Assignment of account type to proto-cuneiform economic texts with Multi-Class Support Vector Machines

Piotr Zadworny
Freie Universität Berlin
p.zadworny@fu-berlin.de

Shai Gordin
Ariel University
Dept. of the Land of Israel and Archaeology
shaigo@ariel.ac.il

Abstract

We investigate the use of machine learning for classifying proto-cuneiform economic texts (3,500-3,000 BCE), leveraging Multi-Class Support Vector Machines (MSVM) to assign text type based on content. Proto-cuneiform presents unique challenges, as it does not encode spoken language, yet is transcribed into linear formats that obscure original structural elements. We address this by reformatting transcriptions, experimenting with different tokenization strategies, and optimizing feature extraction. Our workflow achieves high labeling reliability and enables significant metadata enrichment. In addition to improving digital corpus organization, our approach opens the chance to identify economic institutions in ancient Mesopotamian archives, providing a new tool for Assyriological research.

1 Introduction

Proto-cuneiform is a writing system which emerged in southern Mesopotamia at the end of the 4th millennium BCE.[1] It consisted of over 800 signs representing numbers, goods, and administrative procedures, which were impressed on small clay tablets using a reed stylus. The entire corpus consists of almost 7,000 texts, about 5,500 of which are economic accounts, used alongside other tools, such as small clay "tokens", bullae, and cylinder seals, to control the operations of early cities (Fig. 1).

Figure 1: *MS 4631*. A clay envelope with a seal impression (right) and an array of tokens kept inside it. Artifacts of this kind were the predecessors of writing, and continued to be used by the accountants after writing was invented as well.

The majority (ca. 80%) of proto-cuneiform artifacts originate from the large Eanna area in the city Uruk (modern-day Warka, Iraq). Excavations at the site since the 1920s by the *Deutsche Orient-Gesselschaft* (DOG), unearthed more than 5,000 texts. The focus of the Eanna excavators was, however, architecture, which determined the choice of a less-than-optimal approach towards other finds (Nissen, 2024).

Today, our understanding of those accounts' original use context is limited. On the one hand, this is due to the excavation documentation, where information is constrained to a square coordinate in a 20x20 m excavation grid, and occasional comments. On the other hand, it does not help that the tablets were already discarded in antiquity, and used as construction material within Eanna, so their deposition location is not the one in which they were written or stored. As one may expect in such a situation, they are often severely damaged.

Nevertheless, already Englund suggested that the distribution of tablets across the site echoes the original institutions from where they were taken (Englund, 1998). He observed that despite the secondary character of their deposition, accounts documenting the operations of the same sector of the archaic economy tend to be found together. In recent years, scholarly efforts into learning more about this site and the origins of writing, allowed us to gain a better understanding of the archaeological record of Uruk (Nissen, 2024; Naccaro, 2025).

In this article, we offer a machine learning

[1] The debate whether proto-cuneiform is genuine "writing" or just a mnemotechnical tool similar to other used at the time in the Ancient Near East is open. It is rooted in an exclusive definition of writing, which only allows glottographic systems (like later cuneiform), and not semasiographic ones (like proto-cuneiform).

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 22–30
May 4, 2025.

approach to automatic labeling of those archaic tablets according to the economic sectors they deal with — the account type. Once a trustworthy method of doing this is established, we can measure the accounts' similarity to each other and try to cluster them on that ground to determine local environments — or even offices — they were originally written in. In the future, we may as well try to artificially complete, or at least expand, the damaged artifacts using a model of similar accounts as a template. Those tasks are not innovative on their own — in fact, they are what cuneiform experts traditionally do — however, considering the amount of material to work with, we think that developing an automatic solution is the best approach for measurable results at scale that are also reproducible.

Most importantly, however, automatically generated account labels are an additional metadata point which, together with a method of identifying similar texts in terms of content, allows other researchers to navigate the archaic corpus in a more informed way. As of now, information about account types is dispersed across different publications, which makes exploring the otherwise completely digitized corpus difficult. Account type metadata, either with original citations or an "automated" tag, is available through *4ky* (Zadworny, 2023), an open-source web application.

