

Daniel M. Stelzer*

A recursive encoding for cuneiform signs

<https://doi.org/10.1515/itit-2024-0067>

Received July 3, 2024; accepted March 7, 2025;

published online March 27, 2025

Abstract: One of the most significant problems in cuneiform studies is the process of identifying unknown signs, which often involves a tedious page-by-page search through a sign list. This paper proposes a new “recursive encoding” for signs, which represents the arrangement of strokes in a way that computers can process. A series of new algorithms then offers scholars a new way to look up signs by any distinctive component, as well as providing new ways to render signs and tablets electronically.

ACM CCS: Applied computing → Document management and text processing → Document preparation → Format and notation.

Keywords: cuneiform; Assyriology; Hittite; encoding; sign lists; sign lookup

1 Introduction

Cuneiform is famous – or infamous – for several reasons. It is the oldest deciphered script in the world, the script used to document the oldest known epic poems and the oldest readable accounting records. And it is infamously difficult both to learn and to read. Huehnergard [1, p. xxiv] calls it “very cumbersome” and “unquestionably the most difficult aspect of learning Akkadian”, relegating it to the later parts of his textbook. Cooper [2, p. 55] claims that “[n]either efficiency nor convenience played an important role in the development of Akkadian cuneiform”, and according to Worthington [3, p. 289], its orthographic features “suggest that ancient sight-readers of cuneiform were expected to decipher a line a bit at a time – not to sweep their eyes across it as we do with our script”.

But this difficulty isn’t just a property of the system itself. As Watkins and Snyder [4, p. 2] put it, “the pedagogical tools are, in many cases, non-optimal”. Looking up

unfamiliar words in a dictionary, physical or electronic, is standard practice when learning a new language. But cuneiform in particular involves at minimum hundreds of phonetic signs, and hundreds of logograms on top of those. Even after memorizing the basics, modern students will inevitably encounter signs they’ve never seen before – and on the level of individual signs, there’s no alphabetical order or computerized search to help find them. While a learner of English has a standardized ordering to help their search (if they’re looking for *reverent*, it will be after *revenge* and before *revile*), a student of Akkadian or Sumerian has no such aid.¹

The traditional solution for this problem is a sign list, such as Borger’s [5] *Mesopotamisches Zeichenlexikon* or Labat’s [6] *Manuel d’épigraphie akkadienne*. These typically order the signs based on their strokes, counted from left to right, in the clean, unambiguous Neo-Assyrian form. But most cuneiform is not Neo-Assyrian, and damaged tablets are the rule rather than the exception. Tablets tend to be found in a minimum of ten distinct pieces [7, p. 8], and signs with damage to the left side are common. In these cases, Robson [8] recommends checking other sign indices, and offers a few alternatives if those also fail:

You can make an educated guess at the value the sign ought to have, based on the signs immediately around it, and then look up that value in the relevant index of Borger or Labat. Then you can compare your sign with the entry in Labat’s table or Borger’s paleographic list. The PSL lists of homophones and compounds can often be useful aids in this type of search.

Or you can simply page through Labat’s table, looking for the closest match to your sign in the relevant script. I have done this countless times. It doesn’t seem very clever or efficient, but sometimes it is the only way to find what you are looking for.

The aim of this project is to find a better way. While learning cuneiform has always been a difficult and time-consuming task, we now have technologies the ancient Babylonians never dreamed of. Can we find a better way of looking up an unknown, potentially-damaged sign than scanning through Labat one page at a time?

*Corresponding author: Daniel M. Stelzer, Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, E-mail: stelzer3@illinois.edu

¹ Or rather, a student working with *cuneiform* has no such aid. Alphabetized dictionaries exist for Sumerian, Akkadian, and Hittite *in transliteration*, which are invaluable when translating a text, but are no use until the signs have been identified in the first place!

Section 2 reviews previous work in this area, both with cuneiform and other logographic writing systems. Section 3 draws on this to propose a new encoding system for cuneiform, based on earlier work with Xixia. Section 4 then describes several algorithms using this new system, and the uses to which they can be put.

2 Previous work

2.1 Sign lists

Unfamiliar logograms² in cuneiform are not a new problem – one ancient letter describes a perplexed Kassite administrator receiving shipments of straw (the very common logogram $\text{𒀭} > \text{IN}$) when he'd ordered clay pots (the similar and much rarer $\text{𒀭} > \text{KAN.NI}$) [9, pp. viii, 142]. Ancient lexical lists giving names and meanings for cuneiform logograms are archaeologically common [2], as is evidence of students using them. Babylonian scribes referenced, copied out, and eventually memorized many of these sign lists to prepare for their job – as we can learn from their own complaints!³

If you have learned the scribal art, you have recited all of it, the different lines [of the dictionary], chosen from the scribal art; the [names of] animals living in the steppe to the [names of] artisans, you have written; after that, you hate writing!

However, it's unclear whether scribes would actually memorize every logogram, and texts describe scribes both reciting them and referencing physical tablets (from the same text as above: “all the vocabulary of the scribes in the *eduba* is in your hands”). Sjöberg [10, p. 164] points out that, while scribes would brag about their memorization skills, the surviving lexical lists are often extremely long, and would be infeasible to memorize by rote. The aforementioned list of “names of artisans”, $\text{LÚ} = \text{ša}$, consists of about a thousand lines [11], while the “animals living in the steppe” comes from $\text{UR}_5\text{-RA} = \text{hribullu}$, with around 3,000 [12]. Worthington [3, p. 289] goes further and suggests that reading fluently through a text simply did not happen in ancient Mesopotamia, with scribes frequently needing to pause to figure out an unclear or ambiguous sign.

In modern times, several authors have revived the ancient tradition of cuneiform sign lists, such as Deimel and

Gössman [13] for Sumerian, Rüster and Neu [14] for Hittite, or Borger [5] and Labat [6] for Akkadian. These are generally extensive lists of signs with names and information on each one. While there's no universal order for cuneiform signs that could aid in searching, an informal standard has arisen: sorting starts at the left side of the sign, with — preceding ↖ preceding ↗ preceding ↙ preceding ↑ [5, p. 1, 27, p. 26, 23, p. 183]. Borger set precedent basing his sorting on the Neo-Assyrian style, where the ordering of strokes tends to be fairly clear. But in his own words [15, p. 2]:

Regrettably, it is practically impossible to arrange other versions of cuneiform writing (including the Ur III signs) by the shape of their signs in a consistent and unequivocal way.

Indeed, while Labat's [6] general Akkadian index and Rüster and Neu's [14] Hittite sign list try to follow the same pattern,⁴ it's an imprecise measure at best. The ancients don't seem to have had any better system – the ordering of signs in standard lexical lists comes off as arbitrary at best, and may have generally come down to the whim of the compiler⁵ [2, p. 48]. Similar-looking and similar-sounding signs were generally grouped together, but without a broader sorting order as found in modern dictionaries⁶ [18].

More recently, electronic references such as Šašková [19] and *ePSD* [20] have arisen to make it easier to find information on particular signs. These are significantly more convenient than physical sign lists when looking up signs by name or reading, since they can take advantage of electronic searching. But they're no help when looking at an autograph or an actual tablet: the student needs to already know at least one reading to use these tools.

2.2 Cuneiform encodings

While sign lists and dictionaries remain the most popular cuneiform references, the present author is far from the first to notice the problem. According to Gottstein [16], these sign lists “are very helpful for translation work, but are in most cases extremely impractical to handle”⁷ – in particular, “because of how comparatively time-consuming it is

² Following standard practice in Hittite studies, phonograms (signs indicating sound) are written in *italics*, while logograms (signs indicating meaning) are written in CAPS. CAPS are used when talking about a sign in and of itself, without committing to any particular interpretation.

³ Taken from Sjöberg [10, p. 163].

⁴ Though there are slight variations: Huehnergard [1, p. 563] groups ↖ , ↗ , and ↙ together as a single category, for example, while Gottstein [16] combines ↖ and ↙ but keeps ↗ separate.

⁵ See Cooper [2, pp. 49–52] for an extensive example.

⁶ Though Veldhuis [17, pp. 41–46] argues that there was more method to the madness than Cooper recognized; it just tended to be obscured by corruptions and reinterpretations over the long history of the most popular lexical lists.

⁷ Gottstein [16, p. 127]: *die zwar sehr hilfreich für die Übersetzungsarbeit, in der Handhabung jedoch größtenteils äußerst unpraktisch sind.*

to find particular signs, especially in academic introductory classes”.⁸ In his words:

A precise system for classifying and, more importantly, looking up particular cuneiform signs within the framework of a analytical sign dictionary has long been among the desiderata of Ancient Near Eastern research, but so far has been neither realized nor tackled in any consistent way.⁹

A chemist by profession, Gottstein was inspired by molecular formulae. There’s no obvious way to impose an alphabetical order on three-dimensional chemical structures, either, but reference works on chemistry have found a solution: molecules are indexed by the number of each type of atom they contain. A conventional order for the atoms ensures that these formulae can be sorted in a coherent way: the entry on nitrobenzene would be listed under $C_6H_5NO_2$, after quinone ($C_6H_4O_2$) but before benzene (C_6H_6).¹⁰

A molecular formula isn’t necessarily unique – $C_6H_5NO_2$ can also describe nicotinic acid, or a handful of other chemicals. But there are generally few enough chemicals with a particular formula that a student can scan through them easily to find the one they need. This is the sort of system Gottstein hoped to extend to cuneiform.

Cuneiform across all times, places, and languages is generally analyzed as having five basic types of wedges: vertical, horizontal, downward diagonal, upward diagonal, and the “Winkelhaken” or “hook” [22, p. 9, 2, p. 1]. These became the “elements” of Gottstein’s encoding, with the first four labelled ‘a’, ‘b’, ‘c’, and ‘d’ respectively (Figure 1).¹¹ After









			
			
A	B	C	D

Figure 1: A demonstration of the “Gottstein system”, adapted from Gottstein [16, p. 129] and Homburg [21, p. ii131].

⁸ Gottstein [16, p. 127]: *an dem vergleichsweise hohen Zeitaufwand, den das Auffinden bestimmter Zeichen [...] – gerade auch im akademischen Anfängerunterricht – mit sich bringt.*

⁹ Gottstein [16, pp. 127–8]: *Eine stringente Systematik zur Identifikation und vor allem Auffindung bestimmter Keilschriftzeichen im Rahmen eines analytischen Zeichenkompendiums gehört daher seit Langem zu den Desideraten der altorientalistischen Forschung, wurde bislang jedoch weder realisiert noch konsequent in Angriff genommen.*

¹⁰ Like with cuneiform signs, chemicals tend to have names that can be alphabetized cleanly – but this is little help to a student who doesn’t know the name of a new substance.

¹¹ Gottstein’s ordering of the strokes differs from what’s generally used in sign lists, where horizontals come first and verticals last.

some early experiments, he grouped Winkelhaken into the ‘c’ category as well – these identifiers weren’t meant to be entirely unique, and the distinction between downward diagonals and Winkelhaken is not always obvious on an actual tablet. Still, some users of this system have extended it with a ‘w’ category to capture that difference.¹²

Every sign can then be categorized in three ways: by the total number of strokes it contains (“category”), which *types* of strokes it contains (“designation”), and the number of each type of stroke (“Gottstein code”). The sign EME ‘tongue’ in Figure 2 contains nine strokes total, of the ‘a’, ‘b’, and ‘c’ species, giving it category 9, designation ABC, and Gottstein code a3b5c1.

Gottstein proposed a sign list that would be organized first by category, then by Gottstein code, displaying each sign and variant that could possibly have that code: Figure 3 shows the section for category 3, code a3.

In the same paper, Gottstein describes the broad outlines of an experiment, reporting that “every sign listed according to the ‘Gottstein System’ could be found within


ABC

 a3b5c1
 3+5+1=9

Figure 2: Gottstein’s analysis of the sign EME ‘tongue’, adapted from Gottstein [16, p. 129] and Homburg [21, p. ii131].





	nig2	a3
	a	a3
	eš5	a3
	diš/diš/diš	a3

Figure 3: An excerpt from Gottstein’s [16, p. 133] sign list, showing the signs with the Gottstein code a3.

