

# A distributional analysis of affixal rivalry between *-ed* and *-y* in English

Ngo Trung Hieu  
Zheng Peng \*

## Abstract

The paper *Affixal rivalry and its purely semantic resolution among English derived adjectives* by Nagano (2022) has claimed an interesting finding about the affixal rivalry between the ending *-ed* and *-y* in English. According to Nagano, there exists a purely semantic affixal rivalry between these two endings in the denominalized adjectives, created from English concrete nouns. The authors of this study is interested in looking at this finding from a different perspective, by using the distributional methodology, as demonstrated in Naranjo and Bonami (2023).

## 1 Literature review

### 1.1 Affixal rivalry in English affixes *-ed* and *-y*

### 1.2 Distributional methodology for affixal rivalry

## 2 Research questions

Because Nagano (2022) explained the affixal rivalry between *-ed* and *-y* in English from a purely semantic perspective, therefore, the dataset used in that paper is very limited. We are interested in investigating this finding from a distributional perspective. Therefore, the research questions of this study are as follows:

- Is there similarity between the affixes *-ed* and *-y* in terms of their distributional behavior? To answer this question, this study will look at the distributional representations of the two affixes, by using conventional word embeddings such as word2vec.
- Are there evidence of affixal rivalry between *-ed* and *-y* in English, in a larger and not hand-picked dataset? To answer this question, this study will train a neural network on the dataset from Nagano (2022), and test the classifier on unseen data extracted

---

\*M1 Linguistique Informatique - Université Paris Cité

from UniMorph. If there exists a rivalry between the two affixes, the classifier should not be able to accurately predict the correct affix for the unseen data.

- The conventional distributional representations of words may not be able to capture semantic relationships between words. Therefore, this study will also implement a retrofitting algorithm from Faruqi (2015) to augment the word embeddings. If the similarity between the two affixes is increased after retrofitting, it means that there exists a semantic relationship between the two affixes.

### 3 Methodology and Experiment design

There are three experiments in this project. To answer the first research question, we calculate and compare the difference vectors between the two affixes to see if they are semantically comparable or not. The second experiment is to train a neural network classifier on these difference vectors, then test the classifier on unseen data, in this case the data from UniMorph. The last question augmenting the semantics of the word embeddings using the semantic retrofitting method will be addressed in the third experiment.

#### Experiment 1:

For this experiment, the dataset from Nagano (2022) and a distributional word vectorization system are needed. The idea is to represent the roots of chosen words and their derivations in a vector space, so that we can use computation to calculate the similarity. The chosen vector space is the Google-News-300 dataset from word2vec model, which is accessible from the Gensim package. This vector space is created from a vocabulary of 3 million words from Google News, and each word is represented by a 300-dimensional vector. The dataset from Nagano (2022) is a hand-picked dataset of 38 words, which have both *-ed* and *-y* derivations. The roots of these words are concrete nouns, and the derivations are adjectives, so the process is denominalization. The difference vectors between the roots and the derivations are calculated by subtraction, and the average difference vectors of the two affixes are compared using cosine similarity. If the two affixes are distributionally similar, the difference vectors should be similar as well.

#### 3.1 Experiment 2:

For the second experiment, we will create a neural network classifier with one hidden layer to train on the Nagano dataset, then use the classifier to predict the derivation of the unseen root words during training. The test dataset is extracted from UniMorph, and the words are filtered to only include words that can have *-ed* or *-y* affixes. The classifier will be trained on the vectorized roots and their derivations from the Nagano dataset, and the classifier will be trained for 100 epochs using 10 folds to maximize learning efficiency due to the limited number of training examples. After training, the classifier will be tested on the UniMorph

dataset of 4191 words, which were then vectorized by the word2vec model. The ratio of *-ed* affix is 42.92%, while the ratio of *-y* affix is 57.08%. The classifier will be tested on this UniMorph dataset, and the accuracy will be calculated. If the classifier can predict the derivation of the unseen words with an accuracy of around 50%, it means that the classifier is not able to distinguish between the two affixes, and therefore there exists an affixal rivalry between the two affixes.

### 3.2 Experiment 3:

Semantic retrofitting is a method to improve the quality of the word vectors by using the semantic information (synonyms, antonyms, hyper/hyponyms) from a lexicon.

In this experiment, we will use the retrofitting method from Faruqui et al. (2015) and the WordNet lexicon to improve the quality of the word vectors from the training data. The WordNet lexicon provides the semantic information for the words, and then use the retrofitting method to create a new word vector based on the old word vector and the "neighbors" of the word. The idea of retrofitting is to have a new word vector that is close (has a better cosine similarity) to the old word vector, but is also closer to the neighbors of the word. This new word vector is then said to be capturing the semantic information of the word better than the purely distributional word vector.

Then we will recalculate the difference vectors to see if there are improvements or not, and if these improvements are significant. If the difference vectors are more similar after retrofitting, it means that the two affixes are semantically related.

## 4 Results

### 4.1 Experiment 1:

Using the word2vec model, we first calculate the average similarity between two random words in the lexicon. To do this, we randomly chose 100,000 pairs from the lexicon, and compute the cosine similarity between the pairs. The result of this computation is 0.1308, and this serves as a baseline for the cosine similarity results.

The next step is to calculate the average difference vectors between the roots and the derivations of the Nagano dataset. The vectorization model is used on both the roots and the derivations of the words in the Nagano dataset, and then the difference vector is calculated as difference vector = root vector - derivation vector. The difference vectors between each root and its two derivations are then compared using cosine similarity, then these values were used to calculate the average similarity between the two affixes *-ed* and *-y*. The result of this computation is 0.4162, which is 3.18 times higher than the baseline. This result shows that the two affixes are distributionally similar. The detailed result of this experiment is shown in the attached Jupyter notebook.

## 4.2 Experiment 2:

First, the test data was extracted from the UniMorph database, only finding the denominalized adjectives ending in *-ed* and *-y*. This results in a list of 4191 words, with 42.92% ending in *-ed* and 57.08% ending in *-y*.

Next, a neural network classifier was initialized using PyTorch. This classifier has one hidden layer of 128 neurons, using ReLU as the activation function. The input layer has 300 neurons, which is the same as the dimension of the word vectors. The output layer has 2 neurons, which is the number of classes.

The classifier was then trained on the Nagano dataset of 76 examples. The classifier was trained for 100 epochs, using 10 folds to maximize learning efficiency. When tested on the test dataset from UniMorph, the classifier gave an accuracy of 49.146%, which is slightly lower than the expected result of 50%. This result shows that the classifier trained on the difference vectors of *-ed* and *-y* is not able to distinguish between the two affixes, and therefore distributionally, there exists an affixal rivalry between these two affixes. The detailed result of this experiment is shown in the attached Jupyter notebook. However, this result is still lower than the most frequent class of 57.08%, so there might be room to improve.

## 4.3 Experiment 3:

For the last experiment, we wanted to retrofit the word vectors to see if there are semantic similarity between *-ed* and *-y* or not.

# 5 Discussion

# 6 Conclusion

## References

- Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons, 2015.
- Matias Guzmán Naranjo and Olivier Bonami. A distributional assessment of rivalry in word formation. *Word Structure*, 2023.
- A Nagano. Affixal rivalry and its purely semantic resolution among english derived adjectives. *Journal of Linguistics*, pages 1–32, 2022. doi: 10.1017/S0022226722000147.