A FRAMEWORK FOR ALIGNING HUMAN LINGUISTICS AND AI PERCEPTION

Joseph Bingham
Department of Biology
The Technion – Israel Institute of Technology
jbingham@campus.technion.ac.il

ABSTRACT

Grounding natural language in perceptual representations is central to both human cognition and AI reasoning, yet remains challenging under ambiguity and partial information. We present a computational framework that models key aspects of human referential interpretation by aligning linguistic utterances with perceptual representations derived from large-scale, crowd-sourced imagery. The approach approximates human perceptual categorization using scale-invariant feature transform (SIFT) alignment and the Universal Quality Index (UQI), while lightweight linguistic preprocessing captures pragmatic variability in referring expressions. We evaluate the model on the Stanford Repeated Reference Game corpus (15,000 utterances paired with tangram stimuli), a benchmark designed to probe perceptual ambiguity and coordination in human communication. The system achieves robust grounding, requiring 65% fewer utterances than human interlocutors to establish stable mappings, and correctly identifying targets from a single utterance 41.66% of the time (compared to 20% for humans). These results suggest that relatively simple perceptual–linguistic alignment mechanisms can exhibit human-competitive behavior on a classic cognitive task, offering insights into grounded reasoning, perceptual inference, and cross-modal concept formation.

1 INTRODUCTION

Effective cooperation between co-performers in joint activities depends on their ability to establish, maintain, and update a shared representation of common ground, which encompasses knowledge of the task, environment, and each other's capabilities. Achieving such common ground is cognitively demanding, even for humans, particularly under conditions of partial observability that can produce misinterpretations or coordination errors. Human interlocutors often begin with different conceptualizations of the same object, yet over repeated interactions they tend to converge on shared terminology through a process known as *lexical entrainment* Brennan & Clark (1996).

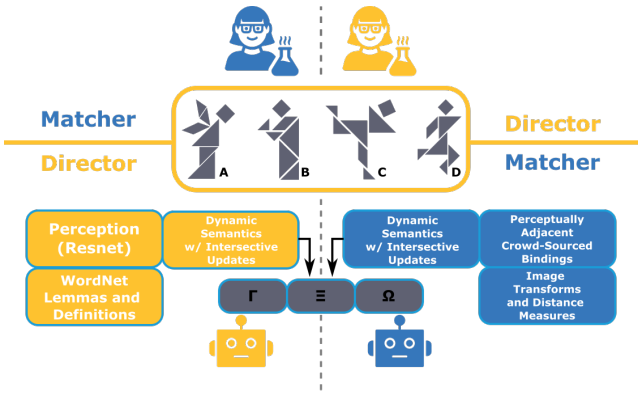

Figure 1: Overview of the repeated reference game and framework for lexical entrainment and common ground formation. The paper focuses on the right side, where a human acts as director and an AI as matcher. The sets Γ, Ξ, and Ω track common ground: finalized conceptual pacts, pacts under negotiation, and rejected pacts, respectively.

Lexical entrainment facilitates the formation of *conceptual pacts*—temporary, partner-specific agreements about how to refer to objects or states in a shared context Brown (1958). Interlocutors establish these pacts by selecting referring expressions that are informative enough to disambiguate

a target from alternative objects while avoiding unnecessary detail, consistent with Grice's Cooperative Principle Grice (1975). This principle posits that humans typically adhere to the maxims of Quality (truthfulness), Quantity (informativeness), Manner (clarity and brevity), and Relation (task relevance). Conceptual pacts are flexible, maintained only with respect to a particular partner; introducing novel expressions outside this established common ground can induce *partner-specific interference*, manifesting as slower responses, confusion, or lower accuracy Trainin & Shetreet (2025).

In this paper, we present a computational framework for automatic lexical entrainment in a machine co-performer (MCP) engaged in the classic repeated reference game. In our experiments, the MCP serves as the matcher, tasked with aligning human-generated referring expressions φ to the intended referent r_φ bound to an object o in the environment. Figure 1 depicts the overall repeated reference game setup and our approach: the left-hand side shows a machine director with a human matcher (future work), while the right-hand side illustrates the current study of a human director and machine matcher. Both director and matcher have access to identical sets of N randomly ordered abstract objects, $O = o_0, o_1, \ldots, o_{n-1}$, consisting of tangram stimuli that are deliberately challenging to describe. The director produces referring expressions $\Phi = \varphi_0, \varphi_1, \ldots$, which are then used to generate a set of referents $R = r_{\varphi_i}, r_{\varphi_{i+1}}, \ldots, r_{\varphi_{i+(n-1)}}$, each linked to a unique object via conceptual pacts, $r_{\varphi_k} \leftarrow o_j$, enabling the matcher to identify the target object o_j. Participants may use any speech acts necessary, but cannot share perceptual information outside of natural language.

