Word Sense Induction with Attentive Context Clustering

Moshe Stekel, Amos Azaria, Shai Gordin

Word Sense Induction with Attentive Context Clustering

Moshe Stekel[1], **Shai Gordin**[2], **Amos Azaria**[1]

[1]Computer Science Dpt., Ariel University, Israel
[2]Land of Israel Studies and Archaeology Dpt., Ariel University, Israel

Corresponding author: Moshe Stekel , mstekel@gmail.com

Abstract

This paper presents ACCWSI (Attentive Context Clustering WSI), a method for Word Sense Induction, suitable for languages with limited resources. Pretrained on a small corpus and given an ambiguous word (a query word) and a set of excerpts that contain it, ACCWSI uses an attention mechanism for generating context-aware embeddings, distinguishing between the different senses assigned to the query word. These embeddings are then clustered to provide groups of main common uses of the query word. We show that ACCWSI performs well on the SemEval-2 2010 WSI task. ACCWSI also demonstrates practical applicability for shedding light on the meanings of ambiguous words in ancient languages, such as Classical Hebrew and Akkadian. In the near future, we intend to turn ACCWSI into a practical tool for linguists and historians.

Keywords

wsi; wsd; nlp; clustering

I. INTRODUCTION

Natural language expresses human concepts, thoughts, emotions and insights. That is, natural language represents a model of extremely high complexity—the human mind (at least, its communication-driven layers). Some researchers believe that natural language is an environment in which compromise is inevitable when projecting the infinite number of dimensions of human thinking onto the much smaller number of dimensions of human speech **?**. Multiplicity of meaning of a single word, such as polysemy (similarity obtained from a common source) or homonyny (accidental similarity), is therefore an expected product of this compromise. Below are two common examples of word sense ambiguity:

- "I can hear *bass* sounds" versus "They like grilled *bass*"
- "We crossed the river to the other *bank*" versus "Mike deposited the money in his *bank* account"

Humans are able to disambiguate the polysemy/homonyny or understand contextual nuances by using the clues that come from the context of the ambiguous word. One of the fundamental tasks of natural language processing is Word Sense Induction (WSI), a task of automatic discrimination of the different senses of words by finding these contextual clues.

It is difficult to overestimate the importance of accurate Word Sense Induction when dealing with common Natural Language Processing (NLP) tasks, such as Information Retrieval or Search Clustering. Furthermore, historical research seeks to correctly induce the meaning of words in order to resolve doubts about many historical issues. As a good example we can refer to the Akkadian lemma "galû", the meaning of which ranges between the negative shade

of "exile" or "deportation", the neutral shade of "relocation" and the positive one of "appointment". Another example is the Hebrew lemma "zakar", which takes on both the meanings of "memory" and "male". Accurate Word Sense Induction is essential for correct understanding of ancient documents.

In this paper, we present an Attentive Context Clustering WSI (ACCWSI). ACCWSI first creates a word-embedding for each word, which is identical for any context that it appears in. ACCWSI uses the cosine similarity between the words in the context and the word in focus to determine the attention that each word should achieve to form a context aware vector representation for each appearance of the word in focus. ACCWSI then clusters the resulting vectors, such that each cluster represents a different meaning of the word. ACCWSI has demonstrated high practical applicability in languages with limited resources and obtained a very high score by the evaluation framework of SemEval-2 2010 Task 14 **?**. ACCWSI achieved a high score not only with the original training dataset, but also with a training dataset reduced to a fraction of 2.6% of the original dataset, which is comparable to the size of the Hebrew Bible.

II RELATED WORK

Word Sense Induction and Word Sense Disambiguation provide fertile ground for researchers, starting from very early attempts to tackle these non-trivial tasks, such as "simulated annealing" according to human-edited dictionary **?** and employing the "conceptual distance" between contexts **?**, going through later unsupervised methods, that use patterns of word co-occurrence **?** or bigrams of web search results **?**, continuing with "hidden concepts" of the contextual words, that not necessarily overlap with the sense of the ambiguous word **?**, and ending with the most recent solutions like **?**, that uses word substitutions of modern Masked Language Models, such as Google BERT MLM.

Our research was inspired by two main works: the context-group discrimination algorithm **?** from the Context Clustering category and the Google BERT language model **?**. Amrami and Goldberg **?** utilize Google BERT for their WSI method. However, their method does not meet our requirement of being able to induce word senses in languages with limited resources, as training Google BERT on small corpora does not provide sufficient accuracy **?**. The high scores achieved by the BertWSI model in the SemEval-2 2010 Task 14 **?** metrics are credited to the fact that the underlying model was pre-trained by Google on a huge corpus of text. Our solution takes advantage of the basic mechanism of attention **?** underlying BERT without applying the complex process of learning attention weights and thus achieves good results when applied to small datasets. The only weight learning process we use is the Word2Vec **?** model training that requires far fewer resources than attention-based learning. Thus, we provide a practical tool in the study of the meanings of words in resource-limited languages, such as ancient dead languages. The Clustering by Committee work **?** gave us the idea to use a threshold of 0.5 as an acceptable proportion of orphan instances when measuring the quality of a clustering solution (see Section 3.4.3). We also explored Lin's algorithm **?**, which uses the word clustering approach by combining words with similar semantics into sense representations, but it was found less effective when it came to discriminating senses of words in resource-constrained languages.

III TASK AND ALGORITHM

3.1 WSI task definition

The general definition of WSI is automatic detection of the set of senses denoted by a word. A simplified version of WSI can be defined as follows: given a list of lemmatized sentences and a query lemma, find all the sentences in the list that contain the query lemma, and group them so that the instances of the query lemma in one group are semantically similar to each other and

noticeably different from the instances in other groups. This is a simplified definition because, when lemmatizing, we ignore some input information, such as the part of speech, tense etc. Note that ignoring the part of speech information of the target word is attractive, especially for ancient genres in which the archaic syntactic forms of words may provide no part of speech information (for instance refer to some hardly explainable verses of the Hebrew Psalms).