¹² This extension appears in Homburg [21], but that may not be its first usage. Wikidata terms it “extended Gottstein encoding”, as seen at <https://www.wikidata.org/wiki/Q119228805>, where one particular sign variant is classified as a13b5c1w2.



Figure 4: Three variants of the sign U_2 ‘plant’. The leftmost is standard, but all three are attested in Hittite.

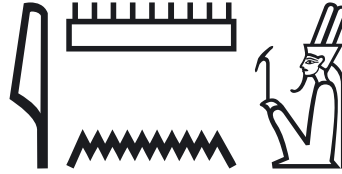
seconds. Even non-specialists could find the signs they were looking for within a few moments.”¹³ While several signs could have the same Gottstein code, this posed little difficulty: “Experience shows that the additional effort only takes a second.”¹⁴ This would seem to be a perfect solution to the problem, and indeed, it forms the basis of various electronic sign recognition systems like CuneiPainter.¹⁵

However, cuneiform signs differ from chemicals in a few crucial respects. For the most part, adding or removing an atom from a molecule creates a different chemical completely: H_2O (water) behaves very differently from H_2O_2 (hydrogen peroxide). But ancient scribes were less consistent. The sign U_2 ‘plant’ is typically drawn with three vertical strokes, as shown in Figure 4, but Rüster and Neu [14, p. 185] report variations with as few as two and as many as seven.

Gottstein’s solution is to list as many variants as possible, with entries for U_2 under a2b2, a3b2, and so on. More crucially, though, it’s very common for a sign to be unclear or damaged. This is something Gottstein’s system fundamentally cannot handle – if damage obliterated the only diagonal wedge in a sign, for example, its category, designation, and code would all be wrong.

A different approach was taken by Homburg [21]. In the study of Egyptian hieroglyphs, the *Manuel de Codage* [23] is a well-established standard for representing the layout of a text; it’s intended to specify where each sign should be placed relative to the signs around it, making it easy to encode complicated arrangements of glyphs. For example, the MdC code in Figure 5 indicates that the game board (‘mn’) is on top of the water (‘n’) and to the right of the reed leaf (‘i’).¹⁶

Homburg’s [21] system, “PaleoCodage”, aims to encode cuneiform signs in the same way: by specifying the position of each stroke relative to its surroundings. The *Manuel*



i-mn:n-C12

Figure 5: The name of the god Amon, encoded in the MdC system.

defines only three operators,¹⁷ which is sufficient for most hieroglyphic texts. But as the name suggests, PaleoCodage is intended for paleography – for analyzing the fine details of handwriting and ductus that go beyond telling one sign from another. For this purpose, knowing that one stroke is to the right of another isn’t enough. The exact distance there could be crucially important!

As a result, Homburg [21, pp. ii133–ii135] extends Gottstein’s four stroke types with the Winkelhaken, two types of reversed diagonal wedges, and two types of “seal wedges” used in archaic number signs, all of which can be modified in three different ways. These can be combined via a staggering array of operators, allowing a paleographer to encode the size, angle, and position of each stroke with perfect accuracy. Three different “to the right of” operators encode subtle differences of spacing, or can be combined for further detail – and if this still proves insufficient, a “factor operator” can be used to adjust them by as little as one percent.

When discussing specific variants of signs, as Homburg [21, p. ii138] puts it, “a very fine-granular modeling is often needed to depict the changes distinguishing the new sign variant from the other standard sign form”. PaleoCodage is thus tuned to specify enough detail to distinguish one scribe’s particular handwriting from another. But this abundance of detail poses a difficulty for our purposes. Homburg [21, p. ii140] proposes a way to look up similar characters by comparing their PaleoCodage encodings – but by these string similarity metrics, a small horizontal stroke (sb) is no more similar to a large horizontal (B) than it is to a small vertical (sa).

In order for PaleoCodage to express this level of precision, its operators don’t quite encode *relationships* (as in the MdC system) between the strokes. Instead, each operator

¹³ Gottstein [16, p. 131]: *jedes nach dem “Gottstein-System” gelistete Zeichen in Sekundenschnelle lokalisiert werden konnte. Sogar Fachfremde haben gesuchte Zeichen in wenigen Augenblicken gefunden.*

¹⁴ Gottstein [16, p. 131]: *Der Mehraufwand beträgt erfahrungsgemäß nur eine Sekunde.*

¹⁵ <https://situx.github.io/CuneiPainter/>.

¹⁶ While it’s not relevant here, the signs within the arrangement are named either by phonetic value or Gardiner code. See Section 2.4.

¹⁷ – separates groups of signs, : stacks signs vertically within a group, and * juxtaposes signs horizontally within a group. Most implementations add a fourth operator, &, for special cases (“ligatures”) that don’t follow any general pattern.

represents a change in state for the parsing automaton [21, p. ii140]. This means that the operators between particular wedges don't necessarily have anything to do with those wedges themselves – two vertical strokes next to each other might be encoded as *a-a* or *a_a*, but the difference has nothing to do with those wedges themselves. Instead, it indicates whether any horizontal strokes crossing those wedges should cross both, or only one! In short, the same features that make PaleoCodage useful for paleography also make it generally unsuitable for a search system.

2.3 Machine learning

In recent years, the problem of cuneiform transcription has attracted more attention from computer scientists, and machine learning algorithms have been applied to a variety of different aspects. Machine learning has been extremely effective at recognizing Han logograms, for example, even with only “off-the-shelf” algorithms [24]. It would seem reasonable to apply these methods to cuneiform in the same way.

Unfortunately, the most effective algorithms for recognizing Han characters (as described by Liu et al. [25] and Liu and Zhou [26], among others) tend to rely on specific details of how those characters are written, such as stroke trajectory. In theory, the concepts behind these algorithms could be adapted to cuneiform wedges. At present, however, the cutting-edge algorithms used in Han lookup tools (such as Pleco) cannot be easily applied to cuneiform. To make them work, some fundamental aspects would need to be redesigned from the ground up.

Unicode attempts to provide a codepoint for each cuneiform sign,¹⁸ and most attempts at cuneiform machine learning, such as Doostmohammadi and Nassajian [27] and Gordin et al. [28], take Unicode-encoded text as their starting point. While remarkable progress has been made in tasks like identifying the language of a text and choosing appropriate readings for each sign, these projects require the signs to have already been transcribed and identified – which can often be the most difficult and time-consuming part. Research in this direction won't help with the actual lookup and identification of logograms.

Other projects, like Dencker et al. [29], attempt to go all the way from tablet photographs to a full transliteration. These projects have also achieved remarkable successes from relatively little training material. But they often suffer

from trying to do too much at once – the system has to learn the entire model at the same time, all the way from recognizing shadows in a photograph to picking out the right reading for a sign. This means that what the system learns about recognizing wedges in Neo-Assyrian cuneiform, for example, can't be easily generalized to another dialect like Old Akkadian, even though the fundamental principles are the same. This becomes a serious issue when there's very limited data for a particular dialect.

One remarkable outlier is Kriege et al. [30], which builds on earlier work by Fisseler et al. [31]. Fisseler et al. aimed to solve a smaller problem: converting scans of tablets into readable autographs. They were remarkably successful at this, and Kriege et al. used the output of that model as the input to theirs, looking at relationships between wedges rather than raw images. By applying graph-based neural networks from Fey et al. [32] to this input, they were able to identify signs with remarkable accuracy.

However, their initial experiment used a very limited selection of signs written by modern scholars, with manual annotation for which wedges belonged to which signs. While their results are still extremely promising, and hint that specialized stroke-detection algorithms along the lines of Liu et al. [25] and Liu and Zhou [26] could be developed for cuneiform, it remains to be seen how well they will generalize to actual ancient tablets with hundreds of distinct signs and no clear marking of sign boundaries.¹⁹

Based on this, it seems that there is currently no effective machine-learning solution to the problem of sign recognition. But note also that one of the most pressing difficulties in applying machine learning to cuneiform seems to be the lack of a good intermediate representation: a concise, useful way of indicating the wedges that make up a sign, which image-recognition models could convert scans and photographs into, and other models could separately convert to readings.²⁰ This would effectively break the problem in half, like how modern machine learning models generally handle OCR (recognition of the graphemes making up written text) separately from the interpretation

¹⁸ Though it has its flaws: Borger [15, p. 2] is quite scathing in his criticism of the implementation. Regardless, it's the most widely-adopted standard for electronic cuneiform.

¹⁹ A notable result in this direction is Stötzner et al. [33], who attempt to automatically partition tablets into signs and recognize the location and direction of each wedge. An experiment in translating the output of their model to the encoding used in this paper is currently in progress.

²⁰ Snyder [34] hoped that Unicode would form this intermediate representation, and as Doostmohammadi and Nassajian [27] and Gordin et al. [28] showed, this is very useful for separating out language detection and sign interpretation. But choosing the right sign is significantly more context-sensitive than Snyder expected, and it's not clear what advantages Unicode encoding has over Romanized sign names or index numbers – *U+1227A* is no less opaque than *pa* or *HZL 174* as a representation of 𐎶.

of that text. The new encoding system proposed in this paper may prove useful for this purpose as well.

2.4 Other logographies

For many languages, ancient and modern, learning the script is only a minor obstacle. Students of Greek generally have no issue learning its alphabet, for instance, and have no need for transliterations. Why is cuneiform any different?

The difference lies in the sheer number of signs. The Greek alphabet uses only 24 letters, while Rüster and Neu [14] have documented 375 signs for Hittite, plus additional variations. 21 of those variations have distinct forms and meanings, raising the total to 396. For Akkadian, Labat's [6] index numbers go up to 567, while Borger's [5] reach 905.

Of the 396 signs used in Hittite, 241 of those are never used phonetically for Hittite words – they're exclusively used for logograms, punctuation, or foreign words.²¹ In some genres and eras,²² it's not uncommon for 85 % of the signs in a text to be logograms [2, p. 53], and some scribes seem to have invented new unique logograms for their own particular uses [7, p. 330]. So while students will quickly master the common phonetic signs, even experienced cuneiform scholars are sure to come across logograms they have never memorized, or perhaps even seen before. This is where a lookup system is essential.

Cuneiform, though, is not the only writing system to require hundreds of logograms. Other scholars have run into this same problem when analyzing Han logograms and Egyptian hieroglyphs, among others, and have needed effective lookup methods long before the advent of machine learning.

For Egyptian hieroglyphs, the standard way to look up a sign originates with Gardiner [35].²³ He created a catalogue of hieroglyphs based on the objects they represent, so that, for example, all signs depicting birds can be found under "G", and all signs depicting ships and parts of ships can be found under "P". This means that a student should only have to look through a few dozen signs to find the one they need, rather than several hundred (Figure 6).

But Gardiner's system depends on students being able to tell easily what a sign depicts, which is sometimes non-trivial (it's not at all obvious to a modern-day student

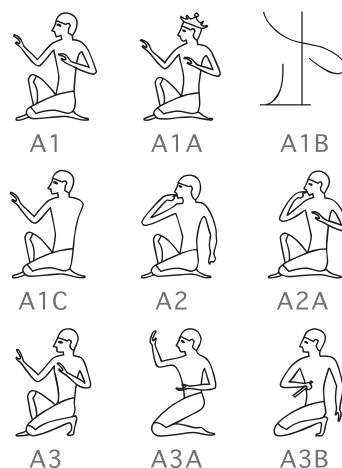


Figure 6: A handful of Egyptian hieroglyphs classified with Gardiner's system. Diagram adapted from work by Gérard Ducher, CC-BY-SA, https://commons.wikimedia.org/wiki/File:Hi%C3%A9roglyphes_-_A_-_Hommes.pdf.

that a speckled circle represents a threshing floor while a lined circle represents an animal placenta²⁴) and sometimes impossible (such as in hieratic writing, a much more stylized variation of hieroglyphic Egyptian). Certain signs could also reasonably fall into multiple categories, increasing the difficulty for the student. Would a sign of a vulture god and a cobra god together be found under G for "birds", I for "reptiles and amphibians", or C for "deities"?²⁵ Cuneiform signs are much less representational than hieroglyphic ones, so this system is infeasible for our purposes.