Our MCP matcher is evaluated using the Stanford Repeated Reference Game corpus, containing over 15,000 director–matcher utterances Hawkins et al. (2020) . While the MCP has access to the same tangram stimuli as human participants, these objects remain challenging, producing a cognitively demanding alignment task. The MCP leverages scale-invariant feature transforms (SIFT) Lowe (2004) to map crowd-sourced images to experimental stimuli and employs the Universal Quality Index (UQI) to quantify image similarity, thereby modeling perceptual alignment in a manner analogous to human visual comparison.

The main contributions of this work are: **(1)** A novel formulation of common ground and conceptual pacts grounded in Update Semantics Goldstein (2019), capturing the dynamic and partner-specific nature of lexical entrainment. **(2)** A procedure for successful machine lexical entrainment based on this common ground representation. **(3)** Methods for improving alignment between human and machine perceptual spaces using sheaves constructed over SIFT features from crowd-sourced images, enabling the MCP matcher to map latent perceptual representations to symbolic referents. **(4)** Empirical evaluation on the Stanford open corpus, showing that the MCP achieves lexical entrainment with 65% fewer utterances, and correctly aligns a single referring expression 41.66% of the time.

2 Motivation, Background, and Related Work

As AI and automated reasoning capabilities have advanced in such fields from genetics Bingham et al. (2022a) and neurology Bingham et al. (2025), to controller state estimation for agriculture Bingham & Helmich (2021) and embedded power infrastructure Bingham et al. (2022b) there is an increasing need to develop the capacity for machines to perform less like automated tools and more like dynamic teammates, capable of taking on complex tasks with interdependence with other machine and human co-performers. This new class of emerging Symbiotic AI Wang et al. (2019) is fundamentally different from prior AI applications, which primarily focused on monolithic AI systems capable of autonomy, and typically operating as a solitary machine in a non-social environment. Next-generation AI systems must have the capability to reason socially and engage with team members in a fundamentally interdependent fashion, which requires the establishment, maintenance, and update/repair of common ground.

This capability is especially desirable with the rise of neurosymbolic AI, which combines the strengths of deep learning, including the ability to learn from experiences, with the ability to reason abstractly. To fully leverage these advances neurosymbolic systems must be capable of linking latent spaces not only to symbolic logic, but also to natural language concepts present in the mind of human co-performers Hamilton et al. (2022).

2.1 COMMON GROUND

Common ground is a cornerstone to joint activities amongst humans. Common ground, in this sense, is from the philosophy of language and refers to propositions, definitions, and other assumptions agreed upon by two individuals involved in discussion. Common ground during co-performance can be taken to be the set of jargon shared by co-performers that impacts task outcomes. It is, in essence, a formalized shared model of communication. A core responsibility of a machine co-performer should be the active establishment and management of this common ground, and a representation of the context set that corresponds to this common ground. We define a common ground C as the context set of possible worlds $\{w_i, w_{i+1}, w_{i+2}, \ldots\}$ in which the conceptual pacts $\Xi = \{r_a \leftarrow o_i, r_b \leftarrow o_j, \ldots\}$ indicating referent-object bindings are true.

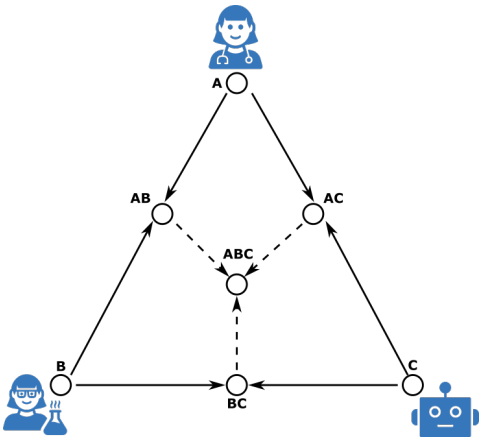

Figure 2: Symmetric simplicial sets of common ground between human co-performers A and B and a machine co-performer C. Pairwise common ground (AB, AC, BC) corresponds to shared understanding about agents, task, and environment, represented at learned Wasserstein barycenters of mutual alignment. Common ground ABC is shared.

The common ground between co-performers can be represented using category theory as symmetric simplicial sets in a barycentric coordinate system Vince (2025) as illustrated in figure 2. For $(n+1)$-performers we define the structure of their existing symbolic knowledge as a standard combinatorial n-simplex defined as the simplicial complex $S(n)$ where $n = 0, \ldots, n$. Each vertex in this simplex represents the labeled symbolic knowledge of one performer, shown as A, B, and C. In between each of these representations at the Wasserstein barycenters of the simplex are the common grounds associated with any subset $m \leq n$ of the $n+1$ performers. These represent a sort of "average" of semantic understandings with ways to express homotopy and homology between concepts, without understanding the underlying topology of the languages in which common ground is established. The process of establishing, maintaining, and repairing common ground is the process of estimating any barycenter of the simplex between co-performers.

