修士論文

万葉集の未解読歌の解読

佐々木 啓晶

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
情報科学

令和 7 年 3 月

Abstract

Man’yoshu, Japan's oldest anthology of poetry, encompasses 4,516 poems, though 42 of them remain undeciphered. Previous studies assumed these compositions, written in Man’yogana (a writing system of using Chinese characters to represent Japanese phonemes), should be interpreted as Japanese. However, Vovin successfully demonstrated that Man’yoshu's poem number 9, which was previously undeciphered, became intelligible when it was read as Old Korean. The goal of this study is to decode the other undeciphered poems of Man'yoshu as Old Korean. We implement Vovin's method as a computational system to automatically or semi-automatically decipher undeciphered poems.

Our proposed method primarily involves two modules. The first module converts sequences of Chinese characters into sequences of phonemes using the “Chinese character phoneme dictionary.” The module then performs morphological analysis on these phoneme sequences using a Korean word dictionary, and finally generates sequences of Korean words. The second module enlarges our Chinese character phoneme dictionary by adding Middle Korean (MK) pronunciations. This is achieved by estimating Middle Korean pronunciations from Chinese pronunciations.

To implement the first module, we begin by creating the Chinese character phoneme dictionary which defines the phonetic symbols associated with each Chinese character. This dictionary compiles five types of pronunciations: Man’yogana, Idu (a writing system of using Chinese character to represent Korean phonemes), Middle Korean pronunciations derived from Late Han Chinese (LHC) pronunciations, Middle Korean pronunciations derived from Early Middle Chinese (EMC) pronunciations, and jeongyong pronunciations that represent Korean phonemes intended to convey the Chinese character’s meaning.

For a given input sequence of Chinese characters, the system converts it to multiple possible sequences of phonemes by looking up the Chinese character phoneme dictionary for each of the input Chinese characters. Next, for each phonetic sequence, morphological analysis is performed to obtain sequences of Middle Korean words. MeCab is employed for this process, since it is the only morphological analyzer for which a Middle Korean word dictionary is available. Specifically, this study utilizes the MkHanDic dictionary with 9,653 words as the MK word dictionary. Then, the most appropriate word sequences are chosen from all generated sequences using the following two-step procedures. (1) Any grammatically incorrect word sequences are removed by the grammatical check. Specifically, we eliminate word sequences that begin with a verbal ending, begin with a

specifier, begin with a dependent noun, end with a numeral, contain consecutive word endings, or include unknown words. (2) From the remaining valid word sequences, those consisting of the fewest words are chosen according to the minimum number count principle, which is a heuristic rule employed in morphological analysis to determine the optimal results. Finally, we manually translate the chosen Middle Korean word sequence, which may represent a possible interpretation of the poem, into Japanese, seeking to decipher the original undeciphered poem.

The second module addresses limitations in our available data. The materials we use to compile our Chinese character phoneme dictionary do not provide Middle Korean pronunciations derived from LHC and EMC for all Chinese characters. Therefore, models to predict Middle Korean pronunciations from Chinese pronunciations are trained. Specifically, we represent phonemes as sequences of International Phonetic Alphabet (IPA) symbols (phonemes) and train a sequence-to-sequence model to convert Chinese phoneme sequences into Middle Korean phoneme sequences. Each IPA symbol is represented as either a feature vector reflecting their phonetic features (called IPA feature vector), or a one-hot vector. Besides, a bidirectional or unidirectional Long Short-Term Memory (LSTM) is used as our sequence-to-sequence model. This results in proposing four models based on various combinations of IPA vector representations and bidirectional/unidirectional LSTMs. These models are trained using the set of Chinese characters associated with both the corresponding LHC or EMC phonemes and the MK phonemes.

Several experiments are carried out to evaluate our proposed method. First, our model to predict Middle Korean pronunciations from LHC or EMC pronunciations is evaluated. The accuracy of the model to convert LHC pronunciations to MK pronunciations is 0.839, while that of the model to convert EMC to MK is 0.800. Notably, the BiLSTM-IPA-IPA model, which is the BiLSTM model handling IPA feature vectors in both the encoder and the decoder, achieves the highest accuracy among our four proposed models, while also requiring the shortest training time.