For the several thousand Han logograms, quite a lot of different lookup systems have been proposed over the years. The most famous and popular are radical-based systems [37]. Most Han logograms contain recognizable smaller parts, known as "radicals". An index of characters by their most distinctive radical, or in newer electronic dictionaries by *all* radicals they contain, can then narrow down the search space tremendously (Figure 7) [37]. Some go a step further, annotating the radicals for their exact positions within the sign, though the reliance on absolute positioning becomes a problem when signs may be obscured or damaged.

To some extent, this sort of system has already been used for cuneiform. The naming rules for signs, as discussed in Tinney [38] and Robson [39], can indicate when a sign is composed of recognizable pieces: the logogram NAG

²¹ 220 of those 241 are used as logograms, one (the "Glossenkeil") only as punctuation, and the remaining 20 for transcribing foreign words.

²² Of Akkadian, at least, though Hittite oracle reports also tend to consist mainly of logograms.

²³ First published in complete form in Gardiner [36], but widely popularized in his 1957 grammar.

²⁴ Or perhaps a woven mat, based on the color it's painted in certain inscriptions – experts disagree. Given how much the paint colors vary, it's likely that ancient scribes did too!

²⁵ The answer is "birds": sign G16, "the Two Ladies".

旭：日 九
宛：夕 卩 宀
謂：月 言 田
韻：音 貝 口 日 立

Figure 7: A selection of Han characters and the radicals they contain, taken from Breen [37, p. 5]. The character can be looked up by one or more of these radicals. (From the top: ‘rising sun’, ‘address’, ‘origin’, and ‘poetic meter’.)

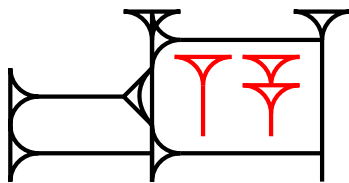


Figure 8: KA (black) enclosing A (red).

‘drink’ (Figure 8) is named KA × A (“KA ‘mouth’ enclosing A ‘water’”), while the logogram *UG* ‘tiger(?)’ is officially named PIRIG&UT (“PIRIG ‘lion’ on top of *ut*”).²⁶ However, most cuneiform signs don’t contain meaningful sub-units larger than single wedges, and signs that were once made of clearly distinct radicals may cease to be so over time: the logogram *MEŠ* ‘[plural]’ was once a transparent compound ME + EŠ (“*me* up against *eš*”), but became a single indivisible unit MEŠ in later eras.

A related system indexes characters by how many strokes they contain; the KANJIDIC database that underlies popular tools like Jisho, for example, includes this type of indexing [40]. However, as shown in Figure 4, the number of wedges in a cuneiform sign is much less consistent than the number of strokes in a Han logogram. Scribes would very frequently leave a wedge off, or include an additional one by mistake. This makes stroke-number systems generally unsuitable for cuneiform.

Another popular system for Han logograms is the “four corners method” [41], [42], which involves dividing all visual elements into ten broad categories, then listing which category each corner of the sign (and sometimes additionally the center) falls under. This gives each glyph a four- or

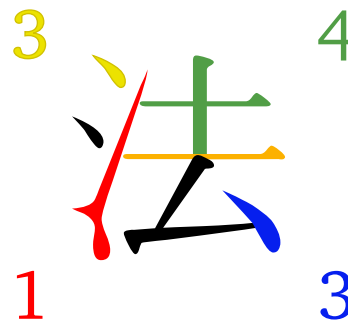


Figure 9: An example of four-corners classification: this sign would be indexed as 3413 in a four-digit system, or 3413-1 in a five-digit system. Diagram by Oona Räisänen, https://commons.wikimedia.org/wiki/File:Four-corner_method.svg.

five-digit number that can be used as an index, as seen in Figure 9. A system broadly similar to this (using the leftmost edge) is already standard in cuneiform sign lists, as discussed above. It can be somewhat helpful; however, it still often leaves hundreds of glyphs to search through, damage to signs is common, and cuneiform scribes were often less careful than Han writers in keeping their stroke types distinct. When the leftmost stroke is used as the main index, uncertainty about that particular wedge – whether caused by scribal carelessness, damage to the tablet, or simply bad lighting in a photograph – can and does render the system unusable.

For Xixia/Tangut, an extinct logography possibly related to Han characters, a different sort of index was proposed by Nishida in 1966.²⁷ Nishida’s system divides all Xixia characters into 319 radicals, then indexes characters based on their structure – three radicals next to each other are structure A3, for example, while two radicals on top of two other radicals are structure M1 (Figure 10). Unfortunately, as discussed above, the dependence on radicals makes this system generally unsuitable for cuneiform.

A related proposal was incorporated into Unicode in 1999, dubbed the “Ideographic Description System” and meant to apply to Han, Xixia, and other similar writing systems; the intent was to allow fonts to synthesize obscure characters that may not have dedicated codepoints. The IDS is similar to Nishida’s index, enumerating a certain number of possible structures that radicals can fit into. But it goes one step further, allowing these structures to be nested recursively, as shown in Figure 11 [44, pp. 689–692].

²⁶ The *ut* here hints at the pronunciation, distinguishing it from, say, PIRIG&ZA for AZ ‘bear’. Compare the use of phonetic complements in Han logograms.

²⁷ The original source is Nishida [43], but as the present author does not speak Japanese, this relies on the synopsis in Downes [42, pp. 12–13].

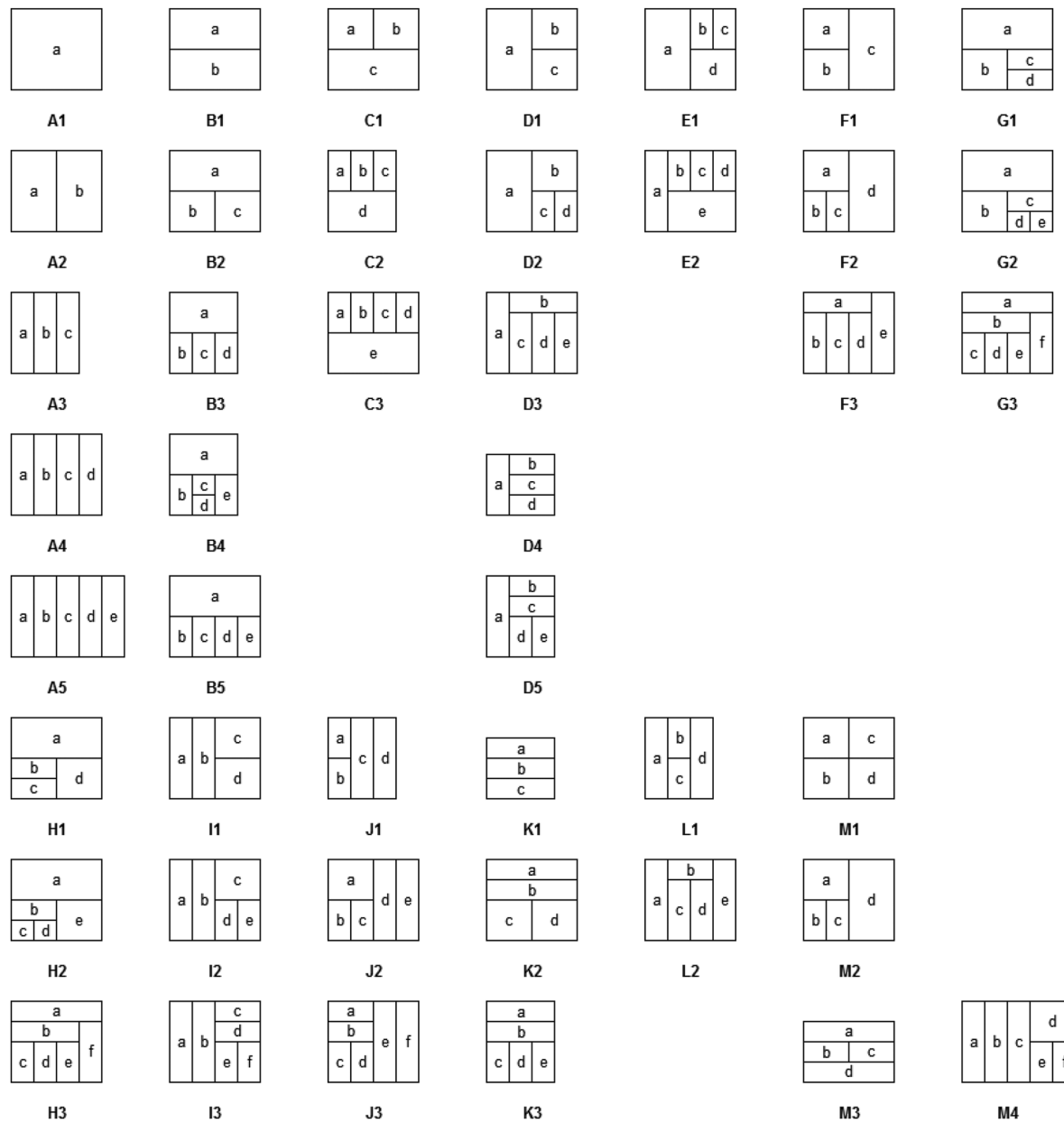


Figure 10: The structures used in Nishida's index, from Nishida [43, p. 246], reproduced in Downes [42, p. 13].



Figure 11: An example of the IDS, from the Unicode consortium [44, p. 691].

While this proposal has potential, the Unicode consortium warns against using it for anything beyond describing unencoded variants, and thus little time and effort has been devoted to it. At the time of writing, no software has

been found that actually uses or even supports it. A similar system was proposed for cuneiform by Snyder [45], but was not accepted by the Unicode consortium.²⁸ Wong, Yiu, and Ng [46] lay out another idea in the same vein, dubbed “HanGlyph”, which uses 41 basic strokes and 12 operations to potentially describe *any* Han character (even one not composed of recognizable radicals). But this proposal is also

²⁸ See <https://unicode.org/mail-arch/unicode-ml/y2004-m02/0012.html> for some discussion of this.

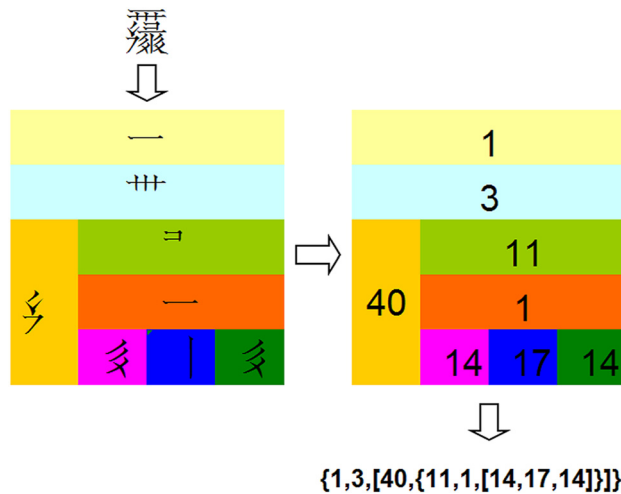


Figure 12: An example of Downes's recursive index for Xixia, reproduced from Downes [42, p. 15].

mostly theoretical, as no actual implementation seems to be available, or any evaluation of its usability.

