# From Rosetta to Match-Up: A Paired Corpus of Linguistic Puzzles with Human and LLM Benchmarks

**Neh Majmudar[1], Anne Huang[2], Jinfan Frank Hu[3], Elena Filatova[1]**
[1]City University of New York (CUNY), [2]Davidson Academy, [3]Phillips Academy
nmajmudar@gradcenter.cuny.edu

**Abstract**

In this paper, we examine linguistic puzzles used in high school linguistics competitions, focusing on two common formats: *Rosetta Stone* and *Match-Up*. We propose a systematic procedure for converting existing Rosetta Stone puzzles into corresponding Match-Up counterparts. Because linguistic puzzle creation is complex and time-consuming, our method provides an efficient way to accelerate the generation of new puzzles. We evaluate the resulting Rosetta Stone–Match-Up pairs with both human participants and large language models (LLMs). Our results show that both expert human solvers and LLMs display an all-or-nothing pattern on Match-Up puzzles, either solving them completely or failing entirely. This work contributes a new dataset of paired puzzles and provides a detailed evaluation of puzzle difficulty across formats, offering insights into both human and machine linguistic reasoning.

**Keywords:** Linguistic Puzzles Formats; Benchmarking; Language Resources; LLM Evaluation

## 1. Introduction

Linguistic puzzles designed for high school–level competitions, such as the International Linguistics Olympiad (IOL)[1] and various national contests, are now used not only to assess the skills of high school students and other linguistics enthusiasts but also as benchmarks for evaluating the performance of Large Language Models (LLMs) (Bean et al., 2024). Thus, studying these puzzles serves a dual purpose: advancing the popularization of linguistics and providing a testbed for both the technical capabilities and the creative potential of LLMs.

One persistent challenge for both human solvers and LLMs is the relatively limited supply of existing puzzles. Creating high-quality puzzles is a creative and engaging process but also labor-intensive, often requiring the expertise of highly skilled linguists to ensure validity. This difficulty is further compounded by the absence of formal, widely accepted criteria for evaluating puzzle quality (Gleason, 1955; Zaliznyak, 1963; Zhurinsky, 1993; Bozhanov and Derzhanski, 2013).

To ensure fairness for all competition participants, puzzles created for both national and international linguistics competitions are typically based on languages unlikely to be familiar to them. The puzzle problem statements are written in the national language of the host country, and participants are not expected to know any foreign languages. In this project, we focus on puzzles prepared for English-speaking participants.

Because generating complex and engaging linguistic puzzles requires a high degree of creativity, the puzzle generation task cannot be fully formalized. This limitation places competition, level puzzle generation beyond the current capabilities of even the most advanced LLMs. Each linguistic puzzle can be characterized along several dimensions, including its difficulty, central linguistic topic, language, and format. In this work, we focus on the last of these—puzzle format—and seek to answer the following question: does each format require its own dedicated generation procedure, or can a single approach accommodate multiple formats?

There are several established formats of linguistic puzzles, including Rosetta Stone, Match-Up (also known as Chaos), Monolingual, Pattern, Computational, and Text. The definitions below are taken *verbatim* from the UKLO website.[2]

- **Rosetta**: The data are sets of corresponding words or phrases between different languages/writing systems, with most of the correspondences given. Parts may be omitted from the data set, leaving gaps to be filled. You must be able to give new correspondences (typically translations) to solve the task.
- **Match-Up**: The data are sets of corresponding words or phrases in multiple languages/writing systems, but with few of the correspondences given. If some words are not part of a set, it can still count as a match-up. You must be able to give new correspondences (typically translations) to solve the task.
- **Monolingual**: The data are texts in an unknown language (or equivalent), with no direct translation ( or transliteration for writing systems) given, with the possible exception of 1-2

[1] https://ioling.org/

[2] https://www.uklo.org/technical-information/#qformat



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6418–6427
11–16 May 2026. 

The Gilbertese puzzle used in UKLO in 2018. This puzzle has two difficulty scores: its score for the Breakthrough participants is $34\%$ and its score for the Foundation participants $59\%$; its linguistic topic is syntax; its language family is Austronesian, Oceanic; its Author is Michael Salter.

| | Gilbertese | English |
|---|---|---|
| 1. | **Ko nakonako ŋkoe** | *You are walking* |
| 2. | **E nakonako te aiine** | *A woman is walking* |
| 3. | **I takaakaro ŋai** | *I am playing* |
| 4. | **E nakonako nakon te titooa Meeri** | *Mary is walking to the store* |
| 5. | **A tekateka irarikin te auti aiine** | *Women are sitting next to the house* |
| 6. | **A tebotebo nakekei n te bong aei** | *People are bathing today* |
| 7. | **I tebotebo inanon te auti ŋai** | *I am bathing in the house* |
| 8. | **A takaakaro inanon te titooa ataei** | *Children are playing in the store* |
| 9. | **Ko tekateka ŋkoe ningaabong** | *You will sit tomorrow* |
| 10. | **E takaakaro irarikin te kawai te ataei n te bong aei** | *The child is playing next to the road today* |
| **Q.5.3** | Translate from English into Gilbertese: | |
| 11. | *Women will play tomorrow.* | |
| 12. | *You are sitting next to the store today.* | |

Table 1: Rosetta Stone Linguistic Puzzle Example

words. You must be able to translate from the language to solve the task.