Formal methods for the establishment of common ground provide new capabilities for MCPs. Establishing common ground means that beliefs, assumptions, and intentions are shared among team members. While common ground needs to be constantly monitored, updated, maintained, and repaired, doing so provides a number of key properties, such as mutual predictability, ensuring that human and machine team members can maintain a shared picture of what's happening in the world, and accurately predict co-performance outcomes. Lexical entrainment is a key part of this process, allowing MCPs to generate symbolic mappings from latent spaces to symbols in knowledge bases, assumptions, beliefs, etc. Lexical entrainment also allows the common ground to be formally inspected for inconsistencies, violated assumptions, or other errors, forestalling potential breakdown of team function Klien et al. (2004).

Automated lexical entrainment will allow machines to form conceptual pacts with human co-performers, reason about joint activity, to notify team members of impending failures Friedenberg & Halpern (2023), to understand and accept joint goals, to align their objective functions Aguirre et al. (2020), and signal when they are unable or unwilling to participate. Lexical entrainment also establishes better predictability allowing for clarifying statements, assertions, and the ability of humans to assert directability. Specifying, it allows for dynamic objectives and commands in ways that can be understood by the machine and entered in an error free manner by humans.

Conceptual pacts resulting from lexical entrainment also allow humans to supplement machine limitations in perception and cognition, providing a model for joint activity and interdependence and supporting co-active design of the joint activity. By representing joint activity in these formal manners, conceptual pacts allow MCPs to understand the goals of the human users and to measure and improve their own AI loyalty Aguirre et al. (2020), by using these conceptual pacts to engage in after action review Dodge et al. (2021) with models of human and machine intent Billings (2018).

2.2 THE REPEATED REFERENCE PROBLEM

The repeated reference problem Hawkins et al. (2020) is an exercise in common ground establishment found frequently in cognitive science and sociological literature. We focus on the variant of the game associated with the Stanford open corpus of more than 15,000 utterances, which we use to evaluate our framework experimentally.

The repeated reference problem involves a game played by two parties, a director and a matcher. Both parties are provided with identical sets of tangram stimuli, but which have no labels and are in randomized, dissimilar, orders. In this variant of the problem the director selects a tangram and produces an utterance, φ, in the form of a natural language sentence. Their goal is to create concise utterances for the matcher that can be uniquely bound to individual objects in the set of tangram stimuli. The matcher on the other hand must take the utterance and produce hypotheses as to which object is intended to be bound to the referent associated with φ. The matcher then can propose a guess, issue a clarifying utterance, or wait for more guidance from the director in the form of another utterance, as seen in 3.

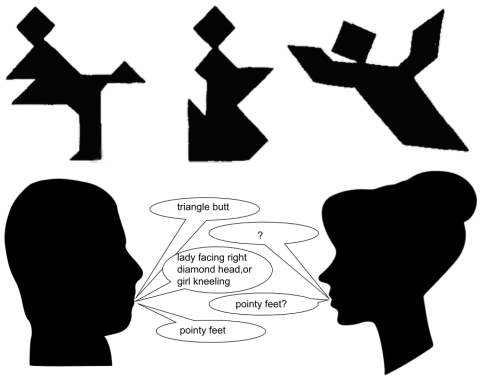

Figure 3: Example of the repeated reference game. The director (right) produces an utterance φ describing selected tangram. The matcher (left) may guess the referent, ask a clarifying question, or wait for information. The example text is drawn from the open corpus.

This is a canonically difficult game for human agents in the field of sociology and communication Hawkins et al. (2020), and to date has not been addressed successfully by machines. The matcher role, the subject of this paper, is particularly difficult due to the need to align with human perceptual spaces to understand the referent-object bindings implied by a referring expression φ. The tangram stimuli used by this exercise is favored primarily because it is difficult for even human performers to align on perceptually. We implement an MCP for the matcher role, aligning human and machine perception by using transformations of φ to produce search terms for web-scraping, resulting in sets of crowd-sourced images related to the provided utterance. We detail this process further in Section 3.

2.3 DYNAMIC SEMANTICS

We model the process of lexical entrainment over humans with a formalization based in dynamic semantics Stalnaker (2002). Dynamic semantics is a perspective on natural language semantics emphasizing the growth of information over time. It asserts that utterances during lexical entrainment are instructions to update an existing context with new information, resulting in an updated context. Common ground, formally, consists of a set of propositions that all interlocutors take to be true, while also taking that all other interlocutors take them to be true. These propositions are the aforementioned conceptual pacts, Ξ.

Under dynamic semantics, assertions and presuppositions interact in a dynamic fashion. Propositions are presupposed if they are part of the pre-existing common ground. New propositions can also be asserted, as a form of instruction to the interlocutor to update shared context with new information. In the case of dynamic semantics, these can be formalized into update semantics using classical propositional logic Goldstein (2019). We borrow Goldstein's notation, using φ to indicate an utterance. Each φ is mapped to a an interpretation function $[\varphi]$ with a context change potential that takes a context as input, and produces a modified context as output. A context C is consistent with φ iff $C[\varphi] = C$.