Another recursive description system for both Xixia and Han was proposed by Downes [42] and elaborated on in Downes [47]. Like the Unicode IDS, Downes' system tries to describe the relationships between radicals in a recursive way. This system uses only three relationships, compared to the IDS's twelve: "stacked horizontally", "stacked vertically", and "enclosed by" (Figure 12). These prove sufficient to describe virtually all Han and Xixia characters; the Xixia characters in fact can be described with only two relationships, since they never enclose one radical with another.²⁹

This system seems the most promising for our purposes.³⁰ Rather than Downes's 176 radicals for Xixia and 420 for Han, we can reduce characters all the way down to their most basic strokes.³¹ As mentioned in Section 2.2, cuneiform across all times, places, and languages is generally analyzed as having five types of wedges – far fewer than Wong, Yiu, and Ng's 41. Empirically, only three types of compositions have proven sufficient to encode 99 % of the signs and variants in Rüster and Neu [14]. Horizontal and vertical stacking are implemented as in Downes [42]; the third form of composition is *intersection* or *superposition*, since this system goes down to the level of individual cuneiform wedges,

while Downes generally calls any intersecting strokes a new radical.

3 Recursive encoding

This new proposal, inspired by Downes [42], is to represent a cuneiform sign as a tree. The leaves of this tree are the five basic types of wedges, and the branches are the three basic ways of combining them ("compositions"): stacked horizontally, stacked vertically, or intersecting (Figure 13). The result is termed the *kadaru encoding*, from Akkadian *kadārum* 'divide up with boundaries'. Informal experiments suggest that Downes's [42] bracket notation ($\{[vv][vv]\}$) is easier for students to read than Snyder's [45] infix notation with precedence ($v+v&v+v$),³² so this forms the foundation of the syntax: wedges are indicated by lowercase letters,³³ and compositions by various types of brackets.

These operations prove sufficient to describe almost all cuneiform signs used in Hittite. Hittite cuneiform is the particular focus of the initial work here for a few reasons: it has a relatively small sign inventory compared to most periods of Akkadian and Sumerian, and these signs were (in the words of Gordin [7, p. 29]) "written with a keen sense of accuracy and symmetry", with very reliable spacing and paragraph breaks.³⁴ At the same time, the Hittite signs defy the straightforward classification of Neo-Assyrian, heightening the need for new indexing tools. And, last but certainly not least, ongoing Hittite classes at the University of Illinois offered a pool of students for testing. However, it's true that the vast majority of cuneiform in existence is not Hittite; Section 3.4 discusses applications to Old Babylonian, Neo-Assyrian, and other styles.

3.1 Aesthetics

The basic *kadaru* system as presented in Table 1 only encodes the five basic types of wedges and their relationships to each other – not any other details of their position or size. This is a notable departure from previous systems

²⁹ Downes [42, pp. 13–14] relates a legend: this ensures that malevolent spirits can always escape and can't get trapped inside the character while it's being written.

³⁰ Though no studies have been found actually putting Downes' work into practice.

³¹ The usual term in cuneiform is "wedges", but since prior work in this direction mostly focuses on Han and Xixia, "strokes" is more common. In this paper, the two will be used interchangeably.

³² Infix notation with precedence is also used by the *Manuel de Codage* itself, though *not* by PaleoCodage, which has too many operators for precedence to be feasible.

³³ Different from the letters used by Gottstein [16] and Homburg [21], for disappointingly mundane reasons – the foundations of this system were built before the author became familiar with Gottstein's work. But the letters used here have proven to be good mnemonics for students, who have occasionally been confused by verticals coming before horizontals in the Gottstein system but after in standard sign lists.

³⁴ For the same reason, Kriege et al. [30] focus on it for their sign recognition experiments.

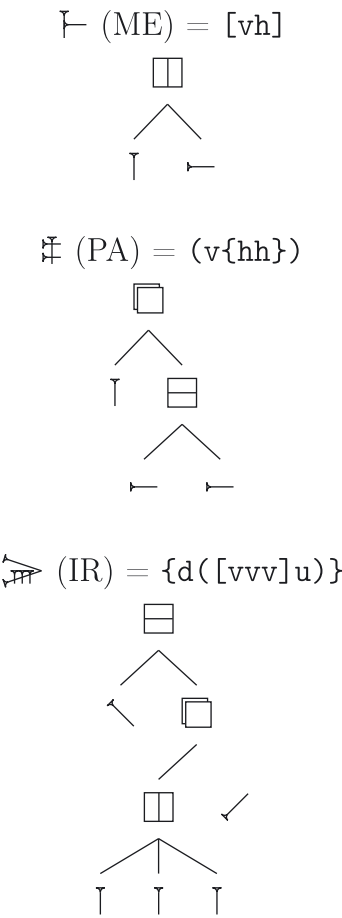


Figure 13: Examples of recursive encoding in the kadamu system.

Table 1: The core of the kadamu encoding.

Name	Example	Syntax
Horizontal		h
Vertical		v
Downward diagonal		d
Upward diagonal		u
Hook		c
Horizontal stack		[]
Vertical stack		{ }
Intersect		()

like the indexing in Ruster and Neu [14], which considers a long horizontal wedge and a short horizontal wedge to be as thoroughly different as a horizontal and a diagonal.³⁵ The four patterns shown in Figure 14 are all separate headings in their sign index.

³⁵ PaleoCodage similarly considers them fundamentally different, but for a more justifiable reason: the difference between them can be vitally important for paleography.



Figure 14: Four consecutive headings from the *Zeichenlexikon*'s sign index.

However, the main effect of this is to index many signs several times over. The sign *pa* \ddagger , for example, is listed separately under the first, third, and fourth of those headings – simply because scribes seldom seemed to notice or care which stroke was longer!

The kadamu encoding is thus designed to ignore the exact size and placement of wedges, focusing only on their relationships to each other. While the parallel wedges in \ddagger may vary in length, they are always horizontal, and one is always placed above the other. This is intended to aid scholars in looking up signs, since they can ignore these details and focus only on the general pattern. During testing, multiple students expressed frustration with this particular aspect of traditional sign lists, saying they were unsure which of the headings from Figure 14 they should be looking at.

It should be noted, however, that the lengths of wedges *can* sometimes carry semantic meaning. The phonograms *ku* and *ma*, for example, are distinguished solely by the lengths of the horizontal strokes, as seen in Figure 15. In the basic system, both would be encoded as $[\{hhh\}v]$. However, cases like this are rare, and even in these instances the stroke lengths can vary significantly. While this is a demonstrable limitation of the system, the practical impact is minimal: a student searching for one of these signs will find both, and decide based on context which is more appropriate.

These five basic wedge types and three basic compositions are sufficient for a search system, but finer control of the details is important for other applications. A typesetting system, for example, *must* be able to distinguish between *ku* and *ma*, and also make its middle stroke shorter than the outer ones (as in Figure 15). And even for searching, scholars will – quite rightly – expect the search results to resemble what they see in the clay, rather than a platonic version that ignores spacing and stroke length.

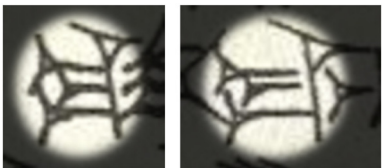


Figure 15: Signs *ku* (left) and *ma* (right), from KBo 23.52.

similar to {vv} than v, two verticals rather than one, but the difference between v2 and {vv} is purely an aesthetic one.

The *tenû* modifier, though, is a fundamental change to the sign: 𐎶^T NINDA ‘bread’ is a different sign from 𐎶 *hi*. Why should this be implemented as a modifier? Wouldn’t it be better to have a “diagonal stacking” composition to capture cases like 𐎶 , encoded perhaps as $\langle \text{ddd} \rangle \text{d}$ – akin to PaleoCodage’s . and , operators?

Notably, though, this “diagonal stacking” has a very restricted distribution. While diagonal strokes are frequently stacked horizontally and vertically (e.g. *ni* 𐎶), horizontal and vertical strokes are never stacked diagonally. And with only one possible exception,³⁷ horizontal and vertical stacks never appear inside or overlap with diagonal ones in Hittite.

Instead, components (or even entire signs) are written “*tenû*”: rotated by 45°. ³⁸ It is likely that this was handled by physically rotating either the hand or the tablet, as opposed to horizontal and vertical wedges, made using different edges of the stylus with the same basic hand position. In other words, these “diagonal stacks” are fundamentally a modification of horizontal and vertical stacks, rather than a separate entity of their own.

For these reasons, diagonal stacking is not considered its own separate composition in the system. Rather, it is treated as either a horizontal or vertical stack, containing only horizontal and vertical strokes, with the *tenû* modifier applied. During normalization, horizontal wedges within a *tenû* composition are replaced with upward diagonals, while vertical wedges become downward diagonals³⁹ – meaning that a search for 𐎶 won’t find 𐎶^T .

3.3 Exceptions

The first true test of this system is its completeness. With the modifiers from Table 2, it can make a satisfactory rendering of all but two signs in the *Hethitisches Zeichenlexikon*. These two exceptions are the logograms GAŠAN ‘Lady’ (a title used in prayers to female deities) and DALĪAMUN₄ ‘Dalhamun’ (an epithet of the Akkadian storm deity Adad).

The latter can be quite reasonably ignored. It consists of the sign NAGA written four times in different directions,

R.s.

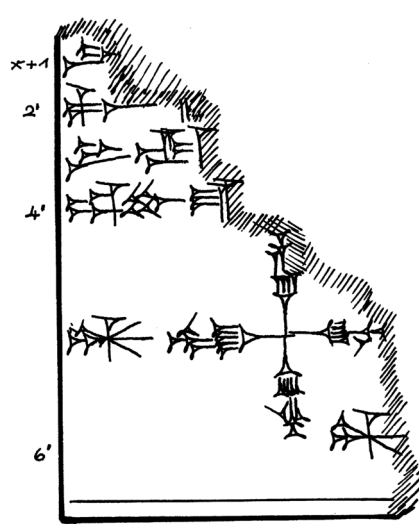


Figure 18: A Kreuzform sign, “ŠIR × 4”, from Torri [48].

converging on a center point to make a cross shape. A similar construction (shown in Figure 18) involves four copies of the logogram ŠIR ‘testicle’ arranged in the same pattern, likely an epithet of the moon-god; this one is a hapax and doesn’t even merit an official name or Unicode codepoint.

These “Kreuzform” signs don’t fit into a standard line of text, and in fact never seem to be used inline on actual tablets in the Hittite era: they appear only in lists of logograms or sketched in colophons [48]. They are not given their own index numbers in Rüster and Neu [14] and are not supported in any common Hittite fonts, so I have no qualms about leaving them unencoded.⁴⁰

The sign GAŠAN poses a larger issue. As presented in the *Zeichenlexikon*, it involves horizontal strokes meeting a diagonal, as shown at the top of Figure 19. This is something the kadamu system currently can’t handle – the best way to do it would be to put downward diagonals inside a *tenû* modifier, which is not allowed.

Curiously, though, the sign as found in tablet autographs tends to look quite different. Several examples can be seen in the middle of Figure 19, and none of them quite match the *Zeichenlexikon*’s form. This makes sense, if diagonal strokes were indeed made by rotating the tablet: having horizontal strokes truly *meet* a diagonal would require significantly more effort than meeting a vertical.

³⁷ See Section 3.3.

³⁸ See Snyder [45] for examples.

³⁹ When talking about entire signs, *tenû* generally means a clockwise rotation. Here it is instead implemented counterclockwise, to avoid needing to reverse the directions of strokes: all diagonal strokes in Hittite point to the right, not to the left.

⁴⁰ DALĪAMUN₄ is considered iconic enough to appear on the cover of the *Zeichenlexikon*, but as a symbol, not as a meaning-bearing logogram.

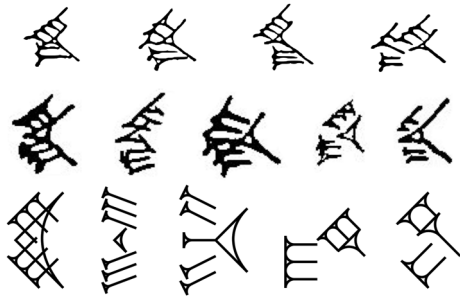


Figure 19: Variants of the sign GAŠAN ‘Lady’. At the top, how it’s presented in the *Zeichenlexikon*; in the middle, how it appears in autographs (KBo 8.110, KBo 24.43, KBo 47.133, KUB 27.1, and KBo 39.159); on the bottom, different ways of representing it in the kadamu encoding.