- **Pattern**: The data are words or sets of forms of words/cognates, conforming to a pattern (possibly with some exceptions). You must be able to give other words conforming to this pattern or identify outliers to solve the task, but (unlike in a Rosetta) there is no translation component.
- **Computational**: The problem data includes a description of a computational or other logical system. To solve the problem, you must be able to analyse and implement this system.
- **Text**: The data are whole texts presented in different languages or scripts, but not subdivided further. To solve the problem, you must use context and other clues to deduce linguistic rules.

Among linguistic puzzle types, Rosetta Stone and Match-Up are the most common (Bozhanov and Derzhanski, 2013). Table 1 presents an example of a Rosetta Stone puzzle, while Table 2 shows a Match-Up puzzle. Both puzzles originally included additional questions; however, for clarity of illustration, we retain here only the parts corresponding to the Rosetta Stone and Match-Up format respectively. The original Rosetta Stone was used in UKLO in 2018.[3] The original Match-Up puzzles was used in UKLO in 2015.[4]

National linguistics competitions, including the United Kingdom Linguistics Olympiad (UKLO) and the North American Computational Linguistics Open Competition (NACLO[5]), are explicitly designed so that no prior knowledge of linguistics or specific foreign languages is required. The UKLO website states that its questions require no prior linguistic training, and archival problem sets demonstrate that puzzles frequently draw on languages that are unfamiliar to most participants, such as Beja, Lezgian, Fur, Saisiyat, and Kavalan. In addition, several UKLO problems are based on constructed languages, including Afrihili, Blazon, Esperanto, Centauri, and Arcutan. Some of these (e.g., Centauri and Arcutan) were created specifically for individual competition problems, while others (e.g., Esperanto and Afrihili) are historically documented attempts to develop regularized international auxiliary languages.

More broadly, linguistic olympiad competitions emphasize the use of low-resource, typologically diverse, or otherwise unfamiliar languages in order to evaluate analytical reasoning rather than memorized linguistic knowledge. This design principle makes linguistics puzzles a particularly valuable benchmark for evaluating large language models (LLMs), which may otherwise rely on parametric knowledge of widely documented languages instead of performing genuine pattern induction from the data provided.

A quantitative survey of UKLO problem sets further illustrates this design choice. Among 235 problems published on the UKLO website between 2010 and 2025, we identified 206 unique languages. Of

[3] https://www.uklo.org/wp-content/uploads/2022/05/2018_5-Gilbertese.pdf

[4] https://www.uklo.org/wp-content/uploads/2022/05/2015_3.-Polish.pdf

[5] https://naclo.org/

The Polish puzzle used in UKLO in 2015. This puzzle has two difficulty scores: its score for the Breakthrough participants is $58\%$ and its score for the Foundation participants $75\%$; its linguistic topic is syntax; its language family is Indo-European, Balto-Slavic; its Author is Daniel Rucki.

| | Polish | | English |
|---|---|---|---|
| A | **Alicja zobaczyła sąsiada.** | 1 | *The cat saw the mouse.* |
| B | **Kot zjadł kiełbasę.** | 2 | *Peter bought the sausage.* |
| C | **Piotr kupił kiełbasę.** | 3 | *Alice bought the cheese.* |
| D | **Mysz zobaczyła sąsiada.** | 4 | *Alice saw the neighbour.* |
| E | **Kot zobaczył mysz.** | 5 | *The mouse saw the neighbour.* |
| F | **Alicja kupiła ser.** | 6 | *The cat ate the sausage.* |

**Q.3.1.** Pair each Polish sentence with its English translation in the table below; for example, if you think Polish sentence A is translated by English sentence 1, write '1' in the box below A.

| Polish | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| English | | | | | | |

Table 2: Match-Up Linguistic Puzzle Example

these, 186 languages appear in only a single puzzle, 16 languages appear in two puzzles, and only 3 languages appear in three puzzles. Modern English is explicitly specified as the primary object of analysis in 8 puzzles; however, only one of these is a Rosetta Stone–style problem, involving the encoding of English using a four-digit cipher. The overwhelming predominance of one-off language usage underscores the deliberate avoidance of repetition and prior familiarity, reinforcing the competition's emphasis on in-problem reasoning rather than accumulated language knowledge.

The UKLO puzzles are fairly balanced across the major linguistic domains of phonology, semantics, morphology, syntax, and writing systems. Computational puzzles and those involving number systems are less frequent. It should be noted that many puzzles span multiple linguistic topics rather than fitting neatly into a single category.