Next, our proposed method is applied to poem number 9, which Vovin successfully deciphered, to verify its validity. Our method generates the same interpretation as Vovin, indicating the effectiveness of our method. This success demonstrates that our method has the potential to decipher other undeciphered poems besides poem number 9. Subsequently, our method is applied to six undeciphered poems to attempt their decipherment. As a result, we obtain the interpretation “Would you eat rice? Would you?” for poem number 3889. The undeciphered section of poem number 3889 constitutes the final part of the poem. The preceding part of the poem is transcribed as “hitodama no sa

aonaru kimi ga tada hitori mo aheri shi ameyo no," which means "I met you, a wandering soul, blue, alone, on a rainy night." The newly obtained interpretation for the undeciphered section is not clearly related to this preceding part, but it is not likely to be completely unrelated. Determining the validity of this interpretation remains a complex challenge.

Moreover, Man'yoshu also contains approximately 500 words of Makurakotoba, whose meanings are largely unknown. Makurakotoba is considered to be a rhetorical word that always modifies a specific word as its head; however, it is generally recognized that its origin remains unclear and that it has no semantic meaning. We consider that, like the undeciphered poems, Makurakotoba might also be represented by Korean phonemes. Therefore, we apply our proposed method to decipher six words of Makurakotoba. As a result, we obtain the interpretation of "father" or "paternal" for the Makurakotoba word "ashihikino." This interpretation aligns reasonably well with its head "mountain." However, since "ashihikino" has 31 written forms of Chinese characters and this feasible interpretation is obtained only from one of them, it is still uncertain that this Makurakotoba has been successfully deciphered.

The future work of this study includes a revision of the input data, an extension of materials for the Chinese character phoneme dictionary, and comprehensive evaluation. In this study, the text based on the Nishihonganji version is used as the input data. However, other Chinese characters from variant versions can also be used. For the Chinese character phoneme dictionary, it is promising to add more Chinese character pronunciations from additional texts, as well as to incorporate the Go-on pronunciations which were prevalent in Japan during the compilation of Man'yoshu. For the evaluation of the method, our method has been applied for only 6 of the 42 undeciphered poems and 6 of the 500 words of Makurakotoba. More comprehensive experiments are required to precisely evaluate the effectiveness of our proposed method. Moreover, augmentation of the Middle Korean dictionary MkHanDic, refinement of the method for selecting appropriate word sequences, and establishment of a methodology to assess the validity of the obtained interpretation of undeciphered poems are other important lines for the future direction.

概要

万葉集は 4,516 首を収録する日本最古の歌集であるが，42 首が未解読のままである．これまでの研究は，万葉仮名で書かれた歌を日本語として解釈することを前提としてきた．一方で，Vovin は未解読の歌である万葉集第 9 番の歌が古い朝鮮語で解読できることを証明した．本研究は，万葉集の未解読歌を古い朝鮮語により解読することを目的とする．そのために，Vovin の手法を計算機処理として実装し，未解読歌を自動的あるいは半自動的に解読することを目指す．

提案手法は主に 2 つのモジュールから構成される．第 1 のモジュールは漢字音辞書を用いて漢字列を音の系列に変換し，次に朝鮮語の単語辞書を用いて形態素解析を行うことで，朝鮮語の単語列を得るものである．第 2 のモジュールは漢字音辞書に中期朝鮮語音(Middle Korean; MK)を追加するために，中国語音から中期朝鮮語音を推定するものである．

第 1 のモジュールを実装するにあたり，まず漢字に対してその音を定義した漢字音辞書を作成する．漢字音辞書には，万葉仮名，吏読，中国の後漢時代の漢字音(Late Han Chinese; LHC)から転じた中期朝鮮語音，中国の隋唐時代の漢字音(Early Middle Chinese; EMC)から転じた中期朝鮮語音，正用(漢字の意味である字義を表すための朝鮮語音)という 5 種類の音を登録する．