3.2 Attention mechanism

Our method uses the following "basic attention" mechanism: given a target word (query) and its "context", either the whole sentence or some "window" of words containing the query word, each element of the context is evaluated by its cosine similarity to the query word. The result is optionally multiplied by a constant factor and eventually softmaxed. We refer to the result as the "weights of similarity" or "weights of relevance". The closer two words are semantically, the greater is the cosine similarity between their embeddings and, therefore, the appropriate weights of relevance are greater. The original word embeddings of the context members are multiplied by the appropriate weights of relevance and thus the power of every context member is improved or worsened according to its relevance to the query word. When these new context-sensitive embeddings are summed into a single vector, this sum represents a context-aware vector of the query word that embeds its "local sense" with respect to this specific context, where the relevance of each context member is taken into consideration. Figure 1 illustrates this mechanism.

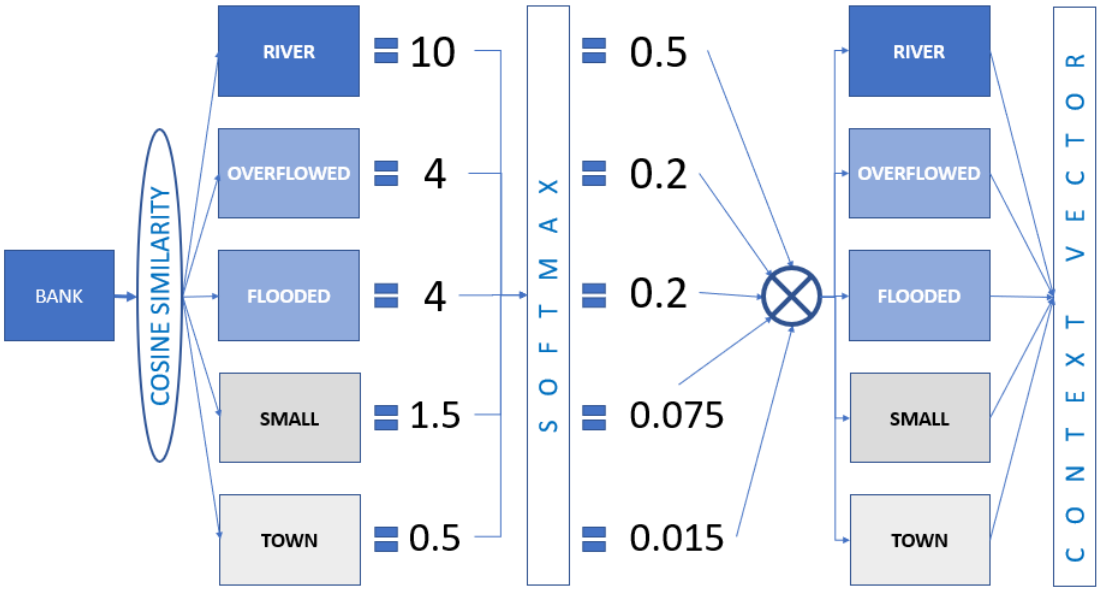


Figure 1: Illustration of the attention mechanism

3.3 The ACCWSI algorithm

We now present our Attentive Context Clustering WSI (ACCWSI) algorithm. The ACCWSI algorithm (see Algorithm III.1) first replaces the lemmas with their Word2Vec embeddings **?**. It then uses the attention mechanism described above (Section 3.2), resulting in context-aware vectors, that are used as input to the DBSCAN clustering algorithm **?**, producing clusters of different "shades of meaning" of the query lemma. Since different contexts are best defined by different most relevant context members, and conversely - similar contexts are defined by similar context members, the result vectors can be easily clustered. Figure 2 illustrates this idea.

Let's explain the algorithm in detail: line 2 runs the language model creation - the process that

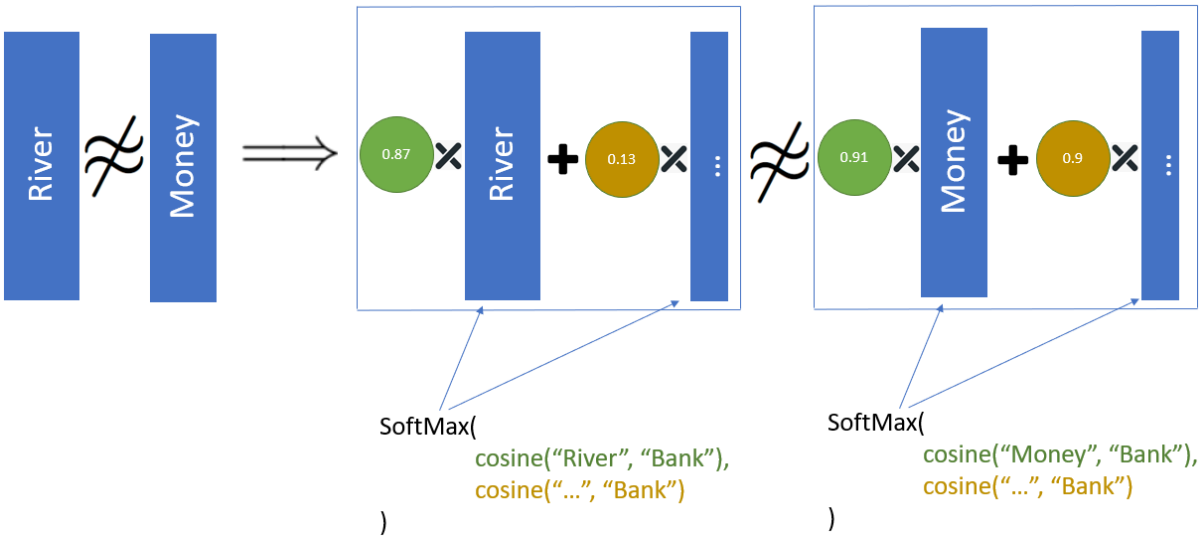


Figure 2: An illustration of separability of context-aware vectors generated by ACCWSI: the most relevant terms (green weights) with respect to the query term "bank" are "river" in the first context and "money" in the second context. They are different and therefore the result context-aware vectors are different. Less relevant terms are multiplied by smaller weights (light brown) and thus have smaller effect on the final context-aware vector.