2 Data

To achieve the goal of trustworthy automatic account type classification, we used an existing edition of a comparable (even if much smaller) collection of archaic tablets by Monaco (2007, 2014, 2016) as the main part of the training dataset. Importantly, the texts edited by Monaco come from the antiquities market, and their archaeological origin is unknown. It is accepted as unlikely that they originate from the Eanna area of Uruk (Lecompte, 2023). This set of accounts was extended with some tablets from Uruk and other smaller collections which were discussed by Englund (1998), and few additional texts which we classified ourselves.[2]

The transcriptions of all the accounts used in this study were sourced from the *Cuneiform Digital Library Initiative* (contributors, 2025). The account type tags assigned by the aforementioned authors to the training dataset were extracted and assigned to the transcriptions manually.

Total number of transcriptions used for training was 596. They were divided into seven account types: **animal husbandry**, **cereal**, and **dairy** texts, accounts of **fields**, **fish**, and **humans**, as well as documents concerning **textiles**. Given that some accounts contain a mix of items from multiple economic sectors, we occasionally assigned more than one tag to one text. This influenced the algorithm design, as explained later.

The composition of the training dataset reflects the archaic corpus as a whole. Most of the accounts are **cereal texts** (323 in the training set), followed by **animal husbandry texts** (125). Together, these two types dominate the corpus, making the development of automatic labeling for these accounts particularly useful.

Automatic labeling of **dairy texts** (23), as well as **fish** (22) and **textile accounts** (24) is an interesting task: although they are easily identifiable for human scholars thanks to well-understood semantic sets of signs, they are relatively rare, making training data scarce.

Field texts (42) are not as common either, and some of them are entirely mathematical in nature, only identifiable as such if we closely follow the accountant's calculations.

Human accounts (58) are challenging for another reason: they usually contain lists of entries understood as individual names and composed of semantically unrelated signs, which may confuse the model. Texts usually assigned to other types, such as grain distributions (a cereal text) or assignments of animal herds (an animal husbandry text) exhibit the same characteristics, adding to the difficulty.

As an additional limitation, we excluded accounts with fewer than 6 signs from the training set, as our experiments showed this led to an improvement of our model's accuracy.

3 Method

The main requirements for the model were its trustworthiness and the ability to assign multiple labels to a single text. Additionally, since we intend for scholars to use our tools online, we aimed for a lightweight implementation.

3.1 Model architecture

Due to the small and unbalanced training dataset, using a neural network was not the optimal solution. Instead, we decided to use support vector machines (SVMs), which are known to perform better in such

[2] The accounts labeled by us have a "manual" tag in *4ky*.

situations.

To allow for assigning multiple labels to each account, we chose a specific type of SVM: a multi-class support vector machine (Wang and Xue, 2014; Zhang et al., 2021). A standard SVM seeks to optimally divide the dataset into exclusive groups, assigning only one account type per text. In contrast, a multi-class SVM treats each account type as a separate 'true or false' question. This allows one account to appear in the 'true' section of multiple account types, and therefore to have multiple labels assigned to it. This approach also allows the accounts to remain unlabeled when they do not meet the criteria of any of the account types. Together with the ability to assess the certainty of each assignment, this improves the quality of the model.

3.2 Approaches to feature extraction

When working with proto-cuneiform accounts, several additional challenges arise that are unique to this writing system. First, the language of those documents is unknown. As the accounts were mainly accounting tools, the writing does not reflect speech. Instead, meaning is encoded through non-linear arrangements of sign sets within *cases*—meaningful subdivisions of the writing surface, similar to text fields in modern forms—making traditional language-oriented methods not applicable.

The second challenge is the non-linearity of the script itself. In Assyriological transcriptions, each proto-cuneiform sign is represented by its *sign name* in Latin script, typically derived from its meaning in later Sumerian cuneiform. The text is also linearized according to the transcriber's intuition, resulting in the loss of information about the original arrangement of the signs (Fig. 2).