In this work, we investigate if the Rosetta Stone and Match-Up formats represent fundamentally the same type of linguistic puzzles viewed from two different perspectives. Answering this question can inform whether a single generation procedure could suffice for both of these formats or whether separate, format-specific procedures are necessary.

For our study, we use a collection of Rosetta Stone puzzles along with their solutions that are listed on the UKLO website. For each puzzle, we take the original problem statement, the associated questions, and the provided answers, and transform them into a corresponding Match-Up puzzle. One outcome of this work is a curated corpus of Rosetta Stone puzzles paired with their corresponding Match-Up versions.

We then select two subsets of the Rosetta Stone, Match-Up puzzle pairs and ask two high school students, both experienced linguistic puzzle solvers, to solve the puzzles from these subsets. Our results show no performance difference (drop) for the solvers when working with original Rosetta Stone versus synthetic Match-Up puzzles. We subsequently provide the complete set of Rosetta Stone/Match-Up pairs to LLMs for evaluation.

We also document the strategies the human solvers use when solving puzzles. We see three key benefits to recording these strategies:

1. They provide insight into what constitutes a well-designed linguistic puzzle.
2. They can inform and guide the development of future puzzle-generation procedures.
3. They can support the analysis of LLMs' chain-of-thought reasoning, helping to better understand differences in decision-making between humans and LLMs.

The rest of the paper is organized as follows:

- Section 2 reviews prior work on linguistic puzzle corpus creation and summarizes research on solving linguistic puzzles by both humans and large language models (LLMs).
- Section 3 outlines the methodology we propose to convert Rosetta Stone puzzles into Match-Up format and details the resulting corpus of paired puzzles.
- Section 4 presents experiments on solving the Rosetta Stone and Match-Up puzzles from Section 3, comparing expert human and LLM performance on a shared subset and evaluating LLMs on additional pairs excluded from the human study.
- Section 5 reports findings from follow-up interviews with the expert human solvers, documenting the strategies they employed to solve puzzles of different types.

- Section 6 summarizes the contributions of this paper and concludes that, overall, a single generation procedure can be used to produce both Rosetta Stone and Match-Up puzzles, achieving varying levels of success across different linguistic topics.

## 2. Related Work

### 2.1. Existing Corpora

The International Linguistics Olympiad (IOL), along with many national linguistic competitions, releases past puzzles together with their official solutions. Examples include NACLO (North America),[6] OzCLO (Australia),[7] UKLO (United Kingdom),[8] etc.

The problems published by national competitions are incorporated into several datasets. For example, the LINGOLY dataset (Bean et al., 2024) consists of puzzles originally created for UKLO. For each competition puzzle, UKLO provides solutions and a set of descriptive attributes, including puzzle difficulty (foundation, intermediate, advanced, etc.), linguistic topic (e.g., writing systems, morphology), question format (e.g., Rosetta Stone, Match-Up), language family, and other metadata.

A recent work, LINGOLY-TOO (Khouja et al., 2025), is an extension of the LINGOLY corpus. The LINGOLY-TOO corpus builds on the LINGOLY dataset by generating new puzzle variants with systematically introduced orthographic obfuscation.

An analysis of IOL puzzles (Bozhanov and Derzhanski, 2013) shows that the Rosetta Stone format is the most frequently used puzzle type. Moreover, IOL results indicate that "experienced solvers are better prepared to handle these [Rosetta Stone puzzles] than problems of other types."

The *PuzzLing Machines* dataset is a carefully curated resource consisting exclusively of **Rosetta Stone** puzzles from linguistic competitions for high school students across various countries (Şahin et al., 2020).[9] The dataset "contains 96 unique puzzles from 81 languages that span 39 different language families from all over the world, as well as two creoles and two artificial languages."

MODELING is another dataset that contains only **Rosetta Stone** puzzles (Chi et al., 2024). However, in contrast to the *PuzzLing Machines* dataset, MODELING includes newly created puzzles authored by expert puzzle writers. The main goal of the MODELING corpus is to create puzzles specifically for low-resource languages.

[6] https://naclo.org/practice.php

[7] https://ozclo.org.au/past-problems/

[8] https://www.uklo.org/past-exam-papers/

[9] https://ukplab.github.io/PuzzLing-Machines/

### 2.2. LLMs and Linguistic Puzzles

Sahin et al. (Şahin et al., 2020) demonstrate that for the *PuzzLing Machines* dataset "both simple statistical algorithms and state-of-the-art deep neural models perform inadequately on this challenge".

Over the past couple of years modern large language models (LLMs) have demonstrated impressive efficiency across a wide range of tasks (Minaee et al., 2024). In text-related domains—such as understanding and analysis, generation and transformation, and conversational interaction—LLMs often outperform traditional pre-trained language models (Zhou et al., 2024).

