Towards Understanding the Relation between Gestures and Language

Artem Abzaliev
University of Michigan
abzaliev@umich.edu

Andrew Owens
University of Michigan
ahowens@umich.edu

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Abstract

In this paper, we explore the relation between gestures and language. Using a multimodal dataset, consisting of TED talks where the language is aligned with the gestures made by the speakers, we adapt a semi-supervised multimodal model to learn gesture embeddings. We show that gestures are predictive of the native language of the speaker, and that gesture embeddings further improve language prediction result. In addition, gesture embeddings might contain some linguistic information, as we show by probing embeddings for psycholinguistic categories. Finally, we analyze the words that lead to the most expressive gestures and find that function words drive the expressiveness of gestures. Our code is available at https://github.com/MichiganNLP/gestures-language.

1 Introduction

Gestures are often referred to as "non-verbal language" and extensive studies in psychology, sociology, and anthropology have demonstrated the important role they play in communication (McNeill, 1992; Iverson and Goldin-Meadow, 1998; Alibali et al., 2000). While language and gesture can occur independently, people often use them together to communicate, suggesting that multimodality plays an important role in understanding gestures. In this work, we consider human gestures together with their corresponding utterances. We jointly learn gesture and word embeddings, and attempt to predict psycholinguistic categories and the language of the speaker from their gesture embeddings.

Even for humans it is very challenging to predict words from gestures alone (or vice versa), due to the many-to-many relationship between words and gestures. Therefore, instead of directly predicting one modality from the other (Desai and Johnson, 2021), we use contrastive pre-training learning to learn a joint embedding space that aligns both modalities (Kiros et al., 2014; Tian et al., 2020; Radford et al., 2021). This allows our model to learn an association between language and gestures, despite a large amount of uncertainty inherent in the task.

The main contributions of this work are as follows:

- First, we explore a multimodal approach to learn gesture embeddings through contrastive learning. Through validation experiments relying on these embeddings, we demonstrate that there is an association between gestures and languages representations.
- Second, we probe gesture embeddings for various psychological and linguistic categories and show that gestures can be predictive of several categories with better-than-random accuracy. We find that function words, such as pronouns, preposition or modal verbs, can be predicted from the gestures. We also show that gesture embeddings can be used to predict discourse markers.
- Third, we show that it is possible to predict the language of a speaker from our learned gestures embeddings. Our findings indicate that the difference in gestures across the languages may be driven by the function words.
- Finally, we conduct several analyses to better understand the learned gesture representations.

2 Related Work

Semi-supervised Multimodal Learning. Our work builds on the idea of multimodal learning, where a model is trained to represent several modalities in a shared embedding space (Chen et al., 2019; Li et al., 2019; Lu et al., 2019; Tan and Bansal, 2019). In particular, we focus on semi-supervised multimodal learning (Yuan et al., 2021; Wu et al., 2021; Zhai et al., 2021), which is effective and useful training strategy for settings where obtaining labeled training data is laborious or prohibitive. We base our model on CLIP (Radford et al., 2021), which uses a large amount of (multi-

modal) unlabeled data combined with efficient pretraining objective leading to strong zero-shot performance in both language and vision tasks.

Pose Estimation and Gesture Understanding. While most of the recent multimodal work has focused on modalities such as vision (Morency et al., 2007), language, or speech (Levine et al., 2009; Ginosar et al., 2019), there are also studies that highlight the importance of gestures for various aspects of human activities; see Kelly et al. (2008) for an overview.

Work in this space has addressed, among other tasks, gesture recognition (Zhang et al., 2020), gesture generation (Kucherenko et al., 2020b; Ferstl et al., 2021; Yoon et al., 2020; Kucherenko et al., 2020a; Alexanderson et al., 2020), gesture-to-gesture generation (Tang et al., 2019). All of these tasks, while close to our problem space, are not directly applicable for gesture representation.

Another line of work focuses on the temporal alignment and interaction of speech and gestures. Loehr (2007), shows that speech and gestures occur synchronously. (Rieser, 2015) shows that utterances and gestures are not always synchronous and suggests using λ-π calculus to model them. (Saint-Amand et al., 2013) provide a comprehensive study of the alignment of speech and gestures in a constraint-based grammar, while (Lücking et al., 2013) show that gestures follow the language but the opposite does not hold.

An important question is how to obtain labels that describe gesture. We use pseudo ground-truth pose estimates from OpenPose (Cao et al., 2019). While even state-of-the-art gesture recognition systems can be noisy, this noise is significantly reduced on videos such as TED talks (Yoon et al., 2018) given that there is only one speaker and have good light conditions. For a comprehensive overview of recent progress in the field of pose estimation see Munea et al. (2020).

3 Data

Our primary source of data is the YouTube Gesture Dataset (Yoon et al., 2018). The dataset consists of over 1,500 TED talk videos of English speakers addressing various topics like science, medicine, society, and others. The camera is usually in front of the speaker, so the gestures are visible. The dataset contains precomputed key points for the head, neck, shoulders, elbows, and wrists, with each pose represented as a 16-dimension vector, with x and y coordinates for each key point. The dataset includes subtitles, auto-generated by YouTube and aligned by gentle [1].We also use an additional dataset that we compiled ourselves, consisting of 600 videos of Spanish speakers. We process videos from the TEDx channel using the playlist "TEDx talks en Español." Subtitles for Spanish data are auto-generated and aligned by YouTube, so we just download the appropriate subtitle file. Figure 1 shows an example of gestures aligned with the corresponding language.