Similar to Stalnaker's original work on assertions Stalnaker (1978), we achieve dynamic semantic capabilities with update semantics. When an utterance φ is issued by a director, we map this utterance to a set of worlds $\{w_i, w_j, \ldots\}$ representing assumptions about the shared environment with the function $[\![\varphi]\!]$. The matcher takes their current understanding of the context C modeled as a set of

worlds representing the current common ground, and intersects it with $[\![\varphi]\!]$ to implement updating C with $[\varphi]$.

2.4 POSSIBLE WORLDS SEMANTICS FOR EPISTEMIC MODALS

Deriving the context change potential function requires alignment of perceptual spaces. Ideally, the MCP would be able to uniquely map an utterance φ to the intended context change potential function using intersecting updates: $C \cap \{(r_\varphi \leftarrow o)\}$ indicating the director intends the referring expression to communicate that object o should be called r_φ. As alignment of perceptual spaces is an unsolved problem, we instead allow the MCP to estimate the function $[\varphi]$ for any utterance φ. To do so, it builds a hypothesized set of potential bindings implied by the referring expression φ as the set $B = \{(r_\varphi \leftarrow o_i), (r_\varphi \leftarrow o_j), \ldots\}$. This indicates the context change potential function $[\varphi]$ is believed to be one of the bindings in B. We utilize traditional possible worlds semantics for epistemic modals might ($\Diamond$) and must($\Box$) to quantify over possible worlds. $\Diamond\sigma$ is true at world w iff σ is true in some world v accessible from w. $\Box\sigma$ is true at w iff σ is true for all worlds v accessible from w

If the estimated meaning of φ is one of the updates in B, then

$$\begin{aligned} C[\varphi] &= C \cap \Diamond B \\ &= C \cap \Diamond\{(r_\varphi \leftarrow o_i)\}_i \\ &= C \cap \{\Diamond(r_\varphi \leftarrow o_i)\}_i. \end{aligned}$$

If $|B| = 1$, this reduces to $C[\varphi] = C \cap \Box(r_\varphi \leftarrow o_j)$ indicating potential lexical entrainment of part of the problem space. If $|B| = 0$ then the MCP matcher has failed to use the utterance to form hypothesized bindings and must wait for another from the director. If $|B| > 1$ then the MCP matcher has a set of hypothesized bindings to choose from and needs additional information. Due to the limitations of working with a prerecorded public corpus our MCP was unable to ask clarifying questions of its own design. In cases where $|B| > 1$, our MCP implementation waits for additional clarifying utterances in the data set to generate further object bindings, resulting in a strictly harder version of the problem.

To implement our MCP, we model the context of estimated common ground as the sets Γ, Ξ, and Ω, discussing update semantics further in section 3.3. The set Γ contains all established conceptual pacts which **must** be true, the set Ξ contains the set of all conceptual pacts which **might** be true, and the set Ω contains the set of all conceptual pacts which have been rejected or disproven, including pacts of the form $\Box\neg(r_\varphi \leftarrow o_i), \forall o_i \in O$ if the MCP fails to perceptually align on a given referring expression φ. These negative pacts are, in essence, an agreement not to use the referent r_φ to refer to any object.

2.5 PERCEPTUAL ALIGNMENT

Our procedure for the MCP matcher relies on building estimates of the context change potential function $[\varphi]$ by estimating the bindings implied by an utterance φ using crowd-sourced images. Because perception of tangram stimuli is varied, we model human perception using crowd sourced data which maps images to semantic labels. Using a set of images outside of the current tangram stimuli that are associated with semantic tags relevant to our corpus we queried popular search engine APIs with transformations of the utterance φ as a means of estimating the meaning of the referring expression. Current works in this field focus on providing faithful results based on query parameters Uzun (2020) and using these results to augment a user's understanding of real-world information Vishwakarma et al. (2019). We utilize the Bing web image scraping API Singh (Feb 10, 2022) to submit our queries drawn from the Stanford corpus.

Given a set of images resulting from φ, called I_φ, these image matching techniques provide us with distance metrics that allow us to reason that $\Diamond(r_\varphi \leftarrow o_j)$ if the images in I_φ are *close* to o_j.

The MCP matcher uses metrics of image similarity to determine closeness of the crowd sourced images to the tangram stimuli. Given some threshold ϵ and a similarity function $g(o_i, I_\varphi)$, if $g(o_i, I_\varphi) > \epsilon$ then $\varphi \Rightarrow \Diamond(r_\varphi \leftarrow o_i)$. This provides a way to quantify how related the crowd sourced images are to our tangram stimuli images. The similarity function, $g(o_i, I_\varphi)$, can be implemented in a number of different ways, such as mean squared error, peak signal-to-noise ratio, struc-

tural similarity index, universal quality image index, spectral angle mapper, etc Veltkamp (2001). In this work, we have found that universal quality image index (UQI) Wang & Bovik (2002) empirically provides the best results, out-performing all other methods tested by approximately 16%. We attribute the performance of UQI due to the fact this method predicts the probability of shared features as the primary similarity metric. This is ideal for determining if two pictures have different shapes but common features that a human may use to identify and label the content.