As solving the issue of transcriptions' linearity is beyond the scope of this paper, we chose to ignore the arrangements of signs within cases altogether. Instead, we alphabetized the order of the signs within each case to ensure that cases containing the same sets of signs are always represented in the same way.

Aware of those challenges, to feed the account transcriptions into our SVMs, we had to choose a way of tokenizing the texts. To our knowledge, no studies have yet determined the optimal way of doing this in the case of archaic Mesopotamian accounts. Thus, we decided to experiment using a TF-IDF tokenizer[3] with two approaches and compare their accuracy: *case-by-case*, treating the entire group of signs within one case of the document as one unit, and *sign-by-sign*, treating each sign separately.

3.3 Adjustments for accuracy

In the course of the study we experimented with other aspects of the dataset as well: we tried to assess the importance of **number signs** and **sign variants**.

The **number signs** were the key that allowed the scholars from the Berlin-based *Archaische Texte aus Uruk* (ATU) project to decipher the archaic accounts in the first place. Through computer-aided statistical analysis, they could show that number signs come in distinct sets depending on what is accounted for, even if some signs are shared across different sets (Green and Nissen, 1987). This observation allowed them to connect the accounts to specific economic sectors, as well as describe sets of semantically similar item signs each sector used.

We were curious if other (non-number) signs alone are enough to make such distinctions. To test this, we prepared for each account an alternative transcription without any numbers, which we then used to train another set of models, and we included the results in the comparison.

Otherwise, when including numbers, we only used the type of the sign, and not its value. For example, the expressions $2N_1$ and $6N_1$ repeat the same sign type — N_1 — to express different values, so we treat them both as just N_1. Our early tests showed preserving the values decreased the accuracy in the preliminary testing phase. This may have to do with the model giving weight to rare tokens. For instance the rare value $5N_{14}$ may appear unique—and thus significant—to some account types, when in fact this is entirely coincidental, and the sign N_{14} is otherwise common.

The **sign variants** are a palaeographic feature of CDLI transcriptions. They are represented using lowercase letters after the tilde sign (Fig. 2). Although in most cases the variants seem not to indicate semantic differences, there is at least one important example to the contrary: the sign DUG_a usually represents beer, whereas DUG_b and DUG_c stand for dairy fats — each a distinctive entry in different types of accounts. Experiments with excluding them invariably led to dips in model qual-

[3]Part of Scikit-learn *TfidfVectorizer*.

