: I was locked out of my hostel [hands raised] and all my things [hands wide apart] were in there clothes [C’s head goes back to notes, P’s hands come back together] I lost everything [extra gesture then hands drop to lap]

Excerpt 2 comes from the same session approximately four minutes later. In the intervening interaction, C has focused on how P might retrieve her belongings. However, C is still having trouble understanding why P is raising the issue about personal belongings and power of attorney, given the protocol item (*Friends and Family*) currently in front of them. C tries to reintroduce this item using a flat hand gesture towards the screen on “family”. P sits forward and directly reformulates this, using a particularly direct form of other-initiated other-repair, accompanied with an emphatic hand gesture. C quickly acknowledges this with a nod and verbal acceptance, but is still showing signs of trouble. He asks a question about an address shared previously, which P answers, partly by taking hold of her bag where she has kept it, but this still doesn’t resolve the issue.

At line 7, C gestures to the screen and then produces a finger pointing up ‘hold’ gesture positioned between them 3. He then produces some filled pauses, makes a small flick of the pointing gesture and explains he needs to remind himself while looking at the screen. At this point, P sits back simultaneously with C dropping his finger point to gesture back to the screen. C’s Line 8 is formatted as reasoning out loud, but the truncated um: and long pause invites a possible response from P. In line 9, C then directly asks why P wants power of attorney. At this point the connection to *Family and Friends* finally becomes clear. P needs the missing power of attorney so she can attend to her (living)



Figure 3: C Points Up ‘Hold’ Gesture, Right Hand

grandmother’s financial affairs.

Excerpt 2: Continued Discussion of Actions Around Relationships and Family.

1 C: now this is don’t forget
 this is all to do with family
 [gesture at screen, P sits
 forward] // umm

2 P: //it’s to do with power of
 attorney [hands apart gesture]
 that I was gettin to =

3 C: = [nods] yeah and its all
 to do with power of attorney
 [hand moves to screen] umm
 because do you remember I give
 an address didn’ I [short eye
 contact]

4 P: yeah [P reaches to bag, C
 looks down touches box] I’ve got
 that here.

5 C: and did you ever do
 anything with ↑ that

6 P: no that’s what we were
 gonna discuss today [pause]

7 C: so:: [3.0s pause while
 gestures to screen than puts
 finger up in the air] cus right
 oka:y let me just remind myself
 now [C point back to screen, P
 sits back] um:: agreeing on
 actions [hand hovers over screen
 pause] um:: [pause]

8 C: cos if you had power of
attorney then you’d be able to:
 um:: [4.0s pause]

9 C: what is it that you need
 power of attorney [P sits
 forward] to achieve now =

10 P: = to sort our my
 grandmother’s affairs

6 Discussion

The examples presented above raise a number of basic points about the organisation of face-to-face interaction and the use of embodiment in shared space as a resource for communication.

The data presented highlight the tension between the use of the protocol as a standardised assessment instrument and its function as a tool to promote effective conversation. Ideally, quantitative assessments of quality of life should be consistent across different participants and different contexts. In practice, the meaning of the different assessment items and even the meaning of the numbers on the assessment scale varies across consultations. Patients and clinicians routinely engage in active, collaborative re-interpretation of the protocol in order to complete the assessment. Standard phrases such as *best case scenario* and *practical help* take on specific meanings depending on individual circumstances. What is ostensibly the same question means different things to different people and can also mean different things to the same people in different sequential contexts.

The observations show the process of detecting and dealing with differences in interpretation is fundamental to effective communication (Healey et al., 2018). Although this recurrent interpretive work means the application of the DIALOG+ protocol is not strictly standardised, it is part of normal interaction and arguably central to the therapeutic effectiveness of the intervention. The assessment items provide prompts that encourage a wider ranging conversation and greater continuity across sessions than observed in the control groups. The more strictly the protocol is applied, the more friction it would cause to the conversation (Drew et al., 2021; NHS Improvement, 2018). The work people do to bridge the gap between the protocol and the details of individual’s lives can play an important role in uncovering the different combinations of practical circumstances and constraints that influence long term outcomes.

Moving to remote delivery directly alters the configuration of resources available to participants in the interaction. One effect will be to change the visibility and control of actions in the application. As noted, input is currently led by clinicians and often not directly visible to the patient. If the ap-

plication is running in a shared window all updates will be immediately visible to both participants—and control of input by one person will not automatically restrict input by the other. This should mitigate, for example, problems with coordinating when a concurrent action (typing notes) starts and finishes—and also what item is currently under discussion. However, the work of (re)interpreting items and actions would still be required. Other possible advantages of remote interaction are the savings in cost and time, as well as the potential for improved access for some patients.

Nonetheless, some important features will be removed by remote interaction. One is the use of the embodiment of the protocol itself as a shared screen positioned in space between the participants. There are two ways in which this matters. First, changes in people’s overall physical orientation, through head movements or posture shifts signal important changes in their focus of attention and engagement. This is especially true of the clinicians, who often use posture changes to and from the device and to and from the patient to mark changes in engagement e.g. to introduce a direct question to the patient (see e.g. the first line of Excerpt 1). Secondly, clinicians (and occasionally patients) have the ability to point at a question and, for example, gesture from the question to a person. These movements, in effect, use the shared space to provide a useful spatial map for the embodied coordination of topics (Deppermann et al., 2021; Guxholli et al., 2021; Kitzinger, 2012).

It is also noticeable from the preceding examples that where a gesture is specifically placed in space is significant. The pointing gestures in Figures 2 and 3 have the same form, but their different speed of execution, orientation and placement give them a different interpretation. In an example from the same session not included in the Extracts, there is a rapid shift by the clinician from a short two finger point at the screen to a two finger point at the patient. The form and speed of the gesture is very similar but its significance is different for the participants because of where it is placed, a phenomenon also noted in other face-to-face contexts (Battersby and Healey, 2009; Özyürek, 2002). It also illustrates the ways in which people can use contrasting head and hand orientation as a means of concurrent triangulation of different reference points (e.g. people and objects) to help coordinate understanding (Battersby and Healey, 2010).

These examples also illustrate how these additional, embodied resources seem to become especially significant at the points where shared understanding is threatened (Healey et al., 2013, 2015; Özkan et al., 2021). The space between participants becomes an important extra resource for detecting and dealing with these misunderstandings.

Remote video-mediated interactions flatten the three-dimensional world of face-to-face interaction in shared space into a two-dimensional window (Mlynář et al., 2018). A shared application and a video window impede the forms of interaction highlighted above. Working out what someone is orienting to requires more effort—and although posture shifts and gestures may be visible on camera, they are attenuated and cannot take advantage of relative position in a shared space. People are able to compensate in these situations and it is an open question what the cost of adapting could be. Shared applications that enable people to see each other’s cursors can partially replicate a sense of the current focus of attention. This may help with the redirection of attention, but provides a significantly reduced set of cues.

One way in which remote interaction can give greater access is to allow asynchronous updates. For example, allowing patients to add notes or modify ratings in response to events outside the clinical context. This could give patients the opportunity to understand the protocol and application better and also feel that they are on a more equal footing with clinicians. It could also provide clinicians with potentially useful additional context updates for use in the face-to-face sessions. This shift in the distribution of control over the tool might impact on the dynamics of clinical interaction and careful design would be needed to avoid it becoming overused as a communication channel.

There are some qualifications to the findings. The observations are selective and are based on a specific population of patients with a diagnosis of psychosis. Although we think the general principles should apply in other contexts, different issues will be encountered with other communication tools and/or other client groups e.g. telephone delivery (Drew et al., 2021). Nevertheless, this study has shed light on some of the concerns and challenges related to designing health communication protocols and some specific issues for remote settings. In the wake of the pandemic, while remote consultations are unlikely to fully replace

face-to-face consultations, it will no doubt become a feature increasingly integrated into current health systems and used alongside conventional practice (Khan et al., 2021). Care needs to be taken to ensure that developments surrounding technologies like DIALOG+ are balanced with appropriate flexibility, as every nuance in communication between patients and clinicians can have a role to play in influencing the quality of therapeutic relationships and the effectiveness of clinical encounters.

7 Conclusion

Structured protocols are increasingly used in community mental healthcare consultations. Detailed analysis of the interactions using one of these protocols (DIALOG+) and its associated tablet application shows some of the advantages and pitfalls of the approach. The application provides a useful tool to support engagement in the consultations, but in practice, deviations from the protocol play an important role in the success of the consultations. The interactions are characterised by collaborative work done to (re)interpret the assessment items in the context of each client's and each clinician's practical circumstances. Participants use embodiment in shared space as an important, flexible interactional resource in doing this. With remote consultations increasingly integrated into healthcare settings, our findings provide a starting point for thinking about how software like the DIALOG+ application can be redesigned for these environments.

Acknowledgments

We are grateful to the National Institute for Health Research (NIHR) for funding Healey and Galindo-Esparza's contribution to this paper through the project "Remote delivery of an app-based intervention (DIALOG+) in community mental health care-development" (Reference: NIHR201680) and to the Wellcome Trust for funding Law's contribution through the programme "Health Data in Practice: Human-centred Science" (Reference: 218584/Z/19/Z).

References

Stuart Battersby and Patrick G T Healey. 2010. Head and hand movements in the orchestration of dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Stuart A Battersby and Patrick G T Healey. 2009. The interactional geometry of a three-way conversation.

In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Janet B Bavelas. 2007. Face-to-face dialogue as a micro-social context. S. Duncan, E. Levy, & Cassell (Eds.), *Language in mind, body, and context*, pages 127–146.

Janet Beavin Bavelas, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and social psychology bulletin*, 21(4):394–405.

Peter E Bull. 2016. *Posture & gesture*, volume 16. Elsevier.

Nicole Chovil. 1991a. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, 25(1-4):163–194.

Nicole Chovil. 1991b. Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15(3):141–154.

Arnulf Deppermann, Lorenza Mondada, and Simona Pekarek Doehler. 2021. Early responses: An introduction.

Paul Drew, Annie Irvine, Michael Barkham, Cintia Faija, Judith Gellatly, Kerry Ardern, JC Armitage, Helen Brooks, Kelly Rushton, Charlotte Welsh, et al. 2021. Telephone delivery of psychological interventions: Balancing protocol with patient-centred care. *Social Science & Medicine*, 277:113818.

Stuart Ekberg. 2021. Proffering connections: Psychologising experience in psychotherapy and everyday life. *Frontiers in Psychology*, page 3686.

Paul Ekman. 1979. About brows. *Emotional and conversational signals in: von Cranach, M./Foppa, K. ua (Hrsg.)(1979): Human Ethology, Cambridge*.

Simeon Floyd, Elizabeth Manrique, Giovanni Rossi, and Francisco Torreira. 2016. Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes*, 53(3):175–204.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, 97:101–121.

Aurora Guxholli, Liisa Voutilainen, and Anssi Peräkylä. 2021. Safeguarding the therapeutic alliance: Managing disaffiliation in the course of work with psychotherapeutic projects. *Frontiers in Psychology*, page 3905.

Patrick G T Healey, Jan P De Ruiter, and Gregory J Mills. 2018. Editors' introduction: Miscommunication. *Topics in Cognitive Science*, 10(2):264–278.

Patrick G T Healey, Mary Lavelle, Christine Howes, Stuart Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker's troubles. In *Proceedings of the annual meeting of the cognitive science society*, volume 35.

- Patrick G T Healey, Nicola Jane Plant, Christine Howes, and Mary Lavelle. 2015. When words fail: Collaborative gestures during clarification dialogues. In *2015 AAI Spring Symposium Series*.
- Christian Heath and Paul Luff. 1992. Collaboration and control/crisis management and multimedia technology in london underground line control rooms. *Computer Supported Cooperative Work (CSCW)*, 1(1):69–94.
- Timo Kaukoma, Anssi Peräkylä, and Johanna Ruusuvaara. 2014. Foreshadowing a problem: Turn-opening frowns in conversation. *Journal of Pragmatics*, 71:132–147.
- Kathrin Kaulard, Douglas W Cunningham, Heinrich H Bülthoff, and Christian Wallraven. 2012. The mpi facial expression database—a validated database of emotional and conversational facial expressions. *PLoS one*, 7(3):e32321.
- Adam Kendon. 2010. Spacing and orientation in copresent interaction. In *Development of multimodal interfaces: Active listening and synchrony*, pages 1–15. Springer.
- Kobin H Kendrick. 2015. Other-initiated repair in english. *Open Linguistics*, 1(1).
- Kobin H Kendrick and Judith Holler. 2017. Gaze direction signals response preference in conversation. *Research on Language and Social Interaction*, 50(1):12–32.
- Abdul Waheed Khan, Nisha Kader, Samer Hammoudeh, and Majid Alabdulla. 2021. Combating covid-19 pandemic with technology: perceptions of mental health professionals towards telepsychiatry. *Asian Journal of Psychiatry*, 61:102677.
- Celia Kitzinger. 2012. Repair. *The handbook of conversation analysis*, pages 229–256.
- Elisa Liberati, Natalie Richards, Jennie Parker, Janet Willars, David Scott, Nicola Boydell, Vanessa Pinfold, Graham Martin, Mary Dixon-Woods, and Peter Jones. 2021. Remote care for mental health: qualitative study with service users, carers and staff during the covid-19 pandemic. *BMJ open*, 11(4):e049210.
- Neda Mahmoodi, G Jones, Tom Muskett, and Sally Sargeant. 2020. Exploring shared decision making in breast cancer care: A case-based conversation analytic approach. *Communication and Medicine*, 16(1).
- Rose McCabe and Patrick G.T. Healey. 2018. Miscommunication in doctor–patient communication. *Topics in Cognitive Science*, 10(2):409–424.
- Rosemarie McCabe, Patrick GT Healey, Stefan Priebe, Mary Lavelle, David Dodwell, Richard Laugharne, Amelia Snell, and Stephen Bremner. 2013. Shared understanding in psychiatrist–patient communication: Association with treatment adherence in schizophrenia. *Patient education and counseling*, 93(1):73–79.
- Chris McVittie, Slavka Craig, and Margaret Temple. 2020. A conversation analysis of communicative changes in a time-limited psychotherapy group for mothers with post-natal depression. *Psychotherapy Research*, 30(8):1048–1060.
- Jakub Mlynář, Esther González-Martínez, and Denis Lalanne. 2018. Situated organization of video-mediated interaction: A review of ethnomethodological and conversation analytic studies. *Interacting with Computers*, 30(2):73–84.
- NHS Improvement. 2018. Spoken communication and patient safety in the NHS. Technical report, NHS.
- C Olwill, D Mc Nally, and L Douglas. 2021. Psychiatrist experience of remote consultations by telephone in an outpatient psychiatric department during the covid-19 pandemic. *Irish journal of psychological medicine*, 38(2):132–139.
- Elif Ecem Özkan, Tom Gurion, Julian Hough, Patrick GT Healey, and Lorenzo Jamone. 2021. Specific hand motion patterns correlate to miscommunications during dyadic conversations. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–6. IEEE.
- Asli Özyürek. 2002. Do speakers design their cospeech gestures for their addressees? the effects of addressee location on representational gestures. *Journal of Memory and Language*, 46(4):688–704.
- Stefan Priebe, Eoin Golden, David Kingdon, Serif Omer, Sophie Walsh, Kleomenis Katevas, Paul McCrone, Sandra Eldridge, and Rose McCabe. 2017. Effective patient–clinician interaction to improve treatment outcomes for patients with psychosis: a mixed-methods design. *Programme Grants for Applied Research*, 5(6).
- Stefan Priebe, Lauren Kelley, Serif Omer, Eoin Golden, Sophie Walsh, Husnara Khanom, David Kingdon, Clare Rutterford, Paul McCrone, and Rosemarie McCabe. 2015. The effectiveness of a patient-centred assessment with a solution-focused approach (dialog+) for patients with psychosis: a pragmatic cluster-randomised controlled trial in community care. *Psychotherapy and psychosomatics*, 84(5):304–313.
- Harvey Sacks. 1992. Lectures on conversation: Volume i. *Malden, Massachusetts: Blackwell*.
- Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.
- Albert E Scheflen. 1973. *Communicational structure: Analysis of a psychotherapy transaction*. Indiana U. Press.
- Mi-Suk Seo and Irene Koshik. 2010. A conversation analytic study of gestures that engender repair in esl conversational tutoring. *Journal of pragmatics*, 42(8):2219–2239.