Linguistic puzzle datasets are increasingly being used as specialized benchmarks for evaluating the performance of large language models (LLMs). These puzzles are deliberately designed to be self-contained, ensuring that no external knowledge is required to solve them. This property makes linguistic puzzles an ideal testbed for assessing the reasoning capabilities of LLMs. Such evaluations are particularly informative when the puzzles involve low-resource languages (Chi et al., 2024).

The performance of LLMs on linguistic puzzle benchmarks depends heavily on the availability of resources for the language featured in the puzzles: "the higher the resource level of the language, the better the scores" (Bean et al., 2024). LLM performance also depends on the linguistic topic at the core of the puzzle. Recent models often surpass human performance on topics such as compounding, number systems, morphology, phonology, semantics, and syntax. However, when puzzles are designed to test the ability to decipher rare or unfamiliar writing systems, humans still consistently outperform LLMs (Majmudar and Filatova, 2025).

In this paper, we use UKLO Rosetta Stone puzzles and convert them into corresponding Match-Up versions. We then evaluate both formats by comparing the performance of human solvers and large language models (LLMs) on each puzzle, examining how performance differs between the original Rosetta Stone puzzles and the generated synthetic Match-Up counterparts.

## 3. Corpus

In this project, we investigate whether Rosetta Stone and Match-Up puzzles represent genuinely distinct formats or simply reflect different perspectives on the same underlying puzzle structure. We focus on these two formats because they are the most frequently used in linguistic competitions. According to data released by UKLO, 45% of all competition puzzles are of the Rosetta Stone type, while 28% are Match-Up puzzles.

In our work, we treat each puzzle as a single unit. This approach differs from that adopted in

the LingOly corpus, where each question within a puzzle is treated as an independent unit of analysis. For example, LingOly would consider the two translation questions shown in Table 1 as separate items, whereas we evaluate overall performance across both questions within the same puzzle. Although it is possible to assess human and LLM performance on individual questions, participants in linguistic competitions cannot solve these questions in isolation. Instead, they must consider all examples (questions) within a puzzle collectively in order to infer the linguistic patterns necessary for solving the puzzle. Because one of the goals of this work is to determine if different puzzle formats require distinct generation procedures, it is essential to analyze each puzzle as a coherent whole.

### 3.1. Conversion Procedure: from Rosetta Stone to Match-Up Format

We begin by collecting the Rosetta Stone puzzles from UKLO. In addition to the problem statements, we also compile the corresponding answers to the translation questions. For example, the information presented in Table 1 is supplemented with the solutions to translation question **Q.5.3** (translate from Gilbertese to English). Specifically,

14. *Women will play tomorrow*
    Solution: **A takaakaro aiine ningaabong**

15. *You are sitting next to the store today*
    Solution: **Ko tekateka irarikin te titooa ŋkoe n te bong aei**

In total, the final version of the puzzle consists of 12 pairs of Gilbertese sentences and their corresponding English translations. We then randomly shuffle the English sentences, producing a set of 12 Gilbertese sentences and a corresponding set of 12 shuffled English sentences labeled A through L. This transformation results in a Match-Up puzzle, where the task is to match each Gilbertese sentence with its correct English translation.

In addition to the aligned text pairs used in Rosetta Stone puzzles, each UKLO puzzle includes a short contextual description (preamble), a brief informational note about the language under analysis. This preamble typically mentions the language family of the target language or highlights particular linguistic features that may or may not be required for solving the puzzle. When generating a Match-Up puzzle corresponding to a Rosetta Stone puzzle, we preserve this preamble information. Consequently, the final version of the Match-Up puzzle corresponding to the 2023 Gilbertese puzzle includes the following preamble:

> The Gilbertese language is an Austronesian language spoken in Kiribati, a country consisting of a number of islands lying to the northeast of Australia.

Thus, the Match-Up Gilbertese puzzle that corresponds to the Rosetta Stone puzzle presented in Table 1 is presented in Table 3.

The solution to the Match-Up puzzle is a table with matched pairs. For example, the answer for the Polish puzzle from Table 2 is the following:[10]

| Polish | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| English | 4 | 6 | 2 | 5 | 1 | 3 |

It must be noted that not all puzzles contain complete sentences. For example, the Rosetta Stone Permyak puzzle used in the 2023 UKLO[11] consists of pairs of strings shorter than full sentences. Here are the Permyak–English text string pairs:

| **Permyak** | *English* |
|---|---|
| **k'erkulan'** | *towards the house* |
| **pɨzannezɨtlən** | *of your (sg.) desks* |
| **ponɨt** | *your (sg.) dog* |
| **purtnɨs** | *their knife* |
| **kəinnezɨs** | *his wolves* |
| **vərələn** | *of my forest* |
| **purtəla** | *for the sake of my knife* |
| **tɨezɨtkət** | *with your (sg.) lakes* |
| **k'erkuezlis'** | *from the houses* |
| **jus's'ezə** | *my swans* |
| **kokɨskət** | *with his foot* |
| **k'ɨɨtlan'** | *towards your (sg.) hand* |

### 3.2. Corpus Description

To construct our corpus, we collected 96 Rosetta Stone puzzles published on the UKLO website as of October 2025, excluding those from 2010 and 2011 due to the absence of participant performance data. From this set, 30 puzzles (with their corresponding Match-Up conversions) were selected for the experiments, based on the design of the human evaluation study (see Section 4.1). The experimental subset is balanced with respect to puzzle difficulty and linguistic topic.