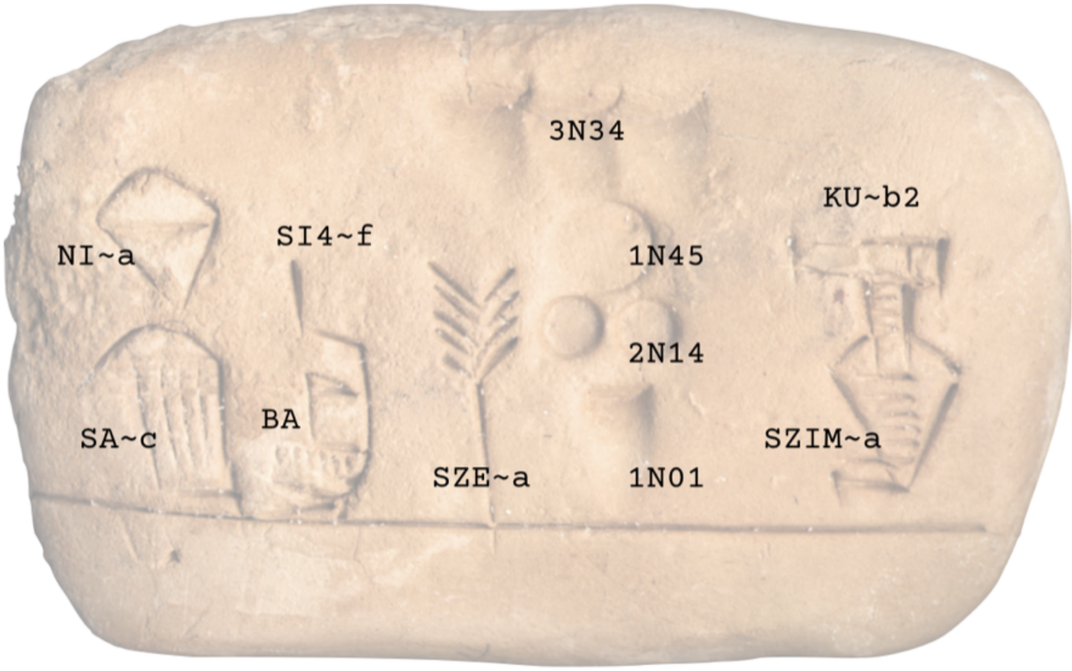

Figure 2: Reverse of *MSVO 3, 64* — metadata section of a cereal account. Here, all signs are inscribed within a single *case*, delineated with a horizontal line.

Transcription in CDLI:
`3(N34) 1(N45) 2(N14) 1(N01), SZE∼a KU∼b2 SZIM∼a SI4∼f BA NI∼a SA∼c`

ity.

Although it did not make much sense from a methodological point of view to allow the tokenizer to use *n-grams* (sequences of *n*-number of signs), since the signs in the transcriptions were reordered in an arbitrary way, we tested this aspect as well. Surprisingly, allowing for *n*-grams improved the model's accuracy, which made us choose to keep this feature. However, it did not matter whether we set the limit of *n* to 2 or more, so again we opted for the lowest value (2) to reduce its complexity.

3.4 Method summary

To summarize, we transformed the original dataset in two ways: through alphabetizing the order of signs within each case, and creating alternative versions of transcriptions without number signs.

Then, we trained four MSVMs to determine which is the most accurate. The variants of tokenization and transcription used were: 1) *line-by-line* with numbers; 2) *line-by-line* without numbers; 3) *sign-by-sign* with numbers; and 4) *sign-by-sign* without numbers.

70% of the dataset was used for training, and 30% for testing the model. The split was random and done once per each model. The models were trained for 10 iterations to see if the outcome changes, and in all cases the results were similar. In the next section, we present the final set of results.

4 Results

For convenience, the results of testing the models will be discussed separately for each account type.

4.1 Animals

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.87	1.00	0.93	0.81	1.00	0.94	148
yes	1.00	0.26	0.41	1.00	0.35	0.52	31
accuracy			0.87			0.89	179
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.93	1.00	0.96	0.93	1.00	0.96	148
yes	1.00	0.61	0.76	1.00	0.61	0.76	31
accuracy			0.93			0.93	179

Table 1: Animal accounts classification

All the values are in the range of 0 to 1, so 0.90 is equal to 90%. The *precision* value shows how many times the model was correct in assigning the label type. So, in the first example, 87% of the negative answers ("the account is not an account of animals") were correct. The *recall* value shows how many matching examples were found—so, together with 87% precision, the 100% value means that although the model assigned the 'not animal' label to too many texts, it caught 100% of the true non-animal texts. The *support* value (last column) is the real number of non-animal and animal accounts in the test set, and it is used as a weight to calculate the *F1 score* — the realistic accuracy measure of the model.

Knowing that, we can see that *line-by-line* tokenization did not work very well in the case of animal accounts: the positive recall scores of 26% (with numbers) and 35% (without numbers) show that the models failed to catch many animal texts in the testing dataset. The performance of the *sign-by-sign* models was slightly better (61% in both). Importantly, neither produced any false positives (the precision score of *yes* is 100%), which is a good sign from the point of view of reliability. Also, it is interesting to see that in this case it did not matter if the numbers were included or not, as the outcome scores were the same.

An additional method of examining the models' performance at this stage is studying **feature importance**. Our models, as they were trained, assigned coefficients (positive and negative) to each token they encountered, to determine how likely each token is to appear in a specific account type. After training, we extracted those coefficients to see what signs or sign combinations models considered as particularly telling for each account type.