David Silverman. 1998. *Harvey Sacks: Social science and conversation analysis*. Oxford University Press on Demand.

Yijin Wu. 2020. Pain talk in hospice care: a conversation analysis. *BMC Palliative Care*, 19(1):1–8.

Norm Participation Grounds Language

David Schlangen

CoLabPotsdam / Computational Linguistics
Department of Linguistics, University of Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

The striking recent advances in eliciting seemingly meaningful language behaviour from language-only machine learning models have only made more apparent, through the surfacing of clear limitations, the need to go beyond the language-only mode and to *ground* these models “in the world”. Proposals for doing so vary in the details, but what unites them is that the solution is sought in the addition of non-linguistic data types such as images or video streams, while largely keeping the mode of learning constant. I propose a different, and more wide-ranging conception of how grounding should be understood: What grounds language is its normative nature. There are standards for doing things right, these standards are public and authoritative, while at the same time acceptance of authority can and must be disputed and negotiated. What grounds language, then, is the determined use that language users make of it, and what it is grounded in is the community of language users. I sketch this idea, and draw some conclusions for work on computational modelling of meaningful language use.

1 Introduction

You fire up the latest language model—a machine learning model induced by guessing words in a very large body of text—and “ask” it: “How would you describe the colour green?”. After a short pause, the following text appears: “*The colour green is refreshing and vibrant. It is often associated with nature and growth.*”. You: “How would you describe the colour red?” – “*The colour red is exciting and full of energy. It is often associated with passion and power.*”. Not bad. You hold two objects up in front of the computer screen: “Which of these is green?”. The reply suddenly is less than satisfying (and also old news in this conversation): “*The colour green is refreshing and vibrant. It is often*

associated with nature and growth.”¹

This of course is a profoundly unfair test. The model has no connection to you other than through what you type, and so can’t observe what “these” refers to, and moreover, it *never* had access to anything other than language data. The question then arises what the consequences of this limitation are. Is it just that those models (unsurprisingly) can’t tell one visually presented object from another, or is there something fundamentally off about their grasp of language?

In an influential recent paper, [Bender and Koller \(2020\)](#) argue that indeed something is off, in that a model that only has access to form cannot learn to connect that form to meaning, which they tie to communicative intention, which in turn can be *about* the world. They also suggest that perhaps giving access to more than linguistic form might make it possible to learn this connection. In this their argument meets with more mainstream views that don’t deny meaning status to language-only (or “internet world scope”) models, but see them as deficient until augmented with additional forms of data ([Bisk et al., 2020](#)).

Here, I will argue that just connecting language with non-language data still misses fundamental properties of language use, along two dimensions. First, the connection between world states and language is only one among several types of connections that must be gotten right (the others being intra-language connections, and connections between language and actions). Secondly, the connections themselves need to be understood as *normative* ones, which again has two kinds of consequences: They effect *commitments* and *entitlements*, but they also are non-necessary and their applicability must be argued for (and can be argued against). Just getting things right occasionally, or even very often, is not enough. The getting it right

¹Output by the GPT-3 model (text-davinci-002) by openAI ([Brown et al., 2020](#)), retrieved on 2022-05-18.

must come from an orientation towards the relevant standards of getting things right. This “orientation towards” shows in the ability to appeal to these standards when challenged, either directly or in the repair of understanding problems, and it forms the difference between what I will call “*norm conformance*” (which is what AI models are trained for), and “*norm participation*” (which is what what grounds language use). In short, there are interactive capabilities that need to underwrite meaningful language use, rendering agents that do not have them deficient language users. As I will argue, this has consequences for the use of NLP systems that can falsely appear as having these capabilities, and it also opens up interesting research directions.

Let us start by looking at simple observational statements, and use this to draw a schematic picture of meaning making in language use.

2 “There is a tiger.”

“There is a tiger”, you say, facing your friend but looking past them at a location behind their back. Leaving aside what these news should do to your friend, let’s think a bit about what I, an overhearer, am justified to think may have been done to you to make you say that, and how that allows me to assign meaning to what you said.

So, why did you say that? There are many possible types of answers to this question, a central class among which (namely those that not also insinuate malicious intent to deceive on your side) will mention in some form the *state of affairs* of there being a tiger, and your stating this as a fact. That is, there is an assumed connection between your expression *e* or more generally your action *a* (of uttering *e*) and a state of affairs *c*. You said “there is a tiger” in part because there is a tiger.

But let’s say you were wrong, and it was just Tibby, my oversized tabby cat which can look, at least for a split second and when she is very hungry, like a (still very small though) tiger. We can’t say anymore that you said “there is a tiger” because there was a tiger, as there was none. But we can fix the description by giving you *a*—potentially misguided—inner life: you said “there is a tiger”, because you believed there to be a tiger, and you believed there to be a tiger, because you misperceived Tibby as one. The chain now goes from *c*, the state of affairs (which in the modified example does not hold, i.e. turns out not to be a fact), to *b*, the belief, to *a*, the action.

This chain can also be used to reconstruct understanding.² How can I get from observing *a* to forming my own belief about *c*? To reverse this chain, I need to see *a* as representative of a *type* of action *A*, and I need to know something about this type’s connection to a *type* of belief *C'*, and its connection to a *type* of states of affairs *C*—and I need to assume that you expected your addressees to know this and to be able to use this knowledge to reason back to the best explanation.³

Let us write out these connections in the form “if *C*, then ___ *A*”, where the underscore shall work as a placeholder for a predicate describing the force of the connection, to be explored presently. If we take the step of seeing the forming of beliefs as an action as well, then this schema covers both parts of the chain described above: “if you are looking at a tiger (and your eyes are open, and you are sighted, etc.), then you ___ form the belief that there is a tiger”, and “if you hold the belief that there is a tiger and you want to inform me of it, then you ___ say (something to the effect of) ‘there is a tiger’ ”.

Looked at from a different perspective, your saying “there is a tiger” has *committed* you to believing that there is a tiger (insofar as that this is the best explanation for why you said that), and to having a good justification for that belief, where the best justification would be there indeed being a tiger (and you having the right kind of epistemic standing). Having this belief further commits you to having other beliefs as well, such as “there is a mammal” or “there is a four-legged animal”, “there is a cat-like creature”, “there is a living entity”, etc.; these are just consequences (material inferences, to be precise) that we can see as being contained in having this belief, or, in other words, as contributing to individuating this belief as the one that it is.

To collect what we have before we move on: This analysis assumes that there are connections between ways the world is and beliefs about it, between beliefs and other beliefs, and between beliefs (and other mental states, such as intentions) and

²Note that the following does not describe a process model. It may very well be that in actual interpretation, shortcuts can be applied that identify the verbal action as part of a larger action type. What matters for the rational reconstruction here is that the constructs described here (beliefs, intentions) are available in reasons you can give for your actions, after the fact.


³The knowledge has to be about types, since *a*, the actual physical event, has happened only now and never before and will never happen again, I cannot previously have known anything about it, other than what I know or learn about the type of which it is a token.

actions. (We leave aside here whether in an ultimate analysis, these beliefs could not be explained away as dispositions to act in a proscribed way.) These connections can figure both in explanations of why you do something (you do *A*, because you are in *C*) and in abductions about states (as you did *A*, the state most likely is *C*). What is open is the exact nature of these connections, which is what we will turn to next.⁴

3 Norm Conformance and Norm Participation

Our task now is to further specify the “if *C*, then ___ *A*” schemas. We want to achieve that they can figure in reasoning about why a speaker said what they said, and, equally importantly, can be offered by the speakers themselves as reasons for why they said what they said. As we will see, these are related, but separate aims: Things can happen for a reason in different ways.

To make the discussion more concrete, let us instantiate the *C* and *A* in this schema, as follows:

- (1) “if *presented with visual features of this kind*  [picture of tiger], then you ___ say ‘there is a tiger’”


Could this conditional feature in reasoning about the behaviour of a human speaker? We would probably hesitate to allow such a description, wanting to qualify the antecedent with something like “and you want to inform your interlocutor about what you see, using the English language”, for otherwise there are many ways in which you could react to the stimulus. Also, there is still the question of how to fill the placeholder, and it seems that it should indeed be filled somehow, as a conditional of the form “if *C*, then you say *E*” seems too strong as description of human linguistic behaviour; even wanting to do something does not unconditionally lead to doing it.

Before we come to that, however, we can observe that something like (1), without any qualification about wanting to inform, is not a bad description of what the training set and learning objective of an image captioning model (e.g., Mitchell et al.

(2012); Vinyals et al. (2015)) realises: To the extent that the model works (as measured by accuracy, or some other metric that measures agreement with a reference), it *conforms* to the *norm* described by (1). To the extent that this type of description fails to characterise the human language use situation, these models remain ungrounded.

What (1) misses, however, is that these regularities, these norms, can feature in self-explanations, and exert a stronger force on language users, which, I propose, is better expressed by making it an element of the norm: one *ought to* behave in this way, given that the conditions are met; and, in reverse, one is *committed to* them being met, if one behaved in this way. This opens up two possible points of contention in the application of such a norm: First, do the conditions indeed hold, that is, can it be applied? Second, is it even a norm, the authority of which I should accept? (“Says who?” as a possible reply.) These are issues that can be, and not rarely are, raised in interaction (not in the artificial situations created by the language use of function-type models such as the aforementioned caption models). To distinguish this kind of actively following norms from just picking up regularities, I will use the label *norm participation* for it.

Before we turn to the ways that this participation process plays out in interactive language use, let us unpack this proposal a bit more. Filling the placeholder and bringing in the intermediate belief state turns (1) into the following:

- (2) a. “if *presented with visual features of this kind*  [picture of tiger], then you ought to *believe that there is a tiger*”
 b. “if you believe that there is a tiger, you ought to *believe that there is a four-legged animal, and that there is a mammal, and that there is a living thing, and ...*”
 c. “if you believe that there is a tiger, and you want to inform your interlocutor about this, using English, you ought to say ‘there is a tiger’”

To anticipate the discussion in the next section, the idea behind stressing the normative force of the connection is to explain why there is a pressure to correct disagreements, even if communicative success may have already been reached. In a very real sense, if you don’t seem to be following these

⁴ What I’ve tried to convey in this short section is my take on some Sellarsian themes (Sellars, 1954, 1969; DeVries, 2005), especially with the three main moves of language-entry, intra-language movement, and language-exit (Sellars, 1954), and a conceptual-role semantics for propositional attitudes (Harman, 1987). This will need to be expanded on in more detail elsewhere.

norms, then to me it seems that there is something wrong with you, at least as a participant in my *system* of norms; or, potentially, there is something wrong with my system.

Let us start with a simple disagreement. Imagine you had pronounced *tiger* as in German (/ˈtiːgə/); I can recognise which norm you were aiming for, but can point out to you that the correct form contains /ˈtaɪgə/, which one ought to use. Or let us assume that I think that what is present is a leopard rather than a tiger; this allows me to spot that there is a deficiency in your belief/belief norm (and your perceptual one), at least compared to how I have it, which I can address by saying something like “they look similar, but have a different coat: tigers have stripes, while leopards have spots”. (As we will discuss presently, I cannot force you to take this on; I can just try to make my claim of authority plausible to you.) Note that this limits the possible misunderstandings— “this is not a tiger, it’s a gazelle” is already odd; “this is not a tiger, it’s a refrigerator” is far too odd, as the belief revision it indicates is too extreme to be plausible.