解読対象となる漢字列に対し，漢字音辞書を用いて個々の漢字を音に変換し，それらを全て組み合わせて複数の音の系列を得る．次に，個々の音の系列に対して形態素解析を行い，中期朝鮮語の単語列を得る．形態素解析には中期朝鮮語辞書を有する唯一の形態素解析器である MeCab を採用し，MkHanDic という 9,653 語を収録する辞書を使用する．そして，2 つの手続きにより，得られた複数の単語列の候補の中から適切なものを選択する．(1)単語列の文法的制約をチェックし，文法的に正しくない単語列を除外する．具体的には，語尾で始まる，指定詞で始まる，依存名詞で始まる，数詞で終わる，連続する語尾を含む，未知語を含むなど，明らかに文法的な誤りを含む単語列を除外する．(2)残された単語列の候補の中から，形態素数最小法の考えに基づき．単語数が少ない単語列を選択する．最後に，得られた中期朝鮮語の単語列(文)を人手により日本語へ翻訳し，未解読歌の解読を試みる．

第 2 のモジュールとして，漢字音辞書の作成に用いた資料では全ての漢字に対して LHC 及び EMC から転じた MK が記載されているわけではないことから，中国語音から中期朝鮮語音を推定するモデルを学習する．具体的には，音を IPA(International Phonetic Alphabet)シンボル(音素)の系列で表現し，中国語音素列を中期朝鮮語音素列に変換する系列変換モデルを学習する．IPA シンボルを IPA の音韻的特徴を反映した特徴量ベクトルもしくは one-hot ベクトルで

表現し，系列変換モデルとして双方向または単方向の Long Short-Term Memory (LSTM)を用いる．IPA シンボルのベクトルと双方向・単方向 LSTM の組み合わせで 4 通りのモデルを提案する．LHC と MK，もしくは EMC と MK の音が既知である漢字の集合を訓練データとしてこれらのモデルを学習する．

実験では，まず中期朝鮮語音の推定モデルを評価した．その結果，LHC から MK への推定で 0.839，EMC から MK への推定で 0.800 の正解率を得た．特に，エンコーダ・デコーダともに IPA 特徴量ベクトル表現を用いた BiLSTM-IPA-IPA モデルは本研究で提案した 4 つのモデルの中で最も正解率が高く，学習に要する時間も短かった．

次に，Vovin が解読に成功した 9 番歌について提案手法を適用し，Vovin と同じ解釈が得られることを確認した．このことは提案手法の有効性を示すものであり，提案手法によって 9 番歌以外の未解読歌を解読できる可能性があることを示唆する．そこで，未解読歌 6 首に対して提案手法を適用し，その解読を試みた．その結果， 3889 番歌について「ご飯を召し上がるのか，あなたは」という解読結果を得ることができた．この 3889 番歌の未解読部分は歌の最後の句にあたる．上の句の部分は「人魂のさ青なる君がただひとりもあへりし雨夜の」と記されており，現代語に訳せば「人魂である君がただひとりさまよっている．その君に私は雨の夜に出会い」という意味になる．解釈された未解読部分が上の句と関連しているとは言い難いものの必ずしも全く繋がらないとも言い切れない．この解釈の妥当性をどのように評価するかは難しい問題として残されている．

また，万葉集には約 500 あると言われる枕詞が含まれ，その意味は不明とされている．枕詞は被枕詞を修飾する修辞語とされるが，生成の由来が不明で，意味はないという認識が一般的である．本研究では，未解読歌と同様に枕詞も朝鮮語音を借りている可能性があると考え，提案手法を用いて 6 個の枕詞の解読を試みた．その結果，「あしひきの」については「父である」「父なる」という新たな解釈を得た．これは被枕詞である「山」に掛かる枕詞として不自然ではない．ただし，「あしひきの」には 31 個の漢字表記があるが，可能な解釈が得られたのはそのうちの 1 つのみであったため，この枕詞の解釈に成功したと断言することはできない．

今後の課題として，入力データ，辞書データ，実験範囲の拡充が挙げられる．入力データについては，西本願寺本を底本とするデータをそのまま使用したが，異本から採取された漢字を入力データとするという選択肢も存在する．漢字音辞書については，さらなる文献からの漢字音の取り込みや，万葉集編纂当時に日本で主流であった呉音の登録も検討に値する．実験範囲については，現時点では 42 首ある未解読歌のうち 6 首のみ，500 に上る枕詞のうち 6 個のみの実験に留まっているため，実験対象のさらなる拡大が必要である．加えて，中期朝鮮語辞

書 MkHanDic を拡張すること，適切な単語列の選択方法を改良すること，そして解読結果の妥当性評価の方法論を確立することも必要である.

