# 楔形文字文献の形態論情報付きコーパス構築の自動化に向けて

山内健二 [1]　森若葉 [2]　山本孟 [3]

[1] 京都大学情報学研究科　[2] 国士舘大学イラク古代文化研究所　[3] 京都大学文学部

yamauchi@nlp.ist.i.kyoto-u.ac.jp

## 1 はじめに

楔形文字とは、古代中近東で約 3,000 年にわたり利用されていた文字であり、主に粘土板に記述された。楔形文字により記述された文献は 2 つの点で、学問的に重要である。まず、楔形文字を利用していた言語が数と系統の面で多様であるため、言語資料として価値がある。また、文献の種類も行政文書から神話まで多岐にわたるため、古代中近東の文化を知る手がかりとなる。

文献に記述された内容の把握や言語学的研究のため、文献の形態論情報付きコーパスの構築が行われているが、発見されている楔形文字文献の数は最低でも約 50 万程度と大規模であるため、計算機による構築の自動化が望まれる。

楔形文字文献の形態論情報付きコーパス構築には、粘土板に書かれている文字種別の特定、翻字 (個々の文字の読みを付与する作業) ・形態論情報付与という作業が必要となる。現状ではこれらの作業は基本的に専門家による手作業で行われているが、前者については近年光学的文字認識 (Optical Character Recognition, OCR) により、粘土板手書きコピー画像から粘土板に記述された文字種別の特定を行うことを目標とした研究が行われている。しかし、現時点では計算機による文字種別の特定の自動化自体は実現できておらず、当然翻字や形態論情報の付与も自動化できていない。そのため、まず OCR による粘土板手書きコピー画像からの文字種別特定を実現する必要がある。

本稿では、粘土板手書きコピー画像から自動で楔形文字文献の形態論情報付きコーパス構築を行うための展望を述べ、さらに、OCR 実現に向けて行なった文字別画像データセットの部分的な構築について述べる。

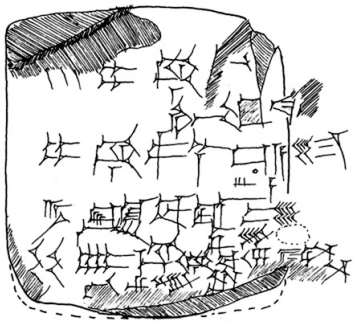

図1: 楔形文字粘土板の手書きコピー画像

## 2 楔形文字の性質と翻字

本節では楔形文字の基本的な性質とその翻字について説明を行う。

### 2.1 楔形文字の外見

各楔形文字は外見上三角形の角 (ヴィンケルハーケン) と直線によって構成され、字数は 600 字程度とされている。また、数千年にわたって使用されていたこともあり、1 文字につき複数の字体が存在している。

### 2.2 楔形文字の読みと翻字

楔形文字はアルファベットとは異なり、日本語における漢字のような読みの特徴を持っている。すなわち、1 文字に対して文脈に応じて読みが多数考えられ、また言語により読み方も異なる。そのため、文献管理を容易にするため、各文字をアルファベットで翻字する。この時、文字そのものを表す場合は英大文字を、文字の読みを表す場合は英小文字を用いる。例えば 𒀭(AN) という文字はシュメール語においては an、diĝir などと読みが文脈で変わり、時には読まれない場合もある。時代が降ると字体も簡略化され 𒀀 となった。

楔形文字と読みは常に一対一の対応ではなく、日本語における熟字訓のような変則的な読み方をする場合がある。例えば PA ・ TE ・ SI の 3 文字がこの順番で並んでいる場合、ensi$_2$ と読む。

### 2.3 手書きコピー画像と翻字

楔形文字が記述される素材は 3 次元的なものであるため、正確な解析を行うには実際の粘土板を直接参照する必要がある場合も多い。写真からのみでは、粘土板の丸みと撮影時の光源の兼ね合いから不明瞭にしか読み取れない部分が残るからである。このため、解析を容易にするため粘土板については図 1のように手書きコピー画像[*1]が作られ、これを参照して翻字を行う場合が多い。手書きコピー画像では文字が側面へ周っている場合は粘土板の枠内から「はみ出すように」書かれる。