3 METHODS

3.1 QUERY CONSTRUCTION FOR WEB-SCRAPING

As previously mentioned, since we do not have direct access to the intent of the director for $[\varphi]$, our MCP matcher utilizes the Bing web-search API to convert the extracted text to a set of images. Submitting the raw utterance φ as the search term yielded poor results, on par with random guessing, with accuracies only a little above 8% in our evaluations. Instead, we explored a number of transformations on φ to improve the results of crowd-sourcing images with higher similarity. By adding cues to the queries, such as appending the text "tangram figure", and by removing stop words, query initial accuracy improved by over 4x.

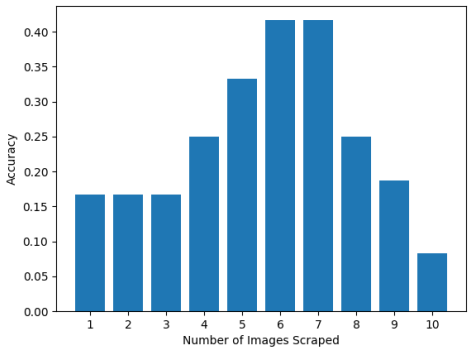

Figure 4: Accuracy of bindings as a function of the number of images scraped by the MCP matcher. Accuracy reflects achieving lexical entrainment on the referent in a single utterance.

3.2 IMAGE MATCHING

For each of the queries formed above, another one of the hyper-parameters that had a large impact on the accuracy was the number of images scraped per query. Estimating the intended context potential change function requires balancing two separate goals: firstly, to collect enough images to find representatives related to natural language utterances, but secondly to avoid scraping non-representative images. As can be seen in Figure 4, the number of images utilized has a strong impact on the matcher. After 7 images, the images provided by the Bing API had a strong destabilizing effect on our decision spaces. This was due to the fact that after 7 or so results, many queries return a generic image of a solved, square tangram, which almost always matches with the same tangram in our object set.

After scraping candidate images I_φ from the web, the MCP matcher then prepares a pipeline computing the similarity of these images to the tangram stimuli objects, shown in figure 6. This pipeline aligns and normalizes the images, and attempts to extract and compare relevant shapes and features from the images to assess the comparative distance of each of the tangram objects to I_φ.

We intentionally employ classical perceptual similarity measures rather than end-to-end learned representations in order to preserve interpretability and to better align the model's internal operations with established accounts of human perceptual comparison.

3.2.1 IMAGE ALIGNMENT

The MCP first attempts to align relevant features in the scraped images, utilizing utilizing SIFT homographies Zhao et al. (2021). Similar to the sliding windows used in CNNs, this method starts by segmenting the images into 3x3 kernels. These kernels are compared and key points are matched based on a Gaussian difference method Lindeberg (2012). This algorithm is scale as well as rotational invariant. This makes it ideal for our application, where images may represent a similar subject from different perspectives, helping to align human and machine perception.

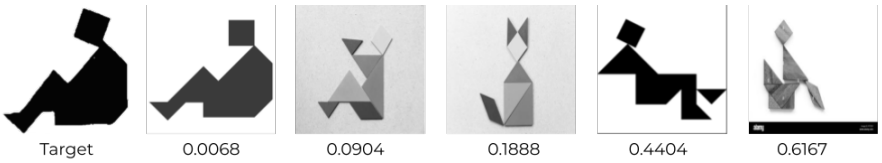

Figure 5: An example of the distances of the closest 5 scraped photos to the target. The query text for this is "tangram figure sitting and looking". *It should be noted that all queries were manually checked to ensure that the scraped images were unique from the tangram figure.*

3.2.2 IMAGE COMPARISON

After processing, the images are subjected to a number of rotational transforms, as well as being grey-scaled and inverted. Doing this helps to generalize our results, and raised the accuracy by approximately 8%. These copies, along with the original picture, are then assessed based on how close they are to the given tangram stimuli by applying UQI Wang & Bovik (2002). UQI normalizes image quality measure which utilizes approximating the noise need to transform one image into another. This measure is applied to each cross product between the scraped images and the tangrams. An example of these outputs can be seen in 5 for a single tangram compared to its top five results.

3.3 FORMALIZING COMMON GROUND ESTABLISHMENT

To model the process of common ground establishment through lexical entrainment, we define the context representing the common ground as the category of sets whose objects are the worlds consistent with three sets we use in the MCP's to model the common ground: Γ, Ξ, and Ω. Γ is the set of conceptual pacts (represented by bindings) that the MCP believes must be true. Ξ represents the sets of bindings the MCP hypothesizes might be true based on utterances and perceptual alignment. Finally, Ω represents the sets of bindings the MCP has determined must be false, including those bindings to referents the MCP was unable to resolve satisfactorily with image matching. The common ground is thus context C of worlds $\{w_0, w_1, \ldots\}$ consistent with $\Gamma \cap \Xi \cap \Omega$.