In terms of animal accounts, for the *sign-by-sign, with numbers* model, the most positively important tokens were KIŠ (an equid sign), UD_{5a} ("ram"), U_8 ("ewe"), and N_2 — a numerical sign used to account for dead animals (Englund, 1998).

Among the negatively important tokens we find signs like N_{19} (a quantity sign for emmer, ca. 150 liters), $KISIM_b$ ("sheep's milk butter"), or $SAL.KUR_a$ (metadata sign used for totals of workers). Those findings suggest that the model correctly identifies which signs belong to the semantic set of animal signs, and used them to label the accounts.

4.2 Cereals

The performance of the model on cereal accounts was significantly better, and we assume it is due to the dominance of those texts in the training dataset. Here, unlike in the animal texts, *line-by-line* models achieved more success, although they were still worse than those using the *sign-by-sign* approach. With those, we see scores similar to the ones above, with the model with numbers performing slightly better than the one without. Unfortunately, every model produced false positives, with the *sign-by-sign, with numbers* model making the fewest mistakes.

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.88	0.72	0.79	0.71	0.79	0.75	78
yes	0.81	0.92	0.86	0.83	0.75	0.79	101
accuracy			0.83			0.77	179
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.89	0.97	0.93	0.81	0.92	0.86	94
yes	0.98	0.91	0.94	0.93	0.83	0.88	85
accuracy			0.94			0.91	179

Table 2: Cereals classification

Feature importance analysis shows the significance of number signs in this account type: although the top-scoring signs are $ŠE_a$ ("barley"; also metadata sign for totals of grain) and DUG_a (most often "beer"), among the top ten positively important signs we have N_{39a} (a quantity of ca. 5 liters of barley), N_{19} (ca. 150 liters of emmer), N_4 (ca. 25 liters of emmer) or N_{24} (ca. 2.5 liters of barley or malt). Included are also bigrams N_1 $ŠE_a$ (ca. 25 liters of barley) and N_{45} N_4 (even larger volumes of emmer).

The significance of number signs stands out among the negative features as well: among the top signs are N_1 and N_{34} (polyvalent number signs; used in different accounting systems to represent different quantities) as well as N_{50} and N_{22} (used in field measurements).

Overall, this is not surprising, as the accounting systems for cereals were the most varied and contained the most unique numerical signs. Like in the case of animal texts, we see that the model could recognize that and use this feature of proto-cuneiform.

4.3 Dairy

The dairy texts are among the most underrepresented in the training set, and this is clearly visible in test scores. In this case, it is difficult to decide which model performed best: the highest score of 22% (*sign-by-sign, without numbers*) is equal to 2 catches, and two other models caught 1 text each. While no model produced false positives, the outcome seems hardly useful. It is also important to

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.95	1.00	0.97	0.96	1.00	0.98	170
yes	0.00	0.00	0.00	1.00	0.11	0.20	9
accuracy			0.96			0.96	179
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.96	1.00	0.98	0.96	1.00	0.98	170
yes	1.00	0.11	0.20	1.00	0.22	0.36	9
accuracy			0.96			0.96	179

Table 3: Dairy classification

acknowledge that the very high compound accuracy scores (96%) are inflated by the overwhelming proportion of true negatives, a phenomenon which repeats for other less common account types, and thus *is not* a meaningful measure of success.

Despite that, the **feature importance** analysis shows that the model was still able to learn the signs which are characteristic for dairy accounts. Among positively important signs we have DUG_c ("dairy fat"), $KISIM_a$ (butter from sheep's milk), as well as bigrams N_1 $KISIM_a$ ("one vessel of the butter from sheep's milk") or N_1 KU_{3a} (a compound number sign representing the quantity of ca. 4 liters of dairy fats).

Interestingly, the negatively important features seem to focus on animal signs, which sometimes do appear in dairy accounts. We find in the report such bigrams as N_1 AB_2 ("one cow") or AMAR $U_4 \times 1N_{57}$ ("one-year-old youngling"), which require further study. However, this "prejudice" against animal signs may cause the low score of the model.