Historically speaking, this sign started as a *gunû* (‘decorated’) form of the sign *u* ‘ten’, a single Winkelhaken. This can be represented in the kadamu system by superposing *tenû* strokes (bottom left). Various other representations are possible too, as shown across the bottom; while none of them are perfect representations of how it’s shown in the autographs, they can approximate it significantly better than they can the *Zeichenlexikon* form. The question of which source best represents the actual tablets, and what the Platonic ideal of the Hittite GAŠAN should truly look like, is left for future work.

3.4 Other languages

While the kadamu system is well-suited for encoding Hittite, this is only a small fraction of the cuneiform in existence. Even if Hittite is the focus of the initial experiments, it’s well worth asking if this system can be generalized to other languages and eras.

While most Akkadian sign lists give between 500 and 1,000 signs, Huehnergard’s [1] textbook provides a short reference list for its exercises, with only 196 signs. Better yet, these 196 signs are provided in three different styles – Old Babylonian lapidary (for carving into stone), Old Babylonian cursive (for impressing into clay), and Neo-Assyrian, as shown in Figure 20. While far from complete, this can be taken as a more-or-less representative sample of the script.

The vast majority of the Old Babylonian cursive signs can be encoded in the kadamu system without problems. Several required one adjustment to the system: unlike Hittite, Old Babylonian signs sometimes make use of upward vertical wedges, as in the sign $\text{†} \text{nu}$. But these are far less common than the five basic types from Table 1 – rare enough that they don’t receive their own headings when categorizing signs, or their own designations in Gottstein’s

system or PaleoCodage. Instead, we represent them with a new modifier, ? , which inverts the direction of a wedge.⁴¹ † can then be represented as $(\text{hv}?)$. Additional modifiers like this are the easiest way to extend the system, without modifying the basic strokes and compositions; the triple-headed wedges in Old Assyrian $\text{†} \text{BAL}$ ‘libate’ could be handled with a new 3 modifier, and the remaining wedge types found in PaleoCodage but not the kadamu system (‘seal wedges’ used in archaic accounting, written with a round stylus instead of a rectangular one) might be implemented the same way.⁴²

With this modifier added, only two Old Babylonian cursive signs posed an issue: one (sign 152, *uk*) because the illustration in Huehnergard is too smudged to make out the individual wedges, the other (sign 136, NA_2 ‘bed’) because of the same issue as Hittite GAŠAN, above. A full text in this style is shown in Figure 21. In Neo-Assyrian, many more signs had the ‘GAŠAN problem’, suggesting that it really does need a solution. However, the lapidary signs ran into issues left and right, with a worrying proportion of them unable to be rendered properly (Figure 20), or in some cases even encoded at all.

As a further experiment, short texts were encoded and rendered in Ugaritic (Figure 22) and Old Persian – two types of cuneiform historically unrelated to the tradition of Sumerian, Akkadian, and Hittite, but designed to be written in the same way with a reed stylus on clay.⁴³ These similarly posed no issue. This suggests that the kadamu encoding is well-suited for the medium of a reed stylus on clay in general, rather than any particular features of Hittite cuneiform. The medium itself puts certain limits on the precision of the strokes, preventing the intricate arrangements of lines and hatches found in the Old Babylonian lapidary style.

3.5 Comparison

The benefits over the Gottstein system are clear. The kadamu encoding includes the same information about the wedge types, plus additional information about how they relate; a database of kadamu-encoded signs can be trivially converted to Gottstein codes, or allow searching by Gottstein code as an alternative.

The benefits over PaleoCodage are less obvious. Both of them fundamentally try to represent the same thing: the

⁴¹ Akin to ! in PaleoCodage; ! was already used in the kadamu system for highlighting.

⁴² Perhaps as vertical, horizontal, and hook strokes with a hypothetical new ‘round stylus’ modifier, like Tinney’s [38] @C. Due to the rarity of these wedges, they have not yet been implemented.

⁴³ The majority of surviving Old Persian inscriptions are on stone rather than clay, but texts written on clay have also been found.


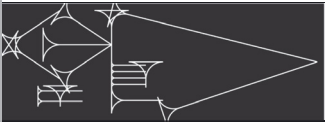

170		ANŠE = <i>imērum</i>
Ident	170	
Tags	bad, lapidary	cursive
Sign		
In Akkadian words		
Determinative		
Sumerogram	ANŠE “imērum”	
Ligatures		
Code	X[{{[ud]h' [du] [θ(v{hh})Eθ]M}v{d[{{[hhh]v}h}uEE]}}EEE]	W[{{[vvv]v}TEh[d[{{[hh]u}h}E]
Notes		
Composition		

Figure 20: A sign from Huehnergard’s [1] sign list, in three different styles, encoded in the kadamu system.

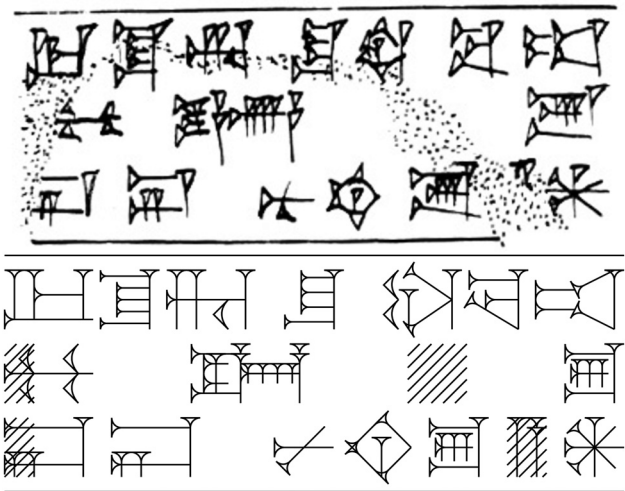


Figure 21: A Sumerian proverb in Old Babylonian cuneiform (CDLI P346305): “The dog knows ‘take it’. It does not know ‘drop it’.” Autograph taken from <https://cdli.mpiwg-berlin.mpg.de/artifacts/346305>. Most sign forms are taken from the Old Babylonian database in Section 3.4; missing and non-matching ones were encoded specifically for this tablet.


relative positions of strokes within a sign. What is there to gain by representing this with a tree, rather than state changes in a finite automaton?

The key is that, in the kadamu encoding, the relationships between the wedges themselves are fundamental. In PaleoCodage, the operator between two strokes doesn’t just express the relationship between them – it conveys how they relate to the sign as a whole, and the other strokes in it. As mentioned in Section 2.2, the difference between a-a and a_a (a vertical wedge next to a vertical wedge) doesn’t necessarily represent anything about those two strokes.



Figure 22: A Ugaritic abecedary tablet (RS 12.063/KTU 5.6). Tablet photo originally from Yon [49, 124, Figure 2a], reproduced at <https://mnamon.sns.it/index.php?page=Esempi&id=30&lang=en>. The 3 (triple-headed) and ? (inverted) modifiers are required here.

Instead, the primary difference is how they interact with other types of strokes overlapping them.

In the kadamu system, on the other hand, the relationship between *any* two strokes can be determined by finding their last common ancestor in the tree. In the sign  ‘wine’, the hook is to the left of the diagonal – and this is reflected in the encoding (Figure 23). The last common ancestor of those two (shown by the blue lines) is a horizontal stack, and the hook precedes the diagonal.

The encompassing algorithm described in Section 4.3 then ensures that a search for [cd] will find this sign, no matter how far apart the hook and diagonal are in the encoded sign. As a result, a student using the kadamu encoding can search for *any part of a sign* – no matter whether other parts of the sign are missing, damaged, badly transcribed, or just difficult to encode. This is something that

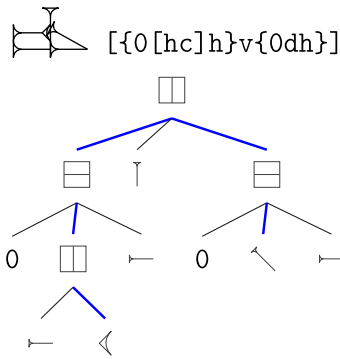


Figure 23: The relationships between certain strokes in the sign GEŠTIN ‘wine’, highlighted in blue.

Table 3: Three signs^a with identical Gottstein codes, but distinct PaleoCodage and kadamu encoding.

Sign	Gottstein	PaleoCodage	Kadamu
𐎶 MAŠ	a1b1	:b-a	{vh}
𐎶 BAR	a1b1	;b-a	{vh}
𐎶 ME	a1b1	a-:b	[vh]

^aThis demonstration is taken from Homburg [21, p. ii131]; in Hittite, the signs MAŠ and BAR have merged.

neither the Gottstein nor PaleoCodage systems currently supports. The kadamu system is designed around the relationships between strokes first and foremost, and as a result is far more resilient to damage (Figure 30) or just simple obscurity than its predecessors.

A comparison of how a few different signs would be encoded in these three systems can be found in Table 3, and a fuller demonstration in Table 4.

4 Algorithms

4.1 Rendering

Now that we have a way of encoding signs into trees, we can write algorithms to manipulate them in various ways. For example, we can render a tree back into an image of a sign. Unlike in PaleoCodage, the encoding is not expected to specify the exact size, position, and angle of each wedge – instead, the rendering algorithm applies various aesthetic principles to arrange the strokes in a clear and pleasing way. This means that, for example, diagonal strokes in the kadamu system only encode whether they’re oriented upward or downward; the exact angle is chosen by the rendering algorithm based on aesthetic constraints.

The main goal of this algorithm is to assign a rectangle of space to each node in the tree. This space is represented as an axis-aligned bounding box, specifying the height, width, and coordinates of the top left corner, but not rotation. Then, each leaf node can be drawn⁴⁴ into the appropriate space.

Since the input to the algorithm is a tree, this is best handled recursively. The root of the tree determines the overall size of the sign: one unit in height, with the width determined by a special character at the beginning of the encoding. Then each node is assigned space by its parent, and assigns an appropriate amount of space to each of its children:


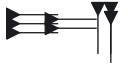






- Intersections assign the full space to each of their children, causing them to overlap.
- Margin adjustments remove an appropriate amount of length (one-fifth the width, one-fifth the height, or one-tenth of a unit, whichever is smallest) from each side of the rectangle, then allot the resulting space to their child.
- *Tenû* adjustments assign the space from (0,0) to $(d\sqrt{2}/2, d\sqrt{2}/2)$ to their child, where d is the minimum of the allotted width and height. Since the axis-aligned bounding boxes can’t represent rotation, these adjustments are handled by transforming the coordinate system, rotating it by 45° and placing the origin at the midpoint of the left side of the *tenû* element.
- All other component modifiers assign their entire space to their child without alteration.

The complicated case is the horizontal and vertical stacks. These require some additional definitions first:

- Horizontal wedges can expand horizontally, but not vertically.
- Vertical wedges can expand vertically, but not horizontally.
- Diagonal wedges and voids can expand in both directions.
- Winkelhaken are always twice as tall as they are wide, and can expand in whichever direction is the limiting factor in that ratio.
- Compositions can expand if any of their children can.
- Allow adjustments can always expand, and restrict adjustments can never expand, regardless of their children.

⁴⁴ As a line, a triangle, the curved wedge shape used in this paper, or something else. This part is currently handled straightforwardly through SVG drawing commands.

Table 4: Various phonetic signs encoded in each system, for comparison.^a

Sign	Name	Gottstein	Kadaru	PaleoCodage	PaleoCodage render
	ya	a3b5	L[{hhh}{hh}vv2]	b:b_b_/b:b_a-a:a	
	bi	b2c2	{[hc'] [hc']}	b:b_w:w	
	ir	a3c2	L{d(u[vvv])}	<;C>>D-::sa-::sa-::sa	
	lu	a3b3	[v{h(hv)Mh}v]	:a-b;b-:::sb:::sa-a	

^aGottstein code, PaleoCodage encoding, and PaleoCodage rendering (the rightmost column) are taken from the demonstration at <https://situx.github.io/PaleoCodage/>. Kadaru encoding and rendering (the leftmost column) are adapted to match this version of the sign if it differs from the Hittite one.

