26

Language technology approach to "seeing" in Akkadian

Aleksi Sahala and Saana Svärd

Introduction

In this chapter we discuss Akkadian verbs that connote "seeing"—more specifically, *amāru*, *naṭālu*, *palāsu*, *dagālu*, *barû*, *ḫiāṭu*, and *ṣubbû*.[1] The theoretical framework for our analysis is based on cognitive linguistics. The basic idea is that one of the few ways we can attempt to understand the perspective of ancient people (emic approach) is by analysing the vocabulary they used. This perspective is based on the work done in cognitive linguistics, which has demonstrated that native speakers of different languages (with differing conceptual categories) can have marked differences in how they see the world.[2]

Our research data was collected in August 2018 from the Open Richly Annotated Cuneiform Corpus (ORACC), one of the largest digital corpora of Akkadian and Sumerian texts (see further below). In total, our collection of data consisted of 9,086 lemmatised Akkadian texts comprising about a million words. In this study, we decided to use all of the lemmatised Akkadian language data that was available to us via ORACC for two reasons. First, some of the seeing verbs were too rarely attested in any one sub-corpus to be statistically analysed. Additionally, the usefulness of the method for analysing rarely occurring words is limited, simply because rarely occurring words can be analysed with qualitative methods just as easily. Second, using all of the data from ORACC helped us identify period-specific and genre-specific uses of words.

For the statistical analysis we applied a word association measure called Pointwise Mutual Information (PMI), which is a well-established method for studying distributional semantics and one of the most important concepts in Natural Language processing (Jurafsky and Martin 2019: 108). Our aim was twofold. First, to establish a semantic field for "seeing" and second, to demonstrate that PMI is a useful tool for analysing Akkadian texts. This second aim was achieved by comparing our results with the results reached by Ainsley Dicks in her 2012 dissertation, "Catching the Eye of the Gods: The Gaze in Mesopotamian Literature." Our results indicate that not only is the proposed methodology viable, but it can also open up new avenues of inquiry for the study of senses in Mesopotamia. Specifically, the results show that PMI was able to highlight several semantic aspects of the verbs of seeing that Dicks (2012) had also observed. The method can be utilised on two levels. First, it can be used to acquire an overall picture of typical collocates and their genre/period distributions. Second, it can be used to sample a small, yet

DOI: 10.4324/9780429280207-27

statistically significant selection of co-occurrences for close-reading that often reveal the most typical semantic aspects of a word.

Here, we begin by briefly discussing the question of semantic fields, then proceed to our data set and the nature of our method. The actual analysis is presented in the last section, followed by a short conclusion.

Semantic fields

In our previous work, we have emphasised the importance of emic research approaches that aim to understand the meanings of words and cultural concepts from an insider's perspective, in other words, from the point of view of the social group that is being studied (Svärd et al. 2018: 226, 229–230). For us, the group being studied is the people who used the Akkadian language. Naturally, the fragmented cuneiform texts preserved from Mesopotamia do not reflect the full complexities of this living language: few people knew how to read and write in Mesopotamia and writing mostly documented elite concerns. Nonetheless, we posit that "behind" the textual evidence there existed a section of the population that used Akkadian as a means of building a common world view and understanding of the world. This hypothesis is based on cognitive linguistics.

Cognitive linguistics is a field of study that encompasses varied approaches, yet the different forms of cognitive linguistics share the hypothesis that our understanding of the world as human beings is mediated by language. In other words, we only make sense of the world through the linguistic categories that we use: "Language, then, is seen as a repository of world knowledge, a structured collection of meaningful categories that help us deal with new experiences and store information about old ones" (Geeraerts and Cuyckens 2010: 5).

This perspective is useful for ancient Near Eastern studies. Although we have a variety of archaeological evidence from ancient Mesopotamia, it is its rich corpus of textual evidence that sets it apart from many other historical periods. We maintain that analysing individual lexemes will help us to better understand the cognitive categories of the people using the ancient language (see also Chapter 28, this volume). Although lexemes do not define such semantic fields comprehensively, we can still use the statistical information regarding individual lexemes and the relationships between them to explore these fields.

The concrete methods of how to map semantic fields for individual lexemes originate from the fields of language technology and corpus linguistics. In general, language technological methods have been rarely applied in the field of ancient Near Eastern Studies, especially in terms of text analysis.[3] For us, these quantitative methods open up new avenues of inquiry that complement the more traditional philological work.

The method chosen for this study is based on our previous work. In our previous research (Alstola et al. 2019; Svärd et al. 2021), we have used three well-established language technological methods for calculating word association measures and word embeddings: Pointwise Mutual Information (PMI) (Church and Hanks 1989), Word2vec (Mikolov et al. 2013), and fastText (Bojanowski et al. 2017).[4] As the work of our research group progressed, however, we realised that these methods produce quite similar results with the Akkadian data set (Svärd et al 2018; 2021). Accordingly, for this study we chose to only use PMI as it is easy to implement and understand, and it produces consistently good results that can be tracked back in the primary sources by using bigram searches.[5] Also, as this study focuses on verbs, PMI provided a way to study their meaning on the basis of their typical arguments (especially subjects and objects). In a nutshell, PMI is based on the old idea from linguistics, that a word's meaning can be better understood by identifying the company that it keeps (Firth 1957).

Researching lexical semantics is of course a topic of research that has been advanced in Assyriology—usually via dictionary work. One of the aims of the most extensive current dictionary—*The Assyrian Dictionary of the Oriental Institute of the University of Chicago* (CAD)—is to catalogue all of the nuances and possible meanings of a given word, in order to facilitate the task of translating new cuneiform texts. What is more, by providing the reader with the original contexts of a given word (sometimes very extensively), the dictionary implicitly subscribes to the idea that the meaning of the word is created in its association with other words. What PMI does, then, is an extension of this goal: PMI provides a complete statistical analysis of these contexts. This provides us with a more complete picture of the semantic field of a word, and enables us to launch studies that engage with linguistics as a field (including translation studies as well as sociolinguistic studies). In the current study, we use the results regarding "seeing" by Dicks (2012) as a point of comparison for our statistical analysis. Her comprehensive analysis made the use of CAD superfluous for these particular verbs.