4.4 Fields

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.91	1.00	0.95	0.93	1.00	0.96	163
yes	0.00	0.00	0.00	1.00	0.25	0.40	16
accuracy			0.91			0.93	169
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.95	1.00	0.98	0.93	1.00	0.96	163
yes	1.00	0.50	0.67	1.00	0.25	0.40	16
accuracy			0.96			0.93	169

Table 4: Field texts classification

The outcome here is similar to the one presented in the previous section. Again, we see overall low scores, with the *sign-by-sign, with numbers* model scoring the highest, with 50% of correctly labeled accounts. Again, we have no false positives.

Among the positively **significant features**, we see GAN_2 ("field"), as well as several number signs from the appropriate accounting system: N_{50} (an area of ca. 65ha) and N_{22} (ca. 2,16ha), alongside bigrams formed of various combinations of number signs. This reflects the often mathematical character of field texts, many of which are area calculations.

Other than one puzzling bigram, GAN_2 $APIN_b$ ("land for ploughing?" or "ploughed field?"), the list of negatively important signs consists of seemingly random entries.

4.5 Fish

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.97	1.00	0.98	0.97	1.00	0.99	173
yes	0.00	0.00	0.00	1.00	0.17	0.29	6
accuracy			0.97			0.97	179
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.98	1.00	0.99	0.98	1.00	0.99	173
yes	1.00	0.50	0.67	1.00	0.33	0.50	6
accuracy			0.98			0.98	179

Table 5: Fish accounts classification

The fish accounts had the lowest support value out of all the account types, which makes the accuracy scores accidental.

Similarly to other rare account types, we see that the *sign-by-sign, with numbers* model scored the highest, though too by a margin of a single account. The small difference, together with the low support value makes it difficult to judge the quality of the models.

Nonetheless, like in the case of dairy tablets, the model was able to learn some fish-specific signs. Among the positively **important signs** we find entries like SUHUR ("dried fish"), $ZATU759 \times KU_{6a}$ (a container with fish?), or $GA_2 \times KU_{6a}$ ("basket with fish"), all typical for this semantic set.

The list of negatively important features consists mostly of numerals belonging to the cereal system, however, the highest scoring entry is N_8 SUHUR: a seemingly valid fish qualification.

4.6 Humans

The accounts of humans is an account type where all models failed entirely and did not catch any texts.

Despite this — and in line with what we have seen in the cases of other underrepresented account types — **feature analysis** shows that the model did identify some signs that are indicative of human accounts. We see the bigram N_1 $SAL.KUR_a$

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.92	1.00	0.96	0.92	1.00	0.96	164
yes	0.00	0.00	0.00	0.00	0.00	0.00	15
accuracy			0.92			0.92	179
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.92	1.00	0.96	0.92	1.00	0.96	164
yes	0.00	0.00	0.00	0.00	0.00	0.00	15
accuracy			0.92			0.92	179

Table 6: Human accounts classification

as well as just SAL.KUR$_a$ (metadata sign for total of male and female workers), accompanied by SAL ("adult female") and N$_1$ AL, a qualification describing groups of laborers.

The list of negatively significant entries is coincidental, though it is interesting to see BA (an administrative qualification) there, as it often features in assignment texts, and is used in personnel lists (Johnson, 2014).

4.7 Textiles

	with numbers, line by line			without numbers, line by line			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.95	1.00	0.97	0.96	1.00	0.98	170
yes	0.00	0.00	0.00	1.00	0.22	0.36	9
accuracy			0.95			0.96	179
	with numbers, sign by sign			without numbers, sign by sign			
	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
no	0.97	1.00	0.99	0.97	1.00	0.98	170
yes	1.00	0.44	0.62	1.00	0.33	0.50	9
accuracy			0.97			0.97	179

Table 7: Textile accounts classification

The results for textiles, as expected, resemble those for dairy, fish, and human texts. The *sign-by-sign, with numbers* model scored the highest, though with a narrow margin. The models produced no false positives.