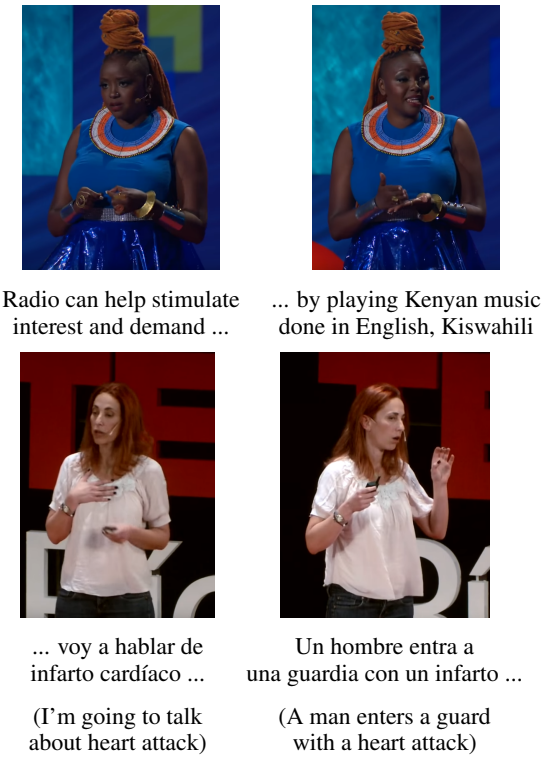

Figure 1: Examples of gestures aligned with the corresponding language. All videos and corresponding utterances are one second in length. Top row: English speaker. Bottom row: Spanish speaker.

We split each video into several clips using PySceneDetect,[2] which detects changes in the camera angle during the talk and splits a single video into several sub-clips. This is necessary so that the pose movements remain continuous during the short clips, even if the camera angle is changed.

Table 1 shows the summary statistics of the dataset. The numbers differ from those reported in Yoon et al. (2018) because we used a more ag-

[1] https://github.com/lowerquality/gentle

[2] https://github.com/Breakthrough/PySceneDetect

	English	Spanish
Videos train	1,349	543
Videos val	167	64
Clips train	127,003	61,810
Clips val	16,813	9,299
Average # words val	3.74	3.4
Average # words train	3.75	3.37
Duration train	35.27h	17.16h
Duration val	4.67h	2.58h

Table 1: Dataset statistics

gressive video filtering strategy,[3] to make sure that only clips with enough gesture variety remain in the dataset.

4 Model

4.1 Input Representation

One of the key considerations for our approach is how to represent gestures and language as input to our model. We start by dividing each clip into a series of gestures using the sliding window approach of Yoon et al. (2018). We cut every clip into 1 second sequences of frames; with 15 frames per second, our input becomes 15 frames. Detailed per-millisecond word alignments are available in our datasets.

We align each one second interval with the corresponding phrase. Our method requires that gestures and language are timely-aligned within some interval of t seconds, where we use t=1 second. It does not require them to be aligned exactly, as the pose encoder and gesture encoder use different positional embeddings. One limitation of such an approach is that longer gestures are truncated, and very short gestures are collapsed together. We experimented with several lengths of the sliding window and found out that the choice of the time interval does not affect the overall performance.

Another possibility would be to split the utterances by word and take all the corresponding frames that fall within the given time interval. For instance, we can take the word 'hello' from the subtitles, and get all the corresponding frames while the word is pronounced. This way we guarantee that there is no overlap between the gestures, and a single word corresponds to a single series of gestures. However, previous results in the literature McNeill (2005) indicate that there are different gesture phases, and they are not necessarily timely aligned with the words. In such a case this approach would be limited. We experimented with such a setting as well, but the resulting performance is only marginally better than random.

[3]We used a threshold of 250 for circular variance, compared to the original value of 150

4.2 Approach

Figure 2 shows the overview of our approach. After preprocessing, poses and phrases are passed through two separate encoders. We use a transformer architecture to separately encode the text and the poses. The pose encoder model very closely follows the CLIP's base image encoder: it is a 12-layer 768-wide model with 12 attention heads. We have to adjust the width from 512 to 768 to match the size of the text model, which is necessary for cosine similarity. The pose encoder is randomly initialized and takes as input a tensor of size $(15, 16)$ where 15 is the number of frames in a 1 sec. clip and 16 is the joint dimension. This input gets transformed to $(15, 768)$ with the fully connected layer and is passed directly to the attention block, bypassing the input embedding layer. This is possible because the pose is already represented as a vector, and does not have to be embedded. We use the last frame as an end-of-sentence token for the prediction. On top of the transformer, there is a 768×768 projection layer. We use the multilingual XLM-RoBERTa (Conneau et al., 2019) as a pre-trained encoder for language with another projection layer on top of the encoder.

After encoding the pose and language into vectors of fixed length, they are normalized and the dot product is taken separately for each modality. The Multi-class N-pair loss (Sohn, 2016) objective is used to learn the match between the poses and the corresponding utterances in a single batch. We selected a batch of size two to make the training task easier. While contrastive learning benefits from large batch sizes (Newcombe, 2018), we found that in our case the higher the batch size, the harder it is for the model to learn. We tried batch sizes 8, 16, and 32, and the results were worse. We hypothesize this is due to a large amount of noise in the data. We use AdamW with a learning rate of $1e-5$ as an optimizer, and cosine schedule as a learning rate schedule.