Data and preparations

As noted in the introduction, our data was collected in August 2018 from ORACC.[6] We used only the lemmatised segments of the Akkadian corpus, because our method was unable to produce meaningful results directly from transliteration due to spelling variation and morphological complexity of the Akkadian language. In total, our collection of data consisted of 9,086 Akkadian texts comprising about a million words. These texts come from all available time periods and genres, excluding those tagged as lexical lists (see Table 26.1). Although the lexical lists would have significantly contributed to the mass of our corpus (over 10,000 texts and 550,000 words), we were more interested in the use of verbs within the language's syntactic constraints.

About 58 per cent of our texts dated to the Neo-Assyrian period, and nearly half of the remaining texts dated to the Seleucid and Hellenistic periods. The Neo-Assyrian texts consisted mostly of administrative letters, royal inscriptions, and legal transactions. In turn, the majority of the texts from the Hellenistic period consisted of legal texts, and the Seleucid texts were mostly omens and incantations. Coincidentally, our data consisted of very few literary texts (43,000 words and 202 texts in total), which makes our data set very different from the data that Dicks (2012) used in her study. Overall, Dicks (2012: 2–3) examined both the Sumerian and Akkadian literary corpus, and the main part of her Akkadian material included Benjamin Foster's volume *Before the Muses* (2005), Andrew George's comprehensive *Babylonian Gilgamesh Epic* (2003), and most of the royal inscriptions from Mesopotamia, including volumes from the series Royal Inscriptions of Mesopotamia (RIMB 2; RIME 1; RIME 2; RIME 3/1; RIME 3/2; RIME 4; RIMA 1; RIMA 2; RIMA 3) and from the series Royal Inscriptions of the Neo-Assyrian Period (RINAP 4).

Table 26.1 Most prominent time periods, their word counts, and most common genres in our data for "seeing" in Akkadian

Neo-Assyrian (565k)	Admin. letter (31%)	Royal inscription (19%)	Legal transactions (11%)
Hellenistic (111k)	Legal (93%)	Unspecified (3%)	Literary (2%)
Seleucid (60k)	Omen (34%)	Incantation-ritual (30%)	Astrological (14%)
Achaemenid (49k)	Legal (40%)	Omen (26%)	Mathematical (11%)
Old Babylonian (45k)	Administrative (38%)	Mathematical (36%)	School (16%)

Note: More comprehensive tables can be found in our Zenodo repository.

We pre-processed the data to increase its usability for the collocation analysis. We filtered out all of the words that had been tagged as pronouns, prepositions, numerals, and different particles directly from our data set. This way, we were left with a text corpus where the remaining words carried more semantic weight. The filtered words, as well as *lacunae,* were replaced with an underscore to preserve their original positions in the texts. We also normalised all proper names by grouping them into categories (royal, divinity, person) according to ORACC's internal metadata. This allowed us to see a more general picture of the verbs' collocates and prevented different proper nouns from overcrowding our results. The pre-processed data set is available in our Zenodo repository, along with the scripts and results.[7]

Methods

In every natural language, individual words can be combined into phrases and sentences only by following a rather complex set of syntactic and semantic rules. Syntactic rules dictate the constraints for order and form of the words, while semantic rules designate which individual words can be used in given positions in order to create understandable and meaningful expressions. When syntactic and semantic rules are applied in practice, they tend to create patterns. The verb "to hit" is typically accompanied by an *object* that receives the impact, and an *instrument* or a *subject* that physically touches the object. Countless different options are possible, but to give a few, we could expect, (A) *a man* and *a car*, (B) *a nail* and *a hammer*, or even, (C) *a bear* and *a steam locomotive*. It is quite obvious to expect that in any real-world corpus of news reports, the option A would be the most frequently attested one, while B would be significantly rarer and C almost certainly unique.

Now, if we count how many times the words of each option are attested close to each other, and we know the size of our corpus, we can calculate the probabilities of their co-occurrence. This probability is called a joint distribution $p(a,b)$, where a and b stand for the co-occurring words. As we speculated already by their assumed frequencies, the joint distribution $p(man, car)$ would be significantly higher than $p(bear, steam\ locomotive)$. The information these probabilities alone give us is not that interesting, unless we also know how likely it would be for the words to co-occur if all the syntactic and semantic information were the same, and all of the words just co-occurred independently. This probability, or rather the chance of independent co-occurrence, can be expressed as the product of the words' marginal probabilities (in other words, the probability that a given word occurs in the corpus). Thus, the chance of *man* and *car* co-occurring independently would be $p(man)p(car)$.

When we divide the joint distribution of our words by the chance of their independent co-occurrence, the quotient of this calculation can be used as an association score, which reveals if the words' co-occurrence is statistically significant or not. If the words co-occur more often than it would be reasonable to expect by chance, the words can be considered as collocates.

If the collocates are ordered by their association score, their order will most likely be quite different to the order of their joint distributions. As the words *man* and *car* are likely a lot more frequent than *hammer* and *nail*, the association score for pair B would be higher than that of pair A. In the case of *bear* and *steam locomotive*, it is possible that these words would be attested very rarely in the corpus. In the extreme case, only once as this very expression. This would mean that this pair would get a perfect association score, because in the context of our corpus, the words would only appear together and never independently.

