Predicting colexification rates through regression of psycholinguistic ratings of concepts

Mani Setayesh

Abstract

Previous work has shown that there is a “goldilocks zone” of semantic relatedness where two concepts are likely to colexify if they are not too unrelated, nor too related in meaning. This project aims to build on this idea by evaluating colexification as a function of psycholinguistic variables (concreteness, imageability, etc). Two regression models were trained on a random subset of the data and then tested on the other using appropriate correlation statistics to evaluate their performance. A total of 300 trials were ran and the significant trials were kept for analysis. Generally, the results of the regression models seem to indicate poor predictive power on the colexification counts for both logistic and linear models. I conclude that there is need for further rigorous investigation into this area and development of more sophisticated models through considering a combination of psycholinguistic factors and their correlation with colexification probabilities.

Introduction

A quote by the famous linguist Wilhelm von Humboldt states that language is a system that “makes infinite use of finite means”, articulating that there are infinitely many thoughts that can be communicated within a finite number of words. It is therefore mandatory for a word or sound to have multiple different meanings - a phenomenon called **colexification**. For example, the word “hand” can be used to refer to a bodily limb, or to a set of cards dealt at a poker table. These two concepts (limb and cards) have colexified - but why? Finding a reason can shed some light on how the human mind categorizes concepts, which is one of the most foundational questions in the domain of cognitive science.

The foundational theory behind colexification relies on the idea of cognitive economy (Rosch & Lloyd, 1978). First, a few findings and ideas on when colexficiation occurs: Colexification can only occur in the absence of lexical ambiguity - i.e. the two distinct concepts should not be confused with each other if they are to share the same term (Karjus, Blythe, Kirby, Wang, & Smith, 2021). Each concept should be relatively distinct enough such that when the colexified word is presented, the listener is able to distinguish the intent of the speaker and the speaker should have enough confidence that there is little room for misinterpretation of the same concept.

However, concepts that are fully different semantically and share nothing with each other run into the risk of creating confusion as well (Xu, Duong, Malt, Jiang, & Srinivasan, 2020). There is then potential for a complete misinterpretation of the conversation, and so the current proposal is that there is a semantical “goldilocks” zone (Brochhagen & Boleda, 2022), where two concepts are distinct and related just enough that they are colexified. The interpretation of this connects back to the notion of cognitive economy - reducing complexity of the lexicon by not having too many words, but also still retaining informativeness by having significant differences in concept articulation.

Preliminary study and its limitations

The study done by Brochhagen and Boleda (Brochhagen & Boleda, 2022) will be the focus of this project - and I will refer to it as “the original study” from here on. This study provides a conclusive account for the role of semantic relatedness in colexification, and provides a good interpretation for why it happens. Due to the inherent vagueness in the idea of semantic relatedness, the original study proposed a breakdown of this notion into 2 quantifiable views: Associativity and distributional similarity (Xu et al., 2020; Brochhagen & Boleda, 2022) - the primer refers to how often two concepts are associated together, and the latter refers to how similar the context of the two concepts are. The study also includes the first principal component (PC1) of the two views as the primary estimate, which highlighted the goldilocks zone and was shown to be fairly reliable in predicting colexification.

The issue that arises is the reliance on the meaning for predicting the colexification of two concepts. The reason why this is an issue is it inherently relies on a shared semantic structure across languages - for example, it is assumed that the concept of “bird” exists and is the same both in English and in Spanish but just referred to with different words. This however can eliminate any role that the culture and background of a language plays in colexifying two concepts, since neither associativity nor distributional similarity (or their first principal component) account for this conceptual bias.

Hypotheses and motivation

To this end, I propose a method that accounts for semantic relatedness in a different manner - through the use of semantic properties of concepts (called semantic variables), rather than their definitions. It has been shown that there are neural correlations between the “concreteness” rating of a concept and its semantic categorization (Li et al., 2021; Ding, Liu, & Yang, 2017). Further, there has been shown to be general correlations between different semantic variables - such as imageability, concreteness, and familiarity (Yao, Wu, Zhang, & Wang, 2017). The benefit of using these psycholinguistic variables is that they occur at a much more primal level that is not reliant on the semantic linguistics of concepts, but rather how humans as a whole perceive concepts in spectrums.