Listing III.1 ACCWSI Algorithm.

```
1   ACCWSI(text,lemma)
2   model := word2vec(text)
3   sentences := filter_by_lemma ( text ,lemma)
4   ctx_aware_vecs := []
5   for each s in sentences
6   ctx_vecs := model. get_vectors (s)
7   lemma_vec := model. get_single_vector (lemma)
8   sim := cosine_sim ( ctx_vecs ,lemma_vec)
9   sim_weights := softmax(sim)
10  new_lemma_vec := ∑_i ctx_vecs_i * sim_weights_i
11  ctx_aware_vecs .push(new_lemma_vec)
12  return DBSCAN(...). fit ( ctx_aware_vecs )
```

generates multi-dimensional vectors that correspond to the words in the text. The distances between the vectors are supposed to express the semantic distances between the corresponding words. Line 3 finds the sentences in which the query lemma occurs. Lines 5-11 iterate over these sentences, compute the similarity of their members to the query lemma (line 8), normalize the similarity values (softmax - line 9) and then use the normalized values to aggregate the weighted values by summing them, resulting in a list of weighted context vectors. These context vectors are eventually clustered in line 12.

3.4 Hyperparameters

Algorithm III.1 uses several hyperparameters: Word2Vec window, the choice of the clustering algorithm and the internal hyperparameters of the latter. The optimal values of these parameters can be found either empirically or by using well-known optimization methods. In this section we explain these hyperparameters, briefly overview the optimization methods, and present the method that achieved best accuracy in our case. Note that in our online tool we will use our

recommendations for these parameters as their default values allowing the endusers to modify these default values according to the characteristics of their datasets.

3.4.1 Word2Vec Window

This parameter determines the size of the context to be scanned from each direction around the target word when training the Word2Vec model to perform the missing word prediction task (CBoW architecture) or the context prediction task (Skip-Gram architecture). The optimal value of this parameter intuitively depends on the native average "density of context" inherent to the target language. We found the optimal value empirically by iterating over the range from 2 to 10 and evaluating the result by manually checking the semantic similarity of words suggested by the model. The best values were 5 for English and 2 for Classical Hebrew. This difference is probably due to the specific syntactic structures of Classical Hebrew verses, which are statistically much shorter than the syntactic structures of typical Modern English sentences.

3.4.2 The choice of the clustering algorithm

We evaluated several clustering algorithms on our task, including KMeans **?**, Gaussian-Mixture model **?** and DBSCAN **?**. DBSCAN, the density-based clustering algorithm, performed slightly better and was therefore selected as our clustering algorithm for this paper. In our online tool, we will present a list of clustering algorithms to choose from and the end users will be able to choose the best clustering algorithm that suits their datasets.

3.4.3 DBSCAN-eps

This parameter is a key one for the density-based clustering proposed by DBSCAN. It defines the maximum distance between two points to be considered as neighbors. There are several methods in the literature for optimizing the value of this parameter, such as the Kneedle algorithm for finding the maximum curvature in the graph of distances, the Silhouette Score for evaluating the clustering quality, and more. Although these optimization methods demonstrated good performance (unsupervised V-Measure of 15.3%), we propose a heuristic that performed better. The rationale behind the heuristic is that text can contain instances of ambiguous words with highly clear context, in addition to other instances with more obscure context. Decreasing the value of eps results in clearer but tighter clusters, filtering out distant "noisy" instances. In our case, narrowing the clusters while keeping the number of the "noisy" instances below 50% gave good results. Algorithm III.2 demonstrates this heuristic.

For other datasets, the heuristic above may be less effective because there should not be a direct correlation between the amount of noise and the distribution of word senses, so it can always be a good idea to let the user adjust the value of this parameter empirically. In our future online tool for researchers, summarizing this research (see Section VI), we plan to add dedicated controls to tune all ACCWSI hyperparameters.

IV EXPERIMENTAL EVALUATION

We ran an experiment to evaluate the algorithm on **Sem-Eval 2010 Task 14 ?**, which aims to objectively measure and compare the quality of WSI systems. Both training and test data are English sentences containing polysemous or homonymous nouns and verbs. The goal of the task is to split the instances of each ambiguous word and their contexts into clusters representing different meanings. The result is assessed by comparison with the "Gold Standard" clustering performed by human experts. In Section 4.1 we present the Unsupervised V-Measure and F-Score metrics of this assessment as well as the Supervised Recall metric.

4.1 SemEval-2 2010 Task 14 Evaluation

In Task 14 of the SemEval-2 2010 workshop **?**, participants were asked to train their models on the corpus of training data provided by the organizers, and then perform word sense induction for a set of sentences containing both ambiguous nouns and ambiguous verbs. The results

Listing III.2 Fine-tuning the DBSCAN eps hyperparameter - the value of eps is iteratively decreased until the noise (the fraction of the orphan instances) becomes greater than $\frac{1}{2}$

```
1
2       BEST_DBSCAN_EPS(cxt_aware_vectors)
3       best_eps := 0.95
4       for each x ∈ range(90, 0, −5) do
5       eps := x/100
6       labels := DBSCAN(eps = eps).fit(cxt_aware_vectors)
7       noise := labels.count(−1)/len(labels)
8       if noise ≤ 0.5 then
9       best_eps := eps
10      else
11      break
12      end if
13      end for
14      return best_eps
```

were assessed against the "Gold Standard" clusters compiled by human experts. The tables below show the metrics achieved with ACCWSI trained on the full training corpus (by training ACCWSI we mean training its internal Word2Vec model), as well as the metrics achieved with the reduced ACCWSI, which was trained on a randomly selected 2.6% of the training data, along with those of the participants with the highest scores in every metric.

V APPLICATION EXAMPLES

In this section we present examples of applying our method to a relatively small Hebrew corpus—the Hebrew Bible. We used the text-fabric version of the BHSA project to generate the appropriate dataset and run the ACCWSI algorithm on it. Figure 3 shows the operation of the ACCWSI algorithm used to obtain two different meanings of "bank" in English. Figure 4 and Figure 5 present the induced classes for two ambiguous Hebrew Biblical lemmas: **khalal** (dead body/desecrate) and **zakar** (male/memory). The instances of the first lemma were split

System	**VM % (All)**	**VM % (Nouns)**	**VM % (Verbs)**
ACCWSI full	17.3	20.7	12.3
Hermit	16.2	16.7	15.6
UoY	15.7	20.6	8.5
KSU KDD	15.7	18	12.4
ACCWSI reduced	15.4	18.8	10.4
Duluth-WSI	9	11.4	5.7
...			
...			
Duluth-WSI-SVD-Gap	0	0	0.1

Table 1: V-Measure (VM) unsupervised evaluation. V-Measure assesses the quality of a clustering solution by explicitly measuring its homogeneity and its completeness. Homogeneity refers to the degree that each cluster consists of data points primarily belonging to a single Gold Standard class, while completeness refers to the degree that each Gold Standard class consists of data points primarily assigned to a single cluster. V-Measure is the harmonic mean of the homogeneity and completeness.