The statistical measure of association described above is called Pointwise Mutual Information (PMI) in modern literature. By formal definition, it is the logarithmic ratio of the joint distribution of two words to the probability of their co-occurrence under the assumption of independence (Church and Hanks 1989):

$$PMI\left(a,b\right)=\log_2 \frac{p\left(a,b\right)}{p\left(a\right)p\left(b\right)}$$

To prevent hapaxes similar to our *bear* and *steam locomotive* example rising to the top ranks of our association scores, it is usually wise to define a frequency threshold that prevents any collocates from having a score unless they are attested at least a certain number of times in the corpus. Additionally, the joint distribution can be squared to reduce the impact of this low-frequency bias even further (Daille 1994). This improvement of PMI is called PMI^2, and it is formally defined as follows:

$$PMI^2\left(a,b\right)=\log_2 \frac{p\left(a,b\right)^2}{p\left(a\right)p\left(b\right)}$$

In PMI^2, the maximum score that a collocate can get is 0, which means that the collocate is only attested close to our keyword, that is, the word for which we are interested in finding collocates. The minimum score is -∞, which in turn means that the words are in a complementary distribution and never attested close to each other.[8] The maximum allowed distance between the keyword and its collocates is referred to as a window. Small window sizes are useful for finding words that occur in fixed expressions or form compound words or idioms, whereas larger ones can be used to detect semantic dependencies.

For this study we chose to use PMI^2 with a symmetric window of seven words, which meant that the collocates could co-occur at a maximum of seven words before or after our keyword (the keyword and the collocate were included in this count). This choice was based purely on trial and error with the Akkadian data: very small windows seemed to leave out interesting collocates, and very large windows tended to overcrowd the results with words from long formulaic expressions and repeating content.[9] The frequency threshold was set to three, and thus all collocates that did not co-occur with our keywords at least three times were discarded. Much higher frequency thresholds were not convenient to use, as the median word frequency in our data set was only four. We took into account the top 50 collocates for each keyword. Naturally, for some rare verbs a top list of 50 collocates entailed searching all occurrences of the verb and examining its use by close-reading. However, we preferred using the same amount of top collocates for all verbs for the sake of consistency. The full lists of collocates, as well as statistics on our data set can be found in our Zenodo repository (see note 7).

Analysing the collocates

We observed semantic relations between words on two levels. At first, the results were examined just as a list of collocates. This high-level analysis provided a quick oversight on the use of certain verbs. For example, the collocate list of *amāru* consisted of several astronomical objects, whereas such objects were completely absent in the results of several other verbs of seeing. Similarly, the collocate list of *barû* contained a lot of scribal terminology, something that the other verbs of seeing did not have. This information alone told us something about the semantic distribution and general usage of the verbs. Second, a lower-level analysis involved actually looking at the contexts where the keyword-collocate pairs were attested. This could be considered as a kind of selective close-reading, where the PMI guided us only to the statistically most relevant contexts where certain words were used. Although the results sometimes comprised dozens or even several hundreds of examples for a given keyword-collocate pair, we found that usually a superficial close-reading of two to five contexts was enough to understand the general connection between the words—if such a connection existed.

In several cases the collocate lists included words that did not necessarily give us any information on the semantic domain of our keyword. Typical cases included segmentation errors in bilingual texts, which caused a few Sumerian words to appear in our results. Problematic collocates also emerged from formulaic and repeating expressions. As the PMI works by counting words inside a given window, it was not uncommon for several adjacent words of a repeated passage to be present in the collocate list concurrently.

To ease the interpretation process of our PMI results, our collocation extraction script generated a search link of every keyword-collocate pair to Korp, which housed "ORACC in Korp" (Jauhiainen et al. 2019).[10] Korp is an online service provided by the Language Bank of Finland that supports complex multi-word search queries and represents data in keyword in context (KWIC) view.[11] From Korp, the search results could be further backtracked to original publications and images of the tablets (if available) through line-by-line aligned links to ORACC. Here, we discuss our results for each verb of interest. We begin by presenting our observations on the Akkadian verbs of seeing and comparing them with Dicks' (2012) philological study.

amāru

With a frequency of 2,769, *amāru* was the most well-attested verb of "seeing" in our data set. It was also the most prominent verb of "seeing" by both absolute and relative frequency in every genre except for treaties.

The semantic field of *amāru* is rather wide and it apparently does not have as specific a function as other verbs of seeing. It can refer to being visible or seeing something concretely or in a more abstract sense. Many of its best ranked collocates came from Neo-Assyrian royal inscriptions, Seleucid and Neo-Assyrian omens and astrological texts/reports, as well as Old Babylonian mathematical problem texts.

In royal inscriptions, the highest ranked collocate *mušarû*, "royal inscription," indicated that the verb is occasionally used to mean "to read," literally "to see an inscription." In omens, *amāru* was most commonly associated with seeing or observing various things that might be interpreted as ominous signs. In several contexts the verb was in the passive voice and referred to something being or becoming visible. Typical collocates included words for different celestial objects and phenomena such as *bibbu*, "planet"; *ṣīt šamši*, "sunrise"; *Šihṭu*, "Mercury"; *Dilbat*, "Venus"; *Makru*, "Mars"; or *kakkabu*, "star"; but also (ominous) animals like *muraššû*, "wild-cat," and *azaru*, "lynx, wild-cat," were found.

Some uses of *amāru* indicated that the verb has a connection to understanding and experience. About one fifth of the collocates came from (especially Old Babylonian) mathematical texts, in which the verb's several collocates consisted of geometrical and mathematical terminology (*igû*, "reciprocal"; *eperu*, "volume"; *šiddu*, "length"; and *mēlû*, "height"). Here, the verb *amāru* seems to have been used as a verb of understanding by seeing. Use as a verb of experiencing emerged from the first millennium omen texts, where someone is expected to see, that is, to personally experience, a financial loss (*ibissû*).