One possible concern for our approach is the size of the training dataset. CLIP uses more than 350 million image-text pairs, while our dataset in-

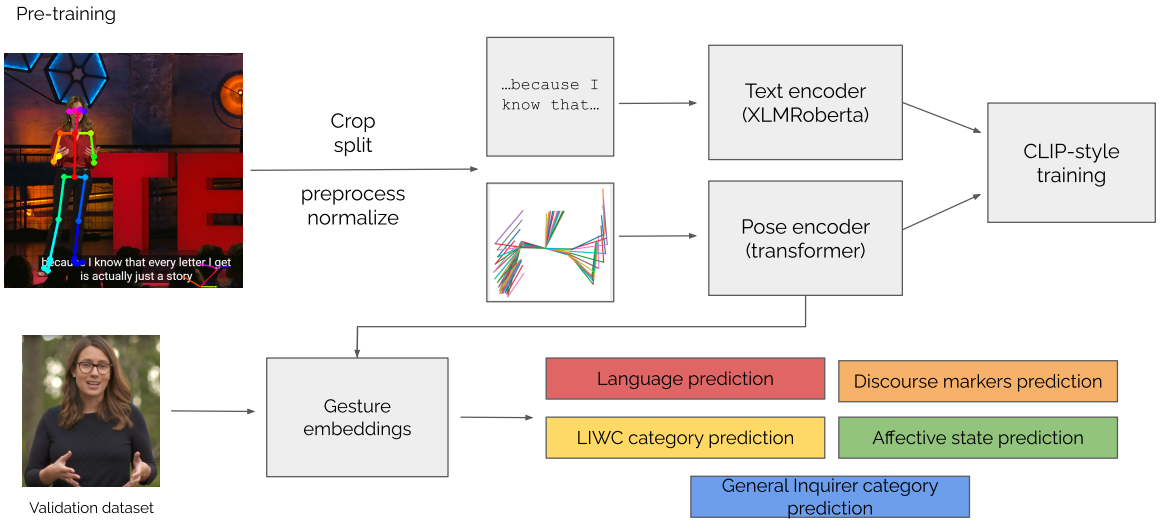

Figure 2: Overview of our approach: we train joint pose and language encoders on training data and produce embeddings. We show we can predict the native language of the speaker and several linguistic categories from the gesture embeddings alone.

cludes only 180,000 1-second clips. To mitigate this, we first conduct an experiment where we show that the proposed pose encoder can predict whether the source of motion was left or right hand. Second, we use a pre-trained text encoder, namely XLMRoberta, as a text encoder. This is in contrast to CLIP, where the authors train the model from scratch. This way we only train a pose encoder, while the text encoder is only fine-tuned. Third, contrary to images that are represented as a 336x336 matrix, poses have much lower dimensionality, namely, our input has a dimension of 15x16. Our intuition is that these factors combined can substantially reduce the required amount of training data.

4.3 Alignment Validation

To make sure that the model is capable of learning gesture–language alignments, we conduct two simple experiments. We aim to verify whether a pose encoder can associate many similar (but with a substantial degree of variety) gestures with a single word/phrase, given the limited amount of training data and proposed model architecture. In other words, before learning a many-to-many mapping (many possible gestures can correspond to the same word, and many possible words can correspond to the same gesture), we want to verify that one-to-many mapping is even possible.

In the first validation experiment, from our dataset, we select only the poses where the source of motion is either the left or the right hand. We do this by calculating the circular variance (Pedregosa et al., 2011) of the angles between joints on the right side and the left side. Only the clips where the circular variance is above 750 on one side and less than 100 on the other side are selected. For these clips, we artificially insert the words ‘left’ or ‘right’ at random position in the existing utterance, depending on which side has is a high variance. This process resulted in 12,287 pose sequences with right-hand movement and 11,846 with left-hand movements.

We set the batch size equal to two and include only one left and one right pose in each batch so that it can be matched with the corresponding text in only one correct way. The resulting accuracy on the validation set is 99.93% and 100% for poses and language respectively. This experiment suggests that learning gesture-text alignments is possible when the gestures have sufficient expressiveness.

In the second validation experiment, given an input gesture and its embedding, we use a simple cosine similarity measure to find the closest text embedding. For each (gesture, text) pair from the validation set, we pick another random pair and

Class	Embeddings	Raw poses	# Observations	Description and Examples
PREPS	51.8 (±2.31)	51.1 (±2.18)	12978.4	Prepositions: to, with, above
PRONOUN	52.5 (±2.94)	52.3 (±2.83)	10235.0	Pronouns: I, them, itself
DISCREP	52.8 (±6.16)	51.6 (±5.79)	2343.2	Discrepancy: should, would, could

Table 2: Accuracy on the LIWC category prediction task for **English**. The accuracy of the majority baseline is 50%. We show the categories where the accuracy of the embeddings model is significantly larger than the majority model at a 0.05 significance level. We use Wilcoxon signed-rank test for significance testing. ± denotes standard deviation

Class	Embeddings	Raw poses	# Observations	Description and Examples
PREPS	52.2 (±3.51)	51.2 (±3.29)	6308.4	Prepositions: sin, con, acerca

Table 3: Accuracy on the LIWC category prediction task for **Spanish**. The accuracy of majority baseline is 50%. We show the categories where the accuracy of the embeddings model is significantly larger than the majority model at a 0.05 significance level, using a Wilcoxon signed-rank test for significance testing.

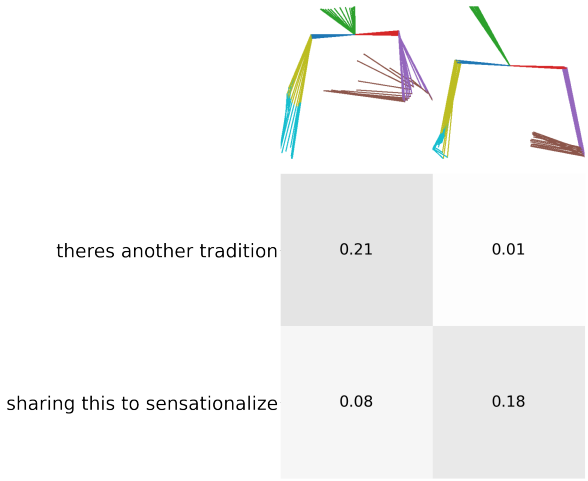

Figure 3: Representation of the input. The main diagonal represents the ground truth.

calculate the cosine similarity between each gesture and text, so we end up with 4 similarity scores, 2 for each gesture (or correspondingly 2 for each text). For each gesture, we take the text with the highest score as a prediction and compare the text we find using the similarity against the correct text paired with the gesture in the data. Figure 3 shows an example.