My hypothesis is that the degree to which two concepts are colexified can be predicted by using four features: the imageability, concreteness, and familiarity ratings of a concept, as well as its age of acquisition (AoA). The additional AoA factor is included to account for cultural differences - where if multiple concepts are learnt at roughly the same age across different cultures, that indicates they are most likely semantically related in some way. The primary hypothesis for this project is that the four features can be used in a linear regression model to predict the degree of colexification between two concepts better than the PC1 model in the original study. An alternate hypothesis is that a logistic regression model using these features can obtain better results than the PC1 model in the original study. Finally, the inter-correlation factor between these features will also be calculated to provide support for the studies that were cited on the relationship between the psycholinguistic variables and semantic relatedness.

Methodology

This project will be using the CLICS dataset (Rzymski et al., 2020) as it was used in the original study, coupled with the psycholinguistic ratings for the earlier mentioned features obtained from the concepticon (List, Cysouw, & Forkel, 2016). The following is an outline of methodology of this project - `https://github.com/manisetayesh/FinalProject` for further python implementation details.

Data Organization

The first step in the preprocessing of the data was obtaining the colexification counts between two concepts. This was done a-priori similar to the original study's methodology. Since many concepts in the dataset had non-applicable values for the features, only concepts that had fully recorded data for all features were kept - resulting in 426 concepts and 3481 colexification counts between 2 distinct concepts.

Since there were multiple rows per concept in the CLICS dataset, the four features were aggregated for each concept through taking their average across the concept's different representations in different languages. Since all the features were rating based and there were a good number of representations per concept, there is likely no significant difference in the aggregation method (e.g. choosing the median instead) so the simplest method was selected. These lists were joined to have two concepts per row, each with its feature averages and their joint colexification count - this joint table will be referred to as "the data" from now on.

Model Training

We will discuss the two parts to training a model and how they were incorporated in this project:

Data Preprocessing The data was randomly split into train-test subsets. Each subset includes one set of 4 features X per concept (i.e. X_1, X_2) and the colexification count represented as $y(c_1, c_2)$ for two concepts c_1, c_2. The feature sets for both concepts were standardized to improve model performance by adjusting for the differing mean and standard deviations - note that each feature set was standardized individually with respect to itself, and the colexification counts were not standardized.

Model Formulation Two models were formulated for learning and prediction - one was a linear regression model, and the other was a logistic regression model that followed a conversion from the linear model. The linear regression model was formulated as:

$$P = \beta_0 + \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \cdot X_1 + \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \cdot X_2$$

where P represents the prediction of the model, β_0 is the intercept, and the beta-vector has the weights for each feature that need to be optimized - same weights for both concepts. The logistic regression model was formulated as:

$$P_2 = \frac{e^{-P}}{1 + e^{-P}}$$

where P_2 is the prediction of the model following that of the linear model's P.

Model evaluation

Linear model Two objective functions were used to evaluate the performance of the model (i.e. its prediction accuracy): the pearson correlation coefficient r and the Spearman rank correlation ρ. The task of the model was to maximize these values (separately) by picking a set of weights for the features and an intercept to add to the sum. Once these parameters were set using the training dataset, they were then tested on the test dataset where both correlation metrics for the model's predictions versus the actual colexification count was recorded. Finally, a separate linear model using the pre-existing sklearn library was used to have a benchmark (with objective function being the residual sum of squares). This model was also trained and tested on the same data split and had the same correlation metrics recorded.

Logistic model Two other objective functions were used for the logistic model. Since the logistic function is non-linear, the root mean squared error (RMSE) and the Coefficient of determination (R^2) were used, similar to that of the original study. Most of the methodology remains the same as the linear model's, including a separate logistic regression model from the sklearn library.

Notes

- The process was repeated for 300 random iterations of the train/test data split. It was ensured that both the logistic model and the linear model worked with the same splits, and did not build on previously obtained parameters from other models (including models with other objectives).

- Alongside the correlation values for the linear model, the p-values were also recorded. $p < 0.01$ was selected (across all models) to filter out inconclusive iterations.

- The entire procedure was done in python, using statistical models and metrics from two dominant libraries: the aforementioned sklearn (Pedregosa et al., 2011) and scipy (Virtanen et al., 2020).