Although not all of the characteristics of *amāru* discussed by Dicks were represented in our results, the general description of the verb seemed to be in line with her findings. Dicks (2012: 115) associated *amāru* with vision as perception instead of action. In our results the perceptual nature of the verb was present in its use in the N-stem, especially in astronomical texts. According to Dicks (2012: 120, 128), *amāru* also has a relation to experience and knowledge, and it can connote a higher degree of cognitive participation. Both of these aspects were also represented in our results: the aspect of knowledge or comprehension was present in mathematical texts and the relationship with experience came from the collocate *ibissû*. However, our results lacked two central characteristics of *amāru* mentioned by Dicks: first, the use of *amāru* as a verb of "seeing" dreams (Dicks 2012: 115), and second, the need for light in order to *amāru*. The words *šuttu*, "dream," and *nūru*, "light," did both appear on the collocate list, but only if we extended it beyond the top 50 results.

naṭālu

In our results, *naṭālu* was used almost exclusively in text types that connect with the highest possible scribal education. A majority of the collocates emerged from the first-millennium omens and literary texts, although texts from the latter genre were underrepresented in our data set. Also, in general, 70 per cent of the occurrences of *naṭālu* emerged from these two genres of texts.

The semantics of *naṭālu* seem to be close to *amāru*: they both can take tangible or intangible objects and be used in semantically similar contexts such as seeing or becoming visible. One-fourth of the collocates came from texts regarding extispicy, which at first glance suggested that *naṭālu* was used when omens were interpreted from intestines. However, a closer look revealed that in these contexts *naṭālu* was merely used to indicate different parts of intestines "looking at" or "facing" each other, rather than the diviner interpreting the ominous marks in them.

In ritual texts and hymns of the Neo-Assyrian, Seleucid, and Achaemenid periods one of the statistically most relevant uses of *naṭālu* was dreaming, or literally "seeing dreams (*šuttu*)." This suggested that the verb denotes vision as perception similarly to *amāru*. Some celestial phenomena such as *qarnu*, "horns of the Moon," and *antallû*, "eclipse," co-occur with *naṭālu* in very similar contexts where *amāru* is used with planets and stars.[12] There is no obvious explanation or pattern why *naṭālu* is sometimes preferred. Additionally, a few collocates of *naṭālu* came from rather dramatic passages in literary texts and Neo-Babylonian royal inscriptions. In these passages seen objects are the cut off (*ru''umu*) wings (*kappu*) of Anzu,[13] someone dying (*dâku*), and a pile of bodies (*pagru*) visible for the birds of prey. The examples were too few to make any serious conclusions, if the dramatic character of the observed event is connected to the choice of the verb.

Dicks (2012: 179) categorises *naṭālu* generally as a verb of vision as perception, but with a higher degree of intellectual involvement than with *amāru*. In our results, only the perceptive aspect as dreaming is clearly represented. Dicks (2012: 185) also mentions the previously discussed use of *naṭālu* as a verb of "facing towards something" in extispicy reports and states that the meaning is purely directional and does not involve a visual component. The speculated use of *naṭālu* in dramatic contexts is not discussed by Dicks (2012: 177), although she mentions that the verb may involve an emotional component.

palāsu

The collocates of *palāsu* (or rather *naplusu*, as it usually is attested in the N-stem) came mostly from Neo-Assyrian royal inscriptions, but a few were also found in ritual texts and incantations of the same period. In total, 55 per cent of the attestations of *palāsu* came from these three genres of texts. Most of the collocates were related to higher beings, especially deities, looking at royal persons and their deeds. Some examples included *epištu*, "deed"; *ilu*, "god"; *bēltu*, "lady"; *šarratu*, "queen"; and *paššuru*, "offering table"; as well as several names of individual deities, which in our data were merged into one categorical keyword. In the royal inscriptions, the action of *palāsu* was always favourable and beneficial to the person being looked at, as seen from collocates such as *balāṭu*, "life"; *damqu*, "good"; *šarāku*, "to grant something"; and *kūnu*, "firmness." In several contexts, the verb was also modified by adverbial *ḫadîš*, "joyfully," although the word did not directly show up in our results. In ritual texts the look could also be malevolent, as in the case of "the evil eye looking into hiding places (*šaḫātu*) in order to empty them." Thus, the beneficial aspect of the action did not inherently come from the verb, but rather from the subject's intentions. In seven instances, *palāsu* described in rituals and incantations was performed through a window (*aptu*) or into a hiding place (*šaḫātu*). Thus, the verb can also have a meaning "to peek in."

In addition to "looking at/being looked at" and "peeking in," *palāsu* was also used as a verb of examination in Seleucid ritual-incantations. The examples were limited to one collocate, *zumru*, "body (of an animal)," which acted as the object of *palāsu*. There was also one collocate, *bakû*, "to weep," that added a sad, emotional aspect to the action of *palāsu* in the Neo-Assyrian period. However, the examples were too few to draw any further conclusions.

Dicks' (2012: 146) interpretation of *palāsu* as a verb of active vision or gaze agrees with our results. The verb was used only with tangible objects and there was no inherent benevolence or malevolence involved, despite the fact that the verb was often found in such contexts (2012: 154). The aspect of joyfulness (2012: 162) was only indirectly discernible in our results as an adverbial modifier *hadîš*. Dicks does not mention the use of *palāsu* as a verb of peeking in, but rather associates a similar function with another verb, *ḫiāṭu* (2012: 213). Some uses of the verb were not represented in our results, namely discovering temple foundations, scrutinising, and looking upwards (2012: 154, 168). We also could not find examples of *palāsu* being used as a verb of active vision (2012: 154–158).

dagālu

Instances of *dagālu* came mostly from royal inscriptions and letters, which constituted about 74 per cent of the verb's total occurrences. It was also the most common seeing verb in treaties, although with only 14 attestations. Most of the collocates in the Š-stem came from Neo-Assyrian royal inscriptions, where waterways (*miṭirtu*), meadows (*tawwertu*), regions (*nagû*), lands (*mātu*), and power (*bēlūtu, šarrūtu*) are transferred and entrusted, that is, given to be looked after, to the local people and their leaders. This use of the verb appeared in the expression *pānu* + *šudgulu*. The verb also appeared with *pānu* in the G-stem: *pānu* + *dagālu*, where its meaning seemed to be more of an abstract kind of looking, namely "looking forward to," that is, waiting for something to happen. In four instances it was used synonymously with its collocate *waqû*, "to wait."

