Restoration of fragmentary Babylonian texts using recurrent neural networks

Ethan Fetaya[a,1,2], Yonatan Lifshitz[b], Elad Aaron[c], and Shai Gordin[c,1,2]

[a]Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002, Israel; [b]Alpha Program, Davidson Institute of Science Education, Weizmann Institute of Science, Rehovot 7610001, Israel; and [c]Faculty of Social Sciences and Humanities, Digital Humanities Ariel Lab, Ariel University, Ariel 40700, Israel

Edited by Emilie Pagé-Perron, University of California, Los Angeles, CA, and accepted by Editorial Board Member Elsa M. Redmond July 7, 2020 (received for review February 27, 2020)

The main sources of information regarding ancient Mesopotamian history and culture are clay cuneiform tablets. Many of these tablets are damaged, leading to missing information. Currently, the missing text is manually reconstructed by experts. We investigate the possibility of assisting scholars, by modeling the language using recurrent neural networks and automatically completing the breaks in ancient Akkadian texts from Achaemenid period Babylonia.

Babylonian heritage | cuneiform script | Late Babylonian dialect | Achaemenid empire | neural networks

Mesopotamian cuneiform is one of the earliest writing systems known. It was probably invented in southern Mesopotamia at the end of the fourth Millennium BCE and initially used to record daily accounting procedures in the Sumerian cities on a clay medium. A good analogy to this earliest phase is the modern "spreadsheets" (1). It was later used to write several languages, including one of the main languages of the ancient world, Akkadian. Over 2,500 y of human activity across most of the ancient Near East have been recorded in documents written in various dialects of Akkadian, which belongs to the Semitic language family (2, 3). The Akkadian language is attested on a limited scale in southern Mesopotamia in the third millennium BCE under the Empire of Sargon, and it spread rapidly to the north and west during the Amorite expansion of the early second millennium. In the Late Bronze Age, it served as a lingua franca for the entire Near East. During the first millennium BCE, Akkadian was gradually displaced by Aramaic, which used an alphabetic writing system, but it retained its prominence during the rise of the Axial Age empires of Assyria, Babylonia, and Persia. In all, more than 10 million words are attested on some 600,000 inscribed clay tablets and hundreds of monumental inscriptions on stone and other materials that are kept in various collections around the world (4).

Clay tablets, although a rather durable medium, are frequently found in fragmentary condition, and once they have been exposed to the elements, they may become brittle and deteriorate if not properly conserved and stored after excavation in museum collections (5). Damage to the written surfaces of tablets, in the form of cracks and small or large patches of flaking or eroded clay, renders it difficult to fully recover the information originally recorded in the inscribed text. This results in a loss of text ranging from a single sign in a line to entire sections (Fig. 1).

The current practice is to reconstruct the missing information manually. This is a time-consuming process carried out by a handful of experts who have mastered both the Akkadian language and the cuneiform writing system. Fortunately, many texts survive in duplicate copies, and by comparing parallel passages on damaged and intact tablets, it is possible to restore many lines of writing in literary, scientific, lexical, and religious texts. Still, this process requires expert knowledge of each genre and corpus of texts, and the restorations proposed by scholars are often subjective. Furthermore, there is no way to quantify the uncertainty in each restoration. The problem is even harder in the case of damaged texts that belong to a well-known genre-archival documents such as contracts or deeds, for example, but do not exist in duplicate form. In such cases, one must resort to predicting the content of damaged passages on the basis of conventions identified by studying intact examples of the genre.

One possible way to ameliorate these difficulties is to design an automatic process that can aid human experts engaged in the task of restoring damaged texts. In this article, we investigate an approach to automatically completing broken passages in Late Babylonian archival texts that relies on modern machine learning methods, specifically recurrent neural networks (RNNs) (6). Due to the limited number of digitized cuneiform texts available at present, it is uncertain whether such data-driven methods would yield plausible restorations in all types of texts, but we hypothesized that for genres with highly structured syntax—such as legal, economic, and administrative Late Babylonian texts—these models should work well, as we will demonstrate here. Furthermore, we are developing an online tool, called Atrahasis (https://babylonian.herokuapp.com/) to make our work available to a wide scholarly community. Our source code is available at GitHub (https://github.com/DHALab/Atrahasis).

Choosing Digital Akkadian Corpora. One challenge in designing an automated text-completion tool is the limited data available

Significance

The documentary sources for the political, economic, and social history of ancient Mesopotamia constitute hundreds of thousands of clay tablets inscribed in the cuneiform script. Most tablets are damaged, leaving gaps in the texts written on them, and the missing portions must be restored by experts. This paper uses available digitized texts for training advanced machine-learning algorithms to restore daily economic and administrative documents from the Persian empire (sixth to fourth centuries BCE). As the amount of digitized texts grows, the model can be trained to restore damaged texts belonging to other genres, such as scientific or literary texts. Therefore, this is a first step for a large-scale reconstruction of a lost ancient heritage.

Author contributions: E.F. and Sh.G. designed research; E.F., Y.L., E.A., and Sh.G. performed research; E.F. and Sh.G. contributed new reagents/analytic tools; E.F., Y.L., E.A., and Sh.G. analyzed data; and E.F. and Sh.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. E.P. is a guest editor invited by the Editorial Board.

[1]E.F. and Sh.G. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ethan.fetaya@biu.ac.il or shaigo@ariel.ac.il.

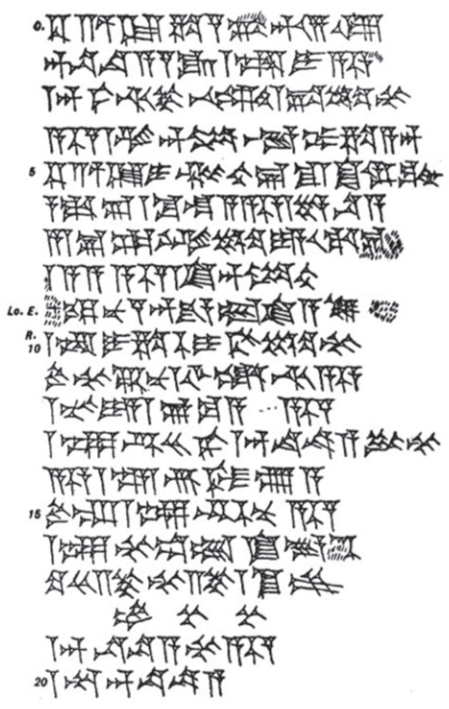
This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2003794117/-/DCSupplemental.

First published September 1, 2020.

Fig. 1. Fragmentary obverse of the house sale contract YBC 7424 (Yale Oriental Series 17 3) from the third year of Babylonian king Nebuchadnezzar II, famous for burning the city of Jerusalem and exiling the Judean elite as described in the Book of Kings. Courtesy of the Yale Babylonian Collection. Image credit: Klaus Wagensonner (Yale University, New Haven, CT).

in digital form. For similar unsupervised language modeling tasks in English, for example, one can collect practically endless amounts of texts online, and the main limitation is the computational challenge of storing and processing large quantities of data (7). For cuneiform texts this is not the case. Automatic optical character recognition cannot be used to reliably identify cuneiform signs, neither in their two-dimensional (2D) representations (hand copies) nor in three-dimensional (3D) scans of actual clay tablets (8–12). Various visual recognition algorithms are being applied to cuneiform, but the results are yet in their infancy (13–16). Therefore, one has to rely on a limited corpus of manually transliterated texts. Although Akkadian cuneiform texts span more than two millennia and the genres available for study are heterogeneous, for many periods, only a limited amount of digital text is available to train the learning algorithm.

