# Evaluating Word Embeddings in Extremely Under-Resourced Languages: A Case Study in Bribri

**Rolando Coto-Solano**
Dartmouth College
rolando.a.coto.solano@dartmouth.edu

## Abstract

Word embeddings are critical for numerous NLP tasks but their evaluation in actual under-resourced settings needs further examination. This paper presents a case study in Bribri, a Chibchan language from Costa Rica. Four experiments were adapted from English: Word similarities, WordSim353 correlations, odd-one-out tasks and analogies. Here we discuss their adaptation to an under-resourced Indigenous language and we use them to measure semantic and morphological learning. We trained 96 word2vec models with different hyperparameter combinations. The best models for this under-resourced scenario were Skip-grams with an intermediate size (100 dimensions) and large window sizes (10). These had an average correlation of r=0.28 with WordSim353, a 76% accuracy in semantic odd-one-out and 70% accuracy in structural/morphological odd-one-out. The performance was lower for the analogies: The best models could find the appropriate semantic target amongst the first 25 results approximately 60% of the times, but could only find the morphological/structural target 11% of the times. Future research needs to further explore the patterns of morphological/structural learning, to examine the behavior of deep learning embeddings, and to establish a human baseline. This project seeks to improve Bribri NLP and ultimately help in its maintenance and revitalization.

## 1 Introduction

Word embeddings are dense vectors that describe the semantics of words (Landauer et al., 1997). They are calculated by collecting the words around a specific word and using them to create a numerical vector that can determine semantic similarity. For example, the words *spinach* and *kale* share neighboring words like *garlic* and *cooked*, which would make them similar to each other and different from words with other neighbors, such as *hammer*. Embeddings can vary in size, or *dimensionality*: Word embeddings from systems like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) can range from size 10 to 300, while the embeddings from GPT-3 (Brown et al., 2020) can have up to 12288 dimensions.

Embeddings are an integral part of a number of NLP tasks. Because these vectors can be used to calculate cosine distance, one could calculate the distance between words like *king* and *queen*, and examine if this is similar to the distance between *man* and *woman*. Learning these patterns entails that the system is learning the gender distinctions in the words and that it is generalizing patterns across words. This would make an embedding a type of language model which could be enlisted to solve tasks like finding the “odd” word in a triad of words. For example, it could determine that *tiger* is the odd word in the set {*man*, *woman*, *tiger*} given its greater cosine distance from the other two. This is known as the *odd-one-out task*. Embeddings can also solve analogies by using vector algebra. The embedding vectors can be used for the operation *king+woman-man*, and the word closest to this result should be *queen*. These two tasks, odd-one-out and analogies, are some of the main tasks used to evaluate the training of embeddings (Schnabel et al., 2015).

The evaluation of embeddings is relevant not just for these specific semantic tasks, but because contemporary deep learning systems like Transformers (Vaswani et al., 2017), which use embeddings for their learning. Embeddings are used as semantically rich input in the training of sequence-to-sequence tasks such as question answering and machine translation. Because embeddings are so widely used it is necessary to understand their behavior in low-resource environments. This would allow researchers to better describe the lexical structures of Indigenous and other minority languages, and it would also help optimize the training of deep learning algorithms for these languages.

*Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467
October 12–17, 2022.

Making NLP for Indigenous languages is a complex task because of the chronic lack of data to train models, so determining the optimal configurations to train embeddings for them can help move these efforts forward. There is relatively little research in this regard (Adams et al., 2017; Stringham and Izbicki, 2020), so this paper seeks to add a case study for how to perform these evaluations in an Indigenous language of the Americas.

## 2 Methodology

This paper will examine the evaluation of word embeddings in extremely under-resourced languages by using a realistic example from an Indigenous language in Central America. The Bribri language (Glottolog `brib1243`) is spoken by approximately 7000 people in Southern Costa Rica (INEC, 2011). It is a Chibchan language and it is vulnerable (Sánchez Avendaño, 2013), which means that some children no longer speak the language. There are efforts to generate NLP tools for Bribri to aid with language maintenance and revitalization efforts, including the construction of digital dictionaries and corpora (Krohn, 2020, 2021; Flores Solórzano, 2017), tools for phonetic analysis (Coto-Solano and Flores-Solórzano, 2016, 2017), machine translation (Feldman and Coto-Solano, 2020; Mager et al., 2021), speech recognition (Coto-Solano, 2021), natural language understanding (Ebrahimi et al., 2021), parsing (Coto-Solano et al., 2021) and morphological analysis (Flores-Solórzano, 2019), as well as tools for the input of the language into digital media (Flores-Solórzano, 2010). The motivation for this paper is to contribute information to analyze the training process of deep learning tools for the language.

### 2.1 Data preparation

It is regularly the case that, when models are trained in low-resource conditions, what is actually presented is a truncated version of a training set from a high-resource language like English as an attempt to simulate low-resource conditions. (Baevski et al., 2020; Sennrich and Zhang, 2019). However this is not realistic because actual under-resourced environments[1] include obstacles in the creation of the training datasets. Not only is data scarce, but it might lack standardization, making the dataset more sparse than it would be for languages with standardized orthographies and numerous speakers. Here we will review the challenges found in creating a dataset to train embeddings for Bribri.

The data collected contains 90,128 words of monolingual Bribri text. It contains 10,328 unique words, 6,071 of which appear as hapax legomena. It has 10,962 sentences, with an average length of 8.2 ± 7.6 tokens. It includes text from two textbooks (Constenla et al., 2004; Jara Murillo and García Segura, 2013), a dictionary (Margery, 2005), a grammar book (Jara Murillo, 2018), several educational books (Sánchez Avendaño et al., 2021b,a), an oral corpus (Flores Solórzano, 2017) and other sources (Ebrahimi et al., 2021). There are numerous challenges in preparing the data, given that the writing has four main sources of variation: (i) Authors use different orthographic conventions, (ii) there are linguistic phenomena that cause variation, (iii) there is dialectal variation, and (iv) there is idiosyncratic variation across authors.

First, there are multiple orthographies in current use. For example, the word ‘tiger’ can be written in at least three different ways: *na̱mù̱*, *ñãmũ̀* and *nąmų̀*. The vowels in this word are nasal, but the nasal mark changes depending on the author. This problem is compounded by the fact that different authors might use different Unicode characters for the diacritics. For example, the line under the vowel can be variously represented as the combinatorial low line (U+0332), the combinatorial minus sign below (U+0320) or the combinatorial macron (U+0331). These variations were standardized in the dataset so that the three orthographic representations of ‘tiger’ could all contribute to the embedding for the word.