On the validation dataset, this experiment leads to 65.4% accuracy. When reversed, i.e., starting with a text embedding, we find the most similar gesture embedding according to a cosine similarity and compare it against the gold standard, we achieve 64.9% accuracy. This performance significantly higher than the random baseline indicates that the learned representations contain useful information.

5 Experiments and Results

To evaluate the strength of the connection between gestures and language, we perform two types of experiments. First, we perform experiments within one single language only, i.e. how gestures and language interact in either English or Spanish. The second type is cross-language, how gestures are different among English and Spanish speakers.

5.1 Single Language Experiments

We aim to predict a psycholinguistic category of utterance from the gesture embeddings obtained with the model described in Section 4. The motivation for this type of analysis is that humans might be able to understand that the person is, for instance, angry from the pose alone. We used Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2007) and General Inquirer categories (Stone et al., 1966) to map the utterances to their categories. Namely, if the text contained any word from LIWC or General Inquirer category, we considered the whole utterance to belong to this category, i.e. label y=1, and y=0 otherwise. Some utterances can have more than one category. In total, we run 146 binary classification problems (65 LIWC categories and 81 General Inquirer categories).[4] We compare our results with a majority baseline, as well as a raw pose baseline, where we fit the logistic classifier on the vector of joints (15 frames with a 16-dimensional vector on each, flattened into the 240-dimensional vector). We use

[4]We discard all the categories that have less than 30 observations in the validation dataset.

30-fold[5] cross-validation on the validation dataset, stratified by the language variable, and grouped by the video id (i.e., several clips from the same video should only be in either the training data or the validation data, to exclude the possibility of data contamination). At each training and test fold, we additionally sub-sample English clips to make the number of English and Spanish clips equal. Therefore the accuracy of the majority baseline is always 50%. For the prediction, we use the same model as Conneau and Kiela (2018), a logistic classifier with the default parameters. To verify that the accuracy of one model is larger, we performed a one-sided Wilcoxon paired signed-rank test (Wilcoxon, 1945) on the accuracy scores from cross-validation. We decided to use this test based on the results from Demšar (2006). The null hypothesis is that the accuracies of two classifiers are the same, and the alternative hypothesis is that the embeddings/raw model is larger than the raw/majority.

We rerun our experiment with 30 different random seeds. Table 2 shows the LIWC categories where the embeddings model significantly outperforms the baseline (majority) model for a 90% of the 30 runs. [6] Interestingly, the resulting categories belong to function words. This finding indicates that gestures accompanying function words may have a more apparent visual appearance, compared to the other words. This finding extends previous work from psychology pointing to the importance of function words in communication (Chung and Pennebaker, 2007). Table 3 shows the same type of analysis for Spanish language. There is only one category overall, prepositions, also part of the function words. Table 4 presents the results for the General Inquirer categories. In addition to the pronouns, active verbs show a strong connection with gestures.

Another type of analysis we conducted is predicting the use of discourse markers in speech from the gestures. We use a list of words from Discsense (Sileo et al., 2020). Table 5 shows the results. The embedding model is significantly better than both the raw poses model and the majority model, suggesting that joint language-vision learning is beneficial for this task. We also attempt to predict Valence-Arousal-Dominance states from (Mohammad, 2018), but neither raw poses nor embedding model could predict better than the majority baseline.

These results are in line with the findings reported in Lücking et al. (2013), where authors conducted the experiments using a dataset with manually annotated alignments between gestures and phrases, and found that prepositional phrases are associated with gestures as well.

An important observation from these results is that in many cases a classifier relying only on the raw poses significantly outperforms the majority baseline, suggesting that gestures by themselves contain information about language. Additionally, this finding is further supported by the improvements obtained with the gesture embeddings, which show that the joint learning of gestures and language is beneficial.

5.2 Experiments with English and Spanish

If gestures are indeed closely related to the corresponding language, we hypothesize that we should be able to predict the language of a speaker (e.g., English or Spanish) from the gestures alone. Table 7 shows the language prediction results using the gesture embeddings. We use the identical cross-validation with the sub-sampling scheme as described in Section 5.1.

To further investigate which gestures lead to better-than-random accuracy on the language prediction task, we use the LIWC lexicon to identify the word categories that have the highest improvement with respect to the majority baseline. This analysis can be interpreted as follows: "While a person is using word category X, the gestures of an English speaker can be more easily distinguished from those of a Spanish speaker." Table 6 shows the accuracy of the language prediction task split by the LIWC categories. We select the poses that have the highest probability to be predicted correctly, and extract their corresponding utterances. From these utterances, we extract the LIWC category for the words, and calculate the accuracy separately for each word category. There are eleven categories where the embedding model outperforms the raw poses model and the raw poses model is better than the majority.

Additionally, we identify the words with the highest mutual information between the occurrence of the word in the utterance and the probability to be predicted correctly. Table 8 shows the top 10

[5] We use more folds than typical in the literature to have more observations for significance testing.

[6] Since such a setting is rather restrictive, we also run Fisher's combined probability test to combine p-values from all the 30 runs. We present these results in the Appendix.