Three temporally and geographically defined corpora in particular are well represented in the digital transliterations available today: the Old Babylonian (approximately 1900 to 1600 BCE; see ARCHIBAB website: http://www.archibab.fr/), Neo-Assyrian (approximately 1000 to 600 BCE; see State Archives of Assyria online website: http://oracc.org/saao/), and Neo- and Late Babylonian (approximately 650 BCE to 100 CE). Our choice of texts, however, was governed by a corpus-based approach so that we might exercise greater control over the diversity of text genres and phrasing. Therefore, we have decided to gather 1,400 Late Babylonian transliterated texts from Achaemenid period Babylonia (539 to 331 BCE) in Hypertext



Obv. 1	1/3 2 1/2 gín kù-babbar níg-ga ^{d}15 *ù*	(1)22.5 Shekels of silver, from the property of Ištar
	^{d}*na-na-a šá* šu^{II} ^{Id}en-i a-*šú šá*	(2)and Nanaya, issued by Bēl-nā'id, son of
	^{Id}dù-*ti*-kám *ina muh-hi* ^{I}*ta-li-mu*	(3)Bānītî-ēreš, are owed by Talīmu,
	a-*šú šá* ^{I}gi-^{d}amar-utu *ina* iti gan kù-babbar *a*	(4-5)son of Mušallim-Marduk.
		He will give the aforementioned silver,
5	1/3 2 1/2 gín *i-nam-din* é-*su ki-šub-ba-a'*	22.5 Shekels, in the Month of Kislimu (9th month).
	šá da é ^{I}*šu-la-a* a-*šú šá* ^{I}sì-*na-a*	The unbuilt plot for his house,
	a ^{I}é-sag-gil-*li-a u* da é	(6-8)which neighbours the house of Šulāya, son of Iddināya
	^{I}*a-a* a-*šú šá* ^{I}ki-^{d}amar-utu-din	of the Esaggilāya Family and the house
Lo.E.	[*maš*]-*ka-nu šá* ^{d}gašan *šá* unug^{ki} *a-di-*[*i*]	of Aplāya, son of Itti-Marduk-balāṭu,
Rev. 10	^{Id}en-i kù-babbar-*šú i-šal-li-mu*	(9-10)serves as pledge of the Lady of Uruk, until
	lú *mu-kin-nu* ^{I}*ši-rik-ti* a-*šú šá*	he will pay Bēl-nā'id his silver.
	^{I}numun-*ia* ^{I}*kal-ba-a* a-*šú šá*	(11-14)Witnesses: Širikti, son of
	^{Id}nà-*re-man-ni* ^{Id}*na-na-a*-šeš-mu	Zērīya. Kalbāya, son of
	a-*šú šá* ^{Id}in-nin-*re-ṣu-ú-a*	Nabû-rēmanni. Nanāya-aḫu-iddin,
15	lú umbisag ^{Id}nà-en-*šú-nu* a-*šú šá*	son of Innin-reṣûa.
	^{Id}nà-mu-gin unug^{ki} iti kin	(15-18)Scribe: Nabû-bēlšunu, son of
	u$_4$-22 kam mu 2-kam ^{I}*ku-raš*	Nabû-šumu-ukīn. Uruk, the month of Ulūlu (6th month),
	lugal kur-kur	22nd day, 2nd (regnal) year of Cyrus,
	^{Id}*na-na-a*-mu a-*šú šá*	king of all the lands.
20	^{I}ir-^{d}*na-na-a*	(19-20)(additional witness:) Nanāya-iddin, son of
		Arad-Nanāya.

Fig. 2. From *Left* to *Right*, the original cuneiform line art, the transliteration, and the translation of the Achaemenid period Babylonian text Yale Oriental Series 7 11.

Markup Language (HTML) format from the Achemenet website (http://www.achemenet.com/).*

The Neo-Babylonian Corpus. The corpus chosen is written in what is commonly termed the Late Babylonian dialect of Akkadian, attested from the rise of the Neo-Babylonian empire in 627 BCE until the end of the use of the cuneiform script, around the first century CE (17). We prefer to describe this corpus as Neo-Babylonian, in recognition of the continuity of the Akkadian dialect of Babylonia over the whole first millennium BCE.† The largest number of texts known from this period are archival documents belonging to economic, juridical, and administrative genres (19–21). The main reason we expect our models to work well on these texts, despite the small amount of data, is that these tablets are official bureaucratic documents, e.g., legal proceedings, receipts, promissory notes, contracts, and so on. They are highly structured, usually short, and prefer parataxis over hypotaxis (see Fig. 2 for an example). These texts are tedious for humans to read and complete, but they display many patterns that are relatively easy for learning algorithms to model, which makes them ideal for our purpose. For a more detailed breakdown into text types and their structure and content, see *Materials and Methods* and the Dataset S6.

Algorithmic Background

In this section, we will give a very brief introduction to techniques for modeling language using RNNs (for a more detailed account, see ref. 6). We can view language as a series of discrete tokens $x_1, \ldots, x_T$, and our goal is to fit a probabilistic model for such sequences; i.e., we wish to find a parametric model that learns the distribution $p(x_1, \ldots, x_T)$ from samples. The first step is to use an autoregressive model, i.e., use the factorization $p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$. What this means is that we can reduce the problem of modeling a full sentence to predicting the next token in a text on the basis of the tokens that precede it.

In RNNs, this autoregressive model $p(x_t | x_1, \ldots, x_{t-1})$ is fitted using a hidden memory. Given the previous hidden memory h_{t-2}, the network first updates the memory based on the new input x_{t-1} and then uses the updated memory to predict the next token and passes the updated memory to the next step. More formally:

$$h_{t-1} = tanh(W_{hh} h_{t-2} + W_{ih} x_{t-1} + b_h), \quad [1]$$

$$prob_t = softmax(W_{ho} h_{t-1} + b_o), \quad [2]$$

where x_{t-1} is a one-hot representation of the input token, W indicates linear mappings, and $prob_t$ is the vector of probabilities for each possible next token. This parametric model is trained by maximizing the training log-likelihood to produce the output model. While simple and effective, due to vanishing gradients simple RNNs have difficulties in modeling long time dependencies, i.e., situations in which the probability of the next token depends on information seen many steps before. To solve this issue, various modifications that introduce a gating mechanism, such as long short-term memory (LSTM), have been proposed (22).

As a baseline model for comparison we trained an n-gram model. The n-gram model is a model that assigns a probability

*Initiated by Pierre Briant of the Collège de France in 2000, this website is entirely dedicated to the history, material culture, texts, and art of the Achaemenid Empire. Since we began our study, the Babylonian text section has grown to include 2,709 texts (accessed 27 May 2020); it is administered by the Histoire et Archéologie de l'Orient Cunéiforme team of Francis Joannès (Unité Mixte de Recherche ArScAn 7041, CNRS, Nanterre, France).

†As shown already by Streck and recently by Hackl, an actual sharp distinction between Neo- and Late Babylonian dialects does not linguistically exist (18) (see also *Materials and Methods*).

Table 1. Loss and perplexity while training the model on Achemenet dataset

	Training loss	Training perplexity	Test loss	Test perplexity
2-Gram	3.26	9.55	3.54	11.60
LSTM	1.41	4.10	1.62	5.05

to each token based on how frequently the sequence of the last $n-1$ tokens in the training set ended in that token. The main limitation of n-gram models is that for small n, the context used for prediction is very small, while for large n, most test sequences of that length are never seen in the training set. We used a 2-gram model, i.e., each word is predicted according to the frequency with which it appeared after the previous one.

Results

In order to generate our datasets, we collected transliterated texts from the Achemenet website, based on data prepared by F. Joannès and coworkers in the framework of the Achemenet Program (National Center for Scientific Research [CNRS], Nanterre, France) (http://www.achemenet.com/fr/tree/?/sources-textuelles/textes-par-langues-et-ecritures/babylonien). We designed a tokenization method for Akkadian transliterations, as detailed in *Materials and Methods*. We trained a LSTM recurrent network and a n-gram baseline model on this dataset (see Datasets S1–S3 for model and training details).

Results for both models are in Table 1. Loss refers to mean negative log-likelihood and perplexity is two to the power of the entropy (in both cases, lower is better).

As expected, the RNN greatly outperforms the n-gram baseline, and despite the limitations of the dataset, it does not suffer from severe over-fitting.