Formally, when we estimate $C[\varphi]$ as $C \cap \Diamond B$ we update our model such that $\Xi = \Xi \cap \Diamond B$. When any referring expression results in an unambiguous context change potential of the form $C[\varphi] = \Box(r_\varphi \leftarrow o_i)$ we update our model such that:

$$\begin{aligned}
&\Gamma \leftarrow \Gamma \cap \Box(r_\varphi \leftarrow o_i),\\
&\Xi \leftarrow \Xi \setminus \{\Diamond(r_\varphi \leftarrow o_j)\}_{o_j \in O},\\
&\Omega \leftarrow \Omega \cap \{\Box\neg(r_\varphi \leftarrow o_k)\}_{o_k \in O,\, o_k \neq o_i}.
\end{aligned}$$

We propose that for the repeated reference game, successful lexical entrainment on a referent r_φ and an object o indicates that the set of agreed upon conceptual pacts, Γ, contains the referent-object binding $(r_\varphi \leftarrow o)$. Common ground is established, and lexical entrainment has succeeded when Γ contains a unique referring expression $(r_i \leftarrow o), \forall o \in O$ with $\Xi = \emptyset$. The state of Ω does not impact establishment, and is merely used to "memorize" unhelpful referents to avoid their use in the future.

The process of establishing common ground between interlocutors uses the theory from dynamic semantics referenced in 2.3. Dynamic semantics provides a framework where the meaning of a sentence is a richer concept than under static semantics (where knowing a sentence provides its meaning) Goldstein (2017). Dynamic semantics approaches the meaning of a sentence as a rule for learning whether or not the sentence is true. These rules take the information state of any interlocutor (or its context, C) applying an update to the context resulting from φ using function $[\varphi]$ from 2.4.

4 EXPERIMENTAL METHODS AND RESULTS

We evaluated our MCP matcher on Stanford's public corpus for the repeated reference game Hawkins et al. (2019) as the basis of our experiments. The data set represents pairs of human deciders and matchers exchanging unique recorded utterances, each with an associated intended target object. The data sets provide the time taken to make an utterance, the utterance itself in text

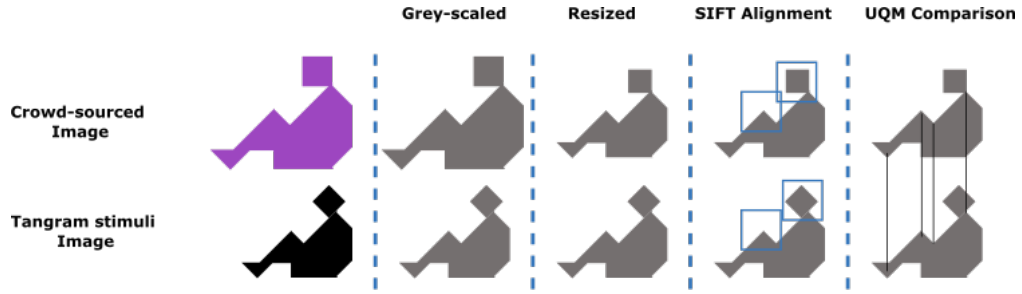

Figure 6: Processing of scraped images to conform with tangram data set from original to comparison measure. The use of grey-scale is to prevent color disagreements from effecting results. *All queries were manually checked that the scraped images were unique from the tangram figure.*

form, and the associated tangram that were shown to the subjects. The MCP matcher attempted to estimate context potential change functions for each utterance through the pipeline detailed in 6.

4.1 Transformations on φ

Because the dataset contains natural language produced by human subjects, many utterance tokens do not contribute to estimating the context potential and can hinder lexical entrainment. This includes common prefixes such as "one that looks like." By removing such stop words, along with tokens not corresponding to nouns, verbs, or conjunctions, we substantially improved the estimated context potential change functions. Modifiers such as "really tall man" do not contribute to alignment relative to "tall man" and often degrade matcher performance. These transformations over stop words and non-contributing speech were implemented with `spaCy` Honnibal & Montani (2017). Finally, subjects in the experiment used informal language and spellings, and occasionally submitted typos. We again used `spaCy` library to automatically normalize these results, improving outcomes.

Like the text, the tangram images needed changes to assist with the matching algorithm. The images of the tangrams were flood filled, removing any artifacts that existed from the copying process. These black and white images were then resized to all consistent with each other, and the tangram stimuli. All samples were down sized to be 300x300, as shown in Figure 6.