翻字の際は、各文字の並びを文法や意味の単位になるようハイフン (-) で接続し、そうでなければスペースで区切る。例えば図 1に示した文献の 3 行目の翻字は以下のようになる。

(1) $\mathrm{nig}_2$-$\mathrm{buru}_{14}$ ba-ra-$\mathrm{du}_8$

ba-ra-$\mathrm{du}_8$ は 3 文字を翻字したものだが、これで 1 つの意味を表す単位であるため、ハイフンで接続されている。また、同音異体字を区別するため下付き文字を用いる。例えば $\mathrm{du}_8$ は du と読む 8 種類目の文字である。

## 3 背景

本節では楔形文字文献の形態論情報付きコーパスの自動構築にむけたこれまでの取り組みを述べる。

### 3.1 楔形文字文献の電子データベース化

楔形文字文献の電子データベース化は多方面で行われており、特に Cuneiform Digital Library Initiative (CDLI)[*2]・ Database of Neo-Sumerian Texts (BDTNS)[*3] ・ The Open Ritchly Annotated Cuneiorm Corpus (ORACC)[*4]・ The Electronic Text Corpus of Sumerian Literature (ETCSL)[*5]といったプロジェクトが代表例としてあげられる。視点や光源を変えて立体的に参照できるように粘土板を 3D データ化する試みもある [1]。

CDLI と BDTNS は粘土板の出土場所、製作年代、手書きコピー画像、翻字などを含むが、形態論情報は付与されていない。

一方、ORACC や ETCSL は各文献について翻字と形態論情報の付与が行われたコーパスである。しかし、構築は手作業で行われておりしかも規模は限定的である。一例として ORACC 中の Digital Corpus of Cuneiform Lexical Texts (DCCLT)[*6]の場合、8,093 の文献のみにアノテーションがなされている。

### 3.2 手書きコピー画像からの文字種別特定

現時点では翻字を含めた楔形文字文献コーパスの自動構築は実現できていないが、手書きコピー画像からの文字種別特定についていくつか試みがある。

粘土板から直接そこに記述された文字の種別の特定を OCR で行おうとする場合、素材が平面的で無い分、紙媒体に記述された言語を読み取る標準的な OCR より複雑な処理が要求される。直接粘土板から文字認識を行うには 3 次元的な処理を行わねばならないため、3D スキャニング用の外部装置が必要だからである。また、粘土板が海外に所蔵してあるといった理由で入手が難しい場合もある。

そのため、粘土板の手書きコピー画像、すなわち 2 次元データを対象とすることを考える。CDLI などで手書きコピー画像は存在するが翻字が完了していない文献も多い (CDLI 中では約 31 万の文献のうち、7 万程度) ため、手書きコピー画像からの文字認識や翻字も十分な有用性がある。

手書きコピー画像からの OCR については、Mara らのグループが積極的に研究を行なっている。彼らは、まず手書きコビー画像について楔形文字が書かれた部分をベクタ化している [2]。その後、各楔形文字をそれを構成するヴィンケルハーケンの位置や向きから得られた特徴ベクトルで表現する手法を考案し、一部の文字について先のベクタ化した手書きコピー画像からの検索を試みている [3] しかし、現時点では手書きコピー画像から個々の文字を特定するまでには行なっていない。

## 4 楔形文字文献の形態論情報付きコーパス構築の自動化

本節では楔形文字文献の形態論情報付きコーパス構築の自動化に向けた展望を述べる。

3.2節で述べた通り、OCR による粘土板に記述された各文字種別の特定に向けた研究は行われている。しかし、翻字や形態論情報の付与を行う場合、2節で述べたように言語や文脈で各文字の読みが変化するため、文法など言語的な特徴の考慮が必要である。そのためには