System	FS % (All)	FS % (Nouns)	FS % (Verbs)
Duluth-WSI-SVD-Gap	63.3	57	72.4
KCDC-PT	61.8	57	72.4
...			
...			
ACCWSI reduced	55.9	51.3	62.7
...			
...			
ACCWSI full	53.8	47.2	63.4
Duluth-WSI-SVD	41.1	37.1	48.2
Duluth-WSI	41.1	37.1	48.2
...			
...			
Duluth-R-110	16.1	15.8	16.4

Table 2: Paired F-Score (FS) unsupervised evaluation: two sets of instance pairs are generated - a set of all possible instance pairs within each induced cluster and a set of all possible instance pairs within each Gold Standard class. Precision is the number of common instance pairs between the two sets to the total number of pairs in the induced clusters, while recall is the number of common instance pairs between the two sets to the total number of pairs in the Gold Standard classes. F-Score is the harmonic mean between precision and recall.

into 2 sense clusters while the instances of the second lemma were split into 5 sense clusters. ACCWSI seems to perform well and provide satisfactory clusters despite the small training corpus.

5.1 ACCWSI performance on ancient Akkadian texts

Another effort we lead these days is Word Sense Induction in ancient Akkadian texts. Between the 9th to the late 7th centuries BCE, the Assyrian Empire deported millions of people across the Near East. By even the most humble estimates, around 1.3 million people were moved around as a result of conquest, labour recruitment or as punishment, just to name the central reasons for this dire process (Sano 2020). However, the records for these deportations are numerous and came down to us in different genres that deal with the act of deportation, or forced migration, from different points of view: contemporaneous Assyrian royal inscriptions, letters and administrative texts, as well as Babylonian historical chronicles, written many years after the events in question. All were written in Assyrian and Babylonian, two close dialects of Akkadian, the oldest known (East-)Semitic language in the world. In all, 19 different verbs deal with various stages of the forced migration, like the capture of people or forced recruitment, their change of location, and resettlement. Even then, there are differences across meanings for specific verbs, sometimes minute ones, but also quite substantial in terms of semantics.

A good example of such a complicated verb is *galû* which the Chicago Assyrian Dictionary (CAD), the most comprehensive dictionary of Akkadian, translates as "1. to go into exile, 2. to deport, to exile (Š-stem, causative)" (CAD Š/3, 201). Its usage is limited to a Babylonian context, either in Assyrian letters dealing with Babylonia or Babylonian chronicles (Sano 2020, 34). As text 1 shows, the usage, much like that of Biblical Hebrew *GLY/H*, is used in consequence of a military conflict. However, a single instance in a letter from the time of Tiglath-pileser III (c. 731-730 BCE), here text 2, shows that under certain political circumstances people could ask for someone to deport them to Assyria, perhaps referring to the safety of being a protected

System	SR % (All)	SR % (Nouns)	SR % (Verbs)
ACCWSI full	63.7	59.6	71.1
ACCWSI reduced	62.7	57.5	69.8
UoY	62.4	59.4	66.8
Duluth-WSI	60.5	54.7	68.9
...			
...			
Duluth-Mix-Uni-Gap	18.7	1.6	43.8

Table 3: Supervised recall (SR) using a test set split with 80% mapping and 20% evaluation. In this evaluation, the testing dataset is split into a mapping and an evaluation corpus. The first one is used to map the automatically induced clusters to Gold Standard senses, while the second one is used to evaluate methods in a WSD setting.

refugee under the direct responsibility of the Assyrian king. This might also be the meaning of certain cases in Aramaic, where *gly* in G-stem active participle means "exile, refugee", or in D-stem means "to emigrate", (Comprehensive Aramaic Lexicon, s.v. gly D and C).

In corpus that is available to us during this experiment (the parts of the Oracc archive that were published as parsable XML files), all cases of *galû* meant either "exile" or "deportation", which are very close to each other and tend to appear in similar contexts. Therefore, ACCWSI produces a single cluster (Figure 6). We therefore demonstrate ACCWSI on the polysemous lemma *kayyānu*, which is found more frequently in Akkadian texts. The results are shown in Figure 7.