Completing Random Missing Tokens. In order to evaluate our models' ability to complete missing tokens, we took random sentences from the test corpus, removed the middle token and tried to predict it using the rest of the sentence. Our model returns a ranking of probable tokens and we report the mean reciprocal rank (MRR). The MRR is the average over the dataset of the reciprocal of the predicted rank of the correct token. It is a very common and useful measure for information retrieval as it is highly biased toward the top ranks, which is what the user is mostly interested in. We also evaluate the "hit@k," which measures the percentage of sentences where the correct completion is in the top k suggestions. For evaluation, we used all test sentences 10 or more tokens in length that contain no breaks, which yielded a total of 520 sentences.

We compared two variations of our model, one that finds the optimal completion based only on the tokens that precede the missing token, denoted "LSTM (start)," and one that takes the full sentence into account, denoted "LSTM (full)." As the "LSTM (full)" model needs to run separately for each candidate for the missing token, we first picked the top 100 candidates using "LSTM (start)." We then generated 100 sentences, one for each possible completion, and reranked them based on the full sentence log-likelihood. If the right completion was not in the top 100, we took the reciprocal rank to be zero.

For comparison, we used two simple 2-gram baselines: one that takes into account only the previous token, denoted "2-Gram (start)," and one that takes into account both the previous and the next token denoted "2-Gram (full)." While this is a relatively weak model, we found it to work surprisingly well, although it was still significantly inferior to the LSTM model in the accuracy (Hit@1) metric.

To further investigate our model's ability to complete various numbers of missing tokens in various locations, we removed up to three tokens in random locations. We ranked possible completions using our model and beam search and show the results in Table 3.

It is clear from the results in Tables 2 and 3 that our algorithm can be of great help in completing a missing token, with an almost 85% chance of completing the token correctly and a 94% chance of including the correct token in the top 10 suggestions. However, as expected, the task becomes much harder and performance is degraded when more tokens are missing. We note that even with two or three missing tokens, however, the model is still useful as the correct completion is present in the top 1 (two missing) or 10 (three missing) completions almost half of the time.

Designed Completion Test. We designed another experiment in order to evaluate our completion algorithm and understand its strengths and weaknesses. We generated a set of 52 multiple choice questions in which the model is presented with a sentence missing one word and four possible completions, and the goal was to select the correct one. Of the three wrong answers, the first was designed to be wrong semantically, the second wrong syntactically, and the third both. This allowed us to track the types of mistakes the algorithm makes. The assumption is that the learning algorithm would be more likely than a human to make semantic mistakes but should be better than a nonexpert in grammar. If this is the case, then the effectiveness of our approach as a way to assist humans should rise, as the strengths of human and machine complement each other.

When we used our model to rank four possible restorations for each of the missing words in the 52 random sentences, it achieved 88.5% accuracy in selecting the one with the highest likelihood (see Dataset S4 for the complete list of questions and answers). Looking at the six failed completions—questions 18, 26, 32, 35, 45, and 50—we see that four are semantically incorrect, one is syntactically incorrect, and one is both, which agrees with our hypothesis.

Discussion

Further study of the different restorations of the designed completion test, taking into account the full ranking of the answers, results in some interesting patterns. This qualitative analysis considers four categories for the answer ranking: 1) correct syntax (i.e., sentence structure), 2) correct semantic identification, 3) poor syntax, and 4) poor semantic identification (see Dataset S5 for the full data analysis).

The majority of the restorations, 44 cases, shows that the algorithm best identifies correct sentence structure (category 1: questions 1 to 23, 25 to 27, 29 to 35, 38 to 42, 46, and 48 to 52). Put more accurately, this means correct syntactic sequences of parts of speech based on the statistical frequency of smaller syntagmatic structures. A total of 34 restorations show correct semantic identification of noun class as well as related verbs in the answer ranking (category 2: questions 1, 3, 6 to 13, 16 to 18, 22, 24 to 27, 29, 31 to 33, 35 to 40, 42, 46 to 49, and 51). In fact, a large subset of these cases, 30 questions, shares also identification of correct sentence structure (i.e., both categories 1 and 2).

Good semantic identification probably derives from paradigmatic relationships between certain classes of words. For example, five cases possibly correctly identified usage of verbal forms based on their context (e.g., in direct speech; questions 3, 6,

Table 2. Completing missing fifth token in sentences

5th index	MRR	Hit@1	Hit@5	Hit@10
2-Gram (start)	0.64	52.0%	78.2%	83.6%
LSTM (start)	0.754	66.1%	86.9%	91.9%
2-Gram (full)	0.80	74.8%	85.5%	90.6%
LSTM (full)	0.89	85.4%	93.2%	94.6%

Table 3. Completing various number of tokens

Missing tokens	MRR	Hit@1	Hit@5	Hit@10
One	0.86	81.5%	92.5%	94.4%
Two	0.55	47.8%	62.2%	70.0%
Three	0.30	24.3%	37.2%	42.6%

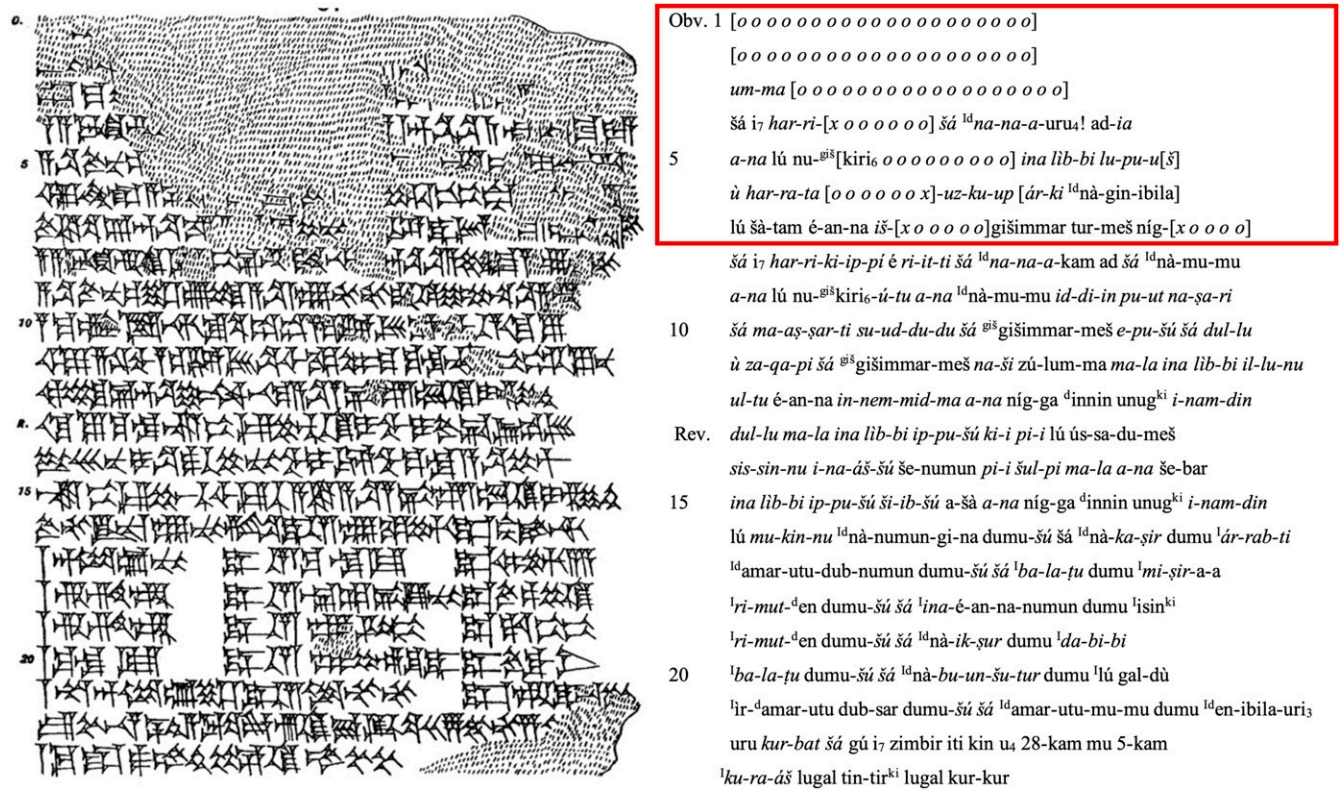
18, 25, and 26). Take, for example, question 3: NAME a*šú šá* NAME *ana* NAME lú *qíipi* ébabbarra *u* NAME lú sanga LOCATION___ *umma*. The model ranked the four possible answers as follows: *iqbi*; *liqbuú*; bar; bán. The example not only shows a correct identification of sentence structure but also a recognition of the relationship between two different forms of the verb *qabû*, "to speak." It does not necessarily reflect an understanding of verbal root form, however; the model's success probably reflects the statistical frequency of *iqbi* in this context and the model's recognition of its similarity to *liqbuú*. This statistical inference emerges more clearly in one of the mistakes made by the model in question 32, where it does not differentiate properly the grammatical person of the verb *nadānu*, "to give, pay" (*taaddinu* vs. *inamdin*).