Class	Embeddings	Raw poses	# Observations	Description and Examples
PRONOUN	52.6 (±2.50)	52.2 (±2.54)	10034.0	Pronouns: you, nobody, us
ACTIVE	51.3 (±2.66)	50.0 (±2.76)	9533.2	Active verbs: do, develop, learn

Table 4: Accuracy on the General Inquirer category prediction task, where the accuracy of the embeddings model is significantly larger than the majority model at a 0.05 significance level, using a Wilcoxon signed-rank test for significance testing.

	Accuracy
Majority	50.0 (± 0.0)
Raw Poses	51.7 (± 3.04)
Embeddings	**52.7** (± 3.01)

Table 5: Accuracy on the discourse marker prediction task for raw poses and gesture embeddings. Some examples of discourse markers include: actually, anyway, so.

unigrams (top row) and bigrams (bottom row) with the highest mutual information. These words can be interpreted as the most expressive, as the corresponding gestures are more clearly distinctive from the other gestures. Once again, it appears that the majority of the expressive unigrams and bigrams represent function words, which further supports the strength of the connection between these groups of words and gestures.

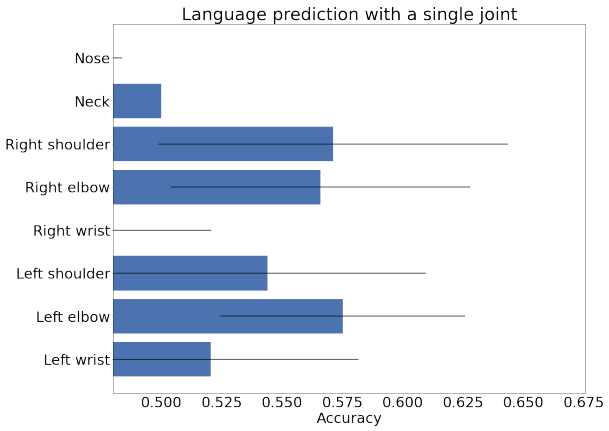

Figure 4: Results from fitting Logistic Classifier on single joint only on language prediction task.

One potential concern is to confirm that there is not any unseen bias (e.g., channel information) that is helping the system to recognize the language of the speaker. We manually inspected a random sample of 100 poses and could not see any differences, such as data artifacts from pose extraction, or any other data processing issues. In addition, we conducted several analyses:

- We analyzed the distribution of the joints' coordinates for the poses that were matched correctly versus incorrect ones. Similarly, we also analyzed the distribution of the joints' coordinates between English and Spanish videos. Maybe some joints are at the very specific position for English/Spanish videos that it makes very easy for the model to distinguish? We could not see any direct difference.

- We analyzed the coefficients of the logistic classifier for the embeddings model. We looked at the magnitude of the coefficients, i.e., whether some features drive the prediction. The motivation for this analysis is the following: if the logistic classifier relies on only a small number of features, instead of the learned representation as a whole, the gesture representation might be suboptimal.

- We fit the model on a **single** joint only. The motivation is the following: can we predict the language of the speaker from the neck (e.g., nose, shoulder) alone. Figure 4 shows the performance with the standard error bars (using 10 fold cross-validation). While the results are still worse than our proposed model, sometimes even one simple joint can lead to very strong performance.

6 Conclusions and Lessons Learned

In this paper, we explored the relation between gestures and language. Using a CLIP-style joint embedding model for gestures and language, applied on a bilingual multimodal dataset consisting of TED talks in English and Spanish, we report several findings:

First, we found that gestures can be used to infer the corresponding language and conversely that language can be used to infer the corresponding gesture. Our proposed model can predict the matching between language and gesture with 65.4% accuracy, compared to the random 50.0% baseline.

Class	Embeddings	Raw poses	Majority	# Observations	Description and Examples
SENSES	65.5 (13.66)	60.6 (16.90)	58.9 (6.62)	47.2	Sensory processes: see, touch, listen
PRESENT	64.6 (11.1)	59.6 (14.63)	54.5 (3.40)	176.0	Present focus: today, now
ARTICLE	64.0 (11.99)	59.3 (14.86)	58.2 (4.59)	161.9	Articles: a, an, the
COGMECH	63.8 (10.84)	59.7 (15.78)	58.1 (4.01)	152.6	Cognitive Processes: cause, know, ought
OTHREF	63.2 (11.86)	59.2 (15.16)	55.6 (4.40)	112.5	Other references: anyone, everyone
SOCIAL	63.6 (11.86)	59.6 (15.40)	58.1 (5.37)	148.6	Social Processes: talk, us, friend
PREPS	63.6 (11.36)	59.6 (14.61)	54.2 (3.57)	231.9	Prepositions: to, with, above
PRONOUN	63.5 (11.56)	59.7 (15.34)	56.4 (4.61)	177.6	Pronouns: you, nobody, us
AFFECT	65.3 (12.94)	62.0 (16.46)	61.0 (7.49)	49.5	Affective Processes: happy, ugly, bitter
INCL	62.9 (11.14)	60.2 (14.97)	58.5 (5.89)	124.9	Inclusive words: and, with, include
COMM	62.8 (14.44)	60.3 (17.42)	59.9 (6.57)	29.8	Common verbs: walk, went, see

Table 6: LIWC categories that drive better-than-random accuracy on the language prediction task.

	Accuracy
Majority	50.0 ($\pm$ 0.0)
Raw Poses	60.1 ($\pm$ 14.28)
Embeddings	**63.8** ($\pm$ 11.00)

Table 7: Accuracy on the language prediction task for raw poses and gesture embeddings

Unigrams	si, ade, y, asking, mas, here, life, po, medicine, ti
Bigrams	ade mas, from the, so that, and it, is to, y con, if your, we just, we can, we 've

Table 8: Top 10 unigrams and bigrams in English and Spanish, with the highest mutual information between the occurrence of the ngram in the utterance and the probability to be predicted correct.