Another commonly found collocate *zāqipu*, "stake," referred to people being forced to watch public impalements of the king's enemies. Because the verb was not often attested in the G-stem in our results, its meaning was difficult to separate from that of *palāsu*. Nonetheless, one could speculate that *dagālu* involves "looking at/watching" or "looking after" something for a longer period of time and with a higher focus than *palāsu*, which seems to refer more generally to looking at something.

Our results agree to a great extent with Dicks' (2012: 194–195, 198) observations. She describes *dagālu* as a verb of gaze connoting fixed attention and prolonged duration, and associates it with waiting. In the Š-stem, she describes the meaning as "forcing someone to watch something (often repellent)," which also matches with our example of watching executions.

Additional meanings of the Š-stem— "to be subject to," "to attend to," and "to belong to"—in expressions with *pānu* (2012: 195) corresponded to our interpretation of the verb as meaning "to entrust something to someone."[14] A very rare instance of *dagālu* taking an intangible object was not present in our result; only one example of this is given in Dicks (2012: 163).

barû, ḫiāṭu, *and* ṣubbû

The verbs *barû*, *ḫiāṭu*, and *ṣubbû* can be summarised as "verbs of examination." They form a well-formed hierarchical semantic range and are connected by contexts where the quality or quantity of something is being monitored visually. Such contexts include, but are not limited to, surveying structures, checking the correctness of written documents, or visually confirming that a certain quantity of valuable resources is present. In addition to the qualitative versus quantitative examination, another key difference observed between these verbs appears to be related to the distance of observation.

The verb *barû* seems to be a verb of close qualitative examination, perhaps in a more profound way than any other verb of seeing. Most of its prominent collocates came from Neo-Assyrian and later periods, where *barû* was often attested in quality assuring formulaic expressions like GIM SUMUN-*šú* SAR-*ma ba-rì up-puš*$_4$ *ṭup-pi*, "written, checked and properly executed according to its original." The collocates consisted mostly of scribal terminology, including different types of tablets and inscriptions, as well as other terminology related to the scribal profession and the writing system itself: *šaṭru*, "written"; *gabarû*, "copy"; *ṭuppu*, "tablet"; *giṭṭu*, "oblong tablet"; *lēʾu*, "writing board"; and *ṭupšarru*, "scribe." Closely associated actions related to *barû* include *ṣaṭāru*, "to write"; *sanāqu*, "to check"; *kunnu*, "to deposit (a tablet) permanently"; and *šeʾû*, "to seek"; as well as the verb of examination, *ḫiāṭu*. When it comes to reading inscriptions and tablets, *barû* seems to have been used when the action is executed very carefully and with a high degree of intellectual involvement. This type of reading is distinct from that of *amāru* in royal inscriptions, perhaps because royal inscriptions were meant to be seen in public places and were not meant to be scrutinised and read, or thoroughly examined and checked, as with other types of text.

Ḫiāṭu shared characteristics of the other two verbs of examination, *barû* and *ṣubbû*, in our data set and can be considered to lie somewhere in the middle of their semantic range. It often referred to distant qualitative examination, but it could also be used for close quantitative examination in the sense of weighing something, that is, visually confirming the amount of *siparru*, "bronze"; *ḫurāṣu*, "gold"; or *kaspu*, "silver"; by placing it on a scale. However, in our results, such usage was restricted to Neo-Assyrian letters. In the sense of distant qualitative examination, *ḫiaṭu* was found in a few Neo-Assyrian royal inscriptions, where typical collocates included such words as *temmēnu*, "foundation," and *libittu*, "mudbrick."

Another, albeit semantically rather obscure usage, was found in Neo-Assyrian and later omen texts, where *ḫiāṭu* refer to demons examining or fixing their eyes upon their unfortunate victims. Collocates indicating this included designations related to demons (*ḫayattu*, "terror"; *lilû*, "Lilu-demon"), body parts (*limittu*, "limbs"; *qātu*, "hand"; *šēpu*, "foot"), diseases (*miqit šamê*, "falling-sickness, epilepsy"), and their symptoms (*ruʾtu*, "flowing saliva"; *emēmu*, "to become feverish"; *zūtu*, "sweat"), as well as verbs of catching diseases and succumbing to them (*ṣabātu*, "to seize"; and *mâtu*, "to die") (see also Chapter 23, this volume). In the sense of demons fixing their eyes on someone, *ḫiāṭu* seemed to have something in common with *palāsu*. The verb *palāsu* did not, however, appear on its collocate list. Instead, the most closely associated verbs to *ḫiāṭu* statistically were *barû* and *šeʾû*, "to seek."

When *ḫiāṭu* was used in the same expression as *barû*, the order of the verbs was fixed. First, the object was examined with the verb *ḫiāṭu*, and then a closer examination was performed with *barû*. The reversed order was not attested in our data set.