The level of the model's semantic knowledge becomes apparent with regard to noun class; 14 questions show possible correct identifications of countable nouns (questions 1, 9, 17, 29, and 33), names of professions (questions 11, 38, 39, and 46), temporal designations (questions 12, 36, and 41), gender (question 31), and even a contextual formulaic legal clause (the so-called *elat*-clause; question 51). Four cases show correct identification of prepositions, particle use, or pronouns (questions 7, 34, 42, and 51). The choice in question 7, between the related prepositions *ina* and *ana*, makes it clear that these choices are again based on frequency in specific contexts. Moreover, a purely statistical grasp of parts of speech seems to be a decisive factor in at least eight cases of restoration, which are best identified in the analysis as those fitting category 3: poor syntax (questions 24, 28, 36 to 37, 43 to 45, and 47). Such statistical inference achieves surprisingly good results—e.g., preferring kurkur over LOCATION after lugal (question 37)—with only one restoration ended up being erroneous (question 45). However, it is clear, by comparing this group to other restorations in the designed completion test, that statistical inference is not a consistently reliable method for choosing completions. For example, it can interfere with contextual identification of the correct restoration—by preferring *ina* igi over *ina* šu^{II} before NAME (question 35).

The model does not seem to identify alternate logographic and phonetic writings of the same words, e.g., Sum. da = Akk. *itti*, or Sum. im.dub = Akk. *tuppi* (questions 14 and 19). It obviously lacks enough examples of such interchangeability in the studied corpus, since it does identify when different logograms have similar usage (e.g., a and dumu, both meaning "son" or "descendant," are the top answers in question 40). Further confusion can occur when the model detects a similarity between the answer and another word close by in the sentence, either a noun or a verb. Especially problematic are cases when there are very few similar sentences to train on, and so the algorithm makes an "educated" guess resulting in a mistake (for example, question 45).

Conclusion

Our model—as far as can be judged by this experiment—is, as expected, good in teasing out sentence structures. However, it was also surprisingly better than we expected in making semantic identifications on the basis of context-based statistical inference (rather than finding underlying grammatical rules and morphology). In order to greatly reduce the number of false identifications based on the statistical frequency of contextual semantic



Obv. 1 [*o o o o o o o o o o o o o o o o o o o o*]
[*o o o o o o o o o o o o o o o o o o o o*]
um-ma [*o o o o o o o o o o o o o o o o o o o o*]
šá i$_7$ *har-ri-*[*x o o o o o o*] *šá* ^{Id}*na-na-a-*uru$_4$! ad-*ia*
5 *a-na* lú nu-^{giš}[kiri$_6$ *o o o o o o o o o*] *ina lìb-bi lu-pu-u*[*š*]
ù har-ra-ta [*o o o o o o x*]*-uz-ku-up* [*ár-ki* ^{Id}nà-gin-ibila]
lú šà-tam é-an-na *iš-*[*x o o o o o*]gišimmar tur-meš níg-[*x o o o o*]
šá i$_7$ *har-ri-ki-ip-pí* é *ri-it-ti šá* ^{Id}*na-na-a-*kam ad *šá* ^{Id}nà-mu-mu
a-na lú nu-^{giš}kiri$_6$*-ú-tu a-na* ^{Id}nà-mu-mu *id-di-in pu-ut na-ṣa-ri*
10 *šá ma-aṣ-ṣar-ti su-ud-du-du šá* ^{giš}gišimmar-meš *e-pu-šú šá dul-lu*
ù za-qa-pi šá ^{giš}gišimmar-meš *na-ši* zú-lum-ma *ma-la ina lìb-bi il-lu-nu*
ul-tu é-an-na *in-nem-mid-ma a-na* níg-ga ^{d}innin unug^{ki} *i-nam-din*
Rev. *dul-lu ma-la ina lìb-bi ip-pu-šú ki-i pi-i* lú ús-sa-du-meš
sis-sin-nu i-na-áš-šú še-numun *pi-i šul-pi ma-la a-na* še-bar
15 *ina lìb-bi ip-pu-šú ši-ib-šú* a-šà *a-na* níg-ga ^{d}innin unug^{ki} *i-nam-din*
lú *mu-kin-nu* ^{Id}nà-numun-gi-na dumu-*šú* šá ^{Id}nà-*ka-ṣir* dumu ^{I}*ár-rab-ti*
^{Id}amar-utu-dub-numun dumu-*šú šá* ^{I}*ba-la-ṭu* dumu ^{I}*mi-ṣir*-a-a
^{I}*ri-mut-*^{d}en dumu-*šú šá* ^{I}*ina*-é-an-na-numun dumu ^{I}isin^{ki}
^{I}*ri-mut-*^{d}en dumu-*šú šá* ^{Id}nà-*ik-ṣur* dumu ^{I}*da-bi-bi*
20 ^{I}*ba-la-ṭu* dumu-*šú šá* ^{Id}nà-*bu-un-šu-tur* dumu ^{I}lú gal-dù
^{I}ir-^{d}amar-utu dub-sar dumu-*šú šá* ^{Id}amar-utu-mu-mu dumu ^{Id}en-ibila-uri$_3$
uru *kur-bat šá* gú i$_7$ zimbir iti kin u$_4$ 28-kam mu 5-kam
^{I}*ku-ra-áš* lugal tin-tir^{ki} lugal kur-kur

Fig. 3. Line art and transliteration of Achaemenid period Babylonian text Yale Oriental Series 7 51 from the Eanna archive in Uruk. Fragmentary upper half of obverse marked by a red square.

relationships, much more training material will be needed. Nevertheless, we have demonstrated that even without access to large amounts of data, we can successfully train LSTM models and use them to complete missing words. In our completion test, we show good results that, while not sufficient for fully automatic completion, prove that the model can be an invaluable tool in helping scholars with text restoration.

Our results with the Late Babylonian corpus are significant because most entry-level scholars or other interested historians and social scientists who focus on the large first-millennium BCE Babylonian archives cannot acquire the very specific knowledge and expertise to understand underlying political, social, or historical structures without reading through hundreds of texts. For this reason, we are in the process of incorporating our model into an online tool, called Atrahasis. It will be of immense help to scholars in the historical sciences, allowing them to overcome the high entry barrier to restoring fragmentary Akkadian texts. Initially, the model will achieve the most success with structured archival documents, but as the dataset grows, one can train the model on more genres, such as scientific or literary texts. Both access to the primary sources in their original state and the ability to restore broken passages are equally necessary for understanding Akkadian corpora on a macroscale.

Related Work

The task of text modeling and its applications in tasks such as text restoration, lies in the intersection of Natural Language Processing (NLP) and Computer Linguistics. The Computational Linguistics models are predominantly rule based, while current NLP models are predominantly statistical and use machine learning. Currently, machine-learning methods achieve state of the art performance on most NLP challenges. In most modern languages, basic text modeling tools can identify spelling errors such as missing, added, transposed, or wrong letters (23–26). These tasks can become more challenging when the language in question has a richer morphology, like Arabic, for example (27), or limited digital corpora (28), as in the case of ancient Near Eastern languages (29, 30).