The full dataset contains 192 puzzle files, corresponding to 96 Rosetta Stone puzzles and their Match-Up counterparts. Each entry includes a preamble, the puzzle questions, and reference answers. In both the human and LLM experiments, only the preamble and questions are provided to participants, while the reference answers are used solely for evaluation. We rely on the puzzle format labels (e.g., Rosetta Stone) provided by the

[10] https://www.uklo.org/wp-content/uploads/2022/05/2015_3.-Polish.pdf

[11] https://www.uklo.org/wp-content/uploads/2023/03/2023_R1_5-Permyak.pdf

The Gilbertese language is an Austronesian language spoken in Kiribati, a country consisting of a number of islands lying to the northeast of Australia. Below are some sentences in Gilbertese, followed by their English translations in a random order.

| | Gilbertese | | English |
|---|---|---|---|
| 1 | **Ko nakonako ŋkoe** | A | *Women will play tomorrow* |
| 2 | **E nakonako te aiine** | B | *You are walking* |
| 3 | **I takaakaro ŋai** | C | *A woman is walking* |
| 4 | **E nakonako nakon te titooa Meeri** | D | *People are bathing today* |
| 5 | **A tekateka irarikin te auti aiine** | E | *You are sitting next to the store today* |
| 6 | **A tebotebo nakekei n te bong aei** | F | *You will sit tomorrow* |
| 7 | **I tebotebo inanon te auti ŋai** | G | *Mary is walking to the store* |
| 8 | **A takaakaro inanon te titooa ataei** | H | *I am bathing in the house* |
| 9 | **Ko tekateka ŋkoe ningaabong** | I | *Children are playing in the store* |
| 10 | **E takaakaro irarikin te kawai te ataei n te bong aei** | J | *Women are sitting next to the house* |
| 11 | **A takaakaro aiine ningaabong n te bong aei** | K | *The child is playing next to the road today* |
| 12 | **Ko tekateka irarikin te titooa ŋkoe n te bong aei.** | L | *I am playing* |

**Q.1** Determine the correct correspondence. (A to L)

| Gilbertese | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| English | | | | | | | | | | | | |

Table 3: Match-Up Pair for the 2023 UKLO Gilbertese Rosetta Stone puzzle.

original puzzle authors. However, not all UKLO puzzles strictly follow the canonical Rosetta Stone format discussed in this paper. As a result, several puzzles cannot be converted into Match-Up format. Any deviations or special structural features are explicitly documented.

## 4. Solving Rosetta Stone and Match-Up Puzzle Pairs

Using the generated corpus (Section 3), we evaluate how converting Rosetta Stone puzzles into the Match-Up format affects human and LLM performance on puzzle-solving tasks.

### 4.1. Human Evaluation Experiment

For the human evaluation, we engaged two accomplished Linguistic Olympiad participants with NACLO experience but no prior exposure to UKLO puzzles.[12] The evaluators were not informed that the Match-Up puzzles were derived from the corresponding Rosetta Stone versions and were simply asked to solve a set of linguistic puzzles.

The number of human annotators was limited due to the specialized expertise required for the task. Linguistic puzzle solving, particularly across multiple languages, linguistic domains, and difficulty levels, demands advanced analytical skills and prior experience with competition-style puzzles (e.g., foundation through advanced levels). Consequently, annotators were selected based on demonstrated proficiency in solving such puzzles rather than general linguistic background alone.

Our evaluation protocol did not aim to measure inter-annotator agreement over multiple competing solutions, as is common in labeling tasks. Instead, the goal was to assess *solvability*. Specifically, we evaluated whether synthetic Match-Up puzzles, automatically converted from Rosetta Stone puzzles, could be successfully solved by qualified human solvers. The outcome measure was binary at the puzzle level: a puzzle was considered solvable if at least one expert solver arrived at a correct and complete solution using only the information provided in the puzzle. Under this framework, difficulty is orthogonal to validity. While some puzzles may require substantial reasoning effort, the existence of a correct solution obtained independently by a qualified solver provides evidence that the generated puzzle is well-formed and solvable.

This evaluation design aligns with the primary objective of the study: to determine whether the conversion procedure preserves logical structure and inferential sufficiency, rather than to assess solution variability across individuals.

[12] It must be pointed out that there is a small set of puzzles that were used in both NACLO and UKLO.