4.1.1 Image Distance Measure

As explained before, we tested a multitude of different metrics for assessing the difference between two images. The measures we tested are mean squared error, mean absolute error, peak signal-to-noise ratio, structural similarity index, Erreur Relative Globale Adimensionnelle de Synthèse, Spatial Correlation Coefficient, Relative Average Spectral Error, Spectral Angle Mapper, Visual Information Fidelity, and Universal Quality Image Index Wang & Bovik (2002).

In order to appraise which distance measure performed best in our application, we repeated the tests documented in Table 1 with all measures previously listed. Further, as some of the aforementioned contain self-aligning algorithms as part of their measure, the test was also repeated with all measures with, and without pre-alignment. From these tests, we found that the Universal Quality Image Index (UQI) with SIFT alignment out performed all others by about 16% accuracy.

4.2 Metrics

In order to understand our success at lexical entrainment, we examine our resulting common ground outcomes from a set of diverse metrics to characterize the quality of our solution. Unfortunately, due to lack of a record of the hypotheses a human holds as a result of an utterance, we are unable to compare our results to human matchers beyond top_1 accuracy, which we estimate on the basis of successfully lexically entrained referrants from a single utterance. Our implementation of the MCP matcher, however, was able to achieve lexical entrainment after one utterance 41.66% of the time, versus humans 20%, as reported by Hawkins et al. (2020).

When lexical entrainment cannot be established with one utterance, we allow the matcher to propose additional hypotheses using a softmax function to convert the latent space of computed distances

	A	B	C	D	E	F	G	H	I	J	K	L	Average
Time (ms)													
Human	31737	21156	15311	27794	16614	50496	21756	26559	37634	37392	60380	42110	32411.58
matcher	1.2	7.8	3.3	0.4	2.9	14.1	2.1	1.8	2.4	2.2	2.9	5.1	3.9
Utterances Needed													
Human	2.5	3.75	2.5	2.4	2.4	2.5	2.4	2.4	2.4	2.4	4.8	2.3	2.73
matcher	1	1	2.3	1	1	2.3	2.5	1	1	2.3	1	5	1.78

Table 1: Comparison between the human matcher from the data set and our matcher given the same input phrases (utterances) from the director. Time includes full pipeline from 6

$g(o_i, I_\varphi)$ into a probability distribution of hypotheses, with our decision criteria ϵ set to provide the indicated number of top hypotheses. Allowing the MCP to utilize three hypotheses improved the rate of lexical entrainment from a single utterance from 41.66% to 63.01%, and utilizing five hypotheses resulted in lexical entrainment 83.56% of the time.

4.2.1 SPEED OF LEXICAL ENTRAINMENT

In addition to our success rate at entrainment with a single utterance by varying our closeness decision criteria ϵ, we also measured the speed at which lexical entrainment was achieved, both in terms of cognition/reasoning time (shown in column 2 of table 1), and the total number of utterances needed to reach our stopping criteria (shown in column 3 of table 1). As can be seen from our results, the MCP matcher out performed the user in terms of wall clock time, but more importantly, requiring on average only 65% of the number of exchanged utterances, reaching entrainment on all tangrams with an average of 1.78 utterances per object, vs. 2.73 for human performers. While actual speed of cognition may seem to be an unfair comparison, it is important to note in this work as it represents the difficulty humans experience at creating utterances, and more importantly, at interpreting utterances. Reports on trials with the Stanford corpus, and other similar data sets place a heavy emphasis on the time taken for a given exchange, and further to establish lexical entrainment for a given tangram. Achieving fast lexical entrainment with machine assistance is especially desirable in critical co-performance, where the lack of established common ground can prove detrimental to safety-critical, and even life-critical joint activities, such as triage, or other crisis decision making situations currently seeking to employ symbiotic AI to improve outcomes.

5 CONCLUSIONS

The lack of automated solutions to the problem of lexical entrainment in the literature makes it difficult to compare our methods to prior work, as indeed this solution is the first in the literature that we, or our colleagues in cognitive and social science are aware of. As such we believe these results are novel, and we present them in the context to the current state of the art achievable by human co-performers. Our current results are also limited in what can be inferred from the available public corpora. Current corpora do not record the space of hypothesized lexical entrainments implied by an utterance that humans possess in their minds, and the intent of director actions are likewise not recorded, yielding no dependable ground truth by which to explicitly compare our estimation of the context change potential function. For example, there are instances where the human matcher asked a question to the director, like "pointy feet?". It is impossible to assert that they asked this because they believed that the correct tangram did, or believed it did not have pointy feet, where as with our MCP, the set of all held beliefs is extant in Ξ, explicitly. Within these limitations, our solution represents a first of its kind capability, showing successful lexical entrainment in less time, with fewer utterances than human performers are achieving, when the MCP is in the matcher role. This provides strong evidence that our matcher is superior when compared to humans.