Second, because the orthography is not standardized, there is much phonological variation that finds its way into the writing. For example, the word *a̱mì̱* ‘mother’ can be pronounced as *mì̱* in rapid speech, and it frequently appears this way in text. This means that half of the information for the ‘mother’ embedding would be contained in the word *a̱mì̱* and half would be contained in *mì̱*, diluting the embedding for both forms.

Third, there is considerable dialectal variation in the dataset. This is present in every human language, but it is magnified in this small dataset. For example, the word ‘road’ is *ña̱là̱* in the Amubri

[1] The standard terminology is to call these languages *low-resource*. However, they are actually *under-resourced*, as are their communities of speakers. The languages are as fully-fledged as a high-resource language, and contain ample knowledge about the world. They also tend to have speakers and experts willing to compile data. What these languages usually lack are economic resources for the datasets to be created.

dialect and *ñolö̀* in the Coroma dialect. The corpus includes 41 occurrences of *ña̱là̱* and 18 occurences of *ñolö̀*, a 69%-31% distribution. This means that a significant portion of the information for ‘road’ will not be included in the embedding of either alternative.

Finally, there are numerous idiosyncratic differences in spelling. Lack of standardization is no obstacle for communication and should not slow down the development of written materials, but it does impact NLP. There are numerous words that have different spellings in the same document. For example, the corpus contains 5 spellings for *ta̱î̱* ‘much’, and other documents include 9 additional variants such as *tá̱i̱* and *táìn*. This variation made it so that some monolingual documents with major internal variation could not be included.

In addition to these sources of variation, Bribri has a different typology of English, which doesn’t lend itself well to the word embedding architecture. Bribri is morphologically fusional. Its words regularly take suffixes, and as a result, each Bribri verb can have 23 different conjugations and can also appear with numerous clitics attached to it (Flores Solórzano, 2017). This makes the data itself more sparse than it would be for an English text with the same word count.

## 2.2 Embedding Evaluation

Once the corpus is in a standardized form, the next challenge is to adapt embedding evaluation techniques to work on Bribri data. There are numerous obstacles to this, including: (i) the paucity of data in the monolingual corpus and (ii) the mismatches between the English and Bribri vocabulary. This subsection will propose four experiments to evaluate Bribri embeddings taking these challenges into account.

### 2.2.1 Technical Aspects of Training

In order to perform the experiment a total of 96 types of models were trained using the Word2Vec algorithm (Mikolov et al., 2013) implemented in Gensim (Rehurek and Sojka, 2011). This paper performs a systematic examination of different hyperparameters. The models could vary in their type of training (Skip-Gram and CBOW), the size of the embedding (10, 25, 50, 100, 150, 200, 250, 300 dimensions), the window of words used to calculate the embedding (2, 5, 10 words) and the minimum frequency of the tokens considered when constructing the embeddings (n=1 or n=2). Each model was trained 60 times to account for potential variations in the training phase. Therefore, the calculations in the results section use a total of 5760 trained models. The models were trained using a CPU-based Google Colab environment, with a runtime of approximately 8.5 hours.

Four evaluations were performed: (i) A basic similarities test, (ii) a correlation between the Bribri similarities and the similarities in the WordSim353 set (Finkelstein et al., 2001), (iii) a series of odd-one-out tests with both semantically and structurally “odd” words, and (iv) a series of semantic and structural analogies.

### 2.2.2 Basic similarities

The first test served as a kind of sanity test in order to verify that the system was producing meaningful results. For each model, four similarities were calculated: two semantic and two structural. For the semantic case, the word *aláköl* ‘woman’ was compared to the hypothetically similar word *wë̀m* ‘man’ and the hypothetically less similar word *na̱mù̱* ‘feline, tiger’. The woman-man pair should be more similar than the woman-tiger pair.

For the structural case, two pairs of verbs were selected. The system compared the perfective active form of the verb (e.g. *yö’* ‘made’, *ña̱’* ‘ate soft food’) with the perfective middle voice form of the same verb (e.g. *yö̀ne̱* ‘was made’, *ñà̱ne̱* ‘was eaten’).

The similarities were calculated for each of the 96 model types, and then the average and standard deviation for each of the four pairs was calculated.

### 2.2.3 WordSim353 Correlations

The second experiment explores the similarities in the Bribri set and examines if they have the same patterns as other well understood datasets. This experiment will use the word pairs in the WordSim353 dataset (Finkelstein et al., 2001), a set that contains 353 pairs of English words and a measure of their similarity. The experiment calculates the correlation between the WordSim353 pair (e.g. *tiger-cat*, similarity=7.35) and its corresponding Bribri translation (e.g. *na̱mù̱-pûs*, similarity=0.87)[2].

Converting the WordSim353 dataset to be usable in Bribri involved several challenges. First, many

[2] The similarities in WordSimPair353 go from 0 to 10, while the similarities in the Gensim Bribri data go from 0 to 1. A WordSimPair353 score of 5 would be roughly equivalent to a Gensim score of 0.5.

of the pairs had words that were not a part of Bribri culture and did not appear in the corpus (e.g. *Harvard-Yale*). Second, many of the words did not have translations into Bribri (e.g. *vodka-gin*). There are some words from foreign languages that have indeed been borrowed into Bribri (e.g. *banco* 'bank'), but this particular word appeared only once in the corpus so it couldn't be used for the experiment. In the end only 19 pairs could be translated; these are included in Appendix A. We calculated the Pearson correlation between the 19 English-Bribri pairs for the 96 model types. This was done 60 times for each model type, and then the correlations from the 60 runs were averaged for each model type.

#### 2.2.4 Odd-One-Out Testing

The third experiment implements two types of odd-one-out testing: semantic and structural. First, 20 triads were constructed where the odd word would be different from the other two because of its semantic properties (e.g. {*alákö̈l*, *wë̀m*, NA̱MÙ̱} 'woman, man, TIGER'). Second, 20 triads of structurally related words were constructed, where the odd word was different because it had a different part of speech, had a different verbal conjugation, or belonged to a different word paradigm. One example of different verbal forms is the triad {*yö'*, *ña̱'*, KÙ̱NE̱} 'made, ate, WAS FOUND', where the third verb is in the middle voice. One example of words that belong to different paradigms is {*e'töm*, *bö̀töm*, MA̱ÑÁ̱L} 'one.flat, two.flat, THREE.HUMAN'. The first two words are counters for nouns in the flat word class, whereas the third word is a counter for human nouns, and therefore should be more salient. All the semantic and structural triads can be found in Appendix A.