[*1] http://cdli.ucla.edu/search/archival_view.php?ObjectID=P101139 から抜粋
[*2] http://cdli.ucla.edu
[*3] http://bdtns.filol.csic.es
[*4] http://oracc.museum.upenn.edu
[*5] http://etcsl.orinst.ox.ac.uk
[*6] http://oracc.museum.upenn.edu/dcclt/

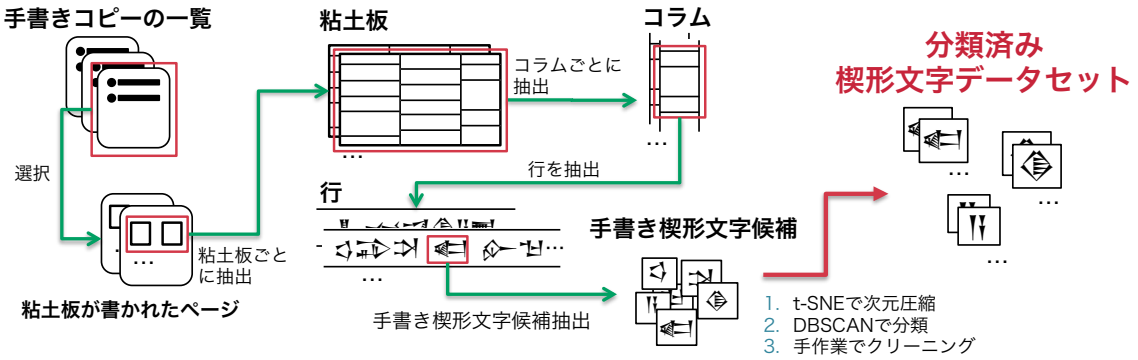


図2: 楔形文字データセットの作成の流れ

言語モデルの構築が必要であるが、小規模ながら 3.1節で述べた言語コーパスが存在するため、これらを利用可能である [4]。

直接コーパスの構築にはつながらないが、このような言語モデルの構築は OCR による文字種別特定の精度向上にも貢献できる。粘土板には物理的な破損が多く起きるため、これの補完を画像処理のみだけでなく、言語モデルを併用することで補完可能であると考えられる。同一文献内でも文脈による欠損部分の推定を行えるが、楔形文献は同一内容の文書が繰り返し書写されたため、外部知識として他文献の内容を参照すればより正確な補完が行える。

さらに、このような補完は同時に文献の復元へも応用できる。同じ内容を表す断片群を特定できれば、これを集約することで、元は 1 つだった文献の内容を復元可能だからである。この種の文献復元への言語処理の応用は、楔形文字を利用していた言語の 1 つであるヒッタイト語の文献について既に試み [5] がある。

## 5 楔形文字データセットの作成

本節では手書きコピー画像からの文字種別特定に向けて作成した、各楔形文字の手書き画像のデータセットについて述べる。

### 5.1 データセット作成の目的

3.2節で述べた通り、手書きコピー画像に記述された個々の楔形文字を特定できるまでには至っていない。個々の文字の特定を行うためには、それぞれの文字についての特徴を得るためのデータセットが必要となる。しかし、楔形文字については、少数のフォントが存在するものの、現時点では十分な量のフォントや手書き画像が存在しない。そのため、今回はまず予備段階として各楔形文字の手書き画像について、限定的なデータセットを作成することを考える。具体的には紀元前 3,000 年期の文献の手書きコピー画像を用いて、よく利用されている楔形文字 100 個それぞれで 200 の手書き画像を用意することを目標とする。

### 5.2 流れ

既存の手書きコピー画像を使い、手書き楔形文字の候補となる部分を抽出し分類を行う。その後、分類結果から手動で最終的な楔形文字データセットを作成する (図 2)。以下、抽出の流れを詳述する。