目次

図目次

表目次

第1章 はじめに

1.1 背景

万葉集は 4,516 首を収録する日本最古とされる歌集である．全ての歌には 1 番歌，2 番歌といったように連番が振られている．中国語の漢字音を借りて上代日本語の音節を表す表記法である万葉仮名によって主に書かれており，これまで日本語で読めるという前提のもとで解釈されてきた[1]．しかし，現代に至るまで 42 首が未解読のままとなっている[2]．

この未解読歌の解読には複雑な課題が存在する．万葉仮名による表記では，一つの漢字に複数の読み方が存在し，同じ音を異なる漢字で表現する可能性がある．さらに，歌の解釈には文脈や韻律も考慮する必要があり，解読の難しさを増している．

これまでの解読は主に人手による試行錯誤に依存してきた．しかし，考えられる読み方の組み合わせは膨大であり，人手での網羅的な検討には限界がある．また，解読の結果は解読者の言語知識や経験に大きく依存するため，客観的な評価が困難であるという問題も存在する．したがって，万葉集の未解読歌を自動的に解読する，もしくは人手による解読を支援する手法の開発が望まれる．

一方，漢字以外の古代文字については機械学習による解読研究が進展している．例えば，ウガリット語や線文字 B など地中海地域のアルファベット系言語については，古代の未解読テキストを計算機で解読する研究が進められてきた[9,10,11,12]．一方で，漢字系文字の機械学習による解読は未開拓の領域として残されている．これは，漢字特有の課題である文字種の多さ，音価の多様性，表意性への対応が必要とされるためである．

このような状況の中で，朝鮮半島では中国語の漢字音を借りて朝鮮語の音節を表す吏読(イドウ)と呼ばれる表記法[3]が古文書に残されている．この吏読の読み方を応用して万葉集の未解読歌を古い朝鮮語で解読しようとする研究が Vovin によってなされた．Vovin は特に万葉集の 9 番歌を解読し[4]，さらに他の 10 首ほどについて古い朝鮮語との関連性を指摘している[5,6,7,8]．

1.2 目的

本研究の目的は，万葉集において未解読とされている歌を古い朝鮮語により解読することである．具体的には，Vovin が示した手法を計算機処理として実装し，未解読歌を自動的あるいは半自動的に解読することを目指す．計算機に

よる網羅的な解釈可能性の探索を実現することで，解の探索が限定的であるといった従来の手作業による解読の問題点を克服する．人手では扱いきれない膨大な読み方の組み合わせを効率的に処理し，形態素解析に基づく単語列の選択基準を設定することにより，解読過程の客観性と再現性を示す．また，これまで特に意味を持たないとされている枕詞に提案手法を適用し，その意味を解釈することを試みる．

また，本研究は漢字系文字の解読の先駆けとなるものである．従来の非漢字系文字の解読手法を拡張し，漢字系文字に適用することで，漢字特有の課題に対する新たな解決アプローチを提案する．

1.3 本論文の構成

本論文の構成は以下の通りである．第2章では，本研究に関連する先行研究について述べる．第3章では，本研究の提案手法を詳述する．まず，漢字音辞書を作成し，次に漢字音の系列の形態素解析を行い，その結果から最適な単語列を選択する手法を提案する．また漢字音辞書作成にあたり中期朝鮮語音を機械学習によって推定する方法について説明する．第4章では，評価実験について述べる．まず関連研究の成果を提案手法で再現するための予備実験について述べる．その上で未解読歌の解読を試みた実験の結果を報告する．第5章では，提案手法の枕詞の解読への適用可能性について論じる．第6章では本研究のまとめと今後の課題を述べる．

第2章 関連研究

2.1 未解読文字の言語学的解読

2.1.1 非漢字系文字の解読

現在世界で使われている文字の系統は，ギリシャ文字，アラム文字，ブラーフミー文字，漢字に大別できる[13]．本研究では，ギリシャ文字，アラム文字，ブラーフミー文字の系統に属する文字を「非漢字系文字」と定義する．現代でも非漢字系文字における未解読の文字は存在し，線文字 A，クレタ聖刻文字，インダス文字などがある[14]．このような非漢字系の古代文字の解読においては，書かれている言語が明確でない場合がありうる．即ち，1)文字と使用言語が共に不明な場合，2)文字は不明だが使用言語は推定できる場合，3)文字は判明しているが使用言語が不明な場合がある[15]．