The percentage of correct responses for the odd-one-out task was recorded for each of the 60 runs of each of the 96 model types, and the average for each of the model types was calculated. Two separate averages were considered: Semantic and structural averages.

#### 2.2.5 Analogies

The final experiment is the evaluation of analogies. The first step was to use the BATS dataset (Gladkova et al., 2016) to examine which analogies in English could be translated into Bribri. Some of the semantic relationships could be translated. For example, the words for man:woman :: boy:GIRL were present in the corpus. There were numerous pairs that couldn't be used because Bribri expresses different English concepts with the same word (e.g. 'sir/madam' = *akë̀kë*), or because the English words are underspecified for their translation. For example, Bribri has words for maternal and paternal uncles, but it doesn't have a single word for 'uncle'. Similarly, there were many English language pairs that were not present in the corpus, such as *euler/mathematician*. A total of 20 analogies (40 pairs) were constructed based on the model from BATS; they can be found in Appendix A. These quartets include 4 hypernymic relationships (e.g. 'quetzal:bird :: corn:PLANT'), 3 "place to live" relationships (e.g. 'fish:water :: man:HOUSE'), 4 antonym relationships (e.g. 'small:big :: white:BLACK'), 3 object/action analogies (e.g. 'ball:play :: book:READ'), one holonymic set ('beam:house :: eye:HEAD') and 5 family relationships. The family relationships are particularly important because they had to be adjusted to the Iroquois family system used in Bribri (Constenla et al., 2004). The language has different words for maternal and paternal relatives, but also for different relationships with siblings (e.g. *ë́l* 'sibling of the same sex; cousin from the same sex as the parent', *kutà* 'sister of man', *akë̀* 'brother of a woman'). Therefore, it would be important to calculate if the models can learn these culturally relevant relationships.

Finding morphological/structural pairs was more complex because of the differences between English and Bribri. The two languages share some morphological phenomena such a plurals (e.g. 'child:children :: he.she:THEY'), but the rest of the pairs had to be constructed using the morphology of the language. Some analogies paired a noun with its numeral class (e.g. *alákö̈l:wë̀m* :: *chìchi*:E'TÖM 'woman:one.human :: dog:ONE.FLAT'), while others used verbal conjugations (e.g. *katö̀k:katèke̱* :: *yawö̀k*:YÈKE̱ 'toEat:eating :: toMake:MAKING'). This method of using grammatical phenomena unavailable in English has been used in other adaptations of BATS (Kang and Yang, 2018).

A total of 20 semantic and 20 structural analogies were constructed; the complete list is in Appendix A. These analogies were tested on 60 trial runs for each of the 96 model types. Three results were calculated: (i) The percentage of times that the expected result was the first result of the analogy, (ii) the percentage of times that the expected result was within the first 10 results of the analogy, and (iii) the percentage of times that the expected

| Pair | Translation | Similarity |
|---|---|---|
| *aláköl - wë̀m* | woman - man | 0.92 ± 0.04 |
| *aláköl - na̱mù̱* | woman - tiger | 0.77 ± 0.10 |
| *yö’ - yö̀ne̱* | made - was made | 0.93 ± 0.07 |
| *ña̱’ - ñà̱ne̱* | ate - was eaten | 0.95 ± 0.03 |

Table 1: Semantic and structural similarities across all trained models

| Type | Size | Window | MinFreq | r |
|---|---|---|---|---|
| SG | 25 | 10 | 2 | 0.33 |
| SG | 50 | 10 | 2 | 0.31 |
| SG | 25 | 10 | 1 | 0.30 |
| SG | 50 | 10 | 1 | 0.30 |
| SG | 100 | 10 | 2 | 0.28 |
| **SG** | **100** | **10** | **1** | **0.28** |

Table 2: Average Pearson correlation (r) between Bribri and WordSim353 similarities for top performing models (SG: Skip-gram, MinFreq: Minimum frequency of words included in the model)

result was within the first 25 results of the analogy. These numbers were averaged for each model type.

## 3 Results

### 3.1 Basic Similarities

The first set of results confirms that the model contains some knowledge of the Bribri grammar and lexicon. Table 1 shows the similarity between two semantically related pairs (‘woman - man’, ‘woman - tiger’) and two structural pairs (‘made - was made’, ‘ate - was eaten’), averaged over the 60 runs and the 96 models. These initial results show that the word for ‘woman’ is more similar to ‘man’ (similarity = 0.92) than it is to ‘tiger’ (0.77). The structural pair shows that the system might be learning the morphological relationships between verbal tokens. Here the middle voice verbs show similar distances to their corresponding perfective active voice verbs (similarity$_{\text{‘made}-\text{wasMade}'} = 0.93, similarity_{\text{‘ate}-\text{wasEaten}'} = 0.95$).

### 3.2 WordSim353 Correlations

After having performed a basic check for learning, the next step is to confirm this with more similarities and check that those resemble the similarities in equivalent English models. Figure 1 shows the average correlation between the Bribri pairs and the WordSim353 pairs for each of the 96 model types. Table 2 shows the average correlation for the top performing model types.

The first visible pattern is that Skip-grams have higher performance than CBOW models. The average correlation for CBOW is r=-0.04±0.09, lower than the r=0.17±0.09 obtained for Skip-grams. The best performing models had correlations ranging between r=0.28 and r=0.33 and all of them were trained as Skip-grams[3]. The second pattern is that the best results were obtained with the largest possible window. It might be the case that the corpus is so small that it needs a large window to be able to learn the relationships between words. A third observable pattern is that small to medium embedding sizes have better performance. The best results occur with embeddings that range between 25 and 100 dimensions; smaller and larger embeddings have relatively lower performance. Finally, for this particular test the minimum frequency of the words does not appear to affect the performance of the models: Models have relatively higher correlations regardless of whether the minimum word frequency is one or two. In general, this experiment shows that the models are learning some Bribri semantics. As an example, figure 2 shows the correlation for one of the top performing models: Skip-gram, 100 dimensions, window=10, minimum word frequency=1. As can be seen in the figure, words with higher similarities in English also have relatively higher similarities in Bribri.