Second, we showed that it is possible to predict several social or psycholinguistic word categories from the gestures with better than random probability. Through extensive probing of gesture embeddings for LIWC and General Inquirer linguistic categories, we were able to identify the categories where gesture embeddings significantly outperform random baselines: the majority of these categories consist of function words, which is a finding that aligns with previous social science findings. We report the results separately for English and Spanish.

In a similar vein, we showed that gestures can be also predictive of discourse markers. Our results indicate that gesture embeddings contain useful information about discourse structure, outperforming both majority and joint-only baselines.

Finally, we reported that gestures by themselves are predictive of the native language of the speaker, and that gesture embeddings further improve this result. Through several analyses, we found that function words are most strongly associated with gestures, which aligns with theories of language evolution that posit that function words are closely connected to the body.

There are several limitations of this work. First, this work focuses on hands and head gestures only, ignoring whole body movements and facial expressions. We also assume that two active hands/arms perform a single gesture, while it is also possible to have two separate gestures for two hands. The gestures are also specific for public presentations.

For future work, we plan to include hand gestures in our dataset. We also consider compiling an 'in-the-wild' gestures dataset, to extend our findings to more forms of communication, and expand beyond the gestures and language for TED talks.

7 Acknowledgement

This material is based in part upon work supported by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. We thank Hanwen Miao for his data analyses and suggestions on an earlier iteration of this project, and our colleagues Laura Biester, Dojune Min, Ashkan Kazemi, Jonathan Kummerfeld, Naihao Deng, Simin Fan, and Siqi Shen for reviewing an initial version of the manuscript and for all the feedback they provided.

References

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 39(2):487–496.

Martha W Alibali, Sotaro Kita, and Amanda J Young. 2000. Gesture and the process of speech production: We think, therefore we gesture. *Language and cognitive processes*, 15(6):593–613.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Universal image-text representation learning.

Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, 1:343–359.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *CoRR*, abs/1803.05449.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. It's a match! gesture generation using expressive parameter matching.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. *Computer Vision and Pattern Recognition (CVPR)*.

Jana M Iverson and Susan Goldin-Meadow. 1998. Why people gesture when they speak. *Nature*, 396(6708):228–228.

Spencer D. Kelly, Sarah M. Manning, and Sabrina Rodak. 2008. Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, 2(4):569–588.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020a. *Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation*, page 242–250. Association for Computing Machinery, New York, NY, USA.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020b. Gesticulator: A framework for semantically-aware speech-driven gesture generation. *Proceedings of the 2020 International Conference on Multimodal Interaction*.

Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. In *ACM Transactions on Graphics*, volume 28, page 172. ACM.

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.

Daniel Loehr. 2007. Aspects of rhythm in gesture and speech. *Gesture*, 7(2):179–214.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: The bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7(1):5–18.

David McNeill. 1992. Hand and mind: What gestures reveal about thought.

David McNeill. 2005. *Gesture and Thought*.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE.

Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. 2020. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348.

Nora S Newcombe. 2018. Three kinds of spatial cognition. *Stevens' handbook of experimental psychology and cognitive neuroscience*, 3:1–31.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Hannes Rieser. 2015. When hands talk to mouth. gesture and speech as autonomous communicating processes. *SEMDIAL 2015 goDIAL*, page 122.

Katya Saint-Amand, Katya Saint Amand, and Katya Alahverdzhieva. 2013. Alignment of speech and co-speech gesture in a constraint-based grammar.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2020. DiscSense: Automated semantic analysis of discourse markers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 991–999, Marseille, France. European Language Resources Association.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers.

Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. 2019. Gesturegan for hand gesture-to-gesture translation in the wild.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2021. Nüwa: Visual synthesis pre-training for neural visual world creation.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2018. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A new foundation model for computer vision.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2021. Lit: Zero-shot transfer with locked-image text tuning.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking.

A Appendix

Here we present the results for LIWC/General Inquirer category prediction task, but instead of selecting categories that were significant at least 90% out of 30 runs with different random seeds, we run Fisher's combined probability test to merge p-values from all the 30 runs into single p-value.

Accuracy on the General Inquirer category prediction task, where the accuracy of the embeddings model is significantly larger at a 0.05 significance level than the majority model.