In the same manner as above, the model captured the **signs specific** to textiles fairly well: among the positively significant signs we see TUG$_{2a}$ ("garment?"), BARA$_{2a}$ (a type of garment), and SIG$_{2b}$ ("wool?"), as well as combinations of those with number signs.

The negatively important features, again, appear coincidental.

5 Corpus-wide experiment and discussion

It seems that the only feasible approach to tokenizing proto-cuneiform is *sign-by-sign, with numbers*, as other approaches consistently scored lower or failed to produce useful results entirely. We see also that the model is conservative, which is good for overall reliability: the precision values for *yes* are almost always 100%, making false positives extremely rare. This meets our basic requirements for experimentally labeling accounts across the entire corpus.

Additionally, feature analysis showed that the model managed to learn semantic sets of signs indicative of all the economic domains, including the underrepresented ones. Despite low scores in the testing phase, we saw this as an optimistic starting point.

To try labeling the entire corpus, we used the *sign-by-sign, with numbers* MSVM to assign labels to the entire corpus of archaic accounts. An advantage of the MSVM architecture is the ability to set a certainty threshold for the model, which allowed us to set the required threshold to a conservative 90%. We also opted to exclude texts containing fewer than 6 signs from labeling entirely, similar to what we did in the training phase. The outcome of this stage of experiment is presented in Table 8.

	animals	cereals	dairy	fields	fish	humans	textiles
training	125	323	23	42	22	58	24
assigned	151	498	68	13	48	31	49
increase (%)	+121%	+154%	+296%	+30%	+218%	+53%	+204%

Table 8: Outcome of applying the model to the entire archaic corpus.

In terms of quantity, the results represent a compound increase of 143% over the training dataset, and together they correspond to 39% of all texts longer than 6 signs (1,454 out of 3,737 texts). It is yet to be determined whether we can decrease the lower limit of account length without sacrificing quality.

Moreover, the results show a large disparity in efficiency: for example, the model found very few new accounts of fields or humans. We can tentatively explain this through the overall scarcity of those texts.

On the other hand, we have large amounts of new dairy, fish, and textile tablets that the model identified, especially interesting due to its low efficiency when dealing with such accounts in the testing phase. To understand the reasons for those disparities, we performed an error analysis of the corpus labels.

5.1 Sample error analysis

Analyzing the errors in all automatically labeled tablets was not feasible, so we opted for a sample-based analysis instead.

For each assigned label, we chose 10 random tablets for evaluation and judged the model's work on a three-level scale: *yes* (the assigned label is correct), *unsure* (we were not able to classify the tablet ourselves), and *no* (the tablet was classified incorrectly). The outcome is presented in Table 9.[4]

	animals	cereals	dairy	fields	fish	humans	textiles
yes	9	5	10	5	9	6	9
unsure	-	1	-	4	1	4	1
no	1	4	-	1	-	-	-

Table 9: Evaluation of assigned labels.

This demonstrates the merit of the conservative approach. Like in the testing phase, we see only few false positives — which again meets our expectations. Also, they are almost entirely limited to one specific type of texts: cereal accounts.

A closer look at the successfully classified texts shows a distinct limitation of our model. Almost all of the texts the it managed to find were inventories, usually containing few signs other than the semantic sets it learned. The few assignments (accounts of distributions of goods to individuals; see the discussion of human accounts in the Data section) the model caught, tellingly, also contained unusually many signs from those semantic sets.

The model's focus on limited sets of important signs is also what helps us understand its failure when dealing with cereal tablets: this account type is particularly diverse: in addition to inventories, it includes assignments of rations, harvest texts, seed texts, etc., each with new sets of tokens to learn. It is likely that those different subtypes of cereal accounts made them less statistically discernible. Although some degree of diversity exists within other types of texts as well, we think the "confusing" accounts in those cases were not numerous enough to have the same effect.

A reflection of the same issue is illustrated by the misclassified "field" account: the model learned too often that the sign GAN_2 is indicative of field texts that it classified an entirely different account, qualified with a known, yet undeciphered administrative term MAŠ GAN_2, as a field account too.