### 3.3 Odd-One-Out

In this next step, the performance of the system will be examined for both semantics and structural relationships by calculating the performance of the odd-one-out task. Figure 3 shows the percentage of correct responses for semantic and structural triads for the 96 model types studied. Table 3 shows the average percentages for the top performing models.

The main pattern in the odd-one-out experiment is that the semantic and structural triads had similar performance. For example, the highest performing model has an average of 76% of correct responses for semantics, and a slightly lower 70% for structure. This is an encouraging result, as it seems that the system is learning not just the meaning of Bribri words, but also some concepts about the grammar of the Bribri language.

[3] Table 5 shows the 19 pairs that were considered in the correlations. One of them is the trivial case ‘tiger-tiger’. This was included because it also appears in WordSim353. If this trivial pair is removed the correlations drop by about half. For example, the model with an average of r=0.33 becomes r=0.17. The better-performing models still have higher correlations.

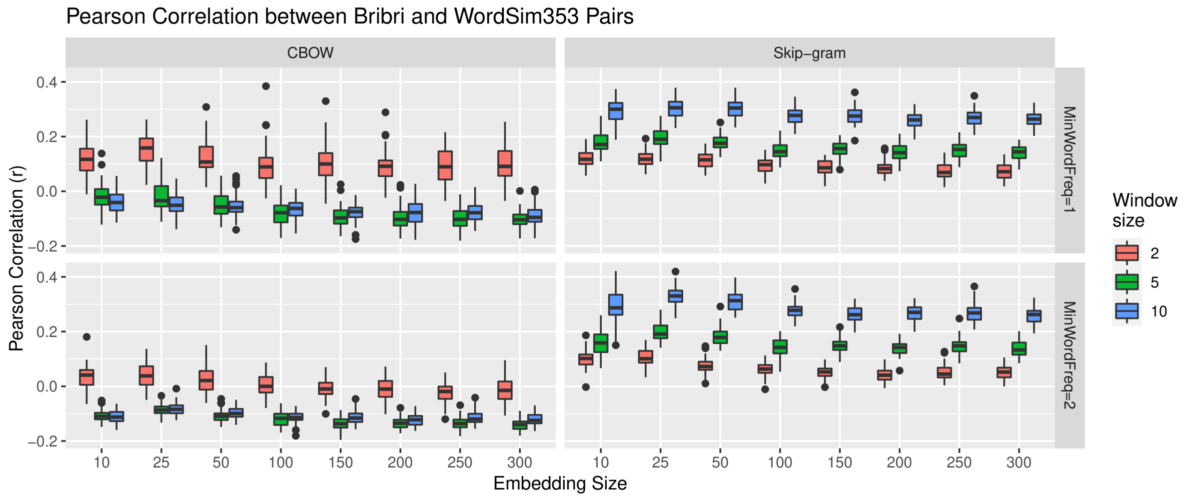

Figure 1: Pearson correlation between WordSim353 similarities and Bribri similarities

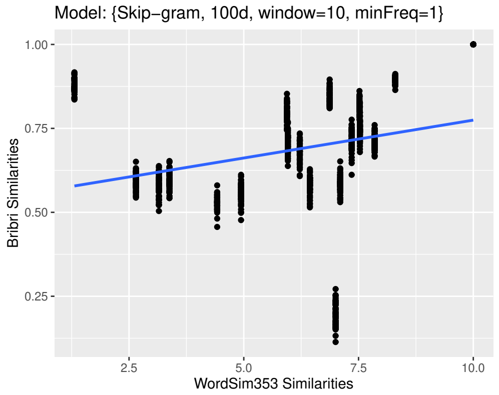

Figure 2: Correlation between WordSim353 similarities and Bribri similarities for a high performing model

| Type | Size | Window | MinFreq | Sem | Str |
|---|---|---|---|---|---|
| **SG** | **100** | **10** | **1** | **76** | **70** |
| SG | 200 | 10 | 1 | 75 | 71 |
| SG | 50 | 10 | 1 | 75 | 68 |
| SG | 25 | 10 | 2 | 75 | 68 |
| SG | 300 | 10 | 1 | 75 | 72 |
| SG | 150 | 10 | 1 | 74 | 71 |

Table 3: Average of correct responses to the Odd-One-Out task for semantic and structural triads (SG: Skip-gram; Sem: % Correct Semantic; Str: % Correct Structural)

Some of the patterns observed in the similarities experiment were also visible here. Skip-grams with a large window of 10 words were the best performing models. In the odd-one-out experiment the dimensionality doesn't seem to determine the performance. On the other hand, it does seem that models that take all the words into account have better performance; almost all high performing models take all the words in the corpus into account (minimum word frequency = 1). Notably, the highest performing model {Skip-gram, 100 dimensions, window=10, minFreq=1} is also amongst the high-performing models for similarities in table 2.

### 3.4 Analogies

The final experiment takes 20 semantic and 20 structural word quartets (e.g. *wë́m*:*alákö̈l* :: *kabë̀*:BÙSI 'man:woman :: boy:GIRL') and tries to calculate the fourth word by performing vector algebra with the first three, in effect performing an analogy. Figure 4 shows the average results for the 96 model types trained. The results indicate how often the target word was found as the first result of the vector algebra operation. They also indicate how often the word was found in the top 10 results and in the top 25 results.

Skip-grams performed best, but even in the best models the results for the structural analogies were very low. For semantic analogies in Skip-gram models the target word appeared 14%±6% of the time as the first result, 40%±9% of the time in the top 10 results, and 53%±8% of the time in the top 25 results. Structural analogies for Skip-grams had a much lower performance: The target word appeared only 0.3%±1% of the time as the top result, 4%±4% in the top 10 results, and 7%±5% of the time in the top 25 results. This indicates that there are limits to how much grammatical structure these systems are learning from such a small corpus. Table 4 confirms this pattern. It shows the results