Lastly, an analysis of this form could also go some ways towards explaining why word uses can be so contentious, even if communicative success is not at issue: Each use makes the implicit claim that this is how one ought to talk, and that it makes the right kinds of distinctions, a claim that addressees may want to disagree with. An analysis of slurs and linguistic interventions (McConnell-Ginet, 2020; Cappelen and Dever, 2019) along these lines might be possible, but is left for future work here.

Again, let us take stock before we move on. I have argued that the right way to connect antecedent and consequent in constructs like (1) is to make direct appeal to their normative status: it is not just that if *C* is the case, one normally or conventionally does *A*, rather one *ought to* do this, and does something wrong or at least something inviting correction when one does not do it. Doing things of these type then commits one in certain ways, and makes one suffer the consequences if these ways turn out to be not warranted. The analysis further has brought out a distinction between (mere) *norm conformance*, which is acting in accordance with a set of norms (for example, as they were realised in a data set of labelled examples) and *norm participation*, which involves treating the norms as possible reasons for acting, which can be offered, requested, and challenged. The interactive

processes in which this is done and which justify the label “participation” will be our topic next.

4 Norm Participation as Interactive Process and Achievement

The idea of the approach sketched here is that the question of which norms hold and how they are to be applied is never fully settled, and can become the overt topic of a conversation. That is, the fact that the connection is via an appeal to what one ought to do has practical consequences, which I will briefly trace in three related domains: language acquisition, conversational grounding, and conceptual disputes. More specifically, it shows in what in the field of conversational analysis is called *repair*, and is rightly assigned a central place in the study of conversation (Schegloff et al., 1977; Hayashi et al., 2013; Jefferson, 2018).⁵

First Language Acquisition Children start out without knowledge of the norms of the language community in which they were born. Hence, they need to rely on the competent speakers around them to initiate them into these norms. The way they do this is by making attempts and observing reactions, which quite frequently involve *repair*. For example, Golinkoff (1986) found that about 50% of attempts by small infants (in their first verbal phase, from 1 to 1.5 years old) resulted in repair. In the light of the schema proposed here, we can understand this as attempts at using a norm, being recognised as such, and then getting demonstrated how the act ought to be performed. As the examples collected by Clark (2020) show, this process can target both the form (that is, schemata of the type of (2-c)) as well as conceptual ones (as in (2-a) and (2-b)); indeed, these levels might often be addressed simultaneously. We can take away from this short review that an orientation towards shared norms seems to play a role already in the acquisition of language abilities.

Conversational Grounding According to H. Clark’s (1996) well-known proposal, it is a constant task in conversation to ensure that common ground is reached, sufficiently for the purposes at hand. We can recognise the stages of

⁵A recent cross-linguistic study by Dingemans et al. (2015) found repair attempts on average about once per 1.4 minutes; studies of task-oriented dialogue found between 4 and 5.8% of turns in the respective corpora to contain clarification requests (Purver et al., 2001; Rodríguez and Schlagen, 2004).

this process (presenting & identifying behaviour and signal; signalling & recognising propositional state; proposing & considering joint project) in the schema in (2). More directly related are the consequences of successful conversational grounding, as discussed by Brennan and Clark (1996) under the label “conceptual pacts”. In the analysis proposed here, these can be understood as “local” norms that are not yet generalised, that is, ways the participants in an interaction mutually have come to think they ought to act with each other.⁶

Meaning Disputes The status of these norms as reasons for acting shows most clearly in those rarer cases where they need to be overtly discussed. Very occasionally, this can even become positive law: In (*Nix v. Hedden*, 149 U.S. 304, 1893), the US Supreme Court judged that for the purposes of taxation, tomatoes are vegetables, despite biologically better fitting under the label fruit. In our framework, this can be understood as an adjudication between a norm that better fits to one type of belief/belief system (tomatoes as fruit, for biological reasons) vs. one that better accords to actual usage (tomatoes as vegetables, for similarity in properties to other vegetables). Further examples are discussed by Ludlow (2014), and more recently, under the label *word meaning negotiation*, by Larsson and Myrendal (2017) and Myrendal (2019), who also provide the beginnings of a formalisation of the dialogue moves that structure this process.

We can take from this very brief review that what is called norm participation here makes up a substantial amount of overt conversational moves, and is something that participants in verbal interactions actively engage in.

5 Some Conclusions for Computational Modelling of Language Use

I contrasted above *norm conformance* from *norm participation*, claiming that current natural language processing systems are only capable of the former, being optimised for *accuracy* and not for systematic engagement in the processes reviewed in the previous section. A possible objection now is to reject that there is a problem—if accuracy can be raised sufficiently high, there would be no need for

⁶A computational model of how such local conventions can reach whole populations has recently been offered by Hawkins et al. (2021).

repair, and norm conformance would be indistinguishable from norm participation. This however presupposes that there is only one set of correct norms, and that this can in principle be found in the source datasets against which accuracy is measured. This is, however, is unlikely to be the case, once one moves outside of the very few domains with authoritative taxonomies (like an outsider may imagine Biology to work; Dupré (2021))—imagine a category like “weed / pest plant”. The “myth of the gold label” is increasingly being noticed as a problem in NLP as well (Basile et al., 2021; Pavlick and Kwiatkowski, 2019).

If the story sketched above is on the right track, it provides a way to understand some ethical issues in the use of NLP systems.⁷ Consumers of computer speech acts will assume that, just like with human speakers, something like the chain in (2) is in place in a captioning system for example, even if in reality there is a more direct and simpler link between visual input and language. A disagreement with a labelling decision or apparent category will need to find an addressee, which the system cannot provide. Organisations deploying such systems will need to take the responsibility for the “commitments” made by the system, as the system cannot do so – as it cannot “suffer the consequences”. Secondly, in the framework sketched above, as discussed, every use of language implicitly contains the claim “this is how one does this”; again, on the principle that the system provider will need to pick up “commitments” made by the system, this is something that seems to argue against the deployment of language generation systems that are wont to reproduce undesirable material (Bender et al., 2021).

As a final example along these lines, consider the application of question answering. In the discussion above, I briefly mentioned the condition of needing to possess the right kind of epistemic standing to form beliefs (discussed in more detail by Goldberg (2015)). This epistemic standing can be “inherited” in knowledge through testimony (Gelfert, 2014). Current search engines indirectly honour these mechanisms, by framing their job only as surfacing source material that provides its own reputational claims towards such epistemic standing. Recent attempts at treating large language models as knowledge bases for question answering (surveyed by AlKhamissi et al. (2022)),

⁷ These will be expanded in a separate paper, which will need to more thoroughly connect to the ongoing discussion in the nascent field of “responsible AI”.

however, break these links without providing others, which renders the status of their replies problematic (a similar point is made by [Shah and Bender \(2022\)](#) and [Potthast et al. \(2020\)](#)).

With these caveats in mind, some potentially productive lines of work can also be motivated from within the framework explored here. A language generating system that is able to maintain a coherent system of norms as described here, can use them to offer self-explanations, and can react to corrections, would go some way towards more grounded, and hence more meaningful, language use. Components of this are already being explored separately. [Zhou et al. \(2022\)](#) show that it is possible to explicate implicit commonsense knowledge from large language models (corresponding to the middle step (2-b)); [Kassner et al. \(2021\)](#) show that a neuro-symbolic system can keep track of corrections to “beliefs” extracted from such models. It seems that combining these approaches in an interactive fashion, adding moves such as discussed by [Larsson and Myrendal \(2017\)](#), would at least go some ways towards systems with more understandable meaning norms.

6 Related Work

The inspiration from the work of Sellars for the ideas explored here has already been mentioned. Beyond the work cited above, the role that *giving and asking for reasons* plays has been noted by [Sellars \(1956\)](#) and expanded upon by [Brandom \(1998\)](#).⁸ The varieties of rule following of course are an important topos from [Wittgenstein \(1984 \[1953\]\)](#) (see [Baker and Hacker \(2009\)](#); [Kripke \(1982\)](#)), as is the necessarily public nature of judgments on the applicability of norms (on this see also [Hegel \(1807\)](#)). The notion of “orienting towards” is central in the field of Conversation Analysis.⁹

On the computational side, [Schlangen \(2016\)](#) makes some related points, although not yet under the normative framework explored here. [DeVault et al. \(2006\)](#) made a similar point, and much

⁸ “[I]n characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says” ([Sellars, 1956, §36](#))

⁹ “CA’s guiding principle is that interaction exhibits ‘order at all points’ [...] This orderliness is normative—it is produced and maintained by the participants themselves in their orientations to social rules or expectations” ([Hoey and Kendrick, 2017, p.2](#))

more carefully (but also more restricted in scope). The forming of conceptual pacts is investigated with modern computational means by [Takmaz et al. \(2022\)](#). Work that could be enlisted for going towards norm participating has already been cited in the previous section.

7 Conclusions

In this paper I have sketched a view of language as the purposeful use of norms for acting (where acting includes the forming of beliefs), where these norms can serve as reasons, can be negotiated, challenged, modified, and locally formed. I have speculated about the consequences of such a view on computational modelling of language use.

No one could mistake this offering here for more than a sketch. To develop this into a fuller proposal, an enormous amount of work remains to be done. How exactly language lends itself to figure in such norms, and how these are composed (note that all examples used full sentences) is an open question (and compositionality is notoriously a problem for conceptual role semantics ([Whiting, 2022](#))), to mention just one technical challenge.

Nevertheless, what I hope to have offered is a potentially productive way to think about how language is grounded, not just in some link to perceptual information, but in the collective uses made of it, which are actively constructed and maintained to be collectively useful. It is my hope that this more interactive perspective on symbol grounding can be informative for computational work on simulating language use.

Acknowledgements Many thanks to the anonymous reviewers for their very detailed and helpful comments. I would have liked to address them in more detail, but for reasons of time and space will need to do so elsewhere and some other time. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 423217434 (RECOLAGE).

References

- [Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.](#)
- [G. P. Baker and P. M. S. Hacker. 2009. Wittgenstein: Rules, Grammar and Necessity: Volume 2 of an Analytical Commentary on the Philosophical Investigations, Essays and Exegesis 185-242. Wiley-Blackwell.](#)

- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. [Toward a perspectivist turn in ground truthing for predictive computing](#). *CoRR*, abs/2109.04270.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2, pages 5185–5198.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. [Experience grounds language](#). *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 8718–8735.
- Robert Brandom. 1998. *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Harvard, MA, USA.
- Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *CoRR*, abs/2005.14165.
- Herman Cappelen and Josh Dever. 2019. *Bad Language*. Oxford University Press.
- Eve V Clark. 2020. [Conversational Repair and the Acquisition of Language](#). *Discourse Processes*, 57(5-6):441–459.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.
- David DeVault, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, USA.
- Willem A. DeVries. 2005. *Wilfrid Sellars*. McGill-Queen’s University Press.
- Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. [Universal Principles in the Repair of Communication Problems](#). *Plos One*, 10(9):e0136100.
- John Dupré. 2021. *The Metaphysics of Biology*. Cambridge University Press.
- Axel Gelfert. 2014. *A Critical Introduction to Testimony*. Bloomsbury Academic.
- Sanford Goldberg. 2015. *Assertion: On the Philosophical Significance of Assertoric Speech*. Oxford University Press.
- Roberta Michnick Golinkoff. 1986. [‘I beg your pardon?’: The preverbal negotiation of failed messages](#). *Journal of Child Language*, 13(3):455–476.
- Gilbert Harman. 1987. (nonsolipsistic) conceptual role semantics. In Ernest LePore, editor, *New Directions in Semantics*, pages 55–81. London: Academic Press.
- Robert X. D. Hawkins, Michael Franke, Michael C. Frank, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2021. [From partners to populations: A hierarchical bayesian account of coordination and convention](#). *CoRR*, abs/2104.05857.
- Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell. 2013. *Conversational Repair and Human Understanding*. OAPEN Library. Cambridge University Press.
- Georg Wilhelm Friedrich Hegel. 1807. *Phänomenologie des Geistes*. Philosophische Bibliothek. Meiner, Hamburg. This Edition 1952.
- Elliott M. Hoey and Kobin H. Kendrick. 2017. [Conversation analysis](#). In *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*, Guides to Research Methods in Language and Linguistics. Wiley.
- Gail Jefferson. 2018. *Repairing the Broken Surface of Talk: Managing Problems in Speaking, Hearing, and Understanding in Conversation*. Foundations of Human Interaction. Oxford University Press, Oxford.
- Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. [BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Saul Kripke. 1982. *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Harvard University Press.
- Staffan Larsson and Jenny Myrendal. 2017. Dialogue Acts and Updates for Semantic Coordination. In *semDial 2017*, pages 59–66, Saarbrücken, Germany.
- Peter Ludlow. 2014. *Living Words*. Oxford University Press, Oxford, UK.
- Sally McConnell-Ginet. 2020. *Words Matter: Meaning and Power*. Cambridge University Press.
- Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. [Midge: Generating image descriptions from computer vision detections](#). In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France. Association for Computational Linguistics.
- Jenny Myrendal. 2019. [Negotiating meanings online: Disagreements about word meaning in discussion forum communication](#). *Discourse Studies*, pages 1–23.
- Nix v. Hedden, 149 U.S. 304. 1893.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Martin Potthast, Matthias Hagen, and Benno Stein. 2020. [The dilemma of the direct answer](#). *ACM SIGIR Forum*, 54(1):1–12.
- Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the means for clarification in dialogue. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.
- Kepa Joseba Rodríguez and David Schlangen. 2004. [Form, intonation and function of clarification requests in german task-oriented spoken dialogues](#). In *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*, pages 101–108, Barcelona, Spain.
- Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organisation of repair in conversation. *Language*, 53(2):361–382.
- David Schlangen. 2016. Grounding, Justification, Adaptation: Towards Machines That Mean What They Say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.
- Wilfrid Sellars. 1969. [Language as thought and as communication](#). *Philosophy and Phenomenological Research*, 29(4):506–527.
- Wilfrid Sellars. 1954. Some reflections on language games. *Philosophy of Science*, 21:204–228.
- Winfrid Sellars. 1956. *Empiricism and the Philosophy of Mind*. Harvard University Press, Cambridge, Mass., USA.
- Chirag Shah and Emily M Bender. 2022. Situating Search. In *CHIIR 2022*, pages 221–232.
- Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. [Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.
- Daniel Whiting. 2022. [Conceptual role semantics](#). In *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002. IEP. Retrieved 2022-05-30.
- Ludwig Wittgenstein. 1984 [1953]. *Tractatus Logicus Philosophicus und Philosophische Untersuchungen*, volume 1 of *Werkausgabe*. Suhrkamp, Frankfurt am Main.
- Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. [Think before you speak: Explicitly generating implicit common-sense knowledge for response generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

Where am I and where should I go? Grounding positional and directional labels in a disoriented human balancing task

Sheikh Mannan

Department of Computer Science
Colorado State University
Fort Collins, CO USA
sheikh.mannan@colostate.edu

Nikhil Krishnaswamy

Department of Computer Science
Colorado State University
Fort Collins, CO USA
nkrishna@colostate.edu

Abstract

In this paper, we present an approach toward grounding linguistic positional and directional labels directly to human motions in a disoriented balancing task in a multi-axis rotational device. We use deep neural models to predict human subjects’ joystick motions and proficiency in the task. We combine these with BERT embeddings for annotated positional and directional labels into an *embodied direction classifier*. Combining contextualized BERT embeddings with embeddings representing human motion and proficiency can successfully predict the direction a hypothetical human participant should move to achieve better balance. Our accuracy is comparable to a moderately-proficient human subject, and we find that our combined embodied model may actually make objectively better decisions than some humans.