- Vertical wedges allow kerning up to half their width from the left and right, and not at all from the top and bottom.
- Horizontal wedges allow kerning up to half their height from the top and bottom, and not at all from the left and right.
- Kern adjustments allow kerning up to half their width/height from every direction.
- Horizontal stacks allow kerning on the left and right based on their leftmost and rightmost children, and on the top and bottom based on the minimum allowed by any of their children.
- Vertical stacks allow kerning on the top and bottom based on their topmost and bottommost children, and on the left and right based on the minimum allowed by any of their children.
- *Tenû* adjustments take the endpoints of all descendant non-void strokes, convert them to the outer coordinate system, and allow kerning to the leftmost, rightmost, topmost, and bottommost of those endpoints.
- Other nodes use the minimum allowed by any of their children on all sides.
- The size factor of an expand adjustment is one plus the size factor of its child.
- The size factor of an allow adjustment is the size factor of its child, allowing it to be wrapped around an expand adjustment.
- The size factor of a cursor (in the interface) is zero.
- The size factor of any other node is one.

Now, to render a horizontal stack:

- As a first pass, divide the width between all the element's children, in proportion to their size factors.
- Then, repeat the following:

- Determine whether any of the element's children can expand horizontally. If not, the process is complete.
- For each pair of adjacent elements, determine if one allows kerning in the appropriate direction, and the other does not.⁴⁵ If so, move them together by the maximum allowed amount.
- Add up how much space has been reclaimed via kerning. If there is none, the process is complete.
- Distribute the reclaimed space between the children that can expand, in proportion to their size factors.

Vertical stacks are the same, substituting “vertical” for “horizontal” and “height” for “width”.

This algorithm is simple – requiring much less detail than the paleologically-focused PaleoCodage – but the end results are remarkably effective. With only two exceptions (see Section 3.3), it has been able to create aesthetically pleasing renditions of every sign and common variant listed in Rüster and Neu [14], across all eras of Hittite. Figure 24 compares the results against the detail-oriented PaleoCodage renderer, and Figure 25 compares this algorithm's rendering of several logograms against a traditional hand-drawn cuneiform font.

While the original goal was only to encode signs for searching purposes, this renderer compares favorably against traditional fonts. It could be used to quickly render unusual glyphs and glyph variants for discussion, much

⁴⁵ If they both allow it, no kerning is applied: two vertical wedges next to each other should not be kerned together, for example.



Figure 24: The phonetic sign *ya*, rendered in PaleoCodage (left) and the kadamu system (right). PaleoCodage's renderer is optimized for precision in detail, as necessary for paleography, while the kadamu renderer is optimized for overall aesthetics and readability at the expense of precision.

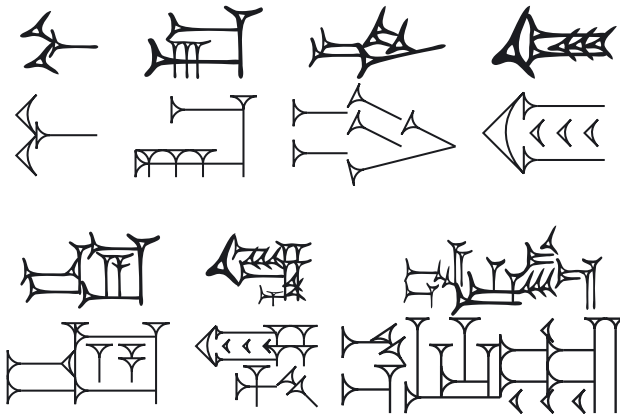


Figure 25: Logograms of varying complexity rendered in a traditional hand-drawn font (lines 1 and 3) and in our new rendering engine (lines 2 and 4). Top row: MUNUS 'woman', IKU 'field', LUGAL 'king', KIŠ 'world'. Bottom row: NAG 'drink', BURU₁₄ 'harvest', UMBIN 'fingernail'. "Ullikummi", created by Sylvie Vanséveren.

like Wong, Yiu, and Ng's [46] proposed HanGlyph synthesizer,⁴⁶ and could potentially render entire documents in a clean, easily-readable style. All the inline glyphs used in this article are rendered in this way. While PaleoCodage can be far more precise in the details – even with aesthetic modifiers, the kadamu system is not designed to capture the paleographic details of a particular scribe's handwriting – the output of the kadamu renderer is meant to be more readable for general usage, equivalent to a typeset edition of a manuscript rather than a facsimile.

Hittite cuneiform tablets in particular have a reputation for being very clean and readable in their layout, even in daily correspondence: words are clearly distinguished, lines of text are more or less justified, paragraphs and sections are marked. Some cuneiform scholars⁴⁷ will even "lovingly" call it "typewriter cuneiform" (*Schreibmaschinen-Keilschrift*). Could the renderer replicate these features,

laying out full tablets on these principles? This would provide, effectively, cuneiform typesetting of a sort seldom attempted before. A prototype can be seen in Figure 26.

4.2 Normalization

One issue with this encoding is that it can be ambiguous. While every recursive code describes exactly one sign, the same sign can sometimes be described by multiple codes. $\overline{\text{ff}}$ *za*, for example, could be described either as $[\{\text{vv}\}\{\text{vv}\}]$ or as $[\text{vv}][\text{vv}]$: a horizontal stack of two vertical stacks, or a vertical stack of two horizontal stacks. This poses a problem for searching. How can a user know which of these two they should search for?

The solution is a recursive algorithm for *normalization*: converting encodings to a form that is unambiguous to compare. The main purpose of this "normalized" or "functional" form is to be used as the input to other algorithms. As a proof of concept, two different "modes" of normalization are implemented; in the standard mode, all five stroke types remain distinct, but in "Gottstein mode", downward diagonals and Winkelhaken are not distinguished, as recommended by Gottstein [16]. This demonstrates that similar adjustments could be made for particular languages or eras as necessary.

In the following algorithms, the word *contains* means "is the parent of", while *indirectly contains* means "is the ancestor of". That is, a node contains its children, and indirectly contains its children, its children's children, and so on.

The normalization algorithm is as follows:

- The normalized form of a double-headed stroke is a stack of two single strokes ($\text{h}_2 \rightarrow [\text{hh}]$) – a vertical stack for vertical and downward diagonal wedges, a horizontal stack for horizontal and upward diagonals.⁴⁸
- If operating in "Gottstein mode", the normalized form of a downward diagonal is a Winkelhaken, or a vertical stack of two Winkelhaken if double-headed. This normalizes away the difference between these two types of strokes for styles of cuneiform where they are not reliably distinguished.
- The normalized form of a void is nothing at all. In other words, voids are discarded.
- The normalized form of any other stroke is that stroke without any modifiers.
- The normalized form of the *tenû* adjustment is the normalized form of its child, with all horizontal

⁴⁶ See Figure 6 in Wong, Yiu, and Ng [46, p. 587] for an example.

⁴⁷ Gordin [7, pp. 27–29] cites Joachim Marzahn.

⁴⁸ Likewise for triple-headed strokes – not part of the system for Hittite, but discussed in Section 3.4 as it pertains to Old Assyrian.

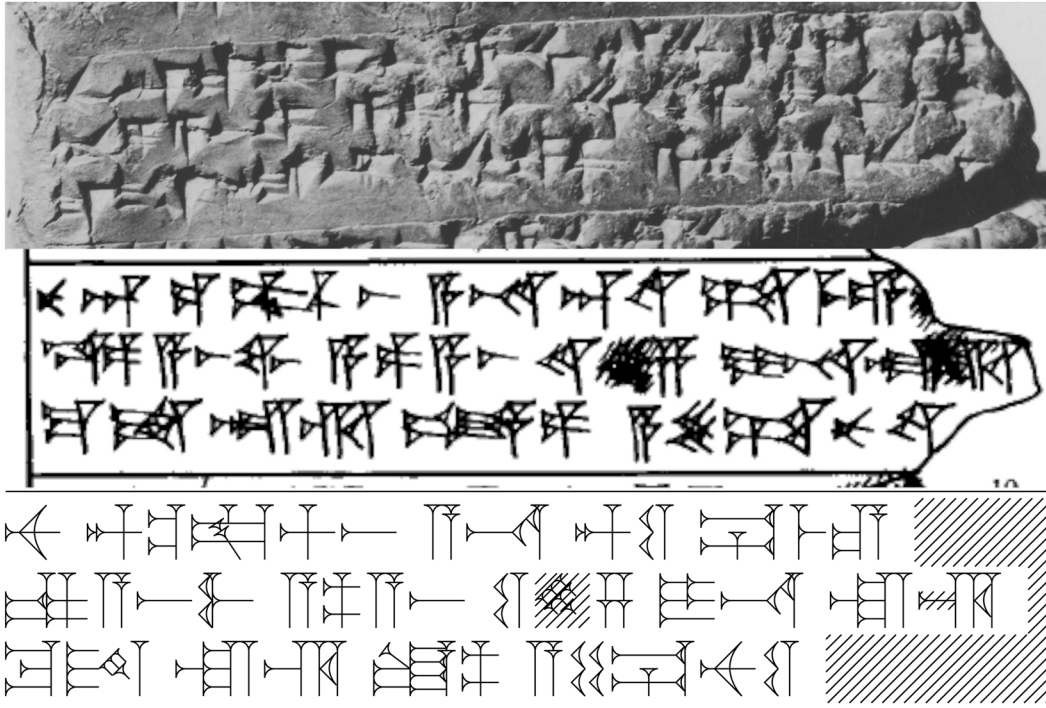


Figure 26: A prototype rendering of three lines from a tablet – in this case, the Hittite Gilgamesh. Autograph taken from https://www.assyrianlanguages.org/hittite/index_en.php?page=textes. Unlike PaleoCodage, the kadaru renderer is optimized for overall readability, sacrificing paleographic detail for this purpose: the scribe of the original tablet left off a vertical stroke in the GIM sign, third on the first line, for example, but the kadaru render uses the standard form.

descendants replaced by upward diagonals, and all vertical descendants replaced by downward diagonals.⁴⁹

- The normalized form of any other component modifier (like E) is the normalized form of its child.
- The normalized form of a composition is based on the normalized forms of all its children. If it's an intersection (where the order of the children doesn't matter), the children are sorted in lexicographic order.⁵⁰ Then, a few special cases are checked:
 - If a composition has only a single child, the normalized form of the composition is the normalized form of the child ($[v] \rightarrow v$).
 - If a composition has no children, the normalized form of that composition is nothing at all.
 - If a composition contains another composition of the same type, the nesting is removed ($[v[vv]v] \rightarrow [vvvv]$).
 - If a vertical stack contains only horizontal stacks, and all those horizontal stacks have the same

number of elements, and the parent of this node is *not* a vertical stack, the normalized form is rearranged into a horizontal stack of vertical stacks. This means that the normalized form of $\overline{\text{ff}}$ is always $[\{vv\}\{vv\}]$, never $\{\{vv\}\{vv\}\}$.

- Conversely, if a horizontal stack contains only vertical stacks, and all those vertical stacks have the same number of elements, and the parent of this node is a vertical stack, the normalized form is rearranged into a vertical stack of horizontal stacks. This reduces the total number of stacks, since a vertical stack inside a vertical stack is removed by the third special case.
- If a vertical stack contains one or more horizontal stacks, and any of those stacks contains a horizontal wedge and a Winkelhaken at the right or left end, remove those Winkelhaken from the ends and rearrange them into vertical stacks of their own. Then combine those vertical stacks together as a horizontal stack: $\{[hc]h\}$ becomes $[\{\}\{hh\}\{c\}]$ (removing one Winkelhaken from the right and none from the left). The empty and singleton stacks are then simplified by the other cases above. This handles the ambiguity of signs

⁴⁹ Internally, this is accomplished by setting a flag that's propagated through the recursive algorithm, using the same mechanism as the mode flag.