The collocates of *ṣubbû* were dominated by formulaic expressions in royal inscriptions from the Neo-Assyrian period, where the verb exclusively referred to distant qualitative examination. The collocates included words denoting parts of structures and their condition or description, as well as tearing down and erecting buildings: *maqittu*, "dilapidation"; *lābiru*, "old"; *temmēnu*, "foundation"; *simtu*, "specifications (of a building)"; *bītu*, "temple"; *raṣāpu*, "to erect"; and *nasāḫu*, "to tear down." Any uses of the verb outside this context were not present in our results.

Although Dicks (2012) does not differentiate between qualitative and quantitative examination, her analysis of these three verbs is very much in agreement with our results. She recognises a similar semantic continuum from *barû* to *ḫiāṭu*, and from *ḫiāṭu* to *ṣubbû* as is suggested by our results (2012: 214, 217), and that each of these actions indicates a high degree of intellectual participation, *barû* requiring the most effort. She also notes that *ṣubbû* denotes a visual supervision or a survey performed from a superior vantage point and from a greater distance (2012: 221).

Dicks (2012: 208) presents *barû* as a verb that denotes "a visual search for an object within an object," while *ḫiāṭu* denotes a close examination of the object itself. Although such a fine nuance escaped us during our initial analysis, a more thorough and careful analysis would have revealed it. In our results, *barû* was extensively used to examine writing or signs on tablets, literally an object within an object. In turn, *ḫiāṭu* was used for visually confirming a given amount of valuables and surveying structures, that is, by definition examining the properties of the object itself. Dicks (2012: 211) also mentions the fixed order of *ḫiāṭu* and *barû*, which she explains by the fact that *barû* connotes more detailed examination than *ḫiāṭu*.

Some uses and properties of *barû* were not represented in our results, namely the cases where deities watch over something or humans observe deities (2012: 202–203). We were also unable to find examples of *barû* as a verb of transferring dreams in the Š-stem (2012: 205). Regarding *ḫiāṭu*, our results did not indicate that the vision expressed by this verb may penetrate barriers (2012: 212–213).

Discussion of the results

Table 26.2 provides a summary of the semantic aspects of the Akkadian verbs of seeing. The column AD represents the observations made by Dicks (2012) and the column PMI describes the various meanings of each verb represented in our statistical analysis within the stated parameters.

Generally, PMI was able to give a good overall picture of the semantic nuances of different verbs of seeing. The results were particularly good in the case of *barû*, *ḫiāṭu, and ṣubbû*, where the method clearly highlighted the semantic continuum from close examination to long distance observation. The most difficult verb to analyse was *naṭālu*. Although its use as a verb of dreaming and pointing direction was apparent from the results, its general meaning and exact semantic difference from *amāru* remained tentative. Overall, our results provided less information about the semantics of the seeing verbs than Dicks' philological work, but in one case, PMI was able to highlight the use of *palāsu* as a verb of "peeking in," even though such a meaning was not mentioned by Dicks.

Although we superficially examined the top 50 collocates for each verb, just analysing the top 15–20 collocates would have yielded very similar results regardless of the verb's frequency. This naturally means that, even though PMI highlights relevant collocates that improve our understanding of the meaning of seeing verbs, a large portion of the collocate lists contain uninformative statistical noise (whose quantity usually increases toward the lower ranks in the

Table 26.2 Summary of the semantic aspects of the Akkadian verbs, comparing Dicks (AD) and PMI

amāru	**AD**	**PMI**
dreaming	X	
impossible in darkness	X	
intangible objects	X	X
associated with knowledge	X	X
stimulates emotional response (happiness, sadness)	X	
experiencing	X	X
visiting	X	
finding, discovering	X	
reading	X	X
(N) becoming visible	X	X
palāsu	**AD**	**PMI**
looking at	X	X
looking after	X	
no intangible objects	X	X
favourable looking	X	X
unfavourable looking	X	X
discovery of temple foundations	X	
looking over, scrutinising	X	
(Ntn) looking upwards	X	
emotion: weep		X
emotion: joyfulness	X	
peeking through windows or into cavities		X
naṭālu	**AD**	**PMI**
requires light	X	
dreaming	X	X
associated with knowledge	X	
high degree of intellectual involvement (watching)	X	
denotes vision, no object	X	
pointing towards (direction)	X	X
(N) being visible	X	X
dagālu	**AD**	**PMI**
fixed focus	X	X
waiting (looking forward to)	X	X
intangible object (only one instance mentioned by Dicks)	X	
(Š) being subject to, belonging to (or to entrust to, to make look after)	X	X
(Š) forcing to watch	X	X
(Š) frightening to see	X	X
barû	**AD**	**PMI**
intangible objects	X	
examining	X	X
scanning an object within an object	X	X
surveillance in order to protect	X	
highly engaged visual perception	X	X
(Š) transferring dreams	X	

(*continued*)

Table 26.2 Cont.

ḫiāṭu	**AD**	**PMI**
close to *barû*	X	X
always precedes *barû* if the verbs are found in same expression	X	X
higher intellectual involvement than *amāru* or *naṭālu*	X	X
examining object	X	X
watchfulness with mischief or care	X	X
can penetrate barriers (peering in)	X	
weighing (in our view: quantitative visual examination)	X	X
ṣubbû	**AD**	**PMI**
surveying structures	X	X
long distance of observation / superior vantage point	X	X

rankings). Sometimes this noise also appears quite high in the rankings; for example, *dagālu* has a third ranking collocate *ḫamšu*, "fifth," which comes from a repeating passage in Sennacherib's inscriptions. The words do fit into the same window, but in fact *ḫamšu* begins a new paragraph in the original inscription while *dagālu* occurs at the end of the previous paragraph. Thus, these words have nothing to do with each other, despite PMI seeing them as closely associated.