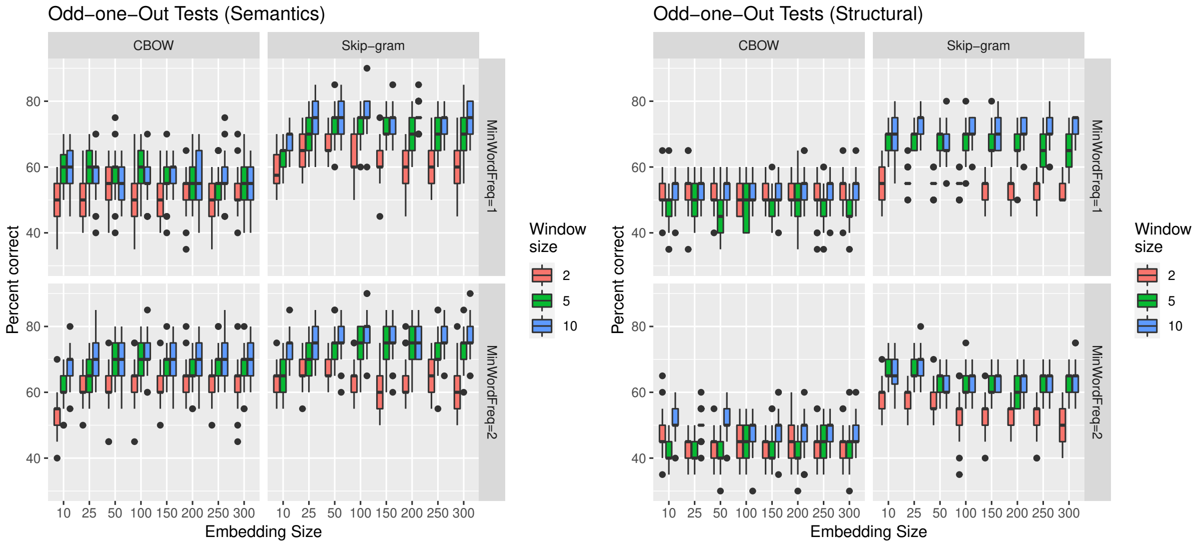

Figure 3: Percentage of correct answers to the Odd-One-Out task for semantic and structural triads

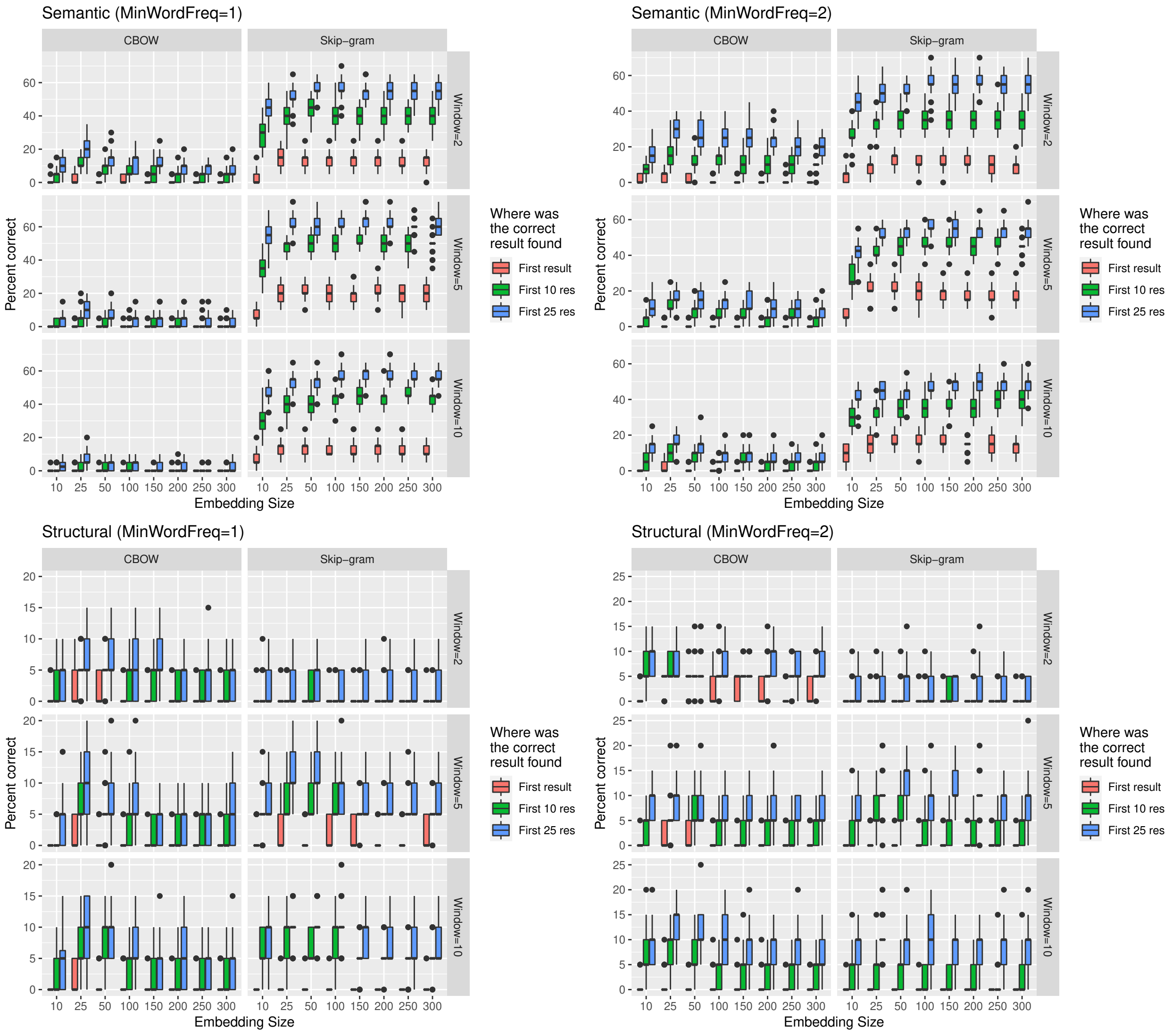

Figure 4: Position of target word in semantic and structural analogies

| Type | Size | Window | MinFreq | Sem | Str |
|---|---|---|---|---|---|
| SG | 100 | 5 | 1 | 63 | 9 |
| SG | 25 | 5 | 1 | 62 | 12 |
| SG | 50 | 5 | 1 | 60 | 12 |
| **SG** | **100** | **10** | **1** | **57** | **11** |
| SG | 150 | 5 | 2 | 56 | 11 |
| SG | 100 | 5 | 2 | 55 | 9 |

Table 4: Percent of analogies that contained the target word in the first 25 results; top-performing models (SG: Skip-gram; Sem: Semantic; Str: Structural)

for whether the target word was contained in the first 25 words. In semantic analogies the target word appears in the top 25 results between 55% and 63% of the time, but in the structural analogies the target never appears more than 12% of the time.