The experiment consisted of two stages distinguished by puzzle difficulty. UKLO defines five difficulty levels: *Breakthrough*, *Foundation*, *Intermediate*, *Advanced*, and *Round 2*, with some puzzles spanning adjacent levels. Stage 1 used easier puzzles labeled *Breakthrough/Foundation* or *Foundation/Intermediate*, while Stage 2 used more complex *Advanced* or *Round 2* puzzles.

To ensure that all parameters other than difficulty remain constant, both stages include puzzles centered on the same linguistic topic or combination of topics. Each stage includes two puzzles per linguistic topic (or topic combination). The lists below present the linguistic topics used in the human evaluation experiments and the number of puzzles selected for each stage.

| **Topic** | *Stage 1* | *Stage 2* |
|---|---|---|
| Morphology | 2 | 2 |
| Syntax | 2 | 2 |
| Syntax and Morphology | 2 | 2 |
| Syntax and Semantics | 2 | 0 |
| Semantics, Morphology, and Syntax | 0 | 2 |

As shown in the table above, for the **syntax and semantics** combination in Stage 1, we were unable to identify puzzles with the exact same topic combination for Stage 2. Therefore, in Stage 2 (the more difficult set), instead of using puzzles focused solely on **syntax and semantics**, we include puzzles that combine **morphology, syntax, and semantics**.

The study includes 16 puzzle pairs (32 puzzles), evenly split across two stages. Evaluators never see both puzzles from the same pair; for each topic per stage, they solve one Rosetta Stone and one Match-Up puzzle generated from different source puzzles on the same topic.

Table 4 shows human and LLM performance on the 16 Rosetta Stone/Match-Up puzzle pairs, reported as the percentage of correctly answered questions per puzzle following UKLO's convention. Details of the LLM evaluation appear in Section 4.2.

Table 4 compares average UKLO performance (**UKLO**) with that of the two human evaluators on the same puzzles. For dual-level puzzles (e.g., Breakthrough/Foundation), the higher UKLO score is used. Because some UKLO puzzles, such as the 2024 Coptic puzzle,[13] include additional questions, UKLO's aggregate scores may reflect tasks beyond the Rosetta Stone format. Columns **HE (RS)** and **HE (MU)** report evaluator accuracy on the original and converted puzzles, respectively.

As expected, average performance for both UKLO participants and our evaluators declines from Stage 1 to Stage 2 due to the higher difficulty of the Stage 2 puzzles.

[13] https://www.uklo.org/wp-content/uploads/2024/03/2024-R2_4-Coptic.pdf

The results in Table 4 show that overall human performance remains stable when solving the synthetic Match-Up puzzles, with no decline relative to the original Rosetta Stone puzzles. Notably, we use a strict scoring procedure for the Rosetta Stone puzzles, counting only answers that exactly match the official UKLO key. Interestingly, the Match-Up puzzles display an all-or-nothing performance pattern among human evaluators.

## 4.2. Large Language Models Experiments

We evaluate two state-of-the-art LLMs—OpenAI's GPT-5 (OpenAI, 2025) and Google's Gemini 2.5-Pro (Comanici et al., 2025), under identical zero-shot settings. For Rosetta Stone puzzles, models receive a preamble and translation strings; for Match-Up puzzles, they are given a preamble and two sets of strings to match.

### 4.2.1. Experiment 1

Table 4 reports the average performance across two puzzles at each stage for each linguistic topic (or topic combination) for OpenAI's GPT-5 (column GPT-5) and Google's Gemini 2.5-Pro (column Gem 2.5-Pro). These are the two SOTA publicly available LLMs at the time of the experiments (October 2025). When evaluating Match-Up puzzles, we follow the strict evaluation procedure described in (Majmudar and Filatova, 2025): if a model's output is presented in a perfectly alphabetical order, we assign a score of 0 to such responses, even if some matches are accidentally correct.

For both LLMs, performance on the Rosetta Stone puzzles, similar to that of human participants, generally decreases from Stage 1 to Stage 2. Overall, Gemini 2.5-Pro outperforms GPT-5 across all linguistic topics except one, namely Syntax. However, for the synthetic Match-Up puzzles, the results across the two stages and linguistic topics are more variable and warrant further investigation.