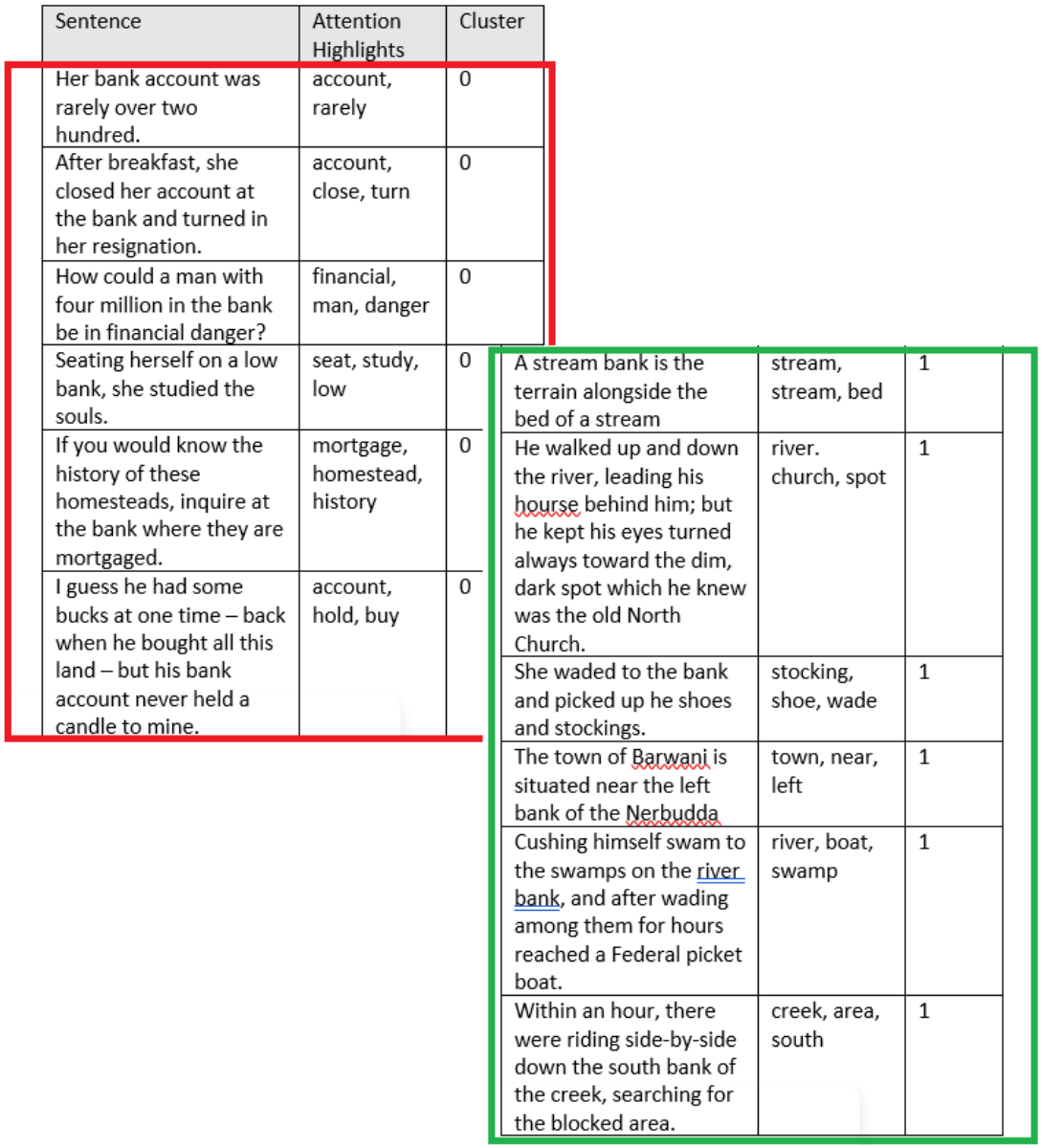
VI FUTURE WORK

Iterating the process of generating context embeddings may improve the accuracy of the clustering. In our future work we plan to develop a method for determining the "center of mass" (or "centroid" for convex clusters) of every cluster. These centers will be treated as new "query" embeddings and the ACCWSI attention-weighted technique will be reapplied within each cluster using its new query (its center). This should provide finer discrimination of meanings. This iterative process can be repeated many times until maximum accuracy is achieved.

The main practical goal of this study is to create an online tool for linguists and historians. This tool will allow the researcher to select a text corpus to scan, a word embedding generation algorithm to use (Word2vec/GloVe/BERT), and specify the values of various hyperparameters that control the ACCWSI flow.

VII CONCLUSION

In this paper we propose ACCWSI, an algorithm to automatically induce various senses of ambiguous words by automatically focusing on the most relevant words from their contexts. After learning generic word embeddings into a Word2Vec model, ACCWSI uses the basic attention technique for determining the most relevant context members and generating context-aware embeddings, each with a semantic direction that aggregates the directions of its context members. Distant meanings imply distant context embeddings and vice versa, and thus standard clustering techniques can be easily applied for grouping the context embeddings by their common semantic directions. ACCWSI has shown excellent performance even when trained on a small subset of the training data in the SemEval-2 2010 task 14. Furtheremore, ACCWSI demonstrated high applicability in disambiguation of word senses in ancient Semitic languages, such as Classical Hebrew and Akkadian.



Sentence	Attention Highlights	Cluster
Her bank account was rarely over two hundred.	account, rarely	0
After breakfast, she closed her account at the bank and turned in her resignation.	account, close, turn	0
How could a man with four million in the bank be in financial danger?	financial, man, danger	0
Seating herself on a low bank, she studied the souls.	seat, study, low	0
If you would know the history of these homesteads, inquire at the bank where they are mortgaged.	mortgage, homestead, history	0
I guess he had some bucks at one time – back when he bought all this land – but his bank account never held a candle to mine.	account, hold, buy	0
A stream bank is the terrain alongside the bed of a stream	stream, stream, bed	1
He walked up and down the river, leading his hourse behind him; but he kept his eyes turned always toward the dim, dark spot which he knew was the old North Church.	river. church, spot	1
She waded to the bank and picked up he shoes and stockings.	stocking, shoe, wade	1
The town of Barwani is situated near the left bank of the Nerbudda	town, near, left	1
Cushing himself swam to the swamps on the river bank, and after wading among them for hours reached a Federal picket boat.	river, boat, swamp	1
Within an hour, there were riding side-by-side down the south bank of the creek, searching for the blocked area.	creek, area, south	1

Figure 3: Two different meanings of **bank**, the financial institute and the geographic terrain, are represented by the clusters in the figure. The "attention highlight" column shows the most relevant context words. The first cluster contains an interesting failure: the fourth sentence is clustered as a financial institute even though a human would cluster it as a geographic terrain. The reason is that the most relevant context words "seat, study, low" are not sufficiently indicative.

The code of the Jypiter notebooks and other utilities we used during our research can be found in the GitHub repository below. The code resources are self-contained and reusable and can be useful in a variety of contexts[1].

[1] https://github.com/mstekel/accwsi

Sentence	Attention Highlights	Cluster
וּמִזִּבְחֲךָ לֹא־תִתֵּן לְהַעֲבִיר לַמֹּלֶךְ וְלֹא תְחַלֵּל אֶת־שֵׁם אֱלֹהֶיךָ אֲנִי יְ־ה־וָ־ה.	נתן, י־ה־ו־ה, עבר	0
לָכֵן אֱמֹר לְבֵית יִשְׂרָאֵל כֹּה אָמַר אֲדֹנָי יְ־ה־וִ־ה לֹא לְמַעַנְכֶם אֲנִי עֹשֶׂה בֵּית יִשְׂרָאֵל כִּי אִם לְשֵׁם קָדְשִׁי אֲשֶׁר חִלַּלְתֶּם בַּגּוֹיִם אֲשֶׁר בָּאתֶם שָׁם.	גוי, ישראל, קדש	0
וְאֶחֱמֹל עַל שֵׁם קָדְשִׁי אֲשֶׁר חִלְּלוּהוּ בֵּית יִשְׂרָאֵל בַּגּוֹיִם אֲשֶׁר בָּאוּ שָׁמָּה.	חמל, גוי, ישראל	0
וַיָּבוֹא אֶל הַגּוֹיִם אֲשֶׁר בָּאוּ שָׁם וַיְחַלְּלוּ אֶת שֵׁם קָדְשִׁי בֶּאֱמֹר לָהֶם עַם יְ־ה־וָ־ה אֵלֶּה וּמֵאַרְצוֹ יָצָאוּ.	עם, גוי, קדש	0
וְנָפְלוּ חַלָּלִים בְּאֶרֶץ כַּשְׂדִּים וּמְדֻקָּרִים בְּחוּצוֹתֶיהָ.	דקר, חוץ, נפל	1
כִּי נָתַתִּי אֶת חִתִּיתִי בְּאֶרֶץ חַיִּים וְהֻשְׁכַּב בְּתוֹךְ עֲרֵלִים אֶת חַלְלֵי חֶרֶב פַּרְעֹה וְכָל הֲמוֹנָה נְאֻם אֲדֹנָי יְ־ה־וִ־ה.	ערל, חרב, נאם	1
אוֹתָם יִרְאֶה פַּרְעֹה וְנִחַם עַל כָּל הֲמוֹנוֹ חַלְלֵי חֶרֶב פַּרְעֹה וְכָל חֵילוֹ נְאֻם אֲדֹנָי יְהוִה.	חרב, נאם, כל	1
שָׁמָּה נְסִיכֵי צָפוֹן כֻּלָּם וְכָל צִדֹנִי אֲשֶׁר יָרְדוּ אֶת חַלָּלִים בְּחִתִּיתָם מְגִבוּרָתָם בּוֹשִׁים וַיִּשְׁכְּבוּ עֲרֵלִים אֶת חַלְלֵי חָרֶב וַיִּשְׂאוּ כְלִמָּתָם אֶת יוֹרְדֵי בוֹר.	ערל, כל, כל	1