1 Introduction

Much of the recent success in AI can be attributed to the meteoric rise of large language models (LLMs), such as BERT (Devlin et al., 2019) and the GPT family (Radford et al., 2019). These language models facilitate coherent, grammatical text generation using high-dimensional representations of words, sentences, and more, that preserve similarity relations across dimensions. Although pretrained on a enormous amount of text, there are many ways in which they fail to demonstrate “understanding” as commonly defined. As argued by, e.g., Bender and Koller (2020), these models lack knowledge of the current situational context, because that context comes from non-textual modalities. Certain multimodal language models, e.g., multimodal BART-large (Lewis et al., 2020) appear to perform better according to certain benchmarks (Moon et al., 2020; Kottur et al., 2021), but there remain many important domains which for the moment appear to be out of reach for state of the art AI.

Consider the problem of human spatial disorientation. During extreme conditions, such as piloting a spacecraft, even expert humans are subject to gravitational transitions where they may not be able to rely on gravitational cues sensed by the vestibular system, leading to fatal accidents (Shelhamer, 2015; Cowings et al., 2018). Even on Earth, the leading cause of fatal aircraft accidents in military pilots is spatial disorientation (Gibb et al., 2011).

Numerical AI models, however, with direct access to quantitative information about position and movement, can potentially determine when a human appears to be losing control and intervene, such as by telling the human what to do in order to right themselves. A successful AI partner that counteracts human disorientation to enhance task performance in real time would need to predict the intent of the human’s motions, make decisions with incomplete information or under environmental uncertainty (Weber, 1987; Talamadupula et al., 2010) and, perhaps most importantly, foster trust in the human (Hengstler et al., 2016).

These are not requirements that even the impressive benchmark performance of modern LLMs can meet. Successful guidance of a human through language requires that the AI “embody” relations between linguistic terms and the human’s situation.

In this paper we combine disambiguated and contextualized linguistic embeddings (Wiedemann et al., 2019) from BERT, with embeddings extracted from numerical AI models that are trained to predict control movements and human performance in a spaceflight-analog disoriented balancing task. Unlike the BERT embeddings, these latter embeddings are “situated,” in that they come from models that are trained to *embody* a human participant’s position in a phase space parameterized by angular position and velocity in the balancing task. This combined model is trained to predict the direction the human should move towards for

better balance given BERT embeddings that represent “thought vectors” about position relative to the balance point, and performance and motion control features extracted from the numerical models. We show that predictions made by our model “agree” on average with those made by a human with a moderate level of proficiency in the balancing task, and a deeper dive into misclassifications suggest that the model may actually be performing better in this task than the raw numerical results indicate.

2 Related Work

This paper brings together research in two distinct and to date largely disjunct areas: multimodal language grounding through human-AI collaboration, and mitigating the effects of spatial disorientation. This section discusses relevant work in these two domains and our goals in synthesizing them.

The Collaborative Research Center’s Situated Artificial Communicator project was a significant early attempt to model the integration of language and sensorimotor skills in a situated context (Rickheit and Wachsmuth, 2006). Recent work in multimodal conversational modeling has continued similar lines of research with multimodal Transformer architectures (Chen et al., 2020; Hu et al., 2020). Other relatively recent work attempts to integrate neurally-encoded robotic arm control with guidance and instruction through dialogue (She et al., 2014; She and Chai, 2017).

Alomari et al. (2017a) use unsupervised learning for concepts such as colors, names and activities by an autonomous robot. Alomari et al. (2017b) combine PCFG trees and visual feature clustering to ground video depictions of actions to linguistic labels. Ilinykh and Dobnik (2022) find that language models in a multimodal task setting learn different semantic information about objects and relations crossmodally and unimodally (text-only).

Importantly, though, these lines of research subsume all grounding and multimodality under combinations of language and *vision*, to the exclusion of other channels, and where AI and humans interact, the interaction focuses on humans guiding AI, not AI assisting humans. Our work brings in modal channels directly related to human motion in a situated environment, to train an AI that ultimately assists humans to mitigate spatial disorientation.

While there is a wide and varied body of research from the neuroscience and biomechanics communities on other modal information channels, such as human spatial awareness, AI has largely not been

applied here.

Rupert (2000) presents a tactile stimulation system that provides intuitive orientation information to aircrew and operators of remote platforms and is compatible with a pilot’s natural sensory system. Intelligent control of such a system could help provide pilots with appropriate cues in disorienting situations, but only if human proclivities in such situations are well-understood and modeled.

Vimal et al. (2016) use a multi-axis rotation system (MARS) device programmed with inverted pendulum dynamics to investigate learning in a dynamic balancing task about an unstable equilibrium point. Subjects attempt to remain balanced by applying joystick deflections to control the motion of the device, and the authors find that subjects improve their performance by making fewer destabilizing joystick movements, and more persistent short-term joystick movements intermittently. Later, they further investigate learning about different roll planes (vertical, horizontal) that disrupt the natural orientational capabilities of humans, combined with the role of gravitationally-dependent otolith and somatosensory cues in the learning of the balancing task (Vimal, 2017; Vimal et al., 2017, 2018, 2019, 2022). They find that absence of gravitationally-dependent otolith and somatosensory cues degrades balancing performance. However, their findings also indicate that balance control can be enhanced in situations lacking gravitationally dependent position cues as in weightlessness, when initial training occurs with such cues present. They also observe that some participants re-learn how to balance themselves in the disorienting condition, demonstrating learning, while others do not. Data from this line of research is used in this paper.

Recent work in this line of research has begun to use machine learning and AI techniques, providing a path forward to integrate the two aforementioned broad areas. Vimal et al. (2020) group subjects performing the balancing task in the horizontal roll plane (HRP), without any gravitational cues, into performance proficiency categories using a Bayesian Gaussian Mixture model. Wang et al. (2022) use the same data to train a stacked gated recurrent unit (GRU) model to predict the occurrence of crashes (where crash boundaries are set to $\pm 60^\circ$ from the balance point) 800ms in advance. Our work extends this line of research toward modeling human behavior in the balancing task so that AI can predict and counteract disorientation.

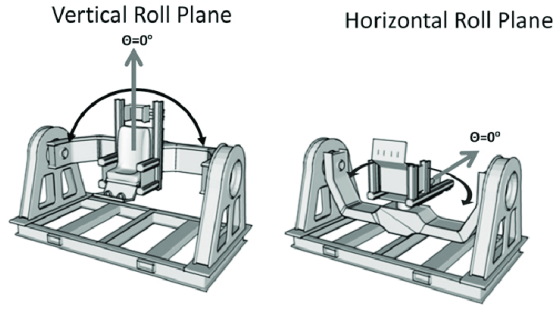


Figure 1: The multi-axis rotation system (MARS), programmed with inverted pendulum dynamics, in the vertical roll axis (left) and the horizontal roll axis (right). Straight grey arrows represent the direction of balance (DOB). Placing participants in the horizontal roll plane disrupts normal gravitational cues, making the balancing task disoriented (figure courtesy of Vimal et al. (2020)).

3 Dataset

We use data and performance proficiency labels from Vimal et al. (2020) which are further explained below. Additionally, we further annotate the data with grounded positional annotations and directional labels for training an embodied AI classifier that predicts optimal direction of movement.

3.1 MARS Data

The data is collected from 34 consenting healthy adult participants (18 females and 16 males, $\mu \approx 20.4$ years old, $\sigma \approx 2.0$ years) with no prior experience in the Multi-Axis Rotation System (MARS).

The MARS was programmed with inverted pendulum dynamics about a horizontal roll axis as shown in Fig. 1 and controlled by a joystick. MARS dynamics were governed by the equation, $\ddot{\theta} = k_P \sin\theta$, where θ is the angular deviation from the direction of balance (DOB) in degrees, and k_P is the pendulum constant. Here, a pendulum constant of $600^\circ \cdot s^{-2}$ ($\approx 0.52\text{Hz}$) was used. Crash limits restricted the angular range of the MARS to $\pm 60^\circ$ from the DOB. Angular velocity was limited to $\pm 300^\circ \cdot s^{-1}$, and angular acceleration to $\pm 180^\circ \cdot s^{-2}$. Every $\sim 0.02\text{s}$, a velocity increment proportional to the joystick deflection was added to the MARS velocity and computed by a Runge-Kutta RK4 solver to calculate the new MARS angular position and velocity. The latency between a joystick deflection and a change in MARS angular velocity was 30ms over the observed range of MARS spectral power of 0 to $\sim 0.75\text{Hz}$ (further experimental details in Vimal et al. (2020)).

Fig. 2 shows a segment of trial data from a representative participant showing changes in angular position (blue), angular velocity (red) and joystick deflection (green). We can see that this participant

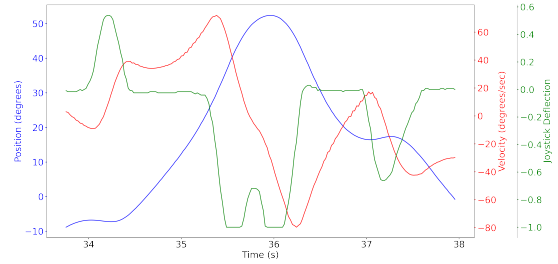


Figure 2: A segment of trial data from a medium proficiency participant showing angular position (blue), angular velocity (red) and joystick deflection (green). The participant just barely prevents a crash as the MARS angular increases to $+50^\circ$ from DOB.

was able to just barely avert a crash as the MARS angular position reached $+50^\circ$, or 10° from the crash boundary.

3.2 Proficiency Labels

Vimal et al. (2020) clusters participants based on their balancing performance using various engineered features, such as:

- **Crash frequency** equals the number of crashes in a trial divided by the trial duration. Higher values correlated with poorer balancing performance. Proficient participants had a mean crash frequency of 0.002Hz and not proficient participants had a mean crash frequency of 0.11Hz.
- **Anticipatory joystick deflections** are those that removed energy from the MARS by decelerating it as it was moving toward the DOB. Anticipatory joystick deflections can help stabilize the MARS; they are often used when poor control leads to high velocities near the balance point. As participants learn to stabilize the MARS the percentage of anticipatory joystick deflections decreases. 0.2% of proficient participants' deflections were classified as anticipatory while not proficient participants used this strategy 14% of the time.
- **Destabilizing joystick deflections** accelerate the MARS away from the DOB. Proficient participants made destabilizing deflections on average 0.0005% of the time and non-proficient participants made them 4.8% of the time.

Vimal et al. (2020) trained a Bayesian Gaussian Mixture model using these features that clustered participants into three distinct groups *Proficient* (or “Good”), *Somewhat Proficient* (or “Medium”), and *Not Proficient* (or “Bad”) based on their balancing performance. Participants were clustered based

on their performance after 2 days of trials, meaning that some proficient participants demonstrated substantial learning in the task over the successive trials, and occasionally some non-proficient participants’ performance actually became worse with repetition. These are the same per-participant proficiency labels we use here.

3.3 Positional & Direction Labels

To ground the situated numerical features from the MARS to a linguistic representation, we annotate the numerical features with sentences that represent position relative to the DOB, or simply put, with possible answers to the question “where am I?” given the numerical features. For example, if they are far off to the right of the DOB, a human may think “I have drifted more towards the right” or if they think they are balanced near the DOB the equivalent thought may be “I think I am somewhere in the center”. These sentence annotations were generated by third-party annotators for each of the three regions; *left* ($< -20^\circ$ from the DOB), *right* ($> +20^\circ$ from the DOB), and *center* (within $\pm 20^\circ$ of the DOB), within a total possible range of $\pm 60^\circ$.

For the direction labels, representing the direction towards which the human should move the MARS (or deflect the joystick) for better balance about the DOB or “where should I go?”, we again divide it into three categories; *left*: deflect the joystick with such amplitude that it prompts the MARS to the left, *right*: deflect the joystick with such amplitude that it prompts the MARS to the right, and *center*: deflect the joystick with as little amplitude as possible such that there is little to no change in the position of the MARS. These are discrete, one-hot vectors depicting the “where I should be going” grounded label, and are assigned using the joystick deflection made after the look-ahead time. The direction labels are defined as *left*: < -0.2 , *right*: $> +0.2$, and *center*: between -0.2 and $+0.2$. $+1$ and -1 represent full deflection.