This work demonstrated the use of dynamic semantics with web-scraped crowd sourced images to estimate the context change potential function to implement an MCP matching agent, which approximates the capacity of a human matcher to achieve lexical entrainment on the repeated reference problem. We outlined the hyper-parameters, techniques, and algorithms required by our matcher and demonstrated that this method achieves the first known example of an automated solution to the repeated reference problem, for the matcher position using publicly available data, achieving comparable or more sample-efficient performance under the same informational constraints as human capability for our evaluation.

ACKNOWLEDGMENTS

This work was done with tutelage of Dr. Eric Davis. Funding provided by Galois Inc.

REFERENCES

Anthony Aguirre, Gaia Dempsey, Harry Surden, and Peter B Reiner. Ai loyalty: a new paradigm for aligning stakeholder interests. *IEEE Transactions on Technology and Society*, 1(3):128–137, 2020.

Charles E Billings. *Aviation automation: The search for a human-centered approach*. CRC Press, 2018.

J. Bingham and S. Helmich. Bonsai: A framework for convolutional neural network acceleration using criterion-based pruning. In *Machine Learning ad Data Mining in Patter Recognition*, volume 17, pp. 221–229. IBAI Publishing, 2021.

Joseph Bingham, Netanel Arussy, and Saman Zonouz. Guide-guard: Off-target predicting in crispr applications. In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 423–431. Springer, 2022a.

Joseph Bingham, Noah Green, and Saman Zonouz. Legonet: Memory footprint reduction through block weight clustering. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, (DASC)*, pp. 1–6, 2022b.

Joseph Bingham, Saman Zonouz, and Dvir Aran. Fine-pruning: A biologically inspired algorithm for personalization of machine learning models. *Patterns*, 6(5), May 2025. ISSN 2666-3899. doi: 10.1016/j.patter.2025.101242. URL https://doi.org/10.1016/j.patter.2025.101242.

Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482, 1996.

Roger Brown. Words and things. 1958.

Jonathan Dodge, Roli Khanna, Jed Irvine, Kin-Ho Lam, Theresa Mai, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Andrew Anderson, Sai Raja, et al. After-action review for ai (aar/ai). *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–35, 2021.

Meir Friedenberg and Joseph Y Halpern. Joint behavior and common belief. *Electron. Proc. Theor. Comput. Sci.*, 379:221–232, July 2023.

Simon Goldstein. *Informative dynamic semantics*. Rutgers The State University of New Jersey-New Brunswick, 2017.

Simon Goldstein. Generalized update semantics. *Mind*, 128(511):795–835, 2019.

Herbert P Grice. Logic and conversation. In *Speech acts*, pp. 41–58. Brill, 1975.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promise in natural language processing? a structured review. *arXiv preprint arXiv:2202.12205*, 2022.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. Characterizing the dynamics of learning in repeated reference games, 2019. URL https://arxiv.org/abs/1912.07199.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6):e12845, 2020. doi: https://doi.org/10.1111/cogs.12845. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12845.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Glen Klien, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, 2004.

T. Lindeberg. Scale Invariant Feature Transform. *Scholarpedia*, 7(5):10491, 2012. doi: 10.4249/scholarpedia.10491. revision #153939.

G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2, 2004.

Guru Prasad Singh. bing-image-downloader 1.1.2, Feb 10, 2022. URL https://pypi.org/project/bing-image-downloader/.

Robert Stalnaker. Common ground. *Linguistics and philosophy*, 25(5/6):701–721, 2002.

Robert C Stalnaker. Assertion. In *Pragmatics*, pp. 315–332. Brill, 1978.

Nitzan Trainin and Einat Shetreet. We do not speak like this here: The role of perceived foreignness in shaping speaker-specific social and linguistic inferences. *Cogn. Sci.*, 49(7):e70086, July 2025.

Erdinç Uzun. A novel web scraping approach using the additional information obtained from web pages. *IEEE Access*, 8:61726–61740, 2020. doi: 10.1109/ACCESS.2020.2984503.

R.C. Veltkamp. Shape matching: similarity measures and algorithms. In *Proceedings International Conference on Shape Modeling and Applications*, pp. 188–197, 2001. doi: 10.1109/SMA.2001.923389.

John Vince. Barycentric coordinates. In *Mathematics for Computer Graphics*, pp. 407–433. Springer, 2025.

Dinesh Kumar Vishwakarma, Deepika Varshney, and Ashima Yadav. Detection and veracity analysis of fake news via scrapping and authenticating the web search. *Cognitive Systems Research*, 58:217–229, 2019. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2019.07.004. URL https://www.sciencedirect.com/science/article/pii/S1389041719301020.

Lihui Wang, Robert Gao, József Váncza, Jörg Krüger, Xi Vincent Wang, Sotiris Makris, and George Chryssolouris. Symbiotic human-robot collaborative assembly. *CIRP annals*, 68(2):701–726, 2019.

Zhou Wang and A.C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9 (3):81–84, 2002. doi: 10.1109/97.995823.

Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep lucas-kanade homography for multimodal image alignment, 2021. URL https://arxiv.org/abs/2104.11693.