There are some ways to reduce the amount of noise. At first, the co-occurrences can be weighted by their contextual similarity, meaning that if words co-occur in partially or fully duplicated contexts, their statistical significance is reduced proportionally to the amount of duplication (Sahala and Lindén 2020). This approach, however, was not yet discovered when this study was carried out, but it was successfully applied in Svärd et al. (2021). Secondly, the texts should be pre-processed in a way that the chance of unassociated words co-occurring within a set window would be minimised. Instead of splitting the corpus into texts and using them as the input for PMI, we should split the texts further into paragraphs according to the translation units given in ORACC (as far as they exist). This would ensure that, at least in most cases, words belonging to completely different but still adjacent parts of the texts would never co-occur within the same window.

Another interesting way of improving the results would be to apply morphological constraints to the PMI results by using a morphological analyser, BabyFST, developed by our team (Sahala et al. 2020).[15] For instance, this would allow us to automatically compare uses of different verbal stems (for example G versus Š), as well as to restrict collocates to certain morphological forms with relevant syntactic function in the sentence (such as subject and object). Such features would bring the statistical approach closer to the philological methodology of Dicks. Naturally, this task may not be as simple as it seems on paper, as case markings in the later stages of the Akkadian language are very inconsistent.

Conclusions

As can be seen from the analysis above, overall the results gained with the PMI method matched the conclusions reached by Dicks in her dissertation, despite the genre differences of our data sets. This confirms that PMI is an interesting approach that makes quantitative analysis of texts quite viable. While statistical, quantitative methods can never replace the traditional methods of philological analysis, they can open up new research avenues and make the research more efficient and reproducible. First of all, instead of analysing every occurrence of a word by hand, PMI can help researchers focus on the most relevant contexts of a word's use: first, by providing an overall image of the semantic field before having to dive into the minute details; and second, the semantic domains created with the help of PMI may provide new philological and technical questions for researchers. For example: why exactly is *naṭālu* preferred instead of *amāru* in certain contexts? How can we improve this method in order to tackle the problem with formulaic expressions? Finally, the use of PMI and statistical measures provides information on the usage of words in a way that previous dictionaries have not. Instead of providing an analysis of every possible nuance of a given word of interest, language technological applications provide information on when, where, and in which genres the word was used and in which contexts it typically occurs.

Although PMI is not technologically a very novel method for semantic analysis, it does have some advantages over the state-of-the-art neural methods. First, PMI does not suffer from having a small corpus size in the same way that many neural network-based methods do. Second, PMI is mathematically very transparent and easy to understand. Also, all word associations detected with PMI can be tracked back to the original text and verified by traditional philological methods, whereas more complex neural models are often black boxes for which the underlying mechanisms and reasoning may be difficult to comprehend at times.

Nonetheless, as discussed above, using PMI on small corpora also requires wariness and careful pre-processing of the text material. The results are typically noisy, and it is often crucial to examine the contexts of the collocates in order to avoid false associations. Additionally, the parameter selection has a major impact on the results. Often setting more constraints and restrictions, like narrowing the window size and increasing the frequency threshold, provide cleaner results, but on the downside such results may be too obvious and thus uninteresting. Less restrictive settings on the other hand tend to produce more noise, but also capture a more vivid selection of associated words.

We hope that the results presented in this chapter, as well as the openly available tools developed by our team, will spark interest in Assyriologists to experiment with the presented methodology to complement and aid their own research.

Notes

1. This chapter is based on the work of a larger research group in Helsinki that focuses on researching semantic domains in Akkadian texts with the help of language technological methods. In addition to the authors, members of this group are Krister Lindén, Heidi Jauhiainen, and Tero Alstola. The authors gratefully acknowledge the support from the group. Furthermore, the research of this group has been made possible by the financial support from the Academy of Finland and the University of Helsinki. The chapter has been a joint research project between the authors, but the idea for the chapter was conceived by Svärd whereas the data-driven analysis was conducted by Sahala. In terms of writing, Svärd wrote sections one and two, Sahala sections three and four and most of sections five and six. Preliminary results using an earlier data set were presented at the ASOR Annual Meeting, November 17, 2017 in Boston, MA.
2. The relationship between language and thought has been an important topic in many different fields of study yet there is no firm consensus on the issue to date. For a good introduction to the topic, see Geeraerts and Cuyckens 2010. This chapter was influenced by the work of Levinson (2003: 14–16) and Fleisch (2007: 41–43, 46–47), which inspired in us a desire to examine the Mesopotamian world through nebulous lexical meanings in Akkadian (see also Svärd et al. 2018: 229–230).
3. For an overview, see Svärd et al. 2018: 227, and the other contributions in the volume.
4. Word2vec-compatible word embeddings can be produced from PMI results through matrix factorisation methods such as truncated Singular Value Decomposition (SVD) (Levy and Goldberg 2014). Our team has experimented on factorising PMI-based term-to-term matrices by SVD with promising results (Sahala 2019).