All of the best results were again Skip-grams. Like in the correlation experiment, medium dimensionality (i.e. around 100 dimensions) seems to be optimal for the analogies. Better-performing windows tend to be shorter than for the odd-one-out task (with more results of window=5), and most high-performing models include all the words in the corpus (mininum frequency=1). Table 4 also contains the model {Skip-gram, 100 dimensions, window=10, minFreq=1}. This model is the only one that appears amongst the top performing models for all of the tests, so it could be considered as the best amongst the examined models. It appears to have an adequate balance for learning all of the tasks presented in this section.

## 4 Discussion

The results indicate that there are certain combinations of hyperparameters that could provide better performance for extremely under-resourced languages. Embeddings with larger windows and low to mid-sized dimensionality (size=100) appear to learn better. Word2Vec also appears to need every word it can get, and its best performance comes when all words in the corpus are included in the training. Skip-grams might be a better alternative for this task because they avoid overfitting for very frequent words, thereby absorbing more of the information from the relatively sparse dataset at hand (Shobana and Murali, 2021). CBOW has been reported to be better at learning morphological relations, but this was not the case in the Bribri dataset, where the CBOW showed very low rates of structural learning.

Indeed, an open question in this paper is the extent to which the model is learning Bribri grammar. The results from the basic similarities and the odd-one-out experiments seem to indicate that the model learned grammar at roughly the same rate as it did lexical relationships. However, this is contradicted by the analogies experiment, where there was almost no evidence of structural knowledge in the language model. It might be the case that large windows are interfering with local structural learning (e.g. learning which counter words go with which nouns) in favor of semantic knowledge (Levy and Goldberg, 2014), and that the good performance in the odd-one-out structural tests might actually have to do with the semantics of the chosen pairs. More research and a larger test set is necessary to fully understand this effect.

Importantly, the corpus provided here appears to be enough for the system to learn general semantic patterns. Future experiments need to expand on this by providing more pairs, including pairs with more culturally specific words. Another future experiment will be to use the data to train other embeddings such as GloVe, fasttext (Bojanowski et al., 2017), BERT-type dynamic embeddings and multilingual embeddings. In the case of BERT, we need to study the effect of their high dimensionality on low-resource semantic learning. The preliminary hypothesis would be that this increase in dimensionality would have a negative impact in semantic learning. This needs to be verified because those larger embeddings are necessary for numerous deep learning techniques.

One important addition to this experiment would be a human baseline. One of the obstacles in working with Indigenous languages is that there are few people who read and write these languages. Moreover, there are few speakers of Bribri that are familiar with tasks like analogies, which are an unusual type of exercise mostly reserved for academic contexts. Therefore, carrying out such an experiment is relatively complex. This is the next step in this project, a necessary one so that the Skip-gram's performance can be placed in context.

The main ethical concern in the project has to do with data sovereignty. The results will be used to train deep learning systems in collaboration with Bribri partners, but there is currently no overarching community organization which controls the access to Bribri data. This paper has restricted itself to data that is publicly available or licensed through Creative Commons, so no private or new data was included. However, an effort needs to be made so that the results of this project benefit the Bribri partners in particular and the Bribri community in general, by using them to produce useful NLP tools.

## 5 Conclusions

This paper presents an evaluation of a word embedding in a truly under-resourced environment. It presents a methodology for embedding evaluation that could be adapted to other Indigenous languages. It provides evidence that some embedding configurations have better performance when dealing with under-resourced scenarios (i.e. Skip-gram trained embeddings, with around 100-dimensions, where the skip-gram attempts to predict words in a window of size 10 and use every word in the corpus to calculate their predictions). The results also confirm the intuition that semantics might be easier to learn than morphology, particularly in morphologically complex languages with little (and sparse) data. Finally, the results here will be used to continue NLP work in the Bribri language with the objective of training deep learning models and understanding their performance and errors, with the ultimate goal of using these to contribute to efforts of language revitalization.

## 6 Acknowledgments

The author wishes to thank Prof. Samantha Wray, Prof. Carla Jara, Dr. Sofía Flores Solórzano and the reviewers of this paper for their help at different stages of the project. The author also wishes to thank the audience at the Universidad Michoacana de San Nicolás de Hidalgo for their comments on a previous version of this project.

## References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the association for computational linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Rolando Coto-Solano. 2021. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Rolando Coto-Solano and Sofía Flores-Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Kánina*, 40(4):175–199.

Rolando Coto-Solano and Sofía Flores-Solórzano. 2017. Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. *CLEI Electron. J.*, 20(1):2–1.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards universal dependencies for bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414.

Sofía Flores-Solórzano. 2010. Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, pages 155–161.

Sofía Flores-Solórzano. 2019. La modelización de la morfología verbal bribri - Modeling the Verbal Morphology of Bribri. *Revista de Procesamiento del Lenguaje Natural*, 62:85–92.

Sofia Margarita Flores Solórzano. 2017. *Un primer corpus pandialectal oral de la lengua bribri y su anotacion morfologica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

INEC. 2011. Población total en territorios indígenas por autoidentificación a la etnia indígena y habla de alguna lengua indígena, según pueblo y territorio indígena. In Instituto Nacional de Estadística y Censos, editor, *Censo 2011*.

Carla Victoria Jara Murillo. 2018. *Gramática de la lengua bribri*. E Digital.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. E Digital.

Hyungsuk Kang and Janghoon Yang. 2018. The Analogy Test Set Suitable to Evaluate Word Embedding Models for Korean. *Journal of Digital Contents Society*, 19(10):1999–2008.

Haakon Krohn. 2020. Elaboración de una base de datos en XML para un diccionario bribri–español español–bribri en la web. *Porto das Letras*, 6(3):38–58.

Haakon S. Krohn. 2021. Diccionario digital bilingüe bribri. `http://www.haakonkrohn.com/bribri`.

Thomas K Landauer, Darrell Laham, Bob Rehder, and Missy E Schreiner. 1997. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez Lugo, Ricardo Ramos, et al. 2021. Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

Enrique Margery. 2005. *Diccionario fraseológico bribri-español español-bribri*, second edition. Editorial de la Universidad de Costa Rica.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Radim Rehurek and Petr Sojka. 2011. Gensim: Python Framework for Vector Space Modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation Methods for Unsupervised Word Embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Rico Sennrich and Biao Zhang. 2019. Revisiting Low-Resource Neural Machine Translation: A Case Study. *arXiv preprint arXiv:1905.11901*.