4 Methodology

Our goal is to combine representations of motion and performance proficiency, which are learned from data directly capturing human embodiment during the MARS balancing task, with linguistic representations of the position and directional concepts involved. A successful model is one which can predict the label for the best direction of motion given the current circumstances by learning correlations between motion, proficiency, and linguistic representation.

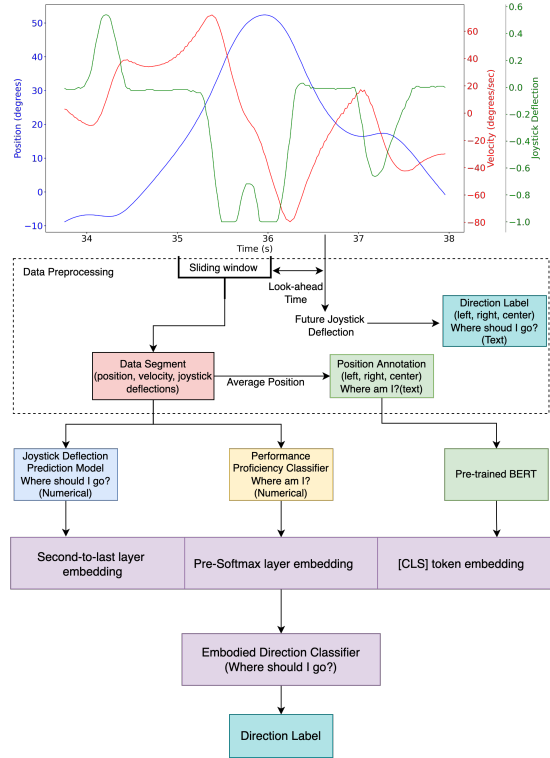


Figure 3: Overview of the embodied model architecture.

The model architecture, shown in Fig. 3, can be divided into five parts: (1) **data preprocessing**; (2) a **joystick-deflection predictor** of immediate future action; (3) a **performance proficiency classifier**, which provides a high-level view of the subject’s task performance; (4) **BERT annotation embeddings**, which provide real-valued semantic representations that the outputs of previous two modules are correlated to, and (5) the combined model, or **embodied direction classifier** (EDC).

4.1 Data Preprocessing

For each trial in the data, we use a fixed sliding window technique to extract segments of joystick deflections, angular velocity and position where the user was in control and no crashes occurred for the given look-ahead time y seconds in the future.

For each viable window extracted, we assign a random sentence annotation for the region corresponding to the user’s average position in the window, e.g., “I think I am somewhere in the center” or “I have drifted more towards the right.”

The processed data has two parts for each sample, (1) the MARS machine features i.e. joystick deflections, position and velocity and (2) the grounded position annotations.

4.2 Joystick-Deflection Prediction Model

Using the processed data on angular position, angular velocity and joystick deflections, we train a

deep feedforward neural network model (see Sec. 5 for hyperparameters) to predict how much the joystick should be deflected to keep the user balanced. Inputs are the 1000ms segments of joystick deflections, positions and velocities, and target values are the joystick deflections made 400ms in the future. Essentially, once operationalized, this model should tell how a user should deflect their joystick to balance themselves¹.

4.3 Performance Proficiency Classifier

To account for how well a user is performing the balancing task, we build a neural performance classifier that is able to tell us the user’s ability to discern and gauge where they are in terms of position and where they should go. The proficiency labels are obtained from Vimal et al. (2020) (described in Sec. 3.2). We train a deep feedforward neural network model (see Sec. 5 for hyperparameters) using the same inputs as those to the Joystick-Deflection Prediction Model (Sec. 4.2). However, here the target labels are discrete proficiency labels of the participant for each sample in turn; *Proficient*, *Somewhat Proficient*, and *Not Proficient*. This model should output a proficiency label for each segment, reflecting how proficient the participant is behaving at that time. The final pre-classification layer of this model outputs embeddings that are situated within the task phase space of the task by preserving high-dimensional similarity relations between actual direction and velocity values and task proficiency.

4.4 BERT Sentence Embeddings

We use pretrained BERT to produce the pooled sentence embedding (the embedding of the [CLS] token) for the the position annotations for each sample. This natural language representation serves as a rather literal “thought vector,” representing the “where am I?” grounded positional label input to our embodied directional classifier.

4.5 Embodied Direction Classifier

Our task is now to take the numerical models learned from embodied human performance, and the linguistic representations from BERT, and train a model, the embodied direction classifier, that grounds the linguistic representation to circumstances described by the numerical data.

We combine the three aforementioned models and build a classification model that has essentially

¹400ms is slightly below the reaction time of average humans (Nagler and Nagler, 1973) and well above the reaction time of trained pilots (Binias et al., 2020).

embodied the operational physics of the disorienting balancing task through human performance data, and has grounding annotations of positional language (“where am I?”). This classifier takes these inputs to predict the grounded directional label, “where should I go?” for better balance.

Input to the EDC is three-fold. **Joystick-Deflection Embeddings** are extracted for each sample from the penultimate layer of the Joystick-Deflection Prediction Model. These vector embeddings represent how much and in which direction the user should deflect their joystick to maintain balance. **Performance Embeddings** are also extracted from the pre-softmax layer of the Performance Proficiency Classifier to represent how well the user can gauge their position and direction. Finally, the **BERT Sentence Embeddings** for the positional thought vectors are extracted. For each sample, these three vector embeddings are concatenated and passed to the model.

The EDC is trained to predict the grounded directional labels, i.e., *left*, *right*, and *center*, which represent the “where should I go?” aspect in the balancing task. In operation, this would be a cue to a guide a human participant through linguistic instruction to either deflect to the left, deflect to the right, or do nothing with the joystick. Here we simply assess the performance of the model and how it compares to humans.

5 Evaluation

We randomly selected 12 participants from the dataset—4 participants of each proficiency. We use 38 of each participant’s 40 trials for the train set and 2 for the test set. As described in Sec. 4.1, we use a sliding window of 1000ms and a look-ahead time of 400ms. After data processing, we end up with about 1.7 million training samples and 80,000 testing samples, for a ~95:5 train-test split.

All neural networks have 3 layers (100 units each, *tanh* activation), and are trained with Adam optimization for 50,000 epochs. The Joystick-Deflection Prediction Model was trained with MSE Loss and both the Performance Proficiency Classifier and EDC were trained with Cross Entropy Loss and a final softmax layer. To evaluate the performance/competence of the EDC we examine:

1. How well the model performs on average and for each proficiency group.
2. Misclassified samples where the model “disagrees” with the apparent ground truth, or the decision the human participant had made.

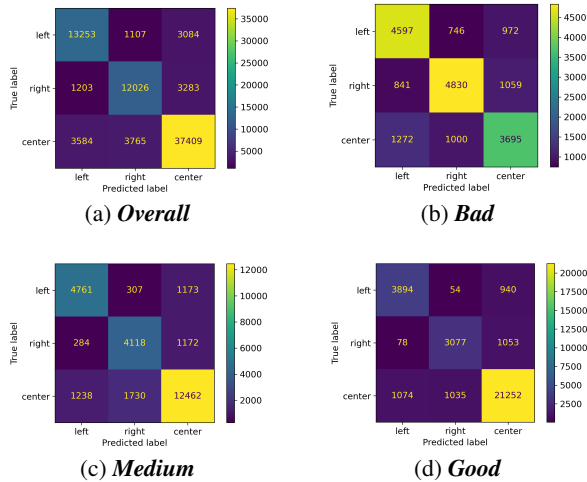


Figure 4: (a) represents the confusion matrix for the full test set of the EDC. (b), (c), and (d) are broken down by proficiency group over the same test set.

6 Results

Table 1 illustrates the performance of the EDC overall and for each of the three proficiency groups. We also show the EDC’s precision, recall, and F1 for the three target labels, i.e., *left*, *right*, and *center*. Here a “correct” answer is one where the human participant made the correct movement choice with respect to their angular position and velocity, and the model predicted the same movement choice.

		<i>Overall</i>	<i>Bad</i>	<i>Medium</i>	<i>Good</i>
Prec.	LEFT	73	69	76	77
	RIGHT	71	73	67	74
	CENTER	85	65	84	91
Rec.	LEFT	76	73	76	80
	RIGHT	73	72	74	73
	CENTER	84	62	81	91
F1	LEFT	75	71	76	78
	RIGHT	72	73	70	73
	CENTER	85	63	82	91
Acc.		80	69	78	87

Table 1: EDC performance as %.

7 Discussion

7.1 Proficiency Breakdown

In Table 1, we can see that the EDC’s performance increases as the proficiency of the participant increases. We see that the Bad proficiency group shows lower performance on correctly grounding the center label, i.e., these participants think they are in the center region, but the model thinks otherwise. They do appear to have a better understanding of whether they are in the left or right region and balance themselves accordingly. The Medium & Good proficiency groups have a better

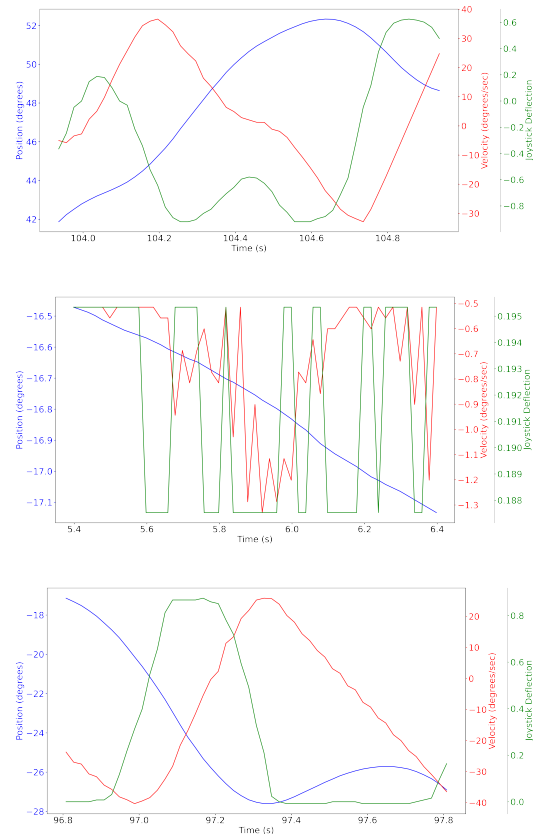


Figure 5: Misclassified test samples from each proficiency group (following conventions from Fig. 2). Top: Bad participant in the right region, truth label *center*, predicted label *left*. Middle: Medium participant drifting toward left region, truth label of *center*, predicted label *right*. Bottom: Good participant in the left region, truth label *center*, predicted label *right*.

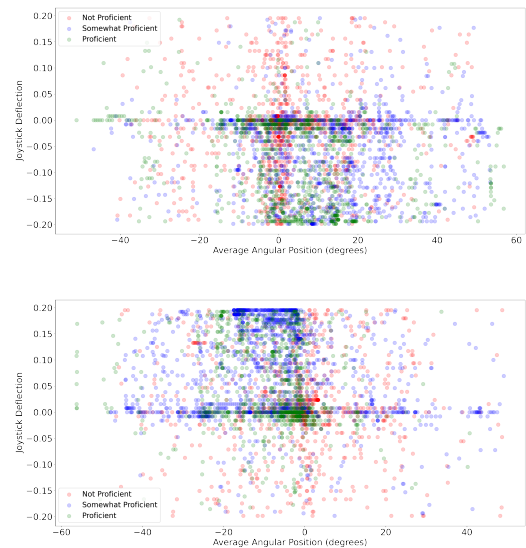


Figure 6: Misclassified test samples where the ground truth labels were center but predicted as *left* (top) and *right* (bottom), showing the spread of actual joystick deflection vs. sample average position when the EDC “disagrees” with the participant’s movement.

understanding of where they are in the problem space than the Bad group, especially when the participants think they are in the center region. For the Good proficiency group, we see that the EDC had an F1 score of 91% for the center label, which means that the model agrees with their decision to do nothing drastic when they are in the center region roughly 91% of the time. This is likely due in part to the fact that many Good (or proficient) participants are able to remain balanced within the center region for most of their trials.

Fig. 4 provides a deeper insight into the what kinds of samples are commonly confused with each other by the EDC. Regardless of proficiency group, the center labels is more often misclassified as left or right than the reverse. This is likely due in part to there being more center labels in the dataset overall (due to Medium and especially Good participants successfully keeping themselves balanced), however the confusion matrices further validate the performance of the model for each of the three proficiency groups: the Bad group has the most confusions and the Good group has the least. The EDC is able to combine the embodied numerical and language representation channels and determine that when a person is in the central region, they should not attempt to move out of it.

Bad participants, meanwhile, are all over the place, and spend $\sim 72\%$ of the time moving either left or right (for the correctly classified samples) whereas Medium and Good participants spend an average of 42% and 25% of their time, respectively, moving left or right. The rest of the time is spent making slight, intermittent movements to remain in the center. They do better at avoiding destabilizing deflections, which the EDC picks up and outputs as directional labels that describe doing just that. Our model, which is trained on data from all proficiency groups, makes decisions that align, in aggregate, with those of a Somewhat Proficient participant.

7.2 Analysis of Misclassified Labels

While the overall metrics for the EDC’s performance are promising, and it performs particularly strongly on Good participants, those numbers do not tell the whole story. Fig. 5 shows one sample from each proficiency group that have a ground truth label of *center* but are predicted as *left* or *right* by the model. Fig. 5 (top) shows a participant from the Bad proficiency group positioned in the right region, closer to the crash boundary, velocity increasing as they deflect the joystick to the right

as well (a destabilizing joystick deflection). The truth label here is *center* as the participant does not move the joystick for 400ms after the end of this sample, but the model predicts that the participant should deflect to the left, which appears to be objectively more correct than the “ground truth” label is. Therefore the training data itself may actually include noise introduced by subpar participants’ suboptimal movements, but the EDC is actually able to learn better intuitive representations from the combination of embodied data and language data from better participants. Fig. 5 (middle and bottom) shows that participants from the Medium and Good proficiency groups respectively, are also occasionally prone to the same situations faced by the participant in top sample, and sometimes make mistakes. Here, the Medium and Good participants are both either in or moving closer to the left region and classifier predicts that the participant should deflect to the right, despite a ground truth label of *center*. This shows that the EDC does learn a better model of both disoriented balancing task performance and in-the-moment guidance through language by learning from multiple participants. If the model were reevaluated against expert/common-sense judgments of optimal human actions, the metrics in Table 1 could rise substantially. In addition, by accurately predicting subpar actions, the EDC may be used to guard against them.