#### 5.2.1 手書きコピー画像の選択

抽出対象の手書きコピー画像が書かれたページの画像を、パブリックドメインとなった 10 の資料資料からスキャニングしたものより選択する。

#### 5.2.2 手書き楔形文字候補の抽出

まず、手書きコピー画像が書かれているページから粘土板と考えられる領域を Suzuki らの手法 [6] により抽出する。

次に、それら粘土板領域から楔形文字の候補となる領域を抽出する。この時、楔形文字を記述する際、書記により粘土板へ一定の間隔で縦横に線が引かれる場合があることを利用する。具体的には、線での分割の単位を**コラム**、横での分割の単位を**行**と呼ぶこととし、まず行単位の抽出を考える。行の抽出は、確率的ハフ変換 [7] により、粘土板領域のコラムと行を分割する線を抽出し、これに沿って行の候補となる領域を切り出す。

最後に切り出した行ごとに再び Suzuki らの手法を用いて楔形文字と考えられる領域を画像として抽出する。

#### 5.2.3 手書き楔形文字候補の分類・クリーニング

ここまでで抽出した手書き楔形文字候補群を 64 ピクセル四方の画像に縮小したのち、t-SNE [8] により 3 次元に次元圧縮し、DBSCAN [9] によりクラスタリングする。その後、各クラスタに分類された画像の内容を

表1: 抽出処理の中間統計情報

| 項目数 | 値 |
|---|---|
| 抽出対象のページ数 | 549 |
| 抽出粘土板画像数 | 2,015 |
| 抽出文字候補数 | 98,357 |

手作業で確認しつつ、対象とする各楔形文字それぞれで 200 ずつ画像を選択し、必要があればノイズも除去する (クリーニング)。

### 5.3 結果

DBSCAN によるクラスタリングまでの抽出処理の中間統計情報を表 1へ示す。抽出元の文献と現時点での抽出状況は`https://github.com/yustoris/cuneiform-image-set` にて 2017 年 2 月 14 日以降公開予定である。

## 6 まとめ

楔形文字文献の OCR に利用することを想定した文字別画像データセットを部分的に作成した。

今後はまず現時点で抽出済みの文字候補画像群をノイズ除去などを行なって整理するとともに、別の時代・文字についても同様に画像抽出を行なってデータセットを拡充することを直近の目標としている。時代別での文字の差異は連続的なものであるため、どの程度離散的に扱うかも今後の課題となる。

また、実際の OCR へ本データセットを用いた場合どの程度の精度を達成できるか、さらに、楔形文字を利用していた主要言語について言語モデルを作成することで、OCR の精度をどの程度向上可能かを検証していきたいと考えている。

## 参考文献


[1] Hubert Mara, Susanne Krömker, Stefan Jakob, and Bernd Breuckmann. “GigaMesh and Gilgamesh: 3D Multiscale Integral Invariant Cuneiform Character Extraction”. In: *Proc. of the 11th VAST*. 2010.

[2] Bartosz Bogacz, Nicholas Howe, and Hubert Mara. “Segmentation free spotting of Cuneiform using part structured models”. In: *ICFHR 2016*. 2016.

[3] Judith Massa, Bartosz Bogacz, Susanne Krömker, and Hubert Mara. “Cuneiform Detection in Vectorized Raster Images”. In: *21th CVWW*. 2016.

[4] Valentin Tablan, Wim Peters, Diana Maynard, and Hamish Cunningham. “Creating Tools for Morphological Analysis of Sumerian”. In: *Proc. of the Fifth LREC*. 2006.

[5] Stephen Tyndall. “Toward Automatically Assembling Hittite-Language Cuneiform Tablet Fragments into Larger Texts”. In: *Proc. of the 50th ACL*. 2012.

[6] Satoshi Suzuki and Keiichi Abe. “Topological structural analysis of digitized binary images by border following”. In: *Computer Vision, Graphics, and Image Processing* 39 (1 1985).

[7] Jiri Matas, Charles Galambos, and Josef Kittler. “Robust detection of lines using the progressive probabilistic hough transform”. In: *Computer Vision and Image Understanding* 78 (1 2000).

[8] Laurens van der Maaten and Geoffrey Hinton. “Visualizing High-Dimensional Data Using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008).

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proc. of 2nd KDD*. 1996.