2.1.2 漢字系文字の解読

本研究では，漢字を起源とした文字体系，あるいは漢字に影響を受けた文字体系を「漢字系文字」と定義する．漢字系文字にも未解読の文字があり，契丹文字，女真文字，そして多くの甲骨文字などが未解読である[15]．漢字をそのまま使用している又は使用していた言語には，中国語諸語，朝鮮語，ベトナム語，日本語がある．朝鮮半島には，中国語の漢字の音価を借りて朝鮮語の音を表現した表記法として古文書に残る吏読があった[3]．日本では，朝鮮半島での用法である吏読に習って，奈良時代に音読み・訓読みによって表音的に記す万葉仮名として漢字を使うようになった[3,17]．このとき，日本語は複数の漢字によって文字化される可能性がある．例えば，英語の spring に相当する概念は，一般には「春」と書くが，万葉集の 1884 番歌「寒過暖来良思(冬過ぎて春来るらし)」では「暖」と書かれている．語を文字化するにあたり文字の書き方が定まっていることが「正書法」であるならば，日本語は正書法がない言葉であると言える[17]．つまり日本語において漢字を読むということは漢字を正書法なしで解読することに等しい．

2.1.3 未解読文書としての万葉集

万葉集は，漢字で表記される万葉仮名で書かれた 4,516 首の和歌を収める歌集である．しかし，この内 42 首は現在でも解読できていない[2]．

万葉集の中で最難訓歌とされるのが 9 番歌(額田王作)である．9 番歌原文全体

は,「莫囂圓隣之大相七兄爪謁氣吾瀬子之射立爲兼五可新何本」となっている. この内, 前半の 12 文字である「莫囂圓隣之大相七兄爪謁氣」については古来 60 以上の解釈が試みられてきたが未だに定説はない[17]. 最も古い解釈は仙覚の残したものである. 仙覚はこれを「ユフツキノアフキテトヒシ」と読んだ[18]. しかし, 仙覚はこの読み方について, どのような根拠に基づいているのか一切説明していない. そのため後代の万葉学が 9 番歌の前半 12 文字の漢字と仙覚が残した日本語の意味を繋ぐ手がかりを得ることはなく, 多数の解釈の中の一つとして埋もれてきた.

これまで万葉集は万葉仮名により書かれ, 日本語で読めるという前提のもとで解釈されてきた[1]. これに対し, Vovin は, この仙覚の日本語の解釈を「古い朝鮮語」を介して解読できることを示した[4]. 仙覚の解釈を正解として, その日本語の意味に相当する「古い朝鮮語」の単語を探り, その単語を構成する漢字の音を示したのである. これにより「莫囂圓隣之大相七兄爪謁氣」は「ユフツキノアフキテトヒシ」と読めると主張している.

Vovin の手続きを具体的に説明する. Vovin は多くの資料を参照しながら検討を重ねているが, 簡単のため漢字の音から日本語の意味に至る手順を説明する. ここで「古い朝鮮語」とは本来「古朝鮮語・Old Korean: OK」と呼ばれる 14 世紀以前の朝鮮語であり, 言語学における解釈では古朝鮮語として再構築することが求められる. しかし実際には Vovin は解読にあたり, 漢字の音から 15〜16 世紀の「中期朝鮮語・Middle Korean: MK」として単語を生成し, その朝鮮語の意味を上代日本語として解釈した. 古朝鮮語によって意味を解釈できなかったのは, 古朝鮮語に関する情報が著しく限定されているためである.

Vovin は, 万葉集で使われていた漢字は 5 種類の音を表すことがあるとしている.

1) 万葉仮名としての音
2) 吏読の音
3) 中国の後漢・六朝時代の漢字音(Late Han Chinese: LHC)から転じた中期朝鮮語音(Middle Chinese: MK)
4) 中国の隋唐時代の漢字音(Early Middle Chinese: EMC)から転じた中期朝鮮語音(Middle Korean: MK)
5) 正用の音(Orthographic use)
   「正用」とは表意性に基づく用法のことで漢字の意味である字義を表す. 対して中国語による漢字の発音を転用する方法を「借用」という. 但し本研究では日本語ではなく朝鮮語の字義の音である.