Figure 4: In the Hebrew Bible, the lemma **khalal** normally takes on the sense of either **dead body**(as a noun) or **desecrate**(as a verb). This figure presents the appropriate clusters generated by ACCWSI. The "attention highlight" column shows the most relevant context words. In the context of **desecrate** (cluster 0), the attention is paid to words like **God**, **sacred**, **nation** etc. while in the context of **dead body** (cluster 1), the attention is paid to **sword**, **stab**, **fall** etc.

Sentence	Attention Highlights	Cluster
וּבֶן שִׁמֹנַת יָמִים יִמּוֹל לָכֶם כָּל זָכָר לְדֹרֹתֵיכֶם יְלִיד בַּיִת וּמִקְנַת כֶּסֶף מִכֹּל בֶּן נֵכָר אֲשֶׁר לֹא מִזַּרְעֲךָ הוּא.	דור, יום, כל	0
פְּקֻדֵיהֶם בְּמִסְפַּר כָּל זָכָר מִבֶּן חֹדֶשׁ וָמָעְלָה פְּקֻדֵיהֶם שִׁבְעַת אֲלָפִים וְחָמֵשׁ מֵאוֹת.	מספר, שבע, פקד	0
בְּמִסְפַּר כָּל זָכָר מִבֶּן חֹדֶשׁ וָמָעְלָה שִׁמֹנַת אֲלָפִים וְשֵׁשׁ מֵאוֹת שֹׁמְרֵי מִשְׁמֶרֶת הַקֹּדֶשׁ.	מספר, מעל, כל	0
וּפְקֻדֵיהֶם בְּמִסְפַּר כָּל זָכָר מִבֶּן חֹדֶשׁ וָמָעְלָה שֵׁשֶׁת אֲלָפִים וּמֵאתָיִם.	מספר, פקד, מעל	0
זְכֹר רַחֲמֶיךָ יְ-ה-וָ-ה וַחֲסָדֶיךָ כִּי מֵעוֹלָם הֵמָּה.	עולם, חסד, רחמים	1
חַטֹּאות נְעוּרַי וּפְשָׁעַי אַל תִּזְכֹּר כְּחַסְדְּךָ זְכָר לִי אַתָּה לְמַעַן טוּבְךָ יְ-ה-וָ-ה.	פשע, חסד, נעורים	1
פְּנֵי יְהוָה בְּעֹשֵׂי רָע לְהַכְרִית מֵאֶרֶץ זִכְרָם.	כרת, ה', רע	1
אֱלֹהִים עַל נַפְשִׁי תִּשְׁתּוֹחַח עַל כֵּן אֶזְכָּרְךָ מֵאֶרֶץ יַרְדֵּן וְהַרְמֹנִים מֵהַר מִצְעָר.	אלהים, נפש, כן	1
וְאִם מִן הַצֹּאן קָרְבָּנוֹ לְזֶבַח שְׁלָמִים לַי-ה-וָ-ה זָכָר אוֹ נְקֵבָה תָּמִים יַקְרִיבֶנּוּ.	נקבה, תמים, ה'	2
אוֹ הוֹדַע אֵלָיו חַטָּאתוֹ אֲשֶׁר חָטָא בָּהּ וְהֵבִיא אֶת קָרְבָּנוֹ שְׂעִיר עִזִּים זָכָר תָּמִים.	חטא, תמים, קרבן	2
לְהַעֲלֹתָם תָּמִים זָכָר בַּבָּקָר בַּכְּשָׂבִים וּבָעִזִּים.	תמים, רצון, כשב	2
וְאִם זֶבַח שְׁלָמִים קָרְבָּנוֹ אִם מִן הַבָּקָר הוּא מַקְרִיב אִם זָכָר אִם נְקֵבָה תָּמִים יַקְרִיבֶנּוּ לִפְנֵי יְהוָה.	נקבה, תמים, ה'	2
וַיָּבֹא אֶלְיָקִים בֶּן חִלְקִיָּהוּ אֲשֶׁר עַל הַבַּיִת וְשֶׁבְנָא הַסֹּפֵר וְיוֹאָח בֶּן אָסָף הַמַּזְכִּיר אֶל חִזְקִיָּהוּ קְרוּעֵי בְגָדִים וַיַּגִּידוּ לוֹ אֶת דִּבְרֵי רַב שָׁקֵה.	ספר, בן, אסף	3
וַיֵּצֵא אֶלְיָקִים בֶּן חִלְקִיָּהוּ אֲשֶׁר עַל הַבַּיִת וְשֶׁבְנָא הַסֹּפֵר וְיוֹאָח בֶּן אָסָף הַמַּזְכִּיר.	ספר, בן, אסף	3
אֲלִיתֹרֶף וַאֲחִיָּה בְּנֵי שִׁישָׁא סֹפְרִים יְהוֹשָׁפָט בֶּן אֲחִילוּד הַמַּזְכִּיר.	ספר, בן, יהושפט	3
וַיִּקְרְאוּ אֶל הַמֶּלֶךְ וַיֵּצֵא אֲלֵהֶם אֶלְיָקִים בֶּן חִלְקִיָּהוּ אֲשֶׁר עַל הַבַּיִת וְשֶׁבְנָה הַסֹּפֵר וְיוֹאָח בֶּן אָסָף	ספר, בן, אסף	3
וְהַיָּמִים הָאֵלֶּה נִזְכָּרִים וְנַעֲשִׂים בְּכָל דּוֹר וָדוֹר מִשְׁפָּחָה וּמִשְׁפָּחָה מְדִינָה וּמְדִינָה וְעִיר וָעִיר וִימֵי הַפּוּרִים הָאֵלֶּה לֹא יַעַבְרוּ מִתּוֹךְ הַיְּהוּדִים וְזִכְרָם לֹא יָסוּף מִזַּרְעָם.	דור, דור, יום	4
יְ-ה-וָ-ה שִׁמְךָ לְעוֹלָם יְ-ה-וָ-ה זִכְרְךָ לְדֹר וָדֹר	דור, דור, עולם	4
אַזְכִּירָה שִׁמְךָ בְּכָל דֹּר וָדֹר עַל כֵּן עַמִּים יְהוֹדוּךָ לְעוֹלָם וָעֶד.	דור, דור, שם	4
וְאַתָּה יְהוָה לְעוֹלָם תֵּשֵׁב וְזִכְרְךָ לְדֹר וָדֹר.	דור, דור, עולם	4