Class	Embeddings	Raw poses	# Observations	Description and examples
AFFECT	51.6 (5.22)	51.9 (4.93)	3014.1	Affective Processes: happy, ugly, bitter
ANX	51.3 (24.35)	52.7 (24.95)	146.9	Anxiety: nervous, afraid, tense
ARTICLE	50.8 (3.07)	50.5 (2.95)	7197.8	Articles: a, an, the
BODY	51.2 (11.73)	50.1 (11.84)	572.3	ache, heart, cough 1
CAUSE	51.3 (8.04)	50.4 (7.42)	1281.8	Causation: because, effect, hence
CERTAIN	51.4 (7.72)	49.1 (7.82)	1279.9	Certainty always, never
COGMECH	51.6 (3.64)	51.3 (3.23)	6801.5	Cognitive Processes: cause, know, ought
COMM	50.7 (6.82)	49.9 (6.92)	1616.3	Common verbs: walk, went see
DISCREP	52.9 (6.17)	51.7 (5.8)	2343.2	Discrepancy: should, would, could
EXCL	51.5 (3.99)	50.0 (3.84)	4510.5	Exclusive: but, except, without
FAMILY	53.5 (22.04)	53.1 (22.45)	261.0	mom, brother, cousin
FEEL	50.6 (15.66)	49.4 (15.04)	319.0	Feeling: touch, hold, felt
HEAR	50.9 (8.25)	50.7 (7.93)	1098.1	Hearing: heard, listen, sound
HUMANS	51.6 (7.4)	51.5 (7.12)	1451.9	boy, woman, group
I	52.5 (6.19)	51.1 (6.19)	2679.6	I, me, mine
INCL	51.1 (3.08)	50.4 (2.95)	7574.8	Inclusive: with, and, include
INHIB	52.5 (17.33)	50.1 (17.93)	247.5	Inhibition: block, constrain
INSIGHT	50.6 (5.87)	50.2 (5.79)	2339.3	Insight: think, know, consider
JOB	52.2 (9.81)	50.7 (10.06)	784.9	benefits, work, board
METAPH	51.4 (19.83)	51.7 (19.39)	253.3	Metaphysical issues: God, heaven, coffin
MOTION	51.0 (7.58)	49.9 (7.65)	1318.5	Motion: walk, move, go
NEGATE	51.9 (8.67)	51.0 (8.27)	1053.7	Negations: no, never, not
NEGEMO	51.3 (8.95)	50.2 (9.15)	914.5	Negative Emotions: hate, worthless, enemy
NUMBER	51.2 (8.36)	50.7 (8.08)	1150.3	Numbers: one, thirty, million
OCCUP	50.7 (6.29)	49.4 (5.95)	2062.9	Occupation: work, class, boss
OPTIM	50.7 (14.0)	49.1 (13.7)	421.5	Optimism and energy: certainty, pride, win
OTHER	50.8 (6.32)	49.7 (6.37)	1946.9	Total third person: she, their, them
OTHREF	51.4 (3.37)	51.3 (3.31)	6134.5	Other references: anyone, everyone
PAST	51.0 (4.66)	51.8 (4.57)	4272.7	Past tense verb: walked, were, had
POSEMO	50.8 (6.16)	51.1 (5.97)	2115.1	Positive Emotions: happy, pretty, good
POSFEEL	50.8 (11.37)	52.6 (10.97)	624.5	Positive feelings: happy, joy, love
PREPS	51.9 (2.32)	51.2 (2.19)	12978.5	Prepositions: on, to, from
PRESENT	51.1 (2.62)	50.2 (2.6)	9455.9	Present tense verb: walk, is, be
PRONOUN	52.5 (2.95)	52.3 (2.83)	10235.1	Total pronouns: I, our, they, you're
SCHOOL	52.4 (14.34)	49.1 (14.07)	409.9	School: class, student, college
SEE	51.0 (10.1)	49.4 (9.45)	769.5	Seeing: view, saw, look
SELF	52.0 (3.99)	52.5 (3.96)	5223.9	Total first person: I, we, me
SIMILES	51.7 (13.45)	51.0 (14.01)	407.9	like
SOCIAL	50.7 (3.21)	51.0 (2.83)	8872.1	Social Processes: talk, us, friend
SPACE	51.8 (5.53)	51.5 (5.17)	2571.3	Space: around, over, up
TENTAT	51.7 (5.95)	50.8 (5.69)	2318.1	Tentative: maybe, perhaps, guess
UP	52.2 (8.31)	51.4 (7.76)	1131.0	up, above, over
WE	52.5 (5.3)	52.8 (5.56)	2567.5	1st person plural: we, our, us
YOU	53.4 (6.47)	50.5 (6.79)	1591.1	Total second person: you, you'll

Table 9: Accuracy on the LIWC category prediction task for **English**. The accuracy of the majority baseline is 50%. We show the categories where the accuracy of the embeddings model is significantly larger than the majority model at a 0.05 significance level. We highlight in bold the categories where the embeddings model is significantly larger than raw poses model. $\pm$ denotes standard deviation

Class	Embeddings	Raw poses	# Observations	Description and examples
ANX	53.4 (25.29)	55.0 (25.38)	94.4	Anxiety: turba, miserable, temer
ARTICLE	51.1 (3.5)	50.4 (3.83)	5442.3	Article: los, la, una
ASSENT	52.3 (26.08)	49.0 (26.52)	156.6	bien, assent, ok
CAUSE	51.3 (10.79)	48.8 (10.84)	740.5	Causation: porque, dependo, recuperaron
COGMECH	50.9 (4.06)	49.9 (3.48)	5119.5	Cognitive Processes: conceder, asombra, pone
EXCL	51.2 (8.76)	50.7 (8.98)	1073.0	Exclusive: sacar, sin, menos
FRIENDS	52.8 (26.9)	48.0 (25.86)	63.5	examiga, comadre*, macho*
FUTURE	52.3 (17.37)	48.5 (17.88)	320.9	empezare*, frotare*, seremos
I	51.0 (8.36)	50.4 (8.36)	1243.1	mi, tuve, yo
INCL	50.6 (4.75)	49.4 (4.86)	3058.5	Inclusive: con, y, junto
LEISURE	52.8 (19.15)	50.9 (20.1)	262.9	trotar, compac, vives
NUMBER	52.7 (14.89)	52.9 (14.59)	365.4	mitad, once, nueve
PHYSCAL	51.0 (10.42)	50.9 (10.51)	736.6	Physical states: cruda, violar, patas
PREPS	52.2 (3.52)	51.1 (3.29)	6308.5	con, para, sobre
PRESENT	51.0 (4.18)	50.8 (3.89)	5028.6	Present tense: coge, entrego, desean
SOCIAL	50.8 (4.61)	49.3 (4.47)	3692.7	entrego, primo, oyes
YOU	51.3 (14.73)	50.8 (13.76)	475.6	estas, vos, tu