Fig. 6 shows samples labeled *center* where the human does not move the joystick but the classifier predicted an optimal movement to the left (top plot) or right (bottom plot). The graphs themselves show the joystick deflection on the Y-axis vs. sample average position on the X-axis. In Fig. 6 (top), many samples are clustered just right of center with joystick deflection to the left (bottom part of the plot). The opposite is true for the bottom plot, with deflections clustered right of center while average position is just left of the DOB.

If we examine these plots by participant proficiency, the Proficient and Somewhat Proficient samples remain mostly in the center region, close to the DOB. These participants make slight joystick deflections to remain within 20° of the DOB, but the model predicts that the best move is a stronger deflection in one direction. These may be cases where the participant is technically within the center region but perhaps close to a left/right boundary. The Not Proficient participants have a much wider spread of average positions where they make close

to no deflection of the joystick. The EDC disagrees with them, demonstrating both the noise in the data when non-proficient participants' actions are taken as ground truth, and the ability of the EDC, despite this, to make objectively "good" decisions in the context of this task. The numerical performance of the model (Table 1) goes up as participant proficiency goes up, but in fact this reveals that the model is already able to make objectively good decisions, and as human performance improves and participants get better at balancing and become more likely to remain in the center region or recover from drifts, the human decisions are more likely to match these. This suggests that a combined embodied-linguistic method as demonstrated here may be suitable for guiding humans in such a task in real time. The EDC appears to actually display some understanding of the correlation between position and velocity in the problem space, and discrete directional labels.

8 Conclusion

The ultimate goal of this work is to train an AI model that can give guiding cues to a human participant in real time to improve their performance in an embodied task such as the MARS balancing or similar. Successful guidance of a human through language requires that the AI "embody" the relation between linguistic terms and the situation inhabited by the human. Here we have presented evidence that an AI model can be trained to ground directional labels to embedding-level representations of angular position and velocity, and can do so in a way that is sensitive to the proficiency level of a participant in this task, if that information is provided as input. These grounded labels can serve as cues to a human participant, as the AI considers the situation and answers "where am I?" with an answer to "where should I go?" (e.g., "I am drifting to the left. I should deflect more to the right.").

Our model, EDC, trained on data from participants of all proficiencies, displays apparent performance on par with a Somewhat Proficient participant, but a deeper dive into misclassifications reveals that even though the training data itself is noisy, as the ground truth is taken to be the actual actions of the participants, even non-proficient ones, our model's apparent mislabels may actually be better decisions than those of study participants.

8.1 Future Work

Given the nature of the task and the need for immediate response by humans, is a linguistic cue really

the best cue to use in this case? While disoriented, humans may not respond as quickly to language cues; perhaps visual or vibrotactile cues are more apt for prompting faster responses. Further experiments need to be carried out in real time human-AI collaboration in this task (e.g., what kind of AI cues help humans perform better?). Nonetheless, the language input seems to be important to the model for predicting directional guidance, regardless of how that guidance is ultimately expressed. Another feature that could improve our situated embodied model is speed of the MARS, i.e., adding thought vectors representing things like "too fast" or "in control" to positional thought vectors could bolster the combined model's effectiveness as a countermeasure to disorientation by factoring in gradations for things like speed or amount of deflection, which would be important for actually guiding humans in the MARS task where continuous joystick deflection is being applied.

In future work, we plan ablation studies to quantify the effect of each type of embedding, in particular the precise role of language. By taking the existing sentence annotations and automatically transforming them into alternate phrasings (e.g., "I think I am somewhere in the center" → "I *am* somewhere in the center"), we can quantify the differences in sentence and contextualized word embeddings, and the resultant predictive power of the EDC. We are also adapting the virtual inverted pendulum environment of Vimal et al. (2020) to facilitate additional high-throughput studies where we can experiment further with language, e.g., by having subjects call out their perceived direction in real time, or having other trained humans give a subject real-time linguistic guidance. The intermediate models themselves—the joystick-deflection predictor and proficiency classifier—can be improved using techniques like LSTMs and GRUs to pick up on time-series patterns. Furthermore, to be an effective partner for an average human, our models would need to be trained to predict directions for lookahead times greater than 400ms to account for different human reaction times.

Acknowledgements

Thanks to Vivekanand Pandey Vimal, Paul DiZio, and James R. Lackner for thoughts, discussion, and providing access to the original data. Thanks to our reviewers for their helpful comments on the paper.

References

- Muhannad Alomari, Paul Duckworth, Nils Bore, Majd Hawasly, David C Hogg, and Anthony G Cohn. 2017a. Grounding of human environments and activities for autonomous robots. In *IJCAI-17 Proceedings*, pages 1395–1402. Lawrence Erlbaum Associates, Inc.
- Muhannad Alomari, Paul Duckworth, David Hogg, and Anthony Cohn. 2017b. Natural language acquisition and grounding for embodied robotic systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.
- Bartosz Biniś, Dariusz Myszor, Henryk Palus, and Krzysztof A Cyran. 2020. Prediction of pilot’s reaction time based on EEG signals. *Frontiers in neuroinformatics*, 14:6.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.
- Patricia S Cowings, William B Toscano, Millard F Reschke, and Addis Tsehay. 2018. Psychophysiological assessment and correction of spatial disorientation during simulated Orion spacecraft re-entry. *International Journal of Psychophysiology*, 131:102–112.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Randy Gibb, Bill Ercoline, and Lauren Scharff. 2011. Spatial disorientation: decades of pilot fatalities. *Aviation, space, and environmental medicine*, 82(7):717–724.
- Monika Hengstler, Ellen Enkel, and Selina Duelli. 2016. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105:105–120.
- Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.
- Nikolai Ilinykh and Simon Dobnik. 2022. Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difrancò, Ahmad Beirami, Eunjoon Cho, et al. 2020. Situated and Interactive Multimodal Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121.
- Charles Arthur Nagler and William Merle Nagler. 1973. Reaction time measurements. *Forensic science*, 2:261–274.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Gert Rickheit and Ipke Wachsmuth. 2006. *Situated communication*. Mouton de Gruyter.
- Angus H Rupert. 2000. An instrumentation solution for reducing spatial disorientation mishaps. *IEEE Engineering in Medicine and Biology Magazine*, 19(2):71–80.
- Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1634–1644.
- Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 89–97.
- Mark Shelhamer. 2015. Trends in sensorimotor research and countermeasures for exploration-class space flights. *Frontiers in Systems Neuroscience*, 9:115.

- Kartik Talamadupula, J Benton, Subbarao Kambhampati, Paul Schermerhorn, and Matthias Scheutz. 2010. Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2):1–24.
- Vivekanand Pandey Vimal. 2017. *The Role of Gravitational Cues in the Learning of Balance Control*. Brandeis University.
- Vivekanand Pandey Vimal, Paul DiZio, and James R Lackner. 2017. Learning dynamic balancing in the roll plane with and without gravitational cues. *Experimental brain research*, 235(11):3495–3503.
- Vivekanand Pandey Vimal, Paul DiZio, and James R Lackner. 2019. Learning and long-term retention of dynamic self-stabilization skills. *Experimental brain research*, 237(11):2775–2787.
- Vivekanand Pandey Vimal, Paul DiZio, and James R Lackner. 2022. The role of spatial acuity in a dynamic balancing task without gravitational cues. *Experimental brain research*, 240(1):123–133.
- Vivekanand Pandey Vimal, James R Lackner, and Paul DiZio. 2016. Learning dynamic control of body roll orientation. *Experimental brain research*, 234(2):483–492.
- Vivekanand Pandey Vimal, James R Lackner, and Paul DiZio. 2018. Learning dynamic control of body yaw orientation. *Experimental brain research*, 236(5):1321–1330.
- Vivekanand Pandey Vimal, Han Zheng, Pengyu Hong, Lila N Fakharzadeh, James R Lackner, and Paul DiZio. 2020. Characterizing individual differences in a dynamic stabilization task using machine learning. *Aerospace medicine and human performance*, 91(6):479–488.
- Yonglin Wang, Jie Tang, Vivekanand Pandey Vimal, James R Lackner, Paul DiZio, and Pengyu Hong. 2022. Crash prediction using deep learning in a disorienting spaceflight analog balancing task. *Frontiers in physiology*, page 51.
- Martin Weber. 1987. Decision making with incomplete information. *European journal of operational research*, 28(1):44–57.
- Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. [Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings](#).

From speed to car and back. An exploratory study about associations between abstract nouns and images

Ludovica Cerini
University of Pisa
ludovica.cerini
@phd.unipi.it

Eliana Di Palma
Sapienza University of Rome
Roma Tre University
eliana.dipalma
@uniroma1.it

Alessandro Lenci
University of Pisa
alessandro.lenci
@unipi.it

Abstract

Abstract concepts, notwithstanding their lack of physical referents in real world, are grounded in sensorimotor experience. In fact, images depicting concrete entities may be associated to abstract concepts, both via direct and indirect grounding processes. However, what are the links connecting the concrete concepts represented by images and abstract ones is still unclear. To investigate these links, we conducted a preliminary study collecting word association data and image-abstract word pair ratings, to identify whether the associations between visual and verbal systems rely on the same conceptual mappings. The goal of this research is to understand to what extent linguistic associations could be confirmed with visual stimuli, in order to have a starting point for multimodal analysis of abstract and concrete concepts.

1 Introduction

In the last years, the debate over abstract and concrete conceptual representations has gained more attention from a cognitive, psycholinguistic and, recently, from a computational point of view too. Explaining the nature of abstract and concrete concepts is very challenging, and a generally agreed definition is still lacking. We can refer to them as internal mental representations (Paivio, 1990), or as units of information and relationships (Payne et al., 2007), or as unit of knowledge for specific categories (Barsalou et al., 2003). Recently, differences between concrete and abstract concepts have been studied in relation to concreteness ratings (Brysbaert et al., 2014; Connell and Lynott, 2012; Ferreira et al., 2015), underling that a dichotomic distinction does not take into account existing relationships among them.

In the context of grounded theories of cognition, the general assumption is that our conceptual representations are strictly linked to our sensorimotor experience. This assumption seems to be fairly explanatory for concrete concepts (e.g., *dog*,

church, *car*), but when it comes to analyse abstract concepts (e.g., *freedom* or *knowledge*), we have prima face no references in the real and physical world that could activate any kind of sensorimotor experience. It is in the process called multimodal simulation (Barsalou et al., 2003) that situated conceptualizations for a given concept arise. One of the open question in the debate, is how a concept could exploit grounding if no direct sensorimotor experience is available. Many studies have been conducted to explore the issue (Kousta et al., 2011; Lakoff and Johnson, 1980; Wilson-Mendenhall et al.). Guenther et al. (2020), for example, assume that through visual experience our cognitive system infers meaning from the linguistic experience, transferring it in a perceptual experience thanks to language-to-vision relations. This means that an abstract concept such as *knowledge* could be linguistically associated to a concrete referent *book* or *university*, or the concept idea could be metaphorically conceptualized as a *light-bulb*. This process is called **indirect grounding**. In fact, in the symbol interdependency hypothesis Louwse (2011, 2018) highlights that language comprehension is symbolic and mediated by interdependencies of amodal linguistic symbols and it is indirectly embodied through linguistic symbols to perceptual representations. In other words, if concrete concepts exploit direct grounding in perceptual representations, abstract concepts anchor their meaning to different referents mediated both by linguistic associations and figurative, metaphorical and analogical associations.

The aim of this study is to investigate what are the mechanisms involved in the indirect grounding of abstract concepts, both in their visual and linguistic anchoring. Furthermore, we are also interested in exploring the figurative interpretation mechanism of concrete images. In particular, our research questions are:

Q1 What kind of linguistic associations emerge between abstract and concrete concepts, accordingly also to the degree of concreteness?

Q2 Are these associations grounded in sensori-motor experience?

Q3 These associations are confirmed also when the stimulus proposed is an image picturing the concrete concept?

Q4 If these associations could be confirmed also in image-abstract noun pairs associations, could we learn something about visual features that contribute to the indirect grounding processes?

In the exploratory study we propose, we conducted five psycholinguistic tests via crowdsourcing – the first three with linguistic data, and the others with multimodal data – to investigate the link between abstract and concrete concepts and images. We selected 130 abstract nouns that we used as stimuli in our tests. In the linguistics tests, we collected **word associations** with three different elicitation methods. Subjects have been instructed to provide respectively only concrete nouns in response to abstract stimuli (**test 1**), both concrete and abstract nouns in response to abstract stimuli (**test 2**), the mental images in response to abstract stimuli (**test 3**). Results of linguistic tests show an interesting response pattern in the three tests, with respect to the concreteness degree of the answers and the elicitation methods.

In the **multimodal association** tests, we presented image and abstract nouns pairs, and we asked subjects to rate the strongest image-nouns association (**test 4**). Finally, we replaced the images with the concrete nouns they represent, and we asked to rate the strongest concrete-abstract nouns association (**test 5**). In many cases, results confirm the associations of the word association tasks, but other interesting issues emerge, since the contextual features of images impact on the multimodal associations.

2 Related work

A large body of studies have been conducted to discuss the relation that interconnects language, non-linguistic meaning, and perception (Siskind, 2001). Abstract concepts have been always at the centre of debate. Due to the high degree of interdependency between concreteness and abstractness, scholars debated about the best method to classify these two semantic macrocategories (Casasanto and Boroditsky, 2008). The aim of this research line is to define

how meaning arises, considering the large variety of internal and external stimuli that humans use to create conceptual associations to understand the world.