Figure 5: In the Hebrew Bible, the senses of the lemma **zakar** are related to either **male** or **memory**. This figure presents the five clusters generated by ACCWSI. The "attention highlight" column shows the most relevant context words. The first cluster represents the sense of **male human**, the second one - **God's memory**, the third one - **male animal sacrifice**, the fourth - **the role of scribe** and the fifth - **chronological memory**

origin	sense	attention_highlight	cluster
P363670:šumma šarru lū šarru x x ēribtu ušaglâ x x $MA x	if king or king (palace-)enterer send into exile	[ēribu, šarru, šarru]	0
P334718:nišēšūni sahhir dīnaššu ina libbi ūme ša iglûni turtānu Zeru-ibni irtugumuniššu issaʾulušu mā issurri ibašši memmēnīka rēhe mā memmēnīyāma lāšu lā rēhe ūmâ atâ	people return give in interior day that be(come) deported commander-in-chief 1 call (out) ask saying perhaps exist somebody remaining saying anybody (there) is not not remaining now why?	[Zeru-ibni, issurri, tartānu]	0
P314272:bēlī lū ūda atâ ina irti Arihi ubbalūni ušaglanāši u šû iqṭibi mā Laqaya x	lord may know why? in breast 1 bring be(come) deported and he say saying from Laqe	[Arihu, wadû, wabālu]	0
P334369:Kummayaʾe dayālī ša issu Kumme ana nāgurtu illikūni udīni lā illakūni ammākamma šunu šarru bēlī lišʾal luṣṣīṣi issurri qanni ammūte ušagalušunu šarru bēlī ina	Kummean scout who from 1 to hire come yet not come there they king lord ask investigate perhaps together with that be(come) deported king lord to	[qannu, issurri, uṣṣuṣu]	0
P237270:ana šar mātāti bēlīya aradka Kudurru Aššur Šamaš Bel u Nabu ana šar mātāti bēlīya likrubū ultu muhhi ūmu ša šarru bēlā ušeglanni ṣabtāk u ašbāk ūmussu ana šarri bēlīya uṣalla Nabu-killanni kî	to king land lord servant 1 1 1 1 and 1 to king land lord bless since skull day that king lord be(come) deported kept in confinement and sitting daily to king lord pray to 1 when	[ūmussu, ašbu, šarru]	0
P334504:nušagli o issurri šarru bēlī iqabbi mā ša bēt Ilumma-taklak Lahiraya šunu x x x	be(come) deported perhaps king lord say saying of house 1 Lahiraean they	[issurri, šarru, qabû]	0
P237257:x x x x akī hīṭu ša ana šarri bēlīya ihṭû ultegeli u	in accordance with crime which against king lord sin deport and	[šarru, haṭû, hīṭu]	0

Figure 6: ACCWSI was unable to separate between different contexts of the **galû** lemma. As it can be seen from the translation column, all instances of the lemma in the texts used by us mean either **exile** or **deportation**. These senses are close to each other and appear in very similar contexts.