Results

After filtering by the p-values, 20 trials were considered to be significant.

Linear Regression

Here are the general statistics on the linear regression models:

Table 1: Linear model statistics

Model	$\overline{\rho}$	$\overline{r}$	σ_{ρ}	σ_{r}
Pearson objective	0.1080	0.1074	0.0267	0.0167
Spearman objective	0.1618	0.0966	0.0215	0.0099
Sklearn Lin.Reg	0.1019	0.1058	0.0239	0.0138

There are two key observations that can be made from this:

- The results of the pre-existing model closely matches the model with the objective to maximize the pearson correlation. This was expected, due to the close relationship between residual sum of squares - which sklearn uses - so it supports the reliability of the methodology.
- The Spearman correlation is likely a better metric for evaluation than the pearson correlation. This can be observed as the model with the Spearman objective is shown to have significantly higher $\overline{\rho}$ while also having relatively the same $\overline{r}$ (well within a standard deviation).

The second observation is particularly relevant as it shows that the relationship between the feature set and the colexification count is more monotonic than linear. This can also be observed through seeing the predictions of the spearman model versus the actual colexification count in the most optimal trial (highest ρ):

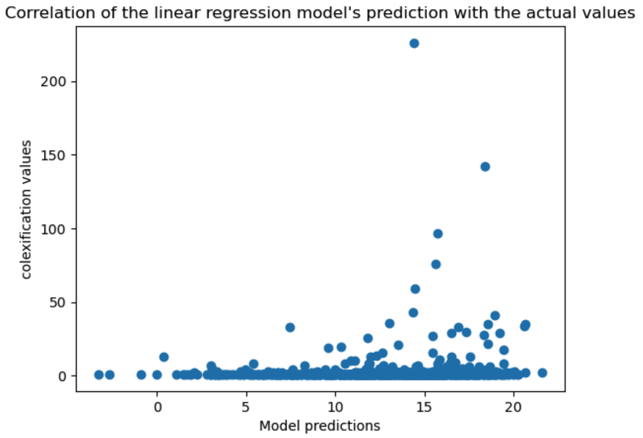

Figure 1: Prediction of the linear regression model using spearman objective - random state = 89

Note that the majority of colexification counts are 1. This can help the interpretation of why the pearson's correlation coefficients are very low - the relationship appears to be non-linear and pseudo-exponential. The models could be adjusted through either a logarithm on the colexification value or setting a threshold to only consider concept pairs that have a colexification count that is higher than 1 or 2 - though this comes with severe limitations on the number of colexified pairs.

Logistic Regression

Here are the general statistics on the logistic regression models:

Table 2: Logistic model statistics

Model	$\overline{R^2}$	$\overline{RMSE}$	σ_{R^2}	σ_{RMSE}
RMSE objective	-0.0268	9.6341	0.0099	2.1718
R2 objective	-0.0272	9.6354	0.0099	2.1698
Sklearn Log.Reg	-0.0393	9.6885	0.0101	2.1678

There are, once again, two key observations that can be made from this:

- The results of the three models are very similar. This suggests that the optimization metric has little to no impact on the logistic regression of these features.
- The R^2 values are roughly 0 - negative values were reported using sklearn metrics, but the valence of the number is not of significance. This suggests that a logistic regression model is not a good fit for predicting colexification values.

Once again, to observe why this would happen, the predictions made by the model in the most optimal trial were considered:

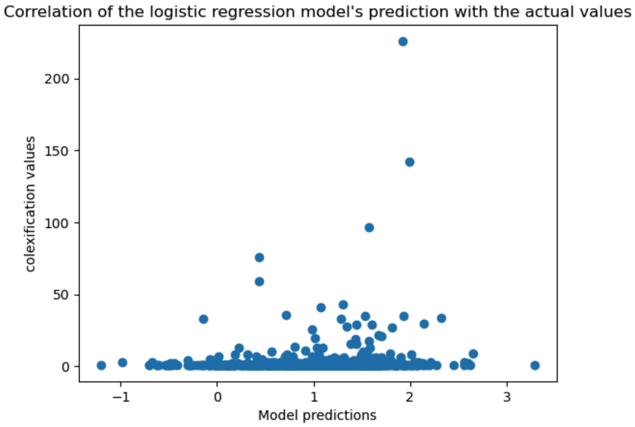

Figure 2: Prediction of the logistic regression model using RMSE objective - random state = 90

Of particular note is the prediction range - centring at 1.