One approach is to first parse the original text and then use this as an input for further tasks. To this end, many studies use rule-based models derived from grammar (31), finite-state machines (32), or lookup in a machine readable dictionary (33) or employ statistical models such as clustering algorithms (k-nearest neighbors or kMeans) (34). Most work on rule-based models has been done in Akkadian (see literature cited in ref. 35). The most recent study dedicated to Akkadian word segmentation used a combination of rule-based, dictionary-based, and statistical algorithms, with best results in dictionary-based models (60 to 80%) (35). Because its algorithms were fitted for East Asian languages like Chinese and Japanese, we concluded that a specific NLP model for Akkadian should be designed. Sumerian, with its simpler syntax, is in the center of the Machine Translation and Automated Analysis of Cuneiform Languages project, which employed dictionary- and rule-based models for annotation of Sumerian (36, 37). A similar study on Hittite designed rule-based models derived from grammar (38). Statistical/machine-learning models achieve better results overall, if they are tailored to the problem at hand. Our project, the Babylonian Engine, recently achieved state-of-the-art results for Akkadian prediction of sign and word transliteration and segmentation using NLP and machine-learning models on Unicode cuneiform: up to 97% using a BiLSTM Neural Network algorithm, see the web-tool Akkademia (https://babylonian.herokuapp.com/). A similar task of automatic phonological transcription of Akkadian (usually termed normalization) based on ORACC material has recently achieved promising results (39).

For the specific task of text restoration and prediction, statistical n-gram Language Models (LMs, e.g., bigrams, trigrams, etc.) are now widely used in NLP tools, including those designed for modern Indian languages (40). A bigram model was successfully applied to a hidden Markov model to restore missing or damaged sign sequences in the ancient undeciphered Indus script (41, 42). However, neural network LMs can perform better in developing meaningful patterns of representations of words and the contexts around them. When these "embeddings" are learned from unsupervised large corpora, they can be transferred to various tasks, retaining a boost in performance (43).

Specifically, RNN language models, like the one employed in this study, have shown success in encoding both semantic and orthographic data in languages of varying levels of morphological complexity (44). The best results in solving the restoration task, so far, have been achieved in studies of machine-reading comprehension, specifically of the cloze-style, in which both the level of character and word are identified (45). In the field of ancient languages, we know of only one other study that used an algorithm with neural network architecture to recover missing letters, in the context of epigraphic inscriptions in ancient Greek (46). Their model, named PYTHIA, uses a sequence-to-sequence NN with LSTM and was trained on the Packard Humanities Institute (PHI) database, the largest digital dataset of ancient Greek inscriptions. It gives best predictions with 30.1% character error rate, compared with the 57.3% error rate of human epigraphists (based on testing the performance of two doctoral students on the training material over 2 h).

Lastly, joining fragments of texts is one of the major challenges of restoring cuneiform manuscripts as close as possible to their original state. Matching fragments are usually only identified by a handful of experts, and the fragments are often so small as to retain only a few signs. An initial study on the Hittite corpus employed matching classifiers, achieving best results with Maximum Entropy Classifiers (47). The Electronic Babylonian Literature project aims to reconstruct the tens of thousands of fragments that make up the remnants of ancient Babylonian and Assyrian literature. A digital corpus of largely inaccessible tablet fragments from museum collections (15,510 fragments as of June 2020) allows users to query these fragments with sequence-alignment algorithms based on the word method Basic Local Alignment Search Tool algorithm. Initial results already show that one can identify new pieces of text as well as many possible text joins. Advances in cost-effective high quality 3D scanning allow exact measurements of inscribed objects that can lead to the joining of broken tablet fragments with a matching algorithm in 3D space, as done for example by the Virtual Cuneiform Tablet Reconstruction Project (48, 49).

Materials and Methods

Neo-Babylonian Archives. Babylonian archives from the end of the sixth to the fourth century BCE are one of the main sources for reconstructing the official and ephemeral heritage of the Achaemenid Empire and its subject peoples in Mesopotamia. These "archives" were not recovered in situ but are artificial constructs imposed by modern scholars. Most Neo-Babylonian texts come from uncontrolled or poorly documented excavations, and the majority are kept in large museum collections (see below). One cannot rely on physical proximity between texts in a given find context to define an archive, since such a context is frequently unavailable or was disturbed in antiquity.

The organization of Neo-Babylonian archives by modern scholars is based mostly on an artificial division between private and institutional ownership (21). Further criteria employed to define an archive include prosopography (i.e., grouping tablets that feature a common core of principal actors engaging in connected activities), document type and content or a common setting in a social or political institution, such as a business firm, temple, or palace. Several studies try to mitigate the lack of archaeological context by employing museum-based archaeology to trace the acquisition history of related texts within a single collection or across different museums (50). The end result, nevertheless, is that for the most part, groupings of tablets according to any of the aforementioned criteria are artificial constructs, with few exceptions.

Table 4. Breakdown of Achemenet dataset used to train our algorithm into archival and administrative text types; top 17 categories (see Dataset S6 for full list)

Text type	Quantity	Akkadian keyword(s)	Reference
Inventory/list/record of transfer	464		19
Business partnership	266	*harrānu*	62
Receipt	236	*eṭir* / *mahir*	19
Purchase	75	*mahīru inbē* (immovable)	20
Promissory note for assessed field rent	57	*imittu*	
Statement in court/deposition	45	e.g., *dabābu*	63
Summon/oath/injunction	40		64
Lease of arable land/orchard	35	*ana nukuribbūti*, *errēšūti*, and *sūti*	
Lease of movable property	33	*ana idī*, *zitti*	
Letter order	19		65
Balanced accounts	18	*nikkassu epēšu*	19
Lease of immovable property	16	*ana idī*, *maddatti*	66
Promissory note	19	*ina muhhi*	67
Fragmentary: legal contract	14		
Correspondence	7		68
Marriage agreements and dowry texts	7		69
Work contract	7	*dullu*	

Fortunately, the three largest text groups recognized as private archives in Achaemenid Babylonia can each be traced back to a building or room where they were deposited in antiquity: the business archives of the Egibi and Nūr-Sîn families from Babylon and of the Murašu "firm" from Nippur, as well as the closely contemporary archive of the Persian governor Bēlšunu from the palace complex of Babylon, known as the Kasr archive (designated Kasr N6; ref. 50).[‡] The Murašu texts especially, along with another cluster of texts written in several rural centers known as the Yahudu "archive," provide significant information on foreign minority communities in the Achaemenid Empire during a period of nearly 200 y and illuminate the fate of the Judean community in Babylonian exile (53). However, the largest textual groups from this period by far are the two multifile archives associated with city temples: the Eanna archive from Uruk and the Ebabbar archive from Sippar. These two institutional archives make up the bulk of the Achemenet dataset, alongside the private Egibi/Nūr-Sîn archive and the Murašu material.

Part of the Egibi/Nūr-Sîn archive in the Achemenet website (80 texts) belongs to the period of the Neo-Babylonian Empire, which technically lies outside of our chosen chronological framework. Most of these date to the reign of the last Babylonian king, Nabonidus (556 to 539 BCE). (In fact, many of the archives mentioned here are attested both in the Neo-Babylonian and [early] Achaemenid period, especially the larger institutional and private archives.) We nevertheless included some of the Neo-Babylonian period texts in our dataset because they are similar in type and content to the early Achaemenid period texts; in any case, they form a negligible fraction of our total dataset. Altogether, the Achaemenid period Babylonian texts on Achemenet are representative of archival groups from almost every large city in Babylonia[§]: Babylon (Ea-eppēš-ilī, Gahal, Nappāhu), Kiš (Eppēš-ilī), Sippar (Bēl-rēmanni, Ea-eppēš-ilī A, Iššar-tarībi, Marduk-rēmanni, Rē'i-sisê), and Uruk (Atû).

The need for text restorations varies from archive to archive, depending either on their method of excavation and preservation in recent times or on the archival selection processes practiced in antiquity (e.g., some archives were regarded as discarded or "dead" archives). The best-preserved tablets found their way into museum collections in Europe and the United States following their discovery in the initial period of exploration during the late 19th and early 20th century. Many came from illicit or clandestine excavations and were acquired through a process of active selection, as curators preferred complete or nearly complete tablets over broken ones. In contrast, tablets from official excavations in Babylon and Uruk, for example, contain a higher percentage of fragmentary texts. Some large archives like Murašu or Kasr (which was already vitrified from an ancient fire) were damaged by poor handling following excavation or suffered from the effects of war.[¶] A large number of Eanna tablets produced before the reign of Darius I were deliberately discarded or smashed already in antiquity after they were no longer needed by the temple administration (56, 57). The obverse of the fragmentary upper half of one of the Eanna tablets, dating to the reign of Cyrus, can be seen in Fig. 3, along with a proposed restoration that is based on known parallels and scholarly study (*SI Appendix*, Fig. S1).