⁵⁰ That is, c comes before h comes before v. This is entirely arbitrary, but consistent.

like $u\check{s}$; the difference between ⌋ and ⌋ is clear in the trees, but not at all clear on the actual clay, and this ensures that the normalized form always has the Winkelhaken outside the vertical stack.

- An intersection composition is *mixed* iff it indirectly contains both horizontal and vertical strokes. If a vertical stack contains only horizontal elements and a mixed intersection, or a horizontal stack contains only vertical elements and a mixed intersection, the normalized form puts the stack inside the intersection. This is the strangest normalization rule, and its purpose is to ensure that patterns like $\{hh(hv)\}$, $\{h(\{hh\}v)\}$, and $(\{hhh\}v)$ have the same normalized form, since they are nearly impossible to distinguish on actual clay: it's not always clear how many of the horizontals the vertical was meant to cross.
- If none of these special cases apply, the normalized form of the composition is the same composition of the normalized forms of its children.

The end result is that aesthetic variations, like the lengths of strokes or the size of the gaps between them, will not affect searching or comparisons.

4.3 Encompassing

The most crucial algorithm is termed *encompassing*. For the search system to be effective, a tree must “encompass” any subset of its strokes, as long as the relationships between those strokes are preserved; the strokes should not have to be contiguous, and the relationships to any other strokes not in the subset should not matter. In other words, a sign should “encompass” any part a user might search for, even if other parts of the sign are damaged or difficult to read.

The algorithm given here is significantly more complex than previous approaches to recursive searching. Downes [42], for instance, uses standard substring matching in Excel, and similar algorithms exist to match subtrees.⁵¹ Homburg [21] likewise uses standard metrics of string similarity. However, subtree matching is not sufficient for our purposes. Consider the sign in Figure 27. It's impossible for both the red component and the blue component to be subtrees. But both are perfectly reasonable ways for a user to search for the sign! Something better is needed.

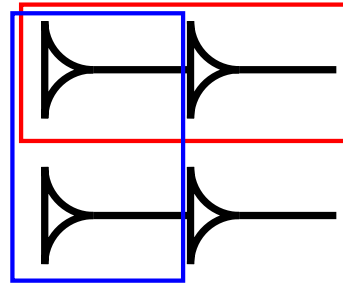


Figure 27: An illustration of why subtree matching is not sufficient for sign identification.

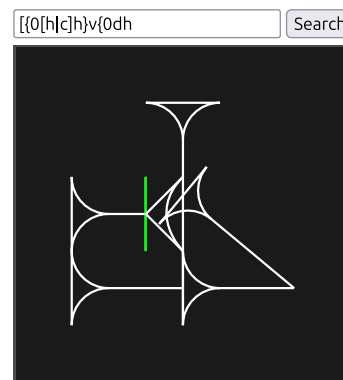


Figure 28: The “canvas” part of the interface, helping users enter kadamu codes by showing the result in real time. The code entered here doesn't need to be complete (note the two brackets opened but never closed), and the position of the cursor is marked with a green bar (between the horizontal and Winkelhaken), indicating where a newly-typed stroke will be inserted. A drop-down at the top changes the rendering style, letting users customize it to their preference.

The encompassing algorithm is the most crucial part of this system. This relation allows users to search for whichever part of a sign is clearest and most readable. They aren't limited to the leftmost part of the sign, or even a complete subtree; *any* visible strokes can be used to narrow the search (Figure 33).

The algorithm for this is, like the others, recursive.

- A stroke encompasses a stroke of the same type.
- Any stroke encompasses a wildcard.
- A composition A encompasses a node B if any child of A encompasses B .
- An intersection A encompasses an intersection B if every child of B is encompassed by some child of A .⁵²
- A stack A encompasses a stack B if:

⁵² Notably, this step does not check that those children of A are distinct. This makes the implementation much simpler, but can lead to the occasional false positives, such as $([vh]c)$ encompassing (vh) . Preventing this would be an easy improvement to the algorithm, at the cost of a significantly higher time complexity.

⁵¹ That is, one particular node of the tree and everything it indirectly contains.

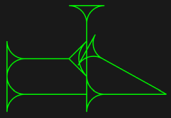
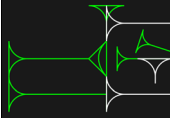
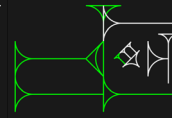
Ident Tags	131	140	144
Sign			
In Hittite words	wi ₅		
In foreign words			
In Akkadian words			
Determinative			
Sumerogram	GEŠTIN (NEŠTIN) “wine”	KIR ₁₄ “nose”	BÚN “thunder”
Code	L[{0[hc]h}v{0dh}]	W[{0[hc]h}v{h{d'hv}Mh}v]	W[{0[hc]h}v]{h[{[vvv]v}TEhvv]Mh}v]W[{0[hc]h}v]{h[{[vvv]v}TE(h[vv])Mh}v]
Notes			
Composition		KA×GAG	KA×IM

Figure 29: The “search” part of the interface, showing all the signs encompassing the code the user created with the canvas. The matching strokes are highlighted in green. Other options allow the user to narrow down signs by name and change the sorting method or normalization mode.



Figure 30: Examples of signs with varying degrees of damage, taken from KBo autographs: KIR₁₆ ‘garden’, TUR ‘small’, NAG ‘drink’, SA₅ ‘red’, and AŠ₃ ‘six’.


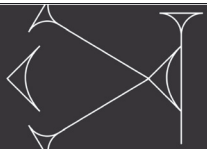



Ident Tags	74				
	old	new	variant	variant, unusual	variant, unusual
Sign					
In Hittite words					
In foreign words	nim				
In Akkadian words	nim, nem, num				
Determinative					
Sumerogram	DÌH “thorn (type of plant)”, ELAM “Elam (place)”, (NIM?)				
Code	L[{du0'}]{cc0'}v]	L[c' {du}c'v]	L[{[cc][hc]0}v]	L[{[ccc][hcc]c}v]	[c{cv}cv]
Notes					
Composition					

Figure 31: The database provides information about both the sign as a whole, such as its various readings, and about its individual forms. In this case, the first form shown is used in Old Script, the second in New Script, and the last is flagged as particularly unusual.

- A and B have the same type (vertical or horizontal), and
- Every child of B is encompassed by some child of A , and
- For any children of B x and y , if x precedes y , then the child of A that encompasses y does not precede the child of A that encompasses x . This ensures that $[vh]$ does not encompass $[hv]$, but

$[hcv]$ does. In other words, it loosely enforces an ordering.

A slight modification of this algorithm can also return a list of the wedges matched in the first bullet point. In the interface, this is used to highlight the matching part of each search result, displaying those wedges in a different color (as in Figures 29, 32 and 33).

4.4 Interface

Finally, some sort of interface is needed for scholars to actually make use of these algorithms. The prototype consists of a *canvas* and a *search engine*, available at <https://dstelzer.pythonanywhere.com/canvas.html> (Figures 28–33).

The canvas is designed to help users input kadamu codes. It has a text box to enter a code, and displays the graphical result next to it, updating in real time. The position of the cursor in the text box is reflected with a green bar in the output, showing where a newly-typed stroke would appear, and any strokes highlighted in the text box are colored green. This is intended to help users develop an intuition for the kadamu system (Figure 28).

The search engine then takes a pattern entered through the canvas and displays a list of signs that encompass that pattern, using the algorithm from Section 4.3 (Figure 29). The database behind this currently contains all signs and major variants from Rüster and Neu [14], and the results can be sorted by Rüster and Neu's index number, sign usage

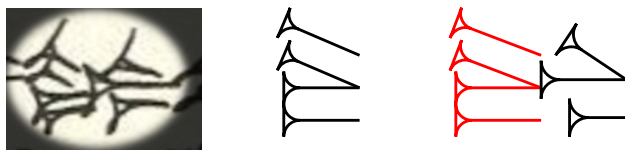


Figure 32: The process of searching for a complete sign: ANŠE ‘donkey’ (left). The user looks for the part of the sign that seems easiest to encode; in this case, that’s the four strokes on the left. Using the interface from Figure 28, they encode this as {ddhh} (middle). The search interface from Figure 29 then shows them a list of all six signs which encompass this component, and the user can identify the correct sign (right) from among them.

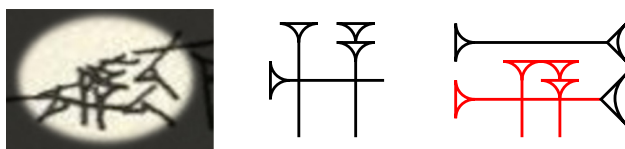


Figure 33: The process of searching for a broken sign: DUG ‘vessel’ (left). Since the left side of this sign is broken, the user encodes the bottom part: the horizontal stroke intersecting an A sign, {h[vv2]} (middle). The program then produces a list of all seven signs encompassing this element, allowing the user to identify the best one (right).

(phonogram, heterogram, logogram, semagram), or number of wedges.⁵³ Since the Hittite script changed significantly over its five centuries of use, individual forms can be “tagged” with notes about their usage (Figure 31); this way, a scholar searching for a damaged sign in a Neo-Hittite tablet can tell if their search results are New Script or Old Script, and common forms or unusual variants.

5 Conclusions

This paper proposes a new method of encoding the shape of a cuneiform sign, as a tree made up of strokes and compositions. As a result, computers can now work with the shape of a sign in a way that wasn’t possible before. At its most basic level, this can be used for rendering, turning this recursive encoding into a picture of the sign.

But rendering is only the tip of the iceberg. Using these tools, scholars can now search for a sign based on any wedges that are visible, regardless of damage or obscurity. Experiments are currently underway to determine if this makes a quantitative difference to the users’ experience, and pilot results are promising.






The kadamu encoding can also serve as an intermediate form for machine learning approaches to cuneiform, allowing problems to be factored into smaller steps than were possible before. I believe these recursive approaches have potential far beyond what’s been tested so far, and may be a significant step forward in computational analysis of cuneiform writing.

5.1 Resources

The code discussed in this article can be found at <https://bitbucket.org/dstelzer/hantatallas>, and can be used at <https://dstelzer.pythonanywhere.com/canvas.html>. A full reference is given in Table 5.

⁵³ Specifically, sign “complexity” is defined as the number of leaves in the normalized form – that is, Gottstein’s “category” number – and users can sort the results by this value. Since different variants might have different numbers of wedges, this value is defined per sign, rather than per form, using the first form listed in Rüster and Neu [14].

Table 5: A complete syntax reference. Canvas size is indicated by the first character; other modifiers and adjustments come after the nodes they modify; commas and whitespace can be used as optional delimiters.