Table 10: Accuracy on the LIWC category prediction task for **Spanish**. The accuracy of the majority baseline is 50%. We show the categories where the accuracy of the embeddings model is significantly larger than the majority model at a 0.05 significance level. We highlight in bold the categories where the embeddings model is significantly larger than raw poses model. $\pm$ denotes standard deviation

Class	Embeddings	Raw poses	# Observations	Description and Examples
ACADEMIC	50.8 (8.68)	50.6 (8.39)	1015.0	academy, dean, coach
ACTIVE	51.3 (2.66)	50.1 (2.76)	9533.3	accost, actor, alarm
BEGIN	50.7 (9.67)	50.3 (9.73)	781.1	bloom, dawn, first
CAUSAL	52.2 (5.17)	52.1 (5.39)	2646.7	order, premise, odds
COLLECTIVITIES	52.1 (6.37)	49.5 (6.04)	2049.3	crowd, cult, family
COMMUNICATION FORM	50.8 (4.27)	50.5 (3.95)	4096.7	ask, assign, discuss
DESCRIPTIVE VERBS	50.7 (3.07)	50.5 (3.49)	6793.7	moan, mumble, pinch
FALL	53.5 (33.31)	57.9 (30.73)	61.4	sunk, drop, collapse
FINISH	51.4 (11.24)	50.2 (10.48)	665.1	cease, expire, lost
FREQUENCY	52.1 (9.4)	49.5 (9.06)	802.3	repeat, weekly, rare
HUMAN'S ROLES	50.5 (4.4)	50.8 (4.42)	3893.7	antagonist, cook, genius
INCREASE	51.0 (8.79)	50.2 (8.94)	1022.7	quicken, run, elaborate
INTERJECTION	51.9 (5.67)	50.0 (5.59)	2339.3	okay, damn, well
INTERPERSONAL	50.5 (4.28)	50.8 (4.2)	4234.0	adversary, hug, recruit
INTERPRETATIVE VERBS	51.0 (2.58)	50.3 (2.77)	10376.2	control, define, educate
KIN	53.3 (18.12)	48.8 (17.87)	297.7	mother, uncle, ma
LEGAL	51.2 (7.46)	50.4 (7.21)	1308.9	convict, crime, unjust
MALE ROLES	52.6 (12.74)	50.5 (12.52)	587.7	salesman, pope, husband
MEANS	50.9 (4.68)	50.0 (4.56)	3393.0	wage, utility, consideration
NEGATION	51.5 (7.09)	50.7 (7.23)	1357.2	aińt, disapprove, no
NEGATIVE	50.8 (4.81)	50.7 (4.93)	2927.1	break, deviation, furious
NUMBER CARDINAL	52.6 (9.21)	51.3 (8.84)	957.0	seven, zero, two
PLACE AQUATIC	54.8 (29.82)	51.3 (31.38)	125.4	bay, swamp, water
PLACE LAND	53.9 (15.02)	53.9 (15.03)	328.7	hilly, desert, cave
PRONOUN	52.7 (2.5)	52.2 (2.55)	10034.0	you, us, those
QUALITY ASSESSMENT	51.3 (6.43)	48.8 (6.18)	1828.9	modesty, hilarious, curve
QUANTITY ASSESSMENT	51.1 (3.46)	49.6 (3.05)	8094.9	considerable, all, another
RELATIONSHIPS	50.8 (5.41)	49.2 (5.37)	2594.5	tie, coherent, unlike
RISE	52.5 (14.33)	48.3 (14.2)	367.3	raise, jump, peak
ROLE	50.9 (6.07)	49.7 (6.39)	2066.7	alcoholic, buddy, mentor
SELF	52.7 (6.66)	51.2 (6.32)	2679.6	me, mine, I
SELF EXPRESSION	50.9 (8.59)	52.8 (8.69)	1121.3	vacation, paint, actor
SPACE	51.9 (4.57)	51.4 (4.8)	3060.5	way, on, nearby
STATE VERBS	50.6 (3.58)	50.9 (3.39)	5919.9	feel, seem, am
STAY	52.2 (17.24)	49.6 (16.45)	282.3	await, locate, set
STRONG	51.0 (2.5)	50.1 (3.03)	10066.9	aptitude, autocratic, defense
SUBMISSION	51.1 (7.55)	51.0 (7.44)	1374.6	respect, kneel, honor
TOOL	52.4 (7.2)	49.8 (7.19)	1341.7	Fork, stove, wheel
TRAVEL	51.1 (6.13)	50.8 (6.15)	1902.4	walk, leave, away
TRY	51.1 (7.92)	51.4 (7.15)	1293.4	bring, attempt, seek
UNDERSTATED	51.2 (3.78)	50.3 (3.93)	4877.8	caution, gamble, rare
VARY	52.1 (9.05)	51.1 (9.1)	887.4	turn, divert, amenable
VICE	51.6 (6.74)	50.9 (6.3)	1784.5	bore, damage, loss
VIRTUE	50.9 (3.92)	50.3 (3.71)	4568.5	invulnerable, free, admirable
WE	53.0 (5.36)	53.0 (5.18)	2488.2	ours, ourselves, we
WEAK	51.7 (5.0)	50.3 (4.76)	3271.2	addict, cheap, sunken
YES	51.9 (10.89)	51.2 (11.04)	665.1	yeah, okay, definitely
YOU	52.9 (6.9)	50.1 (6.5)	1610.5	your, thy, thou

Table 11: Accuracy on the General Inquirer category prediction task. The accuracy of the majority baseline is 50%. We show the categories where the accuracy of the embeddings model is significantly larger than the majority model at a 0.05 significance level. We highlight in bold the categories where the embeddings model is significantly larger than raw poses model. $\pm$ denotes standard deviation