Neo-Babylonian Text Types and Content. The Neo-Babylonian archival texts are divided into text types based on their form and content. Each Neo-Babylonian archival document, such as a promissory note or a contract, has at least three main parts: 1) an operative section made up of one or more formal clauses, usually beginning with a statement on the object(s) in question by the relevant protagonists; 2) a list of witnesses (58) (seldom accompanied by their seals on the tablet; refs. 58 and 60); and 3) a scribal signature. The latter includes not only the name and lineage of the scribe but also the place of issue and precise date given in month, day, and regnal year of the reigning king. Administrative texts, on the other hand, appear mostly in list form detailing involved objects and parties using abbreviated formulae and specific keywords. They are usually dated but do not have a scribal signatures and practically no witness lists (19, 61).

Table 4 shows the numerical breakdown of the Neo-Babylonian texts used to train our algorithm according to their respective archival and administrative text types, based on summaries of their content recorded in the Achemenet database. The division into subcategories of economic, juridical, and administrative genres is not meant to be granular, but rather inclusive, in order to reflect the different thematic elements of the corpus. Overall, there is a higher percentage of different legal archival documents, most of which contain highly structured formulae. On the other hand, the relatively high number of inventory lists, transfer documents, and other administrative material is considerably less standardized in form and content. It remains to be seen if this effected the results of our training. This is not the place to elaborate on individual text typologies, which are usually based on analysis of the main operative section of each document and take into consideration specific legal clauses or keywords, the issuing person or institution, and the prosopographical study of parties, witnesses, and scribes (see ref. 20 and the extensive references therein). Nevertheless, in order to exemplify the structure and consistency of Neo-Babylonian archival and legal formulae, most text types described in the table are also accompanied by relevant Akkadian keywords and primary reference materials.

[‡]Kasr has, in fact, a mixed private and institutional background. See ref. 51 for an overview of cuneiform archives from Achaemenid period Babylonia and their time span. A more detailed discussion of each text group is found in ref. 20.

[§]Designations of archives are listed in parentheses following each city name. Despite being mentioned in the description on the Achemenet website, the Ur archives are not yet represented in that collection. Archives already mentioned above, like Murašu from Nippur, are not included in this list.

[¶]The Murašu texts were damaged during their transport out of Nippur (54), and the Kasr texts partially survived a grim sequence of events triggered by the First World War. Many of them had already suffered ancient fire damage during or after the Achaemenid period (55).

Data Scraping and Transcriptions of the Neo-Babylonian Dialect. We scraped 1,400 Late Babylonian transliterated texts from Achaemenid period Babylonia (539 to 331 BCE) in HTML format from the Achemenet website (http://www.achemenet.com/fr/tree/?/sources-textuelles/textes-par-langues-et-ecritures/babylonien). As the Achemenet website does not have an Application Programming Interface, we received written permission from the project head F. Joannès to scrape the data. We built a scraping script in Python 2.7 to scrape the texts, preprocess and tokenize them. The script uses the "Beautiful Soup" library to remove all of the unnecessary HTML tags and take only the transliterated text itself from the site.

When deciding how to present an Akkadian text to the algorithm, we had to make a choice between (unbiased) transliterated texts and normalized text. A rudimentary distinction may be drawn between the two: a transliterated Akkadian text is a sign-by-sign transcription of the cuneiform text in which the signs that make up each word are separated by hyphens. To mark the necessary contrast between phonetic and logographic writings, they are represented in italic type and roman type, respectively (Fig. 2, middle column). A normalized text eliminates the hyphens between signs and attempts to represent noun and verb morphology correctly; it presents a phonetic approximation of how each word was pronounced in Akkadian.

There are some rules that govern the normalization of Neo-Babylonian. However, in general, it is avoided in most recent publications unless useful for linguistic or pedagogic purposes (20). Neo-Babylonian is the longest consecutive language phase of Akkadian, covering the entire first millennium BCE and ending sometime after the first century CE. The genres and writing conventions of this phase are characterized by their departure from the standardized orthography practiced throughout the second millennium BCE. Many spellings are inconsistent with the actual phonemic renderings of words and can vary to a considerable extent,[#] especially on account of the intensive language contact and interference between Akkadian and Aramaic (70, 71).

For this reason, we have chosen not to train the algorithm in any kind of normalization practices for the time being. In our training corpus, we remained on the level of (unbiased) transliteration, by creating a mechanical (unnormalized) bound transcription: Akkadian phonetic spellings and logographic writings are taken at face value, by simply removing connecting hyphens between syllables and between logograms.

Tokenization. Tokenization is an automatic process in which the text is split into words and each one is replaced by a numeric token. This is an important process that requires language-specific knowledge to prevent the loss of a great deal of semantic content. A classic example in English is tokenizing a word like "aren't." If we do not break it into two tokens, then it is considered a word on its own and loses the connection to "are" and "not." While it might be possible for the learning algorithm to learn that "aren't" is equivalent to "are not," bad tokenization can complicate matters considerably by creating a large number of unnecessary words in our dictionary.

Open source Akkadian tokenizers use the ASCII Transliteration Format (ATF) format that our dataset does not support. Therefore, we created an alternative Akkadian tokenizer. We took into consideration some of the aspects of the current form of the transliterations. We retained the distinction between phonetic and logographic readings (i.e., italic type or roman type): during tokenization, we used the same token for both values of the sign, but we kept the HTML start italics <i> and stop italics <\i> symbols so that the use of the word as a syllable or logogram can be inferred from the context. Although using two tokens to represent the two values of a sign has some advantages, we found that doing so adds a large amount of noise to the preprocessing step and decided to use this method instead.

NUM gín kùbabbar nígga GODNAME *ù* GODNAME *šá* šu^{II} NAME a*šú šá* NAME *ina muhhi* NAME a*šú šá* NAME *ina* MONTH kùbabbar a_4 NUM gín *inamdin ésu kišubbaa' šá* da é NAME a*šú šá* NAME a NAME *u* da é NAME a*šú šá* NAME *maškanu šá* GODNAME *šá* LOCATION *adii* NAME kùbabbar*šú išallimu* lú *mukinnu* NAME a*šú šá* NAME NAME a*šú šá* NAME NAME a*šú šá* NAME lú umbisag NAME a*šú šá* NAME LOCATION MONTH u_4 NUM kam mu NUM kam NAME lugal kurkur NAME a*šú šá* NAME ENDTOKEN

Fig. 4. Mechanical bound transcription of Babylonian text Yale Oriental Series 7 11.

Furthermore, cuneiform uses determinatives, which signify (among others) proper names, such as masculine names, god names, and female names. They are written in the transliterations in superscript as "I" or "Id," "d," and "f," respectively. Since, for our purpose, proper names are of importance only in their syntactical function, we replace the names with a tag, written in capitals, that identifies the particular type of proper name: such as "NAME," "GODNAME," or "FEMALENAME" token. Locations, identified by the superscript determinative "uru" before a toponym, or by the superscript "ki" after a toponym, were replaced by a "LOCATION" token. Month names, with the superscript determinative "iti" before the noun, were replaced by MONTH, and simple numbers were replaced by "NUM." In order to simplify the tokenization of damaged parts of the text, each damaged part was replaced with the token "<BRK>," and words that appeared only two times or less as "<UNK>," since we do not have enough information to determine their meaning.