Among the most promising approaches to address the problem of abstract vs. concrete concepts is to consider them as forming a conceptual continuum, rather than a dichotomic relation. In the experiment conducted by Brysbaert et al. (2014), in fact, abstractness and concreteness are evaluated like a scale, providing evidences about the perception of different degrees of concreteness of 40,000 English words. These data confirm, to some extent, the idea that concrete and abstract concepts rely on different information (Crutch and Warrington, 2005, 2007, 2010). However, despite the fact that abstract and concrete concepts form distinct cognitive domains, when it comes to explore the interconnection among different dimensions and modalities, it is unclear what features emerge and contribute to meaning formation and comprehension. Abstract concepts, in fact, are characterised by an intrinsic complexity related to events, situations, physical and mental states, and they are much more variable in their realization of intra and extra-linguistic meaning (Villani, 2018). Moreover, abstract concepts are also directly connected to metaphorical thinking, events, and affective states (Borghi et al., 2017). Lexico-semantic theories underline the contribution that language brings to the meaning formation, with the respect to the context (Louwerse, 2011). For example, according to the Context Availability Theory (Schwanenflugel et al., 1988; Schwanenflugel, 1992), concrete concepts are associated to more specific contexts, compared to the ones of abstract concepts. The Dual Coding Theory (Paivio, 1990), instead, assumes that all concepts are rooted in the verbal system, while only concrete concepts have a direct connection with images.

Multimodal approaches gained more and more interest both in linguistics, cognitive science, neuroscience and computational studies, and new semantic models combining linguistic and visual information have been proposed. Computational models of semantics make use of linguistic and perceptual information, to obtain complementary data to explore our conceptual system (Andrews et al., 2014). There have been several studies aiming to address the multimodal mechanisms of abstract/concrete grounding (Bruni et al., 2014; Berger et al., 2022). Recently Zablocki et al. (2017) proposed a multi-

modal context-based approach to learn word embeddings, observing that visual surroundings of objects are informative and could be exploited to build word representations, jointly with the visual appearance of the object themselves. Interesting attempt to model abstractness and concreteness with the respect to figurative language and multimodality can be found in [Su et al. \(2021\)](#). While abstractness has mostly been modeled at word-level without paying attention to contextualization, [Su et al. \(2021\)](#) explore the dynamics of meaning interchange between texts and images for visual metaphors, by using different degrees of concreteness.

3 Word association

Data selection We collected 130 English nouns and we divided them as **low** abstract, **medium** abstract and **high** abstract (1).

Abstract Noun	Level	Concreteness rating
belief	High	1,19
democracy	High	1,78
love	Medium	2,07
anxiety	Medium	2,21
sight	Low	3,21
speed	Low	3,62

Table 1: examples of abstract stimulus and concreteness ratings

Since our aim was to explore the interconnection between abstract nouns and images, we performed preliminary investigation to control the availability of images related to our stimuli. We used Unsplash.com as a reference website to collect images, firstly because we want to make sure to use royalty free pictures, and secondly because this portal offers a large variety of User Generated Content labelled with tags chosen directly by users. Then we use the concreteness rating data contained in [Brysbaert et al. \(2014\)](#) to perform a more precise division based on the concreteness ratings. We obtained 47 **low** abstract nouns, 42 **medium** abstract nouns, 38 **high** abstract nouns.

3.1 Word association task

As associative relations are particularly important to organize abstract concepts ([Crutch and Warrington, 2005, 2007, 2010](#)), the 130 nouns have been used as stimuli in 3 **word association** tasks. We exploited this assumption to evaluate what are the

most frequent linguistic associations, given a specific abstract noun. To collect word associations we explored 3 methods. We administered three different tests to 120 native English speakers via Prolific. Stimuli were divided in 6 tests and each test was administered to groups of 20 subjects. Each subject was asked to provide up to three associations for each stimulus. For the three tests we collected respectively 9,032 associations in **test 1**, 6,626 in **test 2**, and 5,212 in **test 3**. Tests were designed as follow:

Test 1 (C) Subjects were explicitly asked to produce concrete nouns as associations.

Test 2 (ND) Subjects were simply asked to produce the first noun that came to their mind, independently of its concreteness.

Test 3 (IMAGERY) Subjects were asked to answer the question “What image comes to mind?”. For each abstract stimulus, participants were instructed to provide mental images of objects, settings or animate beings. In this process they had to simulate the experience of visually perceiving some object, event, or scene.

Analysis For the three tests we calculated the frequency of the answers to obtain data about the most prototypical associations for each of the 130 nouns. Then we selected the word associations with a production frequency ≥ 2 to obtain the subset of data with the strongest prototypical associations. Then, we classified each associate noun in the subset as abstract or concrete, to investigate the distribution of the data in **C**, **ND** and **IMAGERY**.

In this step we were interested in exploring whether the elicitation method affects the type of response and more importantly to understand whether the degree of abstractness of the stimuli affects the type of response in the 3 different elicitation conditions. For each test we performed a Chi-squared test to evaluate the significance of distribution of concrete and abstract associations among the three levels of concreteness (**low**, **medium**, **high**). We obtained a significant (p -value < 0.001) in **test 2 (ND)**). **Test 1** and **3** do not show significant differences (p -value > 0.05), but a predominance of concrete nouns in the **low** level can be observed. (Figures 3.1).

Then, for each level (**low**, **medium**, **high**) we performed a Chi-squared test in order to examine the distribution of abstract and concrete associations

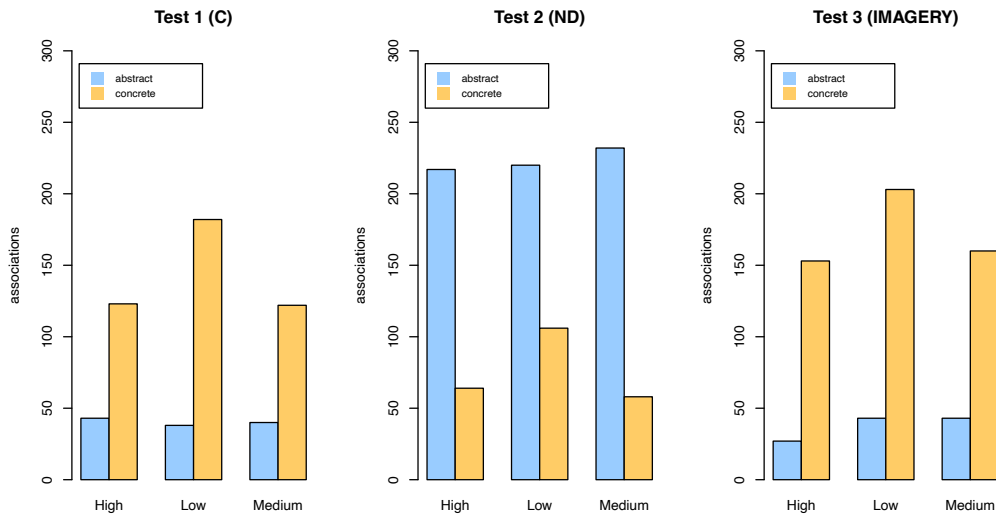


Figure 3.1: Distribution of abstract vs. concrete associates divided by level in **test 1 “C”**, **test 2 “ND”**, **test 3 “IMAGERY”**

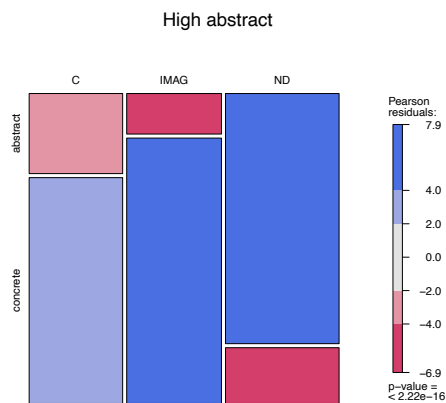


Figure 3.2: Distribution per test of abstract VS concrete according to the degree of concreteness of the stimulus: **High level**

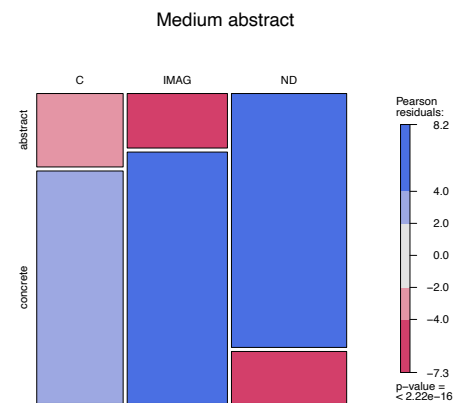


Figure 3.3: Distribution per test of abstract VS concrete according to the degree of concreteness of the stimulus: **Medium level**

among the three tests (**C**, **ND**, **IMAGERY**). For all the three levels, we observed a p -value < 0.001 (Figures 3.2, 3.3 3.4). Results show that abstract stimuli produced more abstract associations in **ND** test, while concrete associations are more common in **IMAGERY** test, with respect to test **C**.

In both analysis what emerges is that if no grounded constrains are given to form the association (meaning the instruction to provide a concrete or imaginable situation in the elicitation method),

subjects do not seem to express a preference for a direct or indirect grounding. Moreover, when we analyse the distribution of abstract-concrete associations considering the level of concreteness, we observed that high abstract stimuli struggle more in finding concrete referents.

4 Image abstract associations

Data collection The next step of the work was dedicated to the multimodal association collection.

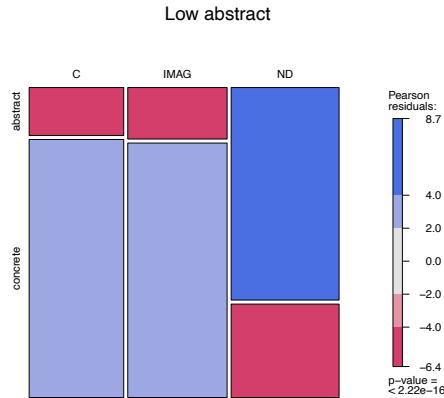


Figure 3.4: Distribution per test of abstract VS concrete according to the degree of concreteness of the stimulus: **Low level**

We identified the most frequent concrete nouns appearing in **C**, **ND** and **IMAGERY** and we selected 44 concrete nouns appearing at least in 2 tests. Each of these concrete nouns was paired with 5 abstract nouns, selected among the most frequent abstract nouns associated to each of 44 concrete nouns. Then, we used Unsplash.com to collect 4 images for each concrete noun.

Concrete noun/ Image	Associated abstracts
BOOK	Knowledge
	information
	learning
	originality
	explanation
CAR	speed
	asset
	future
	risk
	advance
CAT	curiosity
	affection
	instinct
	flexibility
	luck

Table 2: Concrete nouns/image and their most frequent abstract associations.

Image-Abstract nouns rating task We administered to 30 native English speakers a rating test (**Test 4 “IMG”**) in which participants were asked to rate the strongest image – abstract nouns association on a scale from 1 to 5, choosing from a set of 5 abstract nouns. For each image-stimulus we presented 4 different pictures. Participants rated 180 image-noun pairs in total. Only the images were

shown and not the concrete noun they represent, as we wanted to avoid that the human judgments were biased by the linguistic clues.

A second test (**Test 5 “WRD”**) was designed in which participants were asked to rate the strongest concrete – abstract nouns association a scale from 1 to 5, choosing from a set of 5 abstract nouns. In this test, the images of **Test 4** were replaced by the corresponding concrete noun.

Analysis We computed the rating means for the image-noun pairs and for the concrete-abstract noun pairs and the standard deviation for the images in **test 4**. The image - abstract nouns and the word abstract nouns ratings show a very high correlation (Spearman $\rho = 0.77$).

We then computed the correlation between the mean ratings obtained by the **test 4** (e.g., mean rating given to *speed* in relation to the 4 images selected for *car*) and the mean frequencies of the abstract-concrete noun associations (e.g., the value of frequency for the association *speed - car* in **test 1, 2 and 3**) (see table 3). The result was a correlation of $\rho = 0.47$.

The same correlation was calculated between the mean frequencies of the abstract-concrete noun associations and the rating means obtained in the **test 5** (e.g., mean rating given to *speed* in relation to the linguistic stimulus of *CAR*). In this case, the correlation of $\rho = 0.55$ shows that associations like *speed-car* and back (*car - speed*) are quite solid in both directions. The results reveal an higher correlation for the linguistic associations, but still a good correlation within the image-noun pairs.

In order to evaluate if the prototypical associations highlighted by word associations testing phase were confirmed also when images were proposed, we calculated the correlation between the rating means in the **test 4** and the highest frequency value of abstract-concrete association test (e.g., the association *knowledge-book* obtained a maximum frequency value of 11 in word associations tests, see table 3). The result was a quite good positive correlation ($\rho = 0.45$, with Spearman method). We repeated the correlation analysis for the prototypicality with **test 5**. In this case the correlation between the highest frequency value of abstract-concrete association tests and the rating means in the **test 5** was $\rho = 0.52$ (Spearman method).

This suggests that even considering the maximum prototypicality values, linguistic associations show an higher correlation, compared to the image-noun

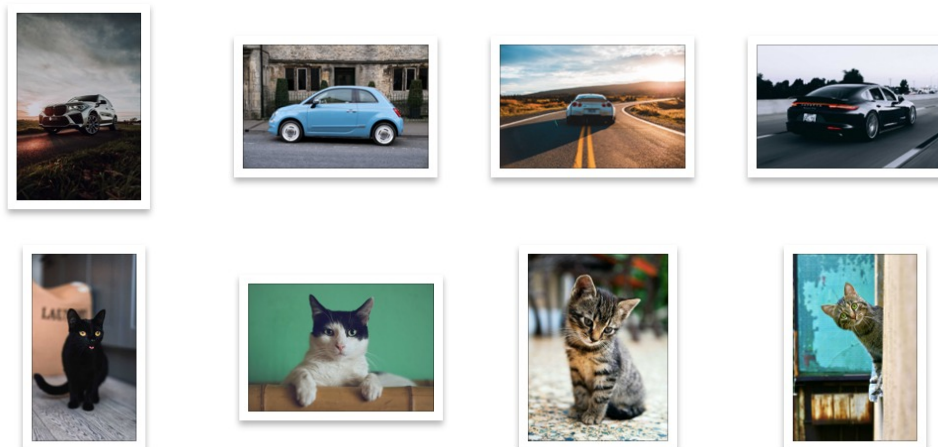


Figure 4.5: Examples of image stimuli in the image-abstract nouns rating task.