The two most noteworthy cases are those involving the Morphology topic and the combination of Semantics, Morphology, and Syntax topics. For Morphology, both models achieve 100% accuracy on Stage 1; however, their performance on Stage 2 drops dramatically to 2% and 0%, respectively. For the Semantics/Morphology/Syntax combination, both models perform relatively poorly on Stage 1 (18.5% and 16%), while on Stage 2, GPT-5's accuracy falls further to 4%, whereas Gemini 2.5-Pro's accuracy increases sharply to 100%. Following (Bean et al., 2024), we assume that these contrasting results are influenced by the specific set of languages used in the puzzles for this combination of linguistic topics.

| Topic | Stage | UKLO | GPT5 | Gem2.5-pro | HE (RS) | GPT5 | Gem2.5-pro | HE (MU) |
|---|---|---|---|---|---|---|---|---|
| | | Rosetta Stone Original Puzzle | | | | Match-Up Conversion | | |
| Morphology | s1 | 76.5 | 97.5 | 100 | 100 | 100 | 100 | 96 |
| | s2 | 39.5 | 77.5 | 93 | 97.5 | 2 | 0 | 50 |
| Syntax, | s1 | 47 | 81.5 | 85.5 | 94 | 100 | 100 | 100 |
| Morphology | s2 | 21 | 71 | 87 | 97 | 22 | 96 | 100 |
| (Morphology), Syntax | s1 | 54.5 | 78.5 | 84.5 | 87.5 | 18.5 | 16 | 100 |
| Semantics | s2 | 37.5 | 67.5 | 72 | 91 | 4 | 100 | 50 |
| Syntax | s1 | 64 | 90 | 81.5 | 81.5 | 89.5 | 89.5 | 100 |
| | s2 | 52 | 64 | 72 | 89 | 56.6 | 63 | 100 |

Table 4: Average Scores by Linguistic Topic and Stage on 16 Pairs of Puzzles. UKLO - The average human performance reported on the UKLO website; GPT5 and Gem2.5-pro - The average performance by GPT5 and Gemini2.5-pro respectively; HE (RS) - The average performance of the project human evaluators on the UKLO Rosetta Stone puzzles; HE (MU) - The average performance of the project human evaluators on the synthetic Match-Up puzzles.

| Topic | Difficulty | UKLO | GPT5 | Gem2.5-pro | GPT5 | Gem2.5-pro |
|---|---|---|---|---|---|---|
| | | Rosetta Stone Original Puzzle | | | Match-Up Conversion | |
| Syntax, Morphology | s1 | 61 | 84.6 | 84.6 | 90 | 15 |
| | s2 | 16 | 98.1 | 96.3 | 100 | 100 |
| | s2 | 80 | 100 | 87.9 | 100 | 100 |
| | s2 | 32 | 56.4 | 64.1 | 0 | 0 |
| | s2 | 12 | 62.5 | 64.6 | 0 | 100 |
| Syntax | s1 | 45 | 100 | 100 | 100 | 100 |
| | s1 | 70 | 87.5 | 91.7 | 100 | 100 |
| | s2 | 96 | 100 | 95.8 | 100 | 100 |
| | s2 | 46 | 90.0 | 100 | 30 | 100 |
| Morphology | s1 | 44 | 62.5 | 95.8 | 0 | 0 |
| | s2 | 27 | 93.7 | 90.6 | 29.4 | 0 |
| | s2 | 31 | 93.3 | 68.9 | 100 | 100 |
| Morphology, Semantics | s2 | 24 | 70.8 | 56.2 | 5.3 | 5.3 |
| Phonology | s2 | 60 | 93.3 | 68.9 | 100 | 100 |

Table 5: Scores by Linguistic Topic and Stage on 14 Pairs of Puzzles. UKLO - The human performance reported on the UKLO website; GPT5 and Gem2.5-pro - performance by Chat-GPT5 and Gemini 2.5-pro respectively.

#### 4.2.2. Experiment 2

In addition to the 16 puzzle pairs evaluated in Table 4, we apply OpenAI's GPT-5 and Google's Gemini 2.5-Pro to an additional set of 14 pairs of original UKLO Rosetta Stone puzzles and corresponding synthetic Match-Up puzzles. The results of running OpenAI's GPT-5 and Google's Gemini 2.5-Pro on the additional set of 14 pairs of puzzles are displayed in Table 5. Due to the nature of the UKLO dataset, it is impossible to create an additional balanced set of linguistic puzzles across the difficulty and the linguistic topic dimensions. Moreover, not all Rosetta Stone puzzles can be easily translated into the corresponding Match-Up puzzles. For example, the 2014 Kairak problem[14] has three types of verbal patterns translated into English and cannot be converted into a Match-Up format following the procedure described in Section 3.

Table 5 reports results for 14 pairs of UKLO Rosetta Stone puzzles and their corresponding Match-Up versions, organized by linguistic topic

[14] https://www.uklo.org/wp-content/uploads/2022/08/2014.6-Kairak.pdf

and stage (Stage 1 or Stage 2; see Section 4.1). For puzzles assigned to two difficulty levels, the higher UKLO score is used. The table includes GPT-5 and Gemini 2.5-Pro performance on both puzzle types. As in Table 4, LLMs outperform UKLO participants on Rosetta Stone puzzles and show comparable results to each other. For Match-Up puzzles, both models exhibit an all-or-nothing pattern, particularly on morphology-related puzzles. We believe it might be due to the shorter text strings typical of this topic (see the 2013 Permyak puzzle example in Section 3.1). The all-or-nothing trait exhibited by both humans and LLMs requires further investigation.