The number of unique words in the vocabulary that was subsequently compiled is 1,549, and the total number of words is 220,926. The number of words that appear only once is 3,175, and 932 words appear twice. For comparison, the Penn treebank dataset, a standard and relatively small English dataset comprising texts from the *Wall Street Journal*, contains 10,000 unique words and a total number of 1,036,580 words. While overfitting is something to be aware of, given the scale of the data, the unique nature of these texts comprised of well-structured bureaucratic information makes them well suited for machine-learning modeling. The resulting Akkadian texts used to train the algorithm look like the example in Fig. 4, which shows the same text as in Fig. 2 but in our mechanical bound transcription.

Data Availability. All study data are included in the article, *SI Appendix*, and Datasets S1–S6. Atrahasis can be accessed on the Babylonian Engine Website (https://babylonian.herokuapp.com/). Our source code is available at GitHub (https://github.com/DHALab/Atrahasis).

ACKNOWLEDGMENTS. This research was supported by the Ministry of Science & Technology, Israel, Grant 89540 for the project "Human-Computer Collaboration for Studying Life and Environment in Babylonian Exile" of Sh.G. and Amos Azaria, as part of Sh.G.'s Babylonian Engine initiative. We thank Eugene McGarry for language editing, Avital Romach for her assistance with the final proofs of this paper, Klaus Wagensonner for tablet photographs, and Moshe Shtekel for designing the web-tool for Atrahasis. We especially thank the anonymous reviewers for their detailed remarks and corrections.

[#]Take, for example, the form of a very common word in the Nippur Achaemenid-period Murašu archive *hatru*. As shown by Stolper (54), the different spellings of this term leave the quality of the middle, dental consonant uncertain: (lú) *ha-ad/t/ ṭ-ru/ri*, its variants range from (lú) *ha-d/ ṭa-ri*, (lú) *ha-dar/tár/ ṭár*, and (lú) *ha-d/ ṭa-ad/t/ ṭ-ri*.

1. N. Veldhuis, "Cuneiform: Changes and developments" in *The Shape of Script. How and Why Writing Systems Changes*, S. D. Houston, Eds. (School for Advanced Research Press, 2012), pp. 3–24.
2. J. Huehnergard, C. Woods, "Akkadian and eblaite" in *The Cambridge Encyclopedia of the World's Ancient Languages*, R. D. Woodard, Eds. (Cambridge University Press, 2004), pp. 218–280.
3. M. P. Streck, "Akkadian in general" in *The Semitic Languages: An International Handbook*, S. Weninger, G. Kahn, M. P. Streck, J. C. E. Watson, Eds. (De Gruyter, 2011), pp 330–339.
4. M. P. Streck, "Großes altorientalistik. Der umfang des keilschriftlichen textkorpus" in *Mitteilungen der Deutschen Orient-Gesellschaft* (Deutsche Orient Gesellschaft, 2010), vol. 142, pp. 35–58.
5. C. Gütschow, *Methoden zur Restaurierung von ungebrannten und gebrannten Keilschrifttafeln* (PeWe-Verlag, 2012).
6. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
7. A. Radford *et al.*, Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
8. J. Cohen *et al.*, "iClay: Digitizing cuneiform" in *VAST 2004: The 5th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, Y. Chrysanthou, K. Cain, N. Silberman, F. Niccolucci, Eds. (The Eurographics Association, 2004), pp. 135–143.
9. H. Mara, S. Krömker, S. Jakob, B. Breuckmann, "Gigamesh and Gilgamesh – 3D multiscale integral invariant cuneiform character extraction" in *VAST: International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, A. Artusi, M. Joly, G. Lucet, D. Pitzalis, A. Ribes, Eds. (The Eurographics Assocation, 2010), pp. 131–138.
10. G. Earl *et al.*, "Reflectance transformation imaging systems for ancient documentary artefacts" in *Electronic Visualisation and the Arts (EVA 2011)*, S. Dunnand, J. P. Bowen, K. C. Ng, Eds. (BCS: The Chartered Institute for IT, 2011), pp. 147–154.
11. H. Hameeuw, G. Willems, New visualization techniques for cuneiform texts and sealings. *Akkadica* **132**, 163–178 (2011).