Strokes	h	Horizontal	
	v	Vertical	
	d	Downward diagonal	
	u	Upward diagonal	
	c	Winkelhaken/Hook	
	Ø	Void	An invisible “stroke” that takes up space
	*	Wildcard	Matches any stroke in the encompassing algorithm
		Cursor	Renders as a line or cross, to show where a new stroke will be inserted in the interface; ignored in searching
Compositions	[]	Horizontal stack	Arrange children from left to right
	{ }	Vertical stack	Arrange children from top to bottom
	()	Intersection	Overlay children onto the same space
Stroke modifiers	'	Shorten head	For most strokes, shorten by bringing the head inward; for a Winkelhaken, reduce size slightly; for a void, prevent expanding horizontally
	”	Shorten tail	For most strokes, shorten by bringing the tail inward; for a Winkelhaken, reduce size greatly; for a void, prevent expanding vertically
	2	Double head	Put an additional head on the stroke
	3	Triple head	Put two additional heads on the stroke
	#	Damage	Draw hatching (diagonal lines) over this stroke, to indicate damage to the tablet in rendering
	!	Highlight	Render this stroke in a different color
	?	Invert	Swap the head and tail of this stroke
	T	Tenû	Rotate a node 45° counter-clockwise
Node adjustments	E	Expand	Ask the arrangement algorithm to give this node twice as much space
	M	Margin	Leave a small amount of empty space on all sides of this node
	R	Restrict	Prevent this node from expanding in any direction
	A	Allow	Allow this node to expand infinitely in all directions
	K	Kern	Allow adjacent nodes to kern into this node from every side
Canvas size	N	Narrow	1:3 (width:height ratio)
	P	Portrait	2:3
	S	Square	1:1 (default)
	L	Landscape	3:2
	W	Wide	2:1
	X	Extra-wide	3:1

Acknowledgments: I would like to thank my thesis advisor, Ryan Shosted, who introduced me to cuneiform studies in the first place and funded my experiments with this system; the user Yellow Sky on StackExchange, who spurred all of this by directing my attention to Downes’ work back in 2020; Daniel Jost, who corrected my translations of Gottstein’s German; the editors and two anonymous reviewers who first gave feedback on this paper for the Cuneiform Digital Library Journal; and the editors and two further anonymous reviewers who then helped refine it here.

Research ethics: Not applicable.

Informed consent: Not applicable.

Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.

Use of Large Language Models, AI and Machine Learning

Tools: None declared.

Conflict of interest: The author states no conflict of interest.

Research funding: None declared.

Data availability: The raw data can be obtained on request from the corresponding author.

References

- [1] J. Huehnergard, *A Grammar of Akkadian. Third. Harvard Semitic Studies*, Winona Lake, Indiana, Eisenbrauns, 2011. Available at: https://www.academia.edu/234695/2011_A_Grammar_of_Akkadian_3rd_edition_.
- [2] J. S. Cooper, “Sumerian and Akkadian,” in *The World’s Writing Systems*, P. T. Daniels and W. Bright, Eds., Oxford University Press, 1996.

- [3] M. Worthington, *Principles of Akkadian Textual Criticism*, Berlin, De Gruyter, 2012.
- [4] L. Watkins and D. Snyder, “The Digital Hammurabi Project,” 2003. Available at: https://www.researchgate.net/publication/247838547_The_digital_hammurabi_project.
- [5] R. Borger, *Mesopotamisches Zeichenlexikon. German. 2., revidierte und aktualisierte Aufl. Alter Orient und Altes Testament; 305*, Münster, Ugarit-Verlag, 2010.
- [6] R. Labat, *Manuel d'épigraphie akkadienne. Signes, syllabaire, idéogrammes*, 6th ed. Paris, Librairie orientalisle P. Geuthner, 1995.
- [7] S. Gordin, *Hittite Scribal Circles: Scholarly Tradition and Writing Habits*, 1st ed. Harrassowitz Verlag, 2015. Available at: <http://www.jstor.org/stable/j.ctvc76zpc>.
- [8] E. Robson, “Using sign lists: Labat, Borger, and PSL,” in *Cuneiform Revealed: An Introduction to Cuneiform Script and the Akkadian Language*, 2010. Available at: <http://kn.prs.heacademy.ac.uk/cuneiformrevealed/learningsigns/usingsignlists/>.
- [9] H. Radau, *Letters to Cassite Kings from the Temple Archives of Nippur*, University of Pennsylvania Department of Archaeology, 1908. Available at: <https://archive.org/details/letters-to-cassite-kings-from-the-temple-archives-of-nippur>.
- [10] Å. W. Sjöberg, “The Old Babylonian eduba,” in *Sumerological Studies in Honor of Thorkild Jacobsen on His Seventieth Birthday*, S. J. Lieberman, Ed., Assyriological Studies 20. University of Chicago Press, 1974, pp. 159–179.
- [11] J. Taylor, “OB Nippur Lu=ša,” in *Digital Corpus of Cuneiform Lexical Texts*, Berkeley, University of California, 2005.
- [12] DCCLT, “Introduction. What is a lexical list?” in *Digital Corpus of Cuneiform Lexical Texts*, University of California, Berkeley, 2003.
- [13] A. Deimel and P. F. Gössman, *Šumerisches Lexikon*, Scripta Pontificii Instituti Biblici. sumptibus Pontificii instituti biblici, 1934. Available at: <https://books.google.com/books?id=2hcdAAAAIAAJ>.
- [14] C. Rüster and E. Neu, *Hethitisches Zeichenlexikon. Inventar und Interpretation der Keilschriftzeichen aus den Boğazköy-Texten*, Studien zu den Boğazköy-Texten: Beiheft. O. Harrassowitz, 1989. Available at: https://books.google.com/books?id=L15pAJsx%5C_y0C.
- [15] R. Borger, *List of Neo-Assyrian Cuneiform Signs: A Practical and Critical Guide to the Unicode Blocks ‘Cuneiform’ and ‘Cuneiform Numbers’ of Unicode Standard Version 5.0*, 2007. Available at: <http://www.sumerisches-glossar.de/download/SignListNeoAssyrian.pdf>.
- [16] N. Gottstein, “Ein stringentes Identifikations- und Suchsystem für Keilschriftzeichen,” in *Mitteilungen der Deutschen Orient-Gesellschaft zu Berlin*, vol. 145, 2013, pp. 127–136.
- [17] N. Christiaan Veldhuis, *Elementary Education at Nippur: The Lists of Trees and Wooden Objects*, Ph.D. thesis, University of Groningen, 1997.
- [18] C. J. Crisostomo, “Introduction to cuneiform sign lists,” in *Digital Corpus of Cuneiform Lexical Texts. The DCCLT Project*, 2019.
- [19] K. Šašková, *Cuneiform Sign List*, 2021. Available at: http://home.zcu.cz/~ksaskova/Sign_List.html.
- [20] ePSD, *The Electronic Pennsylvania Sumerian Dictionary*, University of Pennsylvania Museum of Anthropology and Archaeology, 2006. Available at: <http://psd.museum.upenn.edu/nepsd-frame.html>.
- [21] T. Homburg, “PaleoCodage: enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding,” in *Digital Scholarship in the Humanities 36.Suppement_2*, 2021, pp. ii127–ii154.
- [22] Theo van den Hout, *The Elements of Hittite*, Cambridge, Cambridge University Press, 2011.
- [23] J. Burman, et al., “Inventaire des signes hiéroglyphiques en vue de leur saisie informatique. Manuel de codage des textes hiéroglyphiques en vue de leur saisie informatique,” in *Mémoires de L'académie des inscriptions et belles lettres*, 1988.
- [24] A. Kamate, “Creating a Japanese Handwriting Recognizer: using tensorflow to create a machine learning model for handwriting recognition,” in *Towards Data Science*, 2020.
- [25] X. Liu, B. Hu, Q. Chen, X. Wu, and J. You, “Stroke sequence-dependent deep convolutional neural network for online handwritten Chinese character recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4637–4648, 2020.
- [26] C.-L. Liu and X.-D. Zhou, “Online Japanese character recognition using trajectory-based normalization and direction feature extraction,” in *Tenth International Workshop on Frontiers in Handwriting Recognition*, G. Lorette, Ed., France, Université de Rennes 1. La Baule, Suvisoft, 2006.
- [27] E. Doostmohammadi and M. Nassajian, “Investigating machine learning methods for language and dialect identification of cuneiform texts,” in *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, Ann Arbor, Michigan, Association for Computational Linguistics, 2019, pp. 188–193.
- [28] S. Gordin, et al., “Reading Akkadian cuneiform using natural language processing,” *PLoS One*, vol. 15, no. 10, pp. 1–16, 2020.
- [29] T. Dencker, P. Klinskisch, S. M. Maul, and B. Ommer, “Deep learning of cuneiform sign detection with weak supervision using transliteration alignment,” *PLoS One*, vol. 15, no. 12, pp. 1–21, 2020.
- [30] N. M. Kriege, et al., “Recognizing cuneiform signs using graph based methods,” in *COST@SDM*, 2018.
- [31] D. Fisseler, et al., “Towards an interactive and automated script feature analysis of 3D scanned cuneiform tablets,” in *Scientific Computing and Cultural Heritage*, vol. 2013, 2013.
- [32] M. Fey, et al., “SplineCNN: fast geometric deep learning with continuous B-spline kernels,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 869–877.
- [33] E. Stötzner, et al., “R-CNN based polygonal wedge detection learned from annotated 3D renderings and mapped photographs of open data cuneiform tablets,” in *Eurographics Workshop on Graphics and Cultural Heritage, The Eurographics Association*, A. Bucciero, et al., Ed., 2023.
- [34] D. Snyder, “The initiative for cuneiform encoding,” 2000. Available at: <https://pages.jh.edu/ice/>.
- [35] A. H. Gardiner, *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*, Oxford, Griffith Institute, Ashmolean Museum, Oxford University Press, 1957. Available at: http://web.ff.cuni.cz/ustavy/egyptologie/pdf/Gardiner_signlist.pdf.
- [36] A. H. Gardiner, *Catalogue of the Egyptian Hieroglyphic Printing Type: From Matrices Owned and Controlled by Dr. Alan H. Gardiner: in Two Sizes 18 Point, 12 Point with Intermediate Forms*, Oxford University Press, 1928. Available at: <https://libmma.contentdm.oclc.org/digital/collection/p15324coll10/id/127182>.
- [37] J. Breen, “Multiple indexing in an electronic kanji dictionary,” in *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries. ElectricDict '04*, Geneva, Switzerland, Association for Computational Linguistics, 2004, pp. 1–7.
- [38] S. Tinney, *ATF Inline Tutorial*, Oracc: The Open Richly Annotated Cuneiform Corpus, 2019. Available at: <http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/inlinetutorial/index.html>.

- [39] E. Robson, “Wedges and signs,” in *Cuneiform Revealed: An Introduction to Cuneiform Script and the Akkadian Language*, 2010.
- [40] Electronic Dictionary Research and Development Group, “KANJIDIC project,” 1991–2021. Available at: http://www.edrdg.org/wiki/index.php/KANJIDIC_Project.
- [41] J. Breen, *An Overview of the Four Corner Coding System*, Monash University, 2000. Available at: <http://nihongo.monash.edu//FOURCORNER.html>.
- [42] A. Downes, “The Xixia writing system,” Bachelor of Arts Honours Thesis, Macquarie University, 2008.
- [43] T. Nishida, *Seikago no kenkyū: A Study of the Hsi-hsia Language; Reconstruction of the Hsi-hsia Language and Decipherment of the Hsi-hsia Script*, vol. 2, Japan, Zauho Kaukokai, 1966.
- [44] The Unicode Consortium, “East Asia,” in *The Unicode Standard, Version 9.0*, 2016.
- [45] D. Snyder, “Examples of cuneiform ideographic descriptor usage,” 2004. Available at: <https://pages.jh.edu/ice/basesigns/CuneiformDescriptorUsage.pdf>.
- [46] W. Wong, C. L. K. Yiu, and K. C. F. Ng, “Typesetting rare Chinese characters in LaTeX,” in *TUGBoat*, vol. 24, TeX Users Group, 2003, pp. 582–587.
- [47] A. Downes, *How Does Tangut Work?* Ph.D. thesis, Macquarie University, 2016.
- [48] G. Torri, “Hittite scribes at play: the case of the cuneiform sign AN,” in *Investigationes Anatolicae: Gedenkschrift für Erich Neu*, 2010, pp. 317–327.
- [49] M. Yon, *The City of Ugarit at Tell Ras Shamra*, University Park, Eisenbrauns, 2006.

Bionotes



Daniel M. Stelzer

Department of Linguistics, University of Illinois at Urbana-Champaign, Champaign, USA

stelzer3@illinois.edu

Daniel Stelzer is a doctoral student at the University of Illinois at Urbana-Champaign, specializing in computational analysis of ancient languages. They received their MA and BA in Linguistics and a BS in Computer Science with a minor in Classics from the same institution. Their current focus is digitization of Hittite cuneiform.