J Shobana and M Murali. 2021. Improving Feature Engineering by Fine Tuning the Parameters of Skip Gram Model. *Materials Today: Proceedings*.

Nathan Stringham and Mike Izbicki. 2020. Evaluating Word Embeddings on Low-Resource Languages. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 176–186.

Carlos Sánchez Avendaño, Alí García Segura, et al. 2021a. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. Editorial de la Universidad de Costa Rica.

Carlos Sánchez Avendaño, Alí García Segura, et al. 2021b. *Se' Má Diccionario-Recetario de la Alimentación Tradicional Bribri*. Editorial de la Universidad de Costa Rica.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

## A Experimental Items

| Bribri | Translation | Equivalent English | English Similarity |
|---|---|---|---|
| 01. na̱mù̱ - pûs | big feline - cat | tiger - cat | 7.35 |
| 02. na̱mù̱ - na̱mù̱ | big feline - big feline | tiger - tiger | 10 |
| 03. ya’ - kukuö̀ | drank - ear | drink - ear | 1.31 |
| 04. ya’ - kö̀ | drank - mouth | drink - mouth | 5.96 |
| 05. ya’ - ña̱’ | drank - ate soft things | drink - eat | 6.87 |
| 06. aláala - a̱mì̱ | baby - mother | baby - mother | 7.85 |
| 07. ya’ - a̱mì̱ | drank - mother | drink - mother | 2.65 |
| 08. chakö̀ - kàlwö | food - fruit | food - fruit | 7.52 |
| 09. dù - dakarò | bird - chicken | bird - cock | 7.1 |
| 10. chakö̀ - dakarò | food - chicken | food - rooster | 4.42 |
| 11. dayë̀ - ka̱ñí̱k | sea - jungle | coast - forest | 3.15 |
| 12. na̱mù̱ - íyiwak | big feline - animal | tiger - animal | 7 |
| 13. ajkuö̀ - wö̀bala | skin - eye | skin - eye | 6.22 |
| 14. dalì - inú̱köl | merchandise - money | grocery - money | 5.94 |
| 15. ña̱là̱ - ala’r | road - children | street - children | 4.94 |
| 16. ské̱l - si’ | five.flat - month | five - month | 3.38 |
| 17. ñí̱we - shkè̱na̱ | during the day - to dawn | day - dawn | 7.53 |
| 18. wë́m - aláköl | man - woman | man - woman | 8.3 |
| 19. ña̱là̱ - ká̱ | road - time, space, place | street - place | 6.44 |

Table 5: WordSim353 Correlations and translated Bribri pairs. WordSim353 English similarities go from 0 to 10. A Bribri/Gensim score of 0.5 is roughly equivalent to a WordSim353 score of 5.

| Bribri | Translation |
|---|---|
| 01. aláköl - wë́m - NA̱MÙ̱ | woman - man - TIGER |
| 02. yë́ - a̱mì̱ - Ù̱ | father - mother - POT |
| 03. a̱mí̱la - a̱mì̱ - ALÀ | maternal aunt - mother - CHILD |
| 04. wìm - sàl - DÙ | howler monkey - spider monkey - BIRD |
| 05. cha̱mù̱ - ikuö̀ - DAKARÒ | banana - corn - CHICKEN |
| 06. íyiwak - kàlwak - KUÁ | animal - bug - PLANT |
| 07. skuë̀ - ka̱no̱’ - KALÖ́M | mouse - lowland paca - PLANTAIN |
| 08. tkabë̀ - só - DIKÓ | snake - cockroach - PEJIBAYE DATE |
| 09. átu - ali’ - TABÈ | beans - yucca - KNIFE |
| 10. dalôlô - sarûrû - BITSÎ̱ | black - white - LONG |
| 11. tsuru’ - balo’ - ÀRROS | cocoa - chicha drink - RICE |
| 12. chkö́k - katö́k - YÖ́K | to eat - to eat hard foods - TO DRINK |
| 13. katö́k - ñú̱k - CHKÖ́K | to eat hard food - to eat soft food - TO EAT (GENERIC) |
| 14. ulà - kalö̀ - Ù | hand - foot - HOUSE |
| 15. kàsir - kö̀chi - NI̱MÀ̱ | peccary - pig - FISH |
| 16. wö̀bla - yík - KALÖ̀ | eye - nose - FOOT |
| 17. datsi’ - apàio - SI’ | clothes - shirt - MOON |
| 18. chkì - î̱ñe̱ - DI’ | yesterday - today - RIVER, WATER |
| 19. dë’ - mí̱k - YÖ’ | arrived - went - TO MAKE |
| 20. A̱mù̱bali - Kua’rö - TALÌRI | Town of Amubri - Town of Buenos Aires - SALITRE RIVER |

Table 6: Semantic groups of Odd-One-Out Triads. Words in small caps are the “odd” word.

| Bribri | Translation |
|---|---|
| 01. kàl - íyök - SHKÖ́K | tree - soil - TO GO |
| 02. kuá- íyiwak - SÈ̱RKE | plant - animal - (SOMEONE) LIVES |
| 03. ta̱ - tö - DI’ | with - ergative marker - WATER, RIVER |
| 04. ska - ki̱ - AKË́KËPA | towards - at - OLD PERSON |
| 05. ye’ - be’ - DÙ | I - you - BIRD |
| 06. sa’ - se’ - ÑA̱LÀ̱ | we (exclusive) - we (inclusive) - ROAD |
| 07. tsîr - tsikirîrî - CHKÌ | small - yellow - YESTERDAY |
| 08. bë̂rie - buáala - MÌ̱K | big - beautiful - WHEN |
| 09. yö’ - ña̱’ - KÙ̱NE̱ | made - ate - WAS FOUND (middle voice) |
| 10. kít - ya’ - SÙ̱NE̱ | wrote - drank - WAS SEEN (middle voice) |
| 11. yö́k - i̱nú̱k - YA’ | to drink - to play - DRANK (perfective) |
| 12. shkö́k - sa̱u̱k - DË’ | to walk - to see - WENT (perfective) |
| 13. ché - yawé - SÚ̱ | saying - drinking - SAW (perfective) |
| 14. mi̱’ke̱ - yawèke̱ - DË’ | going - drinking - WENT (perfective) |
| 15. tkër - tulur - DUR | sitting - sitting.plural - STANDING |
| 16. a’r - tkë’ni̱k - TÉ̱N | hanging - hanging.plural - STICKING IN |
| 17. a̱wí̱ - a̱wì̱ - E’ | there (near) - there (far) - THIS ONE |
| 18. diö́ - dià - NE̱’ | there (below, near) - there (below, far) - THIS ONE (ONLY HEARD) |
| 19. e’töm - bö̀töm - MA̱ÑÁ̱L | one.long - two.long - THREE.HUMAN |
| 20. e’köl - bö́l - MA̱ÑÀ̱TÖM | one.human - two.human - THREE.LONG |

