HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Heidi Jauhiainen
University of Helsinki
heidi.jauhiainen@helsinki.fi
Tero Alstola
University of Helsinki
tero.alstola@helsinki.fi

SUOMEN AKATEMIA
FINLANDS AKADEMI
ACADEMY OF FINLAND

# FAST(TEXT) ANALYSIS OF MESOPOTAMIAN DIVINE NAMES


## INTRODUCTION

In the *Semantic Domains in Akkadian texts* project, we are interested in words that belong to the same semantic domain. For example, Nergal, Marduk, and Tašmētu all belong to the lexical semantic domain 'gods'. However, some deities appear in more similar contexts than others and form subgroups of the semantic domain. The words in such a **contextual semantic domain** could:

- be surrounded by similar words
- literally appear in the same texts sequences.

We present here our first experiments on finding semantic similarities of words with FastText, an automated method build to handle large datasets. As FastText is based on Word2vec, we present the latter method as well. We have experimented with Word2vec before but it has proven to be challenging with our rather small corpus. FastText should work better with smaller datasets than Word2vec.

We used a dataset consisting of 5,056 **Neo-Assyrian texts** in the Open Richly Annotated Cuneiform Corpus. For the analysis we used **dictionary forms of words**, as the transliterations of a word differ from each other depending on the number, case, etc, and would be considered separate words by an automated method.

We do not suggest that **automated methods** replace traditional philological research in Assyriology. Rather, they **can be used to handle large datasets and to** find regularities that may not be visible when studying one text at a time. The regularities might in turn **raise new research questions**.

## FASTTEXT

**Word2vec** is one of the methods in the field of natural language processing which can predict semantic similarities. Word2vec:

- plots a **unique vector for each lexeme** in the corpus
- has the vectors **in a multidimensional vector space**
  - vectors of **words appearing in comparable linguistic contexts** have similar coordinates and **cluster near each other** in the space
- uses so-called *artificial neural networks*
- was build to handle large datasets of billions of words faster than previous methods.

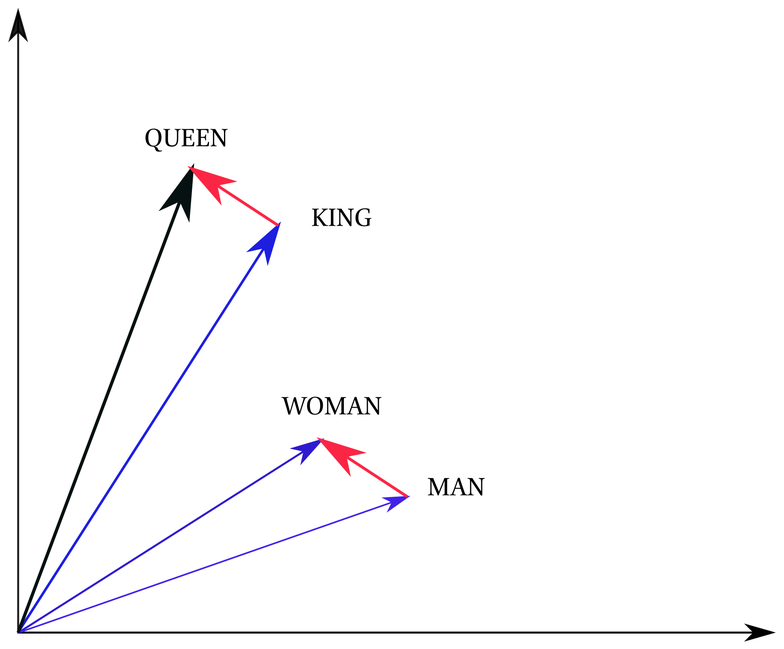
It has been shown that word vectors are better at predicting semantic similarities than traditional methods that count words that appear close to each other.

**FastText** is a method built on Word2vec. While building the word vectors, FastText takes subword information into account by dividing words into smaller sequences of characters. The position of a word in the vector space is determined by adding up the information on all the different sequences of characters derived from that word.

Both Word2vec and FastText have been published as open source toolkits which also contain **scripts for querying semantic similarities of words** one is interested in. We present here the results of some queries we performed to the word vectors of certain Mesopotamian deities in our Neo-Assyrian dataset.

## WHAT IS TO KING AS WOMAN IS TO MAN

There is a famous example of how Word2vec can be used. The user can ask *What is to king as woman is to man*. The program will then take the vector of the word 'king', add to it the distance between the vectors for 'man' and 'woman', and the answer is supposed to be 'queen'.




King + (Man - Woman) = Queen

After building word vectors from our Neo-Assyrian dataset with FastText, one can ask ***What is to Nabû as Zarpanītu is for Marduk***. One would expect the answer to be Tašmētu, the wife of Nabû. And indeed, Tašmētu appears on the list of answers to our query.