6 Conclusions

The original goal of this study was to make the exploration of the archaic corpus easier by enriching its metadata and allowing for more detailed statistical studies of the transcriptions. Using the resulting MSVM models trained, we succeeded in more than doubling the number of labeled accounts, although the error analysis suggests that some of the labels assigned (animals, dairy, fish, textiles) are more reliable than others (cereals).

An open question remains: can we label more accounts in a more detailed way? When we refine our typology of tablets and agree on the expected labels for cereal account subtypes, we may have enough data to process those as efficiently as others. The experiment above showed that the model managed well even with little input, as long as it looked at inventories with fixed semantic sets of signs.

The error check hints at the existence of a more fundamental split of account types than according to economic sectors: one according to their administrative use, dividing the texts into assignments and inventories. Assignments, usually consisting of lists of individuals or institutions, are usually more similar to each other across sectors than to inventories of their own sector, leading to the models' trouble with identifying them. In those texts, we can often only understand the account type through studying the sets of numerals used, or through metadata written at the end of the text. A viable method of approaching assignments in an automated way is yet to be discovered.

As described in the introduction, understanding the typology of accounts in greater detail may help us understand their original institutional environments better — in the Eanna district of Uruk, as well as in other sites. The distinction between inventories and assignments is one that we need to further explore, and having recognized it will help us refine our tools and methods, both digital and traditional.

Additionally, we should see expanding the metadata as an important aspect of developing the digital infrastructure. In an effort to make the archaic corpus more accessible, we used the outcomes of this study to develop a tool which allows scholars to find similar tablets based on their content using the *sign-by-sign, with numbers* tokenizer. This similarity measurement tool is one of the features of **4ky** (Zadworny, 2023), and it is freely available for other researchers interested in using or adapting it to their needs.

All datasets, code, and models created during this study are accessible on GitHub.

[4]Detailed scores are available in our GitHub repository.

Acknowledgments

We wish to thank all the personnel of the *Archaische Texte aus Uruk*, Salvatore Monaco, and CDLI contributors, without whose immense editorial work on the archaic text corpus this paper and future research could not take place.

References

CDLI contributors. 2025. Cuneiform digital library initiativ home. [Accessed: 29 January 2025].

Robert K. Englund. 1998. *Texts from the Late Uruk Period*, pages 15–233. Universitätsverlag Freiburg Schweiz, Vandenhoek Ruprecht Göttingen.

Margaret Green and Hans J. Nissen. 1987. *Archaische Texte aus Uruk: Zeichenliste der archaischen Texte aus Uruk*. CDL Press.

J. Cale Johnson. 2014. Late uruk bicameral orthographies and their early dynastic rezeptionsgeschichte. *Working Paper No. 2/2014*. (http://dx.doi.org/10.17169/refubium-22792).

Camille Lecompte. 2023. Monaco, salvatore f.: Archaic cuneiform tablets from private collections (review). *Orientalische Literaturzeitung*, 118(3):175–186.

Salvatore F. Monaco. 2007. *The Cornell University Archaic Tablets*. CDL Press.

Salvatore F. Monaco. 2014. *Archaic Bullae and Tablets in the Cornell University Collections*. CDL Press.

Salvatore F. Monaco. 2016. *Archaic Cuneiform Tablets from Private Collections*. CDL Press.

Hugo Naccaro. 2025. La mésopotamie du sud á la transition des ive et iiie millénaires avant notre ère: évolution des cultures proto-urbaines. Dissertation, forthcoming.

Hans J. Nissen. 2024. Uruk and i. *Cuneiform Digital Library Journal*, (1).

Zhe Wang and Xiangyang Xue. 2014. *Multi-Class Support Vector Machine*, page 23–48. Springer International Publishing, Cham.

Piotr Zadworny. 2023. 4ky. [Accessed: 29 January 2025].

Zhiqiang Zhang, Zeqian Xu, Junyan Tan, and Hui Zou. 2021. Multi-class support vector machine based on the minimization of class variance. *Neural Processing Letters*, 53(1):517–533.