12. M. Pauzi, M. Asyraf, "Digital preservation of Malaysian historical artefact using 3D scanner: A case study of Mah Meri mask," PhD dissertation, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia (2017).
13. D. Fisseler, F. Weichert, G. Gerfrid. W. Müller, M. Cammarosano, "Extending philological research with methods of 3D computer graphics applied to analysis of cultural heritage" in *Eurographics Workshop on Graphics and Cultural Heritage*, R. Klein, P. Santos, Eds. (The Eurographics Association, 2014), pp. 165–172.
14. B. Bogacz, N. Gertz, H. Mara, "Character retrieval of vectorized cuneiform script" in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (IEEE Computer Society, 2015), pp. 326–330.
15. B. Bogacz, M. Klingmann, H. Mara, "Automating transliteration of cuneiform from parallel lines with sparse data" in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (IEEE Computer Society, 2017), vol. 1, pp. 615–620.
16. B. Bogacz, H. Mara, "Automatable annotations – Image processing and machine learning for script in 3D and 2D with Gigamesh" in *Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4*, H. Busch, F. Fischer, P. Sahle, Eds. (Books on Demand, 2017), pp. 137–149.
17. M. P. Streck, "Babylonian and assyrian" in *The Semitic Languages: An International Handbook*, S. Weninger, G. Kahn, M. P. Streck, J. C. E. Watson, Eds. (De Gruyter, 2011), pp. 359–395.
18. J. Hackl, "Zur Sprachsituation im Babylonien des ersten Jahrtausends v.Chr. Ein Beitrag zur Sprachgeschichte des jüngeren Akkadischen" in *Mehrsprachigkeit vom Alten Orient bis zum Esperanto*, S. Fink, M. Lang, M. Schretter, Eds. (Zaphon, 2018), pp. 209–238.
19. M. Jursa, "Accounting in Neo-Babylonian institutional archives: Structure, usage, and implications" in *Creating Economic Order: Record-keeping, Standardization, and Development of Accounting in the Ancient Near East*, M. Hudson, C. Wunsch, Eds. (CDL Press, 2004), pp. 145–198.
20. M. Jursa, *Neo-Babylonian Legal and Administrative Documents. Typology, Contents and Archives* (Ugarit-Verlag, 2005).
21. M. Jursa, *Aspects of the Economic History of Babylonia in the First Millennium BCE* (Ugarit-Verlag, 2010).
22. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
23. A. A. Akın, M. D. Akın, Zemberek, an open source NLP framework for Turkic languages. *Structure* **10**, 1–5 (2007).
24. N. Gupta, P. Mathur, Spell checking techniques in NLP: A survey. *Int. J. Adv. Res. Comput. Sci. Software Eng.* **2**, 217–221 (2012).
25. B. Kaur, Review on error detection and error correction techniques in NLP. *Int. J. Adv. Res. Comput. Sci. Software Eng.* **4**, 851–853 (2014).
26. S. Singh, S. Singh, "Review of real-word error detection and correction methods in text documents" in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (IEEE Computer Society, 2018), pp. 1076–1081.
27. R. Altarawneh, Spelling detection errors techniques in NLP: A survey. *Int. J. Comput. Appl.* **172**, 1–5 (2017).
28. O. Streiter, E. W. De Luca, "Example-based NLP for minority languages: Tasks, resources and tools" in *Proceedings of the Workshop "Traitement Automatique Des Langues Minoritaires et des Petites Langues", 10e Conference TALN*, O. Streiter, Ed. (Batz-sur-Mer, France, 2003), pp. 233–242.
29. S. Nirenburg, *Language Engineering for Lesser-Studied Languages* (Ios Press, 2009).
30. V. B. Juloux, A. R. Gansell, A. Di Ludovico, *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving* (Brill, 2018).
31. S. P. Singh *et al.*, "Frequency based spell checking and rule based grammar checking" in *2016 International Conference on Electrical, Electronics, and Optimization Techniques* (ICEEOT, 2016), pp. 4435–4439.
32. A. Sahala, M. Silfverberg, A. Arppe, K. Lindén, "BabyFST: Towards a finite-state based computational model of ancient Babylonian" in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (European Language Resources Association [ELRA], 2020), pp. 3886–3894.
33. P. Gakis, C. Panagiotakopoulos, K. Sgarbas, C. Tsalidis, Design and implementation of an electronic lexicon for modern Greek. *Lit. Ling. Comput.* **27**, 155–169 (2012).
34. N. Suguna, K. G. Thanushkodi, Predicting missing attribute values using k-means clustering. *J. Comput. Sci.* **7**, 216–224 (2011).
35. T. Homburg, C. Chiarcos, "Word segmentation for Akkadian cuneiform" in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. Calzolari *et al.*, Eds. (European Language Resources Association (ELRA), 2016), pp. 4067–4074.
36. M. Sukhareva, I. Khait, E. Pagé-Perron, C. Chiarcos, "Machine translation and automated analysis of the Sumerian language" in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics Anthology*, B. Alex *et al.*, Eds. (Association for Computational Linguistics, 2017), pp. 10–16.
37. C. Chiarcos *et al.*, Annotating a low-resource language with LLOD technology: Sumerian morphology and syntax. *Information* **9**, 290 (2018).
38. A. Z. Aktaş, B. Yesiltepe, T. Aşuroğlu, Computerized Hittite cuneiform sign recognition and knowledge-based system application examples. *Eur. Sci. J.* **15**, 32–54 (2019).
39. A. Sahala, M. Silfverberg, A. Arppe, K. Lindén, "Automated phonological transcription of Akkadian cuneiform text" in *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)* (European Language Resources Association [ELRA], 2020), pp. 3528–3534.
40. P. Majumder, M. Mitra, B. B. Chaudhuri, "N-gram: A language independent approach to IR and NLP" in *International Conference on Universal Knowledge and Language* (2002).
41. R. P. N. Rao *et al.*, A markov model of the Indus script. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13685–13690 (2009).
42. N. Yadav *et al.*, Statistical analysis of the Indus script using n-grams. *PLOS One* **5**, e9506 (2010).
43. G. Marra, A. Zugarini, S. Melacci, M. Maggini, "An unsupervised character-aware neural approach to word and context representation learning" in *International Conference on Artificial Neural Networks*, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis, Eds. (Springer, 2018), pp. 126–136.
44. Y. Kim, Y. Jernite, D. Sontag, A. M. Rush, "Character-aware neural language models" in *Thirtieth AAAI Conference on Artificial Intelligence*, D. Schuurmans, M. Wellman, Eds. (AAAI Press, 2016), pp. 2741–2749.
45. Z. Zhang, Y. Huang, P. Zhu, H. Zhao, "Effective character-augmented word embedding for machine reading comprehension" in *Natural Language Processing and Chinese Computing*, M. Zhang, V. Ng, D. Zhao, S. Li, H. Zan, Eds. (Springer International Publishing, Cham, 2018), pp. 27–39.
46. Y. Assael, T. Sommerschield, J. Prag, "Restoring ancient text using deep learning: A case study on Greek epigraphy" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, S. Pad, R. Huang, Eds. (Association for Computational Linguistics, 2019), pp. 6368–6375.
47. S. Tyndall, "Toward automatically assembling Hittite-language cuneiform tablet fragments into larger texts" in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, J. C. Park, Eds. (Association for Computational Linguistics, 2012), pp. 243–247.
48. T. Collins *et al.*, "Computer-assisted reconstruction of virtual fragmented cuneiform tablets" in *2014 International Conference on Virtual Systems Multimedia (VSMM)*, H. Thwaites, S. Kenderdine, J. Shaw, Eds. (IEEE Computer Society, 2014), pp. 70–77.
49. T. Collins, S. Woolley, E. Gehlken, E. Ch'ng, "Computational aspects of model acquisition and join geometry for the virtual reconstruction of the Atrahasis cuneiform tablet" in *2017 23rd International Conference on Virtual System Multimedia (VSMM)*, L. Goodman, A. Addison, Eds. (IEEE Computer Society, 2017), pp. 1–6.
50. M. Jursa, *Das Archiv des Bēl-rēmanni* (NINO, 1999).
51. O. Pedersén, *Archive und Bibliotheken in Babylon: Die Tontafeln der Grabung Robert Koldeweys 1899-1917* (SDV Saarländische Druckerei und Verlag, 2005).
52. C. Waerzeggers, "The network of resistance: Archives and political action in Babylonia before 484 BCE" in *Xerxes and Babylonia: The Cuneiform Evidence*, C. Waerzeggers, M. Seire, Eds. (Peeters, 2018), pp. 89–134.
53. L. E. Pearce, C. Wunsch, *Documents of Judean Exiles and West Semites in Babylonia in the Collection of David Sofer (CDL Press, 2014)*.
54. M. W. Stolper, *Entrepreneurs and Empire: The Murašû Archive, the Murašû Firm, and Persian Rule in Babylonia (NINO, 1985)*.
55. M. W. Stolper, "Kasr texts: Excavated–but not in Berlin" in *Studies Presented to Robert D. Biggs*, M. T. Roth, W. Farber, M. W. Stolper, P. von Bechtolsheim, Eds. (The Oriental Institute of the University of Chicago, 2007), pp. 243–283.
56. E. Frahm, M. Jursa, *Neo-Babylonian Letters and Contracts from the Eanna Archive* (Yale University Press, 2011).
57. K. K. Uruk, *"The fate of the Eanna archive, the Gimil-Nanāya B archive, and their archaeological evidence" in Xerxes and Babylonia: The Cuneiform Evidence, C. Waerzeggers, M. Seire, Eds. (Peeters, 2018), pp. 73–87*.
58. E. von Dassow, "Introducing the witnesses in Neo-Babylonian documents" in *Ancient Near Eastern, Biblical, and Judaic Studies in Honor of Baruch A. Levine*, B. A. Levine, R. Chazan, W. W. Hallo, L. H. Schiffman, Eds. (Eisenbrauns, 1999), pp. 3–22.
59. L. B. Bregstein, "Seal Use in Fifth Century B.C. Nippur, Iraq: A Study of Seal Selection and Sealing Practices in the Murašû Archive" (Doctoral dissertation, University of Pennsylvania, Philadelphia, PA, 1993).
60. E. Ehrenberg, *Uruk: Late Babylonian Seal Impressions on Eanna-Tablets* (Philipp von Zabern, 1999).
61. E. Gehlken, "Uruk: Spätbabylonische Wirtschaftstexte aus dem Eanna-Archiv" in *Ausgrabungen in Uruk-Warka. Endberichte 5* (Philipp von Zabern, 1990).
62. H. Lanz, *Die Neubablonischen Harrânu-Geschäftsunternehmen* (Schweitzer, 1976).
63. S. Holtz, *Neo-Babylonian Court Procedure* (Brill, 2009).
64. B. Wells, F. R. Magdalene, C. Wunsch, The assertory oath in Neo-Babylonian and Persian administrative texts. *Revue Internationale des Droits de l 'Antiquité* **57**, 13–29 (2010).
65. J. MacGinnis, *Letter Orders from Sippar and the Administration of the Ebabbara in the Late-Babylonian Period* (Bonami, 1995).
66. S. Zawadzki, *The Rental Houses in the Neo-Babylonian Period (VI-V Centuries BC)* (Wydawnictwo Agade, 2018).
67. C. Wunsch, "Debt, interest, pledge and forfeiture in the Neo-Babylonian and early Achaemenid period: The evidence from private archives" in *Debt and Economic Renewal in the Ancient Near East*, M. Hudson, M. Van De Mieroop, Eds. (CDL Press, 2002), pp. 221–255.
68. J. Hackl, M. Jursa, M. Schmidl, *Spätbabylonische Privatbriefe* (Ugarit-Verlag, 2014).
69. M. T. Roth, *Babylonian Marriage Agreements: 7th-3rd Centuries BC* (Butzon u. Bercker, 1989).
70. M. P. Streck, Orthographie. B. Akkadisch im II. und I. Jt. *Reallexikon der Assyriologie* **10**, 137–140 (2003).
71. M. P. Streck, "Innovations in the neo-Babylonian lexicon" in *Languages in the Ancient Near East: Proceedings of the 53e Rencontre Assyriologique Internationale,* L. E. Kogan *et al.*, Eds. (Eisenbrauns, 2010), pp. 647–660.