IMG\ WRD Concrete Noun	IMG\ WRD Association	IMG Rating mean	WRD Rating mean	Freq.mean C, ND, IMAGERY	Max freq
BOOK	knowledge	4,38	4,66	7	11
BOOK	Learning	4,41	4,4	7	9
CAR	Speed	3,92	4,2	10	10
CAR	Asset	3,65	3,06	3,33	7
FLOWER	Apology	2,28	2,2	3,66	8
FLOWER	Beauty	4,17	4,86	2	3
CHILD	Honesty	3,48	3,53	3,33	6
CHILD	Hope	3,68	3,93	1	3

Table 3: Examples of associations image/concrete nouns - abstract nouns and their rating means in tests 4 and 5 with respect to Freq. means among Test 1-3 and Max freq.

pairs maximum prototypicality values. However, the data show a quite good correlation in image-noun pairs.

Since we analysed the correlation values among the three word association tests, we found that the frequency values in IMAGERY test show the highest correlation with the rating means both in test 4 and 5 (test 4 $\rho = 0.40$ and test 5 $\rho = 0.47$). This result may suggest that the multimodal and grounded connection between abstract and concrete concepts rely more in mental images associations rather than in the mere abstract-concrete associations.

5 Discussion

In our exploratory study we were interested in understanding Q1 What kind of linguistic associations emerge between abstract and concrete, accordingly also to the degree of concreteness. Q2 Whether these associations are grounded in sensorimotor experience. To answer this first two questions, we conducted three word associations tasks, by adopting three different elicitation methods. A1 We observed that interesting prototypical answers were provided by subjects, confirming tendencies in association paths. These prototypical answers were mainly derived from processes grounded in sensorimotor experiences (*anxiety* -

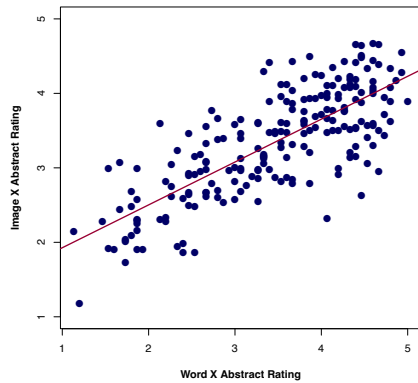


Figure 4.6: Scatterplot rating means in **test 4** and **5**

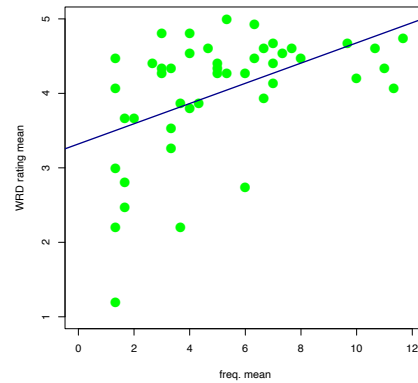


Figure 4.8: Scatterplot showing the rating means of higher frequency means of the 44 concrete nouns (**test 5 “WRD”**)

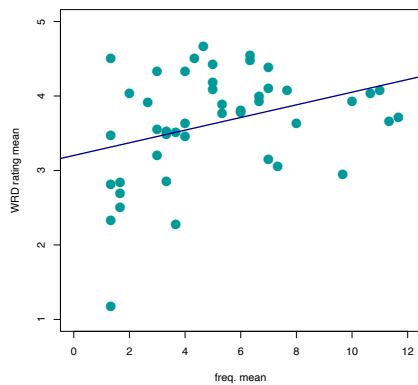


Figure 4.7: Scatterplot showing the rating means of higher frequency means of the 44 images in (**test 4 “IMG”**)

sweat), metaphorical and conventionalized association (*idea - lightbulb*), and also by cultural references (*freedom - America*). Beside the prototypicality of answers we also observed that the distribution of abstract and concrete linguistic associations vary based on the concreteness degree and the elicitation methods. In fact, we can observe a variability in the typology of the responses: if the elicitation method does not specify to provide concrete nouns/association or mental representations to the abstract stimulus, subjects tend to produce an abstract-to-abstract association, especially providing associations of similar linguistic distribution (e.g., synonyms). On the other hand, elicitation methods that clearly specify to produce concrete association or mental images associations provide concrete and grounded answers. Interest-

ingly, we observed also that based on the degree of concreteness of abstract stimuli, a low abstractness is associated more with concrete concepts, and high abstractness is associated more with abstract concepts. **A2** Since we observed that in **Test C** and **Imagery** more concrete associations were provided, we found a great variety of associations referable to different situational context and sensorimotor experiences. (e.g., *apology - flower*; *flow - water*; *attention - eyes*; *consideration - nurse*).

Once we obtained these data, we wanted to understand whether **Q3** These associations are confirmed also when the stimulus proposed is an image picturing the concrete concept; **Q4** If these associations could be confirmed also in image-abstract noun pairs associations, could we learn something about visual features that contribute to the indirect grounding processes? **A3** In the image-abstract association tasks we explored a reverse schema of associations, and we found out that the prototypicality in general is confirmed also when the stimulus is an image. Notwithstanding the predominant prototypicality also with the visual stimuli, some exceptions arise. There are cases in which the strongest linguistic association is not confirmed in the preference of image-abstract nouns (e.g., *apology* is strongly associated to *flower* in the linguistic tasks, and strongly associated to *beauty* in visual tasks). This could be explained by the degree of prototypicality of the associations: the mean association of *apology* and *flower* (3.68) is weaker than the mean association of *book* and *learning* (7). **A4** Since we showed 4 images for the same

concrete concept, we observed that the semantic of the visual scene have an effect on the preference of the abstract concept associated. These results show indeed a strong connection between visual context and the scene interpretation. Grounding of the anchoring mechanisms has to be found especially in the situation depicted. If the visual context change (e.g.: from a picture in which a car is moving to a parked car) different conceptual features are taken into account to define the strongest association between the abstract noun and the visual stimulus. In fact, the correlation between frequency and rating means in the linguistic associations is higher than the correlation of image-abstract associations. This result could be justified by the variety of the visual stimuli proposed. Moreover, the highest correlation values among linguistic and visual norming can be found in **IMAGERY** test, demonstrating also that contextual clues may help in finding better associations between abstract and concrete concepts, as subjects are asked to imagine situations/events to connect two conceptual domains.

6 Conclusions

The study we proposed aimed to explore the role of linguistic associations in the mechanisms that link them to images and abstract concepts. We were particularly interested in studying the indirect grounding processes that connect an abstract concept such as *speed* to a concrete such as *car*, and in the same way how *car* could lead to the idea of *speed*. In order to gain data about these indirect connections of indirect grounding, we collected norming data of word-to-word associations and image-to-word associations. In the word association tasks, we exploited different elicitation methods to first understand what kind of associations emerge if an abstract noun is provided as a stimulus, considering also the degree of concreteness. Our analysis reveals that the degree of concreteness of the abstract stimulus impacts on the distribution of abstract vs. concrete concepts in the three different elicitation methods. Furthermore, the elicitation method also impacts on the kind of the produced associations (abstract/concrete). To some extent this confirms that conceptual systems do not rely only on linguistic information, but context plays an important role in defining the link between concrete vs. abstract concepts (e.g., in the **IMAGERY** test, where subjects were asked to “imagine” the abstract stimuli, mostly concrete

association arise).

With regards to the image-abstract tasks, we observed that when an image is provided, the grounded anchoring offered by the visual scene confirms similar linguistic associations with abstract concepts. In general, our results show that linguistic associations correlate with image-abstract noun ratings. In this view, concepts such as *speed* leads both via linguistic associations and visual associations to the concept *car*. However in the case of image-abstract noun pairs we observed differences of preference in image-abstract pairs, confirming that even the visual content has an important role in the construction of multimodal meaning and the indirect, figurative grounding of images and abstract concepts. For example, despite the association *car - speed* is confirmed in verbal and visual data, the association *car - asset* arises in one of the 4 images proposed to subjects. In this case the car was parked and not moving on the road.

Since this work has been conducted as an exploratory study, we are interested in analysing more in depth the semantic associations that arise from both perspectives. In fact, we qualitatively observed that situational semantic relations are the most prominent in the prototypical associations (McRae et al., 2012), but the degree of concreteness. Furthermore, other information may have an impact on the abstract-concrete association both in visual and verbal systems. In the future we would like to explore the differences arising from different languages to detect cultural and linguistic influences in the grounding processes. If it is true that conventionalizations in linguistic associations may occur in cases such as *idea - lightbulb* across several languages, it is not clear whether association such as *curiosity - cat* could be confirmed in other languages, or if the connection that brings together *belief - church* make use of cultural or linguistic influences, both in word associations and in image-abstract association.

Despite the explorative nature of our study, we think that our results could bring new insights about the many modes in which verbal and visual system continuously interact. More studies are needed to investigate this interconnection, due to the large variety of information that our conceptual systems uses to build meaning. This is particularly important also for computational approaches aiming in exploiting more features of multimodal data.

References

- Mark Andrews, Stefan Frank, and Gabriella Vigliocco. 2014. Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, 6(3):359–370.
- Lawrence W. Barsalou, W. Kyle Simmons, Aron K. Barbey, and Christine D. Wilson. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7:84–91.
- Uri Berger, Gabriel Stanovsky, Omri Abend, and Lea Frermann. 2022. A computational acquisition model for multimodal word categorization.
- Anna M. Borghi, F. Binkofski, C. Castelfranchi, F. Cimatti, C. Scorolli, and L. Tummolini. 2017. The challenge of abstract concepts. *Psychological Bulletin*, 143(3):263–292.
- Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.
- Daniel Casasanto and Lera Boroditsky. 2008. Time in the mind: Using space to think about time. *Cognition*, 106(2):579–593.
- Louise Connell and Dermot Lynott. 2012. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, 125(3):452–465.
- Sebastian J. Crutch and Elizabeth K. Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.
- Sebastian J. Crutch and Elizabeth K. Warrington. 2007. Semantic priming in deep-phonological dyslexia: Contrasting effects of association and similarity upon abstract and concrete word reading. *Cognitive Neuropsychology*, 24(6):583–602. PMID: 18416510.
- Sebastian J. Crutch and Elizabeth K. Warrington. 2010. The differential dependence of abstract and concrete words upon associative and similarity-based information: Complementary semantic interference and facilitation effects. *Cognitive Neuropsychology*, 27(1):46–71. PMID: 20658386.
- Roberto A. Ferreira, Silke M. Göbel, Mark Hymers, and Andrew W. Ellis. 2015. The neural correlates of semantic richness: Evidence from an fmri study of word learning. *Brain and Language*, 143:69–80.
- F. Guenther, T. Nguyen, L. Chen, C. Dudschig, B. Kaup, and A. Glenberg. 2020. Immediate sensorimotor grounding of novel concepts learned from language alone. *PsyArXiv*.
- Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P. Vinson, Mark Andrews, and Elena Del Campo. 2011. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140:14–34.
- G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. University Of Chicago Press.
- M. M. Louwerse. 2018. Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, 10(3):573–589.
- Max M. Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.
- Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy.
- Allen Paivio. 1990. *Mental Representations: A dual coding approach*. Oxford University Press.
- Philip R.O. Payne, Eneida A. Mendonça, Stephen B. Johnson, and Justin B. Starren. 2007. Conceptual knowledge acquisition in biomedicine: A methodological review. *Journal of Biomedical Informatics*, 40(5):582–602.
- Akin C. Luh W.M. Schwanenflugel, P.J. 1992. Context availability and the recall of abstract and concrete words. *Memory Cognition*, 20(1):96–104.
- Paula J Schwanenflugel, Katherine Kip Harnishfeger, and Randall W Stowe. 1988. Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, 27(5):499–520.
- J. M. Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15:31–90.
- Chang Su, Weijie Chen, Ze Fu, and Yijiang Chen. 2021. Multimodal metaphor detection based on distinguishing concreteness. *Neurocomputing*, 429:166–173.
- Caterina Villani. 2018. L’embodied cognition e la sfida dei concetti astratti. un approccio multidimensionale. *Rivista internazionale di Filosofia e Psicologia*, 9(3):239–253.
- C. D. Wilson-Mendenhall, W. K. Simmons, A. Martin, and L. W. Barsalou. Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience*, 25.
- Éloi Zablocki, Benjamin Piwowski, Laure Soulier, and Patrick Gallinari. 2017. Learning multi-modal word representation grounded in visual context.

Author Index

Beinborn, Lisa, 11
Bernardi, Raffaella, 1
Brandl, Stephanie, 11

Cerini, Ludovica, 80
Cetoli, Alberto, 24
Cooper, Robin, 30

Di Palma, Eliana, 80
Dobnik, Simon, 30

Ek, Adam, 30

Frank, Stella, 1

Galindo Esparza, Rosella, 51
Greco, Claudio, 1

Hagström, Lovisa, 45
Healey, Patrick, 51
Hollenstein, Nora, 11

Ilinykh, Nikolai, 30

Johansson, Richard, 45

Krishnaswamy, Nikhil, 70

Larsson, Staffan, 30
Law, Jing Hui, 51
Lenci, Alessandro, 80

Mannan, Sheikh, 70
Maraev, Vladislav, 30
Morger, Felix, 11

Noble, Bill, 30
Norlund, Tobias, 45

Schlangen, David, 62
Somashekarappa, Vidya, 30

Testoni, Alberto, 1