**Query triplet (A - B + C)? Zarpanītu Marduk Nabû**

*Bēl*
*Tašmētu*
*Bēlet*
*Bēlet-Ninua*
*kāribu*

It can also be explained how Bēl, Bēlet, and Bēlet-Ninua are to Nabû something similar as Zarpanītu is to Marduk. The presence of the word *kāribu*, 'he who blesses', in our list might be due to Nabû being mentioned in a certain formula of blessing in letters.

## NERGAL IN GOOD COMPANY

From the word vectors built with FastText, we queried for the closest words to the word Nergal. Our dataset is rather small for FastText and the results differ from query to query. We, hence, **built new word vectors 1,000 times** and every time queried for various names of deities. We then **counted the average position of all words** in the lists of words returned by the program. Nergal's closest words, i.e. the words that have the most similar contexts with Nergal according to FastText, are:

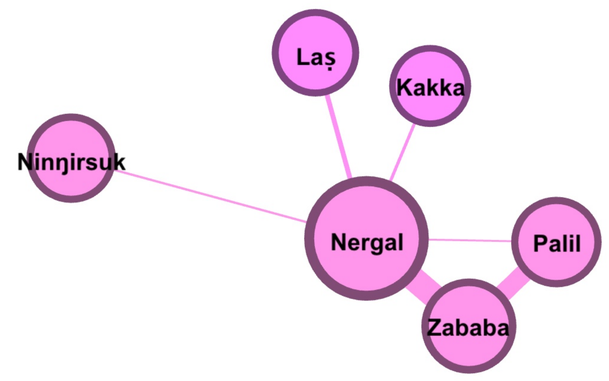
| Word | Average position | |
|---|---|---|
| Zababa | 2 | (warrior god identified with Ninurta) |
| Laṣ | 3.5 | (spouse of Nergal) |
| Kakka | 7.4 | |
| Ningirsu | 9.3 | (warrior god identified with Ninurta) |
| Palil | 10.4 | |
| Gula | 15.9 | |

Laṣ obviously appears in similar contexts as Nergal, being his wife. Zababa and Ningirsu have a connection to Ninurta, who was a god of war like Nergal. Kakka is rarely attested, but his likeness to Nergal in the FastText results derive from the literary text "Nergal and Ereškigal" in which he functions as the messenger of Anu. Gula was the wife of Ninurta, and she was identified with Baba, the wife of Zababa and Ningirsu. Since Zababa, Ningirsu, and Gula had a connection to the war god Ninurta, the contexts where they are attested in our dataset could, indeed, be similar to those of Nergal.

Palil is a rarely attested god. His name is written IGI.DU and this is how Nergal's name is occasionally written. In RINAP 3 153, U.GUR (Nergal) and IGI.DU (Palil?) are attested in consecutive lines so it appears that these words at least sometimes refer to two different gods. The exact connection between the two gods remains unclear. FastText seems, however, to find some similarities between the contexts in which Nergal and Palil appear.

## VISUALISING NERGAL'S COMPANY

We, furthermore, visualised the results we got for various deities using FastText as described above. The small graph here, Nergal's closest "friends", shows the words connected to Nergal, with their connections to each other. When making the graph, each connection between words was given a **weight according to how high** in average **the word was on a list** of the other word. The weight was counted by extracting the average position from 51. Since Zababa had position 2 in *Nergal*'s list, he was given weight 49. *Nergal* was in Zababa's list on position 7.8 and this connection was given weight 44. When the graph was built, the weights of connections in both directions between two words were added up. This way we can see which words are in fact close to each other as word vectors. The **thickness of the lines indicates the added up weight of the connection**.




Nergal's closest "friends" visualised as a graph

From Nergal's graph we can see that the other words (deities) are not as close to Nergal as Zababa, whose connection to him is very "thick". The presence of Palil in Nergal's list might be more due to his closeness to Zababa than to Nergal himself. The connection of Palil, Nergal and Zababa would be an interesting topic to research further.

## CONCLUSIONS

FastTexts gives promising results in connection with deities. **It is not easy to interpret those results** as the contexts where gods appear in the dataset are not self evident. The results may, however, give rise to new avenues of research.

In our analysis, we left out all words that appear in the dataset less than five times. The presence of very infrequent words, such as *Palil*, in the results suggests, however, that **one needs to always consider the number of attestations** of words when evaluating the results. **Visualising the results can**, furthermore, **give new insights to the interpretation**.

As FastText divides words into smaller sequences than a word while analysing, it should work better than Word2vec with highly inflecting languages such as Akkadian. We intend to experiment next with transliterated text rather than dictionary forms.

### References

**Bojanowski** et al., 2017 "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics*, 5:135–146.

**Mikolov** et al., 2013 "Distributed Representations of Words and Phrases and their Compositionality." *Advances in Neural Information Processing Systems* 26.