Table 7: Structural groups of Odd-One-Out Triads. Words in small caps are the “odd” word.

| Bribri | Translation | Type |
|---|---|---|
| 01. wë́m:aláköl :: yë́:A̱MÌ̱ | man : woman :: father : MOTHER | Family relation |
| 02. kabë̀:bùsi :: wë́m:ALÁKÖL | boy : girl :: man : WOMAN | Family relation |
| 03. a̱mí̱la:na̱ù̱ :: aláköl:WË́M | maternal aunt : maternal uncle :: woman:MAN | Family relation |
| 04. akë̀:kutà :: wë́m:ALÁKÖL | brother of woman : sister of man :: man:WOMAN | Family relation |
| 05. talà:yë́ :: wì̱ke:A̱MÌ̱ | paternal grandfather : father ::<br>maternal grandmother : MOTHER | Family relation |
| 06. balo’:yà̱ne̱ :: ali’:ÑÀ̱NE̱ | chicha drink : was drunk :: yucca : WAS EATEN | Object/Action |
| 07. wë́m:dur :: dù:TKËR | man : stands :: bird : RESTS ON A SURFACE | Object/Action |
| 08. bola:i̱nù̱k :: uyë̀jkuö:À̱RITSÖK | ball : to play :: book : TO READ | Object/Action |
| 09. ni̱mà̱:di’ :: wë́m:Ù | fish : water :: man : HOUSE | Place to live |
| 10. ni̱mà̱:di’ :: buà:ÚK | fish : water :: iguana : BURROW | Place to live |
| 11. aláköl:ù :: dù:KÀL | woman : house :: bird : TREE | Place to live |
| 12. pulí:kàl :: pú̱:DÙ | ceiba tree : tree :: eagle : BIRD | Hypernym |
| 13. kabék:dù :: ikuö̀:KUÁ | quetzal bird : bird :: corn : PLANT | Hypernym |
| 14. kaë̀:dù :: átu:KUÁ | pava negra bird : bird :: beans : PLANT | Hypernym |
| 15. dakarò:dù :: chìchi:ÍYIWAK | chicken : bird :: dog : ANIMAL | Hypernym |
| 16. së́:ù :: wö̀bla:WÖ́KIR | beam : house :: eyes : FACE | Holonym |
| 17. tóttô:darë̂rë̂ :: bua’:SULÛ | easy : hard :: good : BAD | Antonym |
| 18. ké̱we:ukö̀ki̱ :: wé̱shke̱:ÛRIKI̱ | before : after :: inside : OUTSIDE | Antonym |
| 19. ká̱jke̱:dikì̱ :: wö́ni̱k:TSÌ̱ | above : below :: in front : BEHIND | Antonym |
| 20. tsîr:bë̂rie :: sarûrû:DALÔLÔ | small : big :: white : BLACK | Antonym |

Table 8: Quartets for the semantic analogies. Words in small caps are the target words.

| Bribri | Translation |
|---|---|
| 01. aláköl:e’köl :: chìchi:E’TÖM | woman : one.human :: dog : ONE.LONG |
| 02. cha̱mù̱:e’töm :: dawás:E’K | banana : one.long :: year : ONE.ROUND |
| 03. dur:ië’te̱n :: tkër:TULUR | stand.sg : stand.pl :: sit.sg : SIT:PL |
| 04. a’r:tkë’ni̱k :: té̱n:TULUR | hanging.sg : hanging.pl :: stuck.sg : STUCK:PL |
| 05. alà:ala’r :: ie’:IE’PA | child : children :: 3sg : THEY |
| 06. awá:awápa :: yë́ria:YË́RIAPA | healer : healers :: hunter : HUNTERS |
| 07. tsîr:tsítsi :: buáala:buàmbuáala | small.sg : small.pl :: beautiful.sg : BEAUTIFUL.PL |
| 08. bë̂rie:wî̱wî̱ :: wá̱wá̱n:WÂ̱NWÂ̱N | big.sg : big.pl :: few.sg : FEW.PL |
| 09. e’köl:e’töm :: bö́l:BÖ̀TÖM | one.human : one.flat :: two.human : TWO.FLAT |
| 10. e’k:e’tökicha :: bö̀k:BÖ̀TÖKICHA | one.round : one.time :: two.round : TWO.TIMES |
| 11. ma̱’tk:má̱shma̱sh :: si̱ê̱:SI̱É̱LSI̱E̱L | red : reddish :: blue : BLUEISH |
| 12. dalôlô:dalóshdalosh :: sarûrû:SARÚLSARUL | black : blackish :: white : WHITEISH |
| 13. yö’:yö̀ne̱ :: sú̱:SÙ̱NE̱ | made : was made :: saw : WAS SEEN |
| 14. shka’:shkàne̱ :: stsë’:STSË̀NE̱ | walk : [someone] walked [there] :: heard : WAS HEARD |
| 15. a̱wí̱:a̱wì̱ :: aí:AÌ | that (near) : that (far) :: that (above, near) : THAT (ABOVE, FAR) |
| 16. a̱wì̱:aì :: a̱wí̱:AÍ | that (far) : that (above, far) :: that (near) : THAT (ABOVE, NEAR) |
| 17. a̱wí̱:a̱wì̱ :: diö́:DIÀ | that (near) : that (far) :: that (below, near) : THAT (BELOW, FAR) |
| 18. sú̱:sa̱ú̱ :: yö’:YAWÖ́ | saw : see! :: made : MAKE! |
| 19. të’:tèke̱ :: yö’:YAWÈKE̱ | hit : hitting :: made : MAKING |
| 20. katö̀k:katèke̱ :: yawö̀k:YÈKE̱ | to eat (hard things) : eating (hard things) :: to make : MAKING |

Table 9: Quartets for the structural analogies. Words in small caps are the target words.