5. Locating results gained from Word2vec or fastText from primary sources is not straightforward, which makes interpretation of the results more challenging.
6. ORACC can be accessed at http://oracc.org. We thank the ORACC steering committee, in particular Niek Veldhuis, who provided access to all of the ORACC data in JSON format. We are indebted to everyone who has been involved in making this research data available, including the authors of the original publications, but also the people who have made the data ORACC-compatible and enriched it through lemmatisations and by adding other metadata. As the total number of people involved would amount to hundreds (the current number of individual subprojects in ORACC is approximately 70), it would be impractical to list them all individually. However, as can be seen from Table 26.1, much of our data is Neo-Assyrian, and much of the Neo-Assyrian linguistic data was created in the State Archives of Assyria (SAA) project. We gratefully acknowledge SAA, created by Simo Parpola and his team and later developed into the State Archives of Assyria online (http://oracc.museum.upenn.edu/saao) by Karen Radner and her team. Finally, we also thank Heidi Jauhiainen, who converted the JSON into Korp compatible format.
7. Our Zenodo repository can be accessed at https://doi.org/10.5281/zenodo.4424188.
8. Note that PMI^2 does not explicitly show if the co-occurrence is independent or not. The score that indicates independent co-occurrence equals $\log_2 p(a,b)$, which means that this value is not fixed and depends on the co-occurrence frequency of the words. We chose not to normalise the scores here, because our normalisation method was not yet published at the time of writing this chapter and it was safer to use a well-established measure instead. For normalised PMI^2, see Sahala and Lindén 2020.
9. We also used a window size of seven words in our previous article (Svärd et al. 2018).
10. The current version (May 2019) of ORACC can be visited at www.kielipankki.fi/corpora/oracc/. Please note that at the time of writing this chapter, Korp contained an earlier version of ORACC. This data set is downloadable from our Zenodo repository (see note 7) to make this study reproducible. Due to a major version change and related fixes to the ORACC in Korp, some search links may yield a different number of hits than is indicated in the results.
11. Korp can be viewed at http://korp.csc.fi.
12. In most instances *amāru* is written logographically IGI or IGI-*ma*. The interpretation as *amāru* comes from ORACC's transcription.
13. Anzu was a monstrous mythological bird, half-lion, half-eagle.
14. This is a matter of translation: "he made the Y belong to X" versus "he entrusted the Y to X."
15. The morphological analyser is already fully functional, but at the time of writing this chapter it was not possible to disambiguate the morphology reliably due to the lack of a gold standard for Akkadian morphology. Such a gold standard was, however, developed while this volume was being edited (see Luukko et al. 2020). The morphological disambiguation is currently being developed.

Bibliography

Alstola, T., S. Zaia, A. Sahala, H. Jauhiainen, S. Svärd, and K. Lindén. 2019. "Aššur and His Friends: A Statistical Analysis of Neo-Assyrian Texts." *JCS* 71: 159–180.

Bojanowski, R., E. Grave, A. Joulin, and T. Milokolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5: 135–146.

Church, K., and P. Hanks. 1989. "Word Association Norms, Mutual Information, and Lexicography," in *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, 76–83.

Daille, B. 1994. "Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques." Ph.D. Dissertation, Université Paris 7.

Dicks, A. A. 2012. "Catching the Eye of the Gods: The Gaze in Mesopotamian Literature." Ph.D. Dissertation, Yale University.

Firth, J. R. 1957. *Studies in Linguistic Analysis*. Oxford: Blackwell.

Fleisch, A. 2007. "How Cognitive Semantics Relate to Comparative Linguistics: A Case Study from Nguni," in T. Machalík and J. Záhořík, eds., *Viva Africa 2007: Proceedings of the 2nd International Conference on African Studies*. Pilsen: University of West Bohemia, 39–53.

Foster, B. R. 2005. *Before the Muses: An Anthology of Akkadian Literature*. Third Edition. Bethesda, MD: CDL Press.

Geeraerts, D., and H. Cuyckens. 2010. "Introducing Cognitive Linguistics," in D. Geeraerts and H. Cuyckens, eds., *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, 3–22.

George, A. R. 2003. *The Babylonian Gilgamesh Epic: Introduction, Critical Edition and Cuneiform Texts*. Oxford: Oxford University Press.

Jauhiainen, H., T. Alstola, and A. Sahala. 2019. *Open Richly Annotated Cuneiform Corpus, Korp Version*, May. Kielipankki. www.kielipankki.fi/corpora/oracc/.

Jurafsky, M., and J. H. Martin. 2019. *Speech and Language Processing*. Third Edition draft of October 2. https://web.stanford.edu/~jurafsky/slp3.

Levinson, S. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language, Culture, and Cognition 5. Cambridge: Cambridge University Press.

Levy, O., and Y. Goldberg. 2014. "Neural Word Embedding as Implicit Matrix Factorization," in Z. Ghahramani, M. Welling, and C. Cortes, eds., *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2*. Cambridge, MA: MIT Press, 2177–2185.

Luukko, M., A. Sahala, S. Hardwick, and K. Lindén. 2020. "Akkadian Treebank for Early Neo-Assyrian Royal Inscriptions," in K. Evang, L. Kallmeyer, R. Ehren, S. Petitjean, E. Seyffarth, and D. Seddah, eds., *Proceedings of the 19th Workshop on Treebanks and Linguistic Theories*. Stroudsburg, PA: The Association for Computational Linguistics, 124–134.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv:1301.3781.

Sahala, A. 2019. "PMI+SVD and Semantic Fields in Akkadian Texts." Poster at HELSLANG Summer Conference in Helsinki, May 27. https://github.com/asahala/pmi-embeddings.

Sahala, A., and K. Lindén. 2020. "Improving Word Association Measures in Repetitive Corpora with Context Similarity Weighting," in A. Fred and J. Filipe, eds., *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management – Volume 1*. SCITEPRESS, 48–58.

Sahala, A., M. Silfverberg, A. Arppe, and K. Lindén. 2020. "BabyFST: Towards a Finite-State Based Computational Model of Ancient Babylonian," in N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, Hélène Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the 12th Language Resources and Evaluation Conference*. Paris: European Language Resources Association, 3886–3894.

Svärd, S., T. Alstola, H. Jauhiainen, A. Sahala, and K. Lindén. 2021. "Fear in Akkadian Texts: New Digital Perspectives on Lexical Semantics," in S.-W. Hsu and J. Llop-Raduà, eds., *The Expression of Emotions in Ancient Egypt and Mesopotamia*. CHANE 116. Leiden: Brill, 470–502.

Svärd, S., H. Jauhiainen, A. Sahala, and K. Lindén. 2018. "Semantic Domains in Akkadian Texts," in V. Juloux, A. Gansell, and A. di Ludovico, eds., *Cyber Research on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*. Digital Biblical Studies 2. Leiden: Brill, 224–256.