## 5. Approaches Towards Solving Linguistic Puzzles

After completing the experiment described in Section 4.1, the human evaluators were interviewed about their experience with puzzles of different formats and the strategies they used to solve them. At this stage, the evaluators were not aware that the Match-Up puzzles had been derived from the corresponding Rosetta Stone puzzles. Both evaluators confirmed that the synthetic Match-Up puzzles appeared to be plausible linguistic puzzles.

The approaches used to solving puzzles varied depending on the puzzle format and the linguistic topic involved. Below we summarize several strategies commonly employed by the evaluators when solving Match-Up puzzles.

**Character and word count** While languages may encode the same meaning using strings of different lengths, human solvers often rely on the heuristic that longer strings correspond across languages. This approach can be helpful but not always reliable. Match-Up puzzles with few tokens (e.g., the UKLO 2013 Permyak puzzle in Section 3.1) are especially difficult and reward familiarity with the language or its family. Such puzzles frequently center on morphology, the category that notably produces the "all-or-nothing" performance pattern observed in both humans and LLMs.

**Repeating Matching Strings** Another effective strategy involves counting occurrences of identical substrings that are likely to correspond to the same English words in translation.

**Use of proper names** Proper names frequently serve as anchors for identifying correspondences between languages. For instance, in Table 1, English *Mary* corresponds to Gilbertese *Meeri*; in Table 2, English *Alice* corresponds to Polish *Alicja*, and *Peter* corresponds to *Piotr*. Proper names thus often provide initial clues in both Rosetta Stone and Match-Up puzzles, an observation that should be taken into account when designing automatic methods for linguistic puzzle generation.

## 6. Conclusion

We test the hypothesis that two linguistic puzzle formats, Rosetta Stone and Match-Up, represent complementary views of the same underlying structure. To investigate this, we develop a systematic conversion procedure and apply it to create a corpus of 96 paired puzzles, each consisting of an original Rosetta Stone puzzle and its Match-Up counterpart.

We evaluate the quality of the generated Match-Up puzzles through a series of experiments on 30 puzzles. Sixteen of the 30 puzzle pairs are solved by human evaluators, and all 30 are solved by large language models (LLMs). Based on the experimental results and follow-up interviews with the evaluators, we conclude: the proposed conversion procedure can effectively generate Match-Up puzzles from existing Rosetta Stone puzzles. However, further research is required to assess its applicability across a broader range of linguistic topics.

The experiments reveal an all-or-nothing performance pattern in both humans and LLMs, particularly on morphology-focused puzzles. This pattern suggests that Match-Up puzzles capture a distinct type of linguistic reasoning. Together, the dataset and methodology introduced here provide a foundation for automated puzzle generation and for future studies comparing human and machine problem-solving across languages and puzzle formats.

## 7. Ethics Statement

This study involved two experienced high-school lingusitic puzzle solvers. Written informed consent (and parental/guardian consent for minors) was obtained; no personal data were collected or stored. Tasks used publicly available competition puzzles. No sensitive attributes were elicited. The released dataset contains only problem statements, preambles, and answer keys from public sources, with metadata derived from UKLO and converted formats; it contains no personal data.

## 8. Limitations

Human study (N=2) limits inference; language/topic coverage is imbalanced; some Rosetta puzzles (e.g., multi-template verb systems) cannot be converted without additional heuristics; LLM scores may vary across versions and decoding settings; and our strict evaluation for Rosetta (exact-match only) may undercount partial progress.

## 9. Data and Code Availability

The data is realized in:
https://github.com/ef2020/lrec2026-data

## 10. Bibliographical References


Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*.

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: A novel dataset for testing linguistic reasoning in language models. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119, St. Julian's, Malta. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Henry Allan Gleason. 1955. *Workbook in descriptive linguistics*. Publisher Holt, Rinehartand Winston.

Jude Khouja, Karolina Korgul, Simi Hellsten, Lingyi Yang, Vlad Neacsu, Harry Mayne, Ryan Kearns, Andrew Bean, and Adam Mahdi. 2025. Lingolytoo: Disentangling reasoning from knowledge with templatised orthographic obfuscation.

Neh Majmudar and Elena Filatova. 2025. Can LLMs Solve and Generate Linguistic Olympiad Puzzles? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Suzhou, China.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

OpenAI. 2025. GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. [Accessed 13-10-2025].

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254.

Andrey A. Zaliznyak. 1963. Linguistics puzzles (in Russian). In Tatyana N. Moloshnaya, editor, *Structural Typology Research*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2024. A comprehensive survey on pretrained foundation models: A history from BERT to Chat-GPT. *International Journal of Machine Learning and Cybernetics*, abs/2302.09419:1–65.

Alfred N. Zhurinsky. 1993. *Word, Letter, Number: A discussion of self-sufficient linguistic problems with an analysis of a hundred samples of the genre (in Russian)*.