origin	sense	attention_highlight	cluster
P348683:Šamaš eddēšû nūr ilī kayyānu Marduk mušīm šīmāti murrik ūmī kurâti mupahhir nišī saphāti muqerrubū ahāmeš rūqūti attunuma	1 constantly self-renewing light god steady 1 one who decrees one who lengthens day short one who gathers people scattered one who brings together far apart one you	[šīmāti, eddēšû, ilu]	0
Q004232:ÉN bēl bēlī šar šarrī Šamaš Šamaš dayyān šamê u erṣeti bēl mātāti eddēšû nūr ilī kayyānu muštēšer elâti u šaplâti rē'û kīnu ša tenēšēti attā Šamaš dayyānu ṣīru ša qibīssu lā uttakkaru annašu ilu mamman lā	incantation (label) lord lord king king 1 1 judge sky and earth lord land constantly self-renewing light god constant good guide upper world and lower world shepherd true of people you 1 judge exalted whose command not be(come) changed approval god anybody not	[eddēšû, ilu, nūru]	0
Q004220:x eddēšû nūr ilī kayyān muštēšer elâti u šaplâti rē'û kēnu ša tenēšēti attā Šamaš dayyānu ṣīru ša qibīssu lā uttakkaru annašu mamma lā	constantly self-renewing light god constant good guide upper world and lower world shepherd true of people you 1 judge exalted whose command not be(come) changed approval anybody not	[eddēšû, ilu, nūru]	0
Q005421:tēmēqu Šamaš bēliya ina supêšu ša ilī rabûti puhur mārī ummâni temmēn labīri ippalsūma papāhī u di'ī ihīṭū zamar itūrūnim yâti iqbû appalisma temmēn labīri ša Naram-Sin šar ullu papāhi Šamaš kayyānu mūšab ilūtišu libbī ihdema immeri pānūya papāhi bēlūtišu u di'ī uaṣbbima ina	prayer 1 lord through prayer of god great assembly citizen craftsman foundation original discover cella and (deity's) throne-platform check quickly return me say discover foundation original of king of Agade king distant past cella 1 permanent residence divinity heart be(come) joyful be(come) bright face cella lordship and (deity's) throne-platform inspect during	[hadû, ilūtu, nawāru]	1
Q005490:Esagil u Ezida kaqdâ kayyānāk ašrāt Nabium u Marduk bēlēya aštene''a kayyānam isinātišunu damqātim akīssunu rabîtim in gumâhī paglūti alpū šuklulūti zuluhê damqūtim immer mīru gukkallam ušummu pelâ nūnum iṣṣūrum simat appārim tibik sīrāš lā nībi māmiš karānam šattišam in nuhši u hegallê in maharišunu ētettiq nišim rapšāti ša	temple of Marduk at Babylon and temple of Nabû at Borsippa constantly constant shrine 1 and 1 lord constantly seek out constantly festival good (a cultic festival) great with prize bull massive ox perfect (a kind of long-fleeced sheep) good sheep fattening (process) fat-tailed sheep (a rodent) egg fish bird appropriate symbol marsh outpouring (a kind of) beer not number like water wine yearly in abundance and plenty in front parade people extensive that	[kaqdâ, ešertu, še'û]	1
Q005600:Nabium-kudurri-uṣur šar Babili rē'û kīnim migir Marduk iššakku ṣīri narām Nabu rubâm na'dam ša alkakāt Marduk bēlu rabi'um ilu bānīšu u Nabium apilšu kīnim narām šarrūtišu ištene''û kayyānam mūda'u telê ša irammu puluhti ilūtišunu ana ṭēmu ilūtišunu bāšâ uznāšu eršu itpēšu ša ana	Nebuchadnezzar II, king of Babylon king Babylon shepherd true favourite 1 city-ruler exalted loved one 1 prince attentive of way 1 lord great god creator and 1 heir true loved one kingship constantly seek out constantly one who knows very able one of love reverence divinity for decision divinity existing wisdom wise one expert of for	[še'û, puluhtu, mūdû]	1
Q003704:rabêya aštê''â ašrāt ilāni rabûti šangûtī ihšuhū irammū nadān zībīya Adad zunnīšu umaššera Ea upattira nagbīšu 5 ammat ê išqu ina absinnišu ērik šubultu 5/6 ammat ešēr ebūri napāš Nisaba kayyān ušahnabū gipāru ṣippāti šummuhā inbu būlu šutēšur ina tālitti ina palêya nuhši ṭuhdu ina šanātiya kummurū hegallum 10 imēr ê 3	adulthood constantly seek out shrine god great priestly office desire love giving food offering 1 rain release 1 loosen underground water a unit of length grain be(come) high in furrow be(come) long ear of corn a unit of length success harvest abundance 1 constantly make grow luxuriantly pastureland fruit orchard very luxuriant fruit livestock success in offspring during reign (of a king) abundance abundance during year piled up plenty a unit of capacity grain	[napāšu, ebūru, šummuhu]	2
Q003707:seheriya adi rubêya aštê''â ašrāt ilāni rabûti šangûtī ihšuhū irammū nadān zībīya Adad zunnīšu umaššera Ea upattira nagbīšu 5 ammat ê išqu ina absinniša ērik šubultu 5/6 ammat ešēr ebūri napāš Nisaba kayyān ušahnabū gipāru ṣippāti šummuhā inbu būlu šutēšur ina tālitti ina palêya nuhši ṭuhdu ina šanātiya kummurū hegallum 12	childhood to ruler constantly seek out shrine god great priestly office desire love giving food offering 1 rain release 1 loosen underground water a unit of length grain be(come) high in furrow be(come) long ear of corn a unit of length success harvest abundance 1 constantly make grow luxuriantly pastureland fruit orchard very luxuriant fruit livestock success in offspring during reign (of a king) abundance abundance during year piled up plenty	[napāšu, ebūru, šummuhu]	2
Q003703:absinnišu ērik šubultu 5/6 ammat ešēr ebūri napāš Nisaba kayyān ušahnabū gipāru ṣippāti šummuhā inbu būlu šutēšur ina tālitti ina palêya nuhši ṭuhdu ina šanātiya kummurū hegallu 10 imēr ê 1 imēr karānī BANMIN šamnī 1	furrow be(come) long ear of corn a unit of length success harvest abundance 1 constantly make grow luxuriantly pastureland fruit orchard very luxuriant fruit livestock success in offspring during reign (of a king) abundance abundance during year piled up plenty a unit of capacity grain a unit of capacity wine oil	[napāšu, ebūru, šummuhu]	2
Q003705:nišī āšibūti Qirbit kayyān ihtanabbatū hubut Yamutbali ālu šuātu ina tukulti Aššur Sin Šamaš Bel Nabu Ištar ša Ninua Ištar ša Arba-ili akšud aššula šallatsu Tandaya itti	people inhabitant 1 constantly constantly loot plunder 1 city that through help 1 1 1 1 1 1 of Nineveh 1 of Arbela conquer carry off booty city ruler of Qirbit with	[habātu, nišu, Yamutbal]	3
Q006597:ṭūdāt mātišu ušharrirūma U₂ x ṢU x x x x alāk gerri kayyān ihtanabbatū iprusū x	path land lay waste course road constantly constantly loot block	[parāsu, habātu, mātu]	3
Q003702:qereb Harehasta lū allik ša Tandaya ana šarrāni abbīya lā iknušu ana nīri u nišī āšibūti Qirbit kayyān ihtanabbatū hubut Yamutbali ālu šuātu	interior 1 may go of city ruler of Qirbit to king ancestor not bow down to yoke and people inhabitant 1 constantly constantly loot plunder 1 city that	[habātu, nišu, Yamutbal]	3

Figure 7: In the Oracc corpus, containing ancient Akkadian texts, the senses of the lemma **kayyānu** occur in various contexts related to **constant** or **permanent**. This figure presents the four clusters generated by ACCWSI. The "attention highlight" column shows the most relevant context words. The first cluster represents the sense of the stable expression of **self-renewing light god**, the second one means **permanent place of divinity/shrine**, the third one is also a stable expression **constantly make grow**, the fourth is also a stable notion of **permanent loot**.

ACKLOWLEDGMENTS
This research is supported by the Ministry of Science & Technology, Israel, Grant 3-16464.