Discussion of results

The obtained results do not provide good support for the proposed hypotheses. While there is some evidence of a significant non-linear correlation between the chosen features and the colexification counts, the generated models seem to be unable to fit very well to the existing data. There is strong evidence to reject the primary hypothesis of a linear correlation, since even the best performing models had $r < 0.13$ which indicates that a linear fit is not appropriate.

For the logistic models, there is some evidence that can support the use of a logistic model. The main argument here is that the predictions of the model were mostly 1 - meaning it was able to locate the center of colexfication count density. With slight adjustments by removing the outliers and limiting the model's guesses (bounded below by 1, so no prediction should be 0 or negative) there can be a significant difference in the R^2 and RMSE values. In the top trial shown in figure 2, the Mean Absolute Error (MAE) was roughly 1911 for a total of 871 items whereas the RMSE was roughly 9, indicating that perhaps the mean absolute error might be a better pick for an objective function on a sigmoid curve.

To conclude the discussion of the results, one final piece of data should be considered: the inter-correlation factors between the 4 features (labelled ACFI for age of acquisition, concreteness, familiarity, and imageability). This can be calculated as a spearman's correlation matrix like so:

Table 3: Feature Inter-correlations (ρ)

Feature	A	C	F	I
A	1	-0.2785	-0.6713	-0.3932
C		1	-0.0499	0.7958
F			1	0.1102
I				1

Of particular note is that the AoA and familiarity ratings of a concept seem to have an negative monotonic relationship, while the concreteness and imageability ratings of a concept have a positive monotonic relationship. This can provide an idea for extension of this research, where the feature set is divided into 2 halves based on the high correlation values above. This gives support to the earlier hypotheses of inter-correlations between psycholinguistic variables.

Conclusion

Generally, after analysis of the model predictions using the psycholinguistic concept ratings as the feature space, there are some key limitations and problems that can be observed. It has been reliably shown that a simple linear regression model is not reliable enough to predict colexification counts given the feature set. This is due to a significant number of concept pairs having $y(c_1,c_2) = 1$ while some outliers have $y(c_1,c_2) > 20$. A normalization technique on the number of colexification counts might prove to be beneficial in increasing model accuracy and prediction power.

In the original study a logistic regression model was used and it performed very well - with $RMSE = 0.34, R^2 = 0.53$. The main explanation for this is the different formulation of the logistic model in this study. Here, the model is trained and tested to predict the actual colexification count between two concepts, summed up across languages. As such, the regression model fits a large range of values (1 to 150) into a binary-like distribution that is centered at the most common colexification count:1. This is why the model has predictions that make little sense - e.g. predicting a negative number for the colexification count. An alternative methodology should be considered where the model instead operates on each individual word in different languages, and treats each case as a bernoulli 0/1 similar to the original study. Then, select this psycholinguistic feature space for the model to consider. The reason why this was not done in the current study was the lack of computational power and capacity for a feasible runtime over all the different words and conceptual combinations.

References

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition*, *226*, 105179. doi: https://doi.org/10.1016/j.cognition.2022.105179

Ding, J., Liu, W., & Yang, Y. (2017). The influence of concreteness of concepts on the integration of novel words into the semantic network. *Frontiers in Psychology*, *8*, 2111. doi: 10.3389/fpsyg.2017.02111

Karjus, A., Blythe, R. A., Kirby, S., Wang, T., & Smith, K. (2021). Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, *45*(9), e13035.

Li, H., Liang, Y., Qu, J., Sun, Y., Jiang, N., & Mei, L. (2021). The effects of word concreteness on cross-language neural pattern similarity during semantic categorization. *Journal of Neurolinguistics*, *58*, 100978. doi: 10.1016/j.jneuroling.2020.100978

List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon: A resource for the linking of concept lists. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 2393–2400).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rosch, E., & Lloyd, B. B. (1978). Principles of categorization. *Cognition and categorization*.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., … others (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, *7*(1), 13.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280.

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 chinese words. *Behavior Research Methods*, *49*(4), 1374–1385. doi: 10.3758/s13428-016-0793-2