Vovin は上記 1), 2), 3), 4)という 4 種類の「借用」の音と, 5)の「正用」の音を用いたことになる. なお LHC と EMC から転じた MK の音は Vovin の推

定によるものである．これにより既存研究のある吏読の音に加え，異なる読みの音を漢字に与えることができるようになった．その上で「莫囂圓隣之大相七兄爪謁氣」のそれぞれの漢字について 5 種類の漢字音の組み合わせを考慮し，中期朝鮮語の単語を生成している．そして，中期朝鮮語の単語に対し，中期朝鮮語から現代韓国語を介し日本語の意味を解釈している．

図 2.1 は Vovin による解釈の処理の流れを示している．まず万葉集の西本願寺本を底本としながらも漢字に校正を施す．次にその漢字音を用意する．その上でその音の列から OK の単語列を推定する．さらに OK から MK へ単語列を置き換えることにより意味を推定する．最後に日本語への翻訳を行い，「ユフツキノアフキテトヒシ」「夕月の仰て問ひし」を得る．

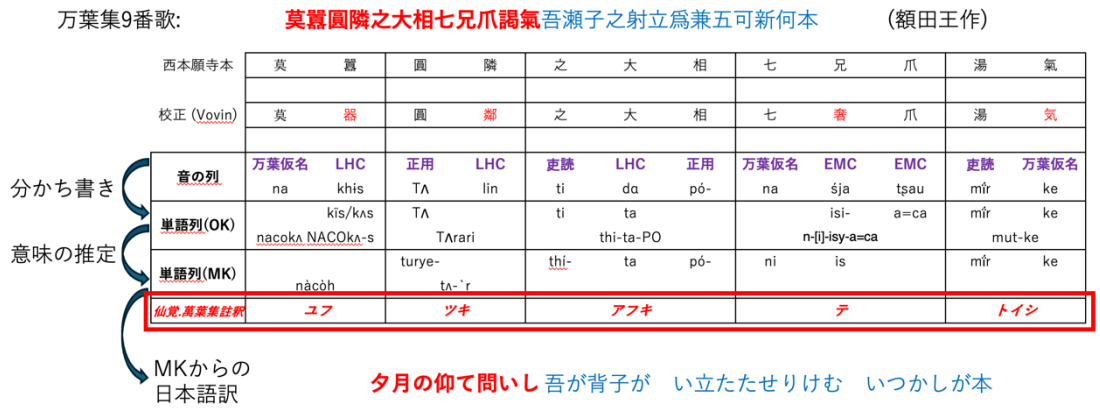

図 2.1 Vovin による万葉集 9 番歌の解釈

Vovin は 9 番歌に関し,「万葉仮名で日本語として読める」ことを当然とせず，一旦解読にあたり文字と使用言語が共に不明であることを仮定した．そして使用言語がアイヌ語であることを排除した上で，これは古朝鮮語からの訳であるに違いないと推測した．その上で 5 つの漢字音の組み合わせを見出し，中期朝鮮語として読めることを証明した．

一方で，未解読歌と認識されている和歌以外にも万葉集には意味が不明な言葉が多く含まれている．それは枕詞である．枕詞はその生成の由来が不明で[19]，訳したくても訳せない[20]とされている．約 500 語あるとも言われる枕詞は調子を整える修辞語であるが意味はないという認識で合意がある[21]．このため枕詞の意味の解釈を試みる研究は進んでいない．

2.2 機械学習による未解読文書の解読

2.2.1 統計的アプローチ

Snyder らは，ウガリット語の単語について，その正しい古ヘブライ語の同根語を推定する手法を提案した [9]．この研究は文字解読研究に機械学習による現代的なアプローチを採用した最初の論文であった．

彼らの方法は，訓練データとして既知の関連言語の非並列コーパスを用意し，アルファベットのマッピングと対応する同根語への単語の翻訳の両方を生成するモデルを学習する．低レベルの符号(アルファベット)のマッピングと高レベルの形態素の対応の両方を同時に捉えるためにノンパラメトリックベイズの枠組みを採用している．実験では，ウガリット語の 30 個の符号のうち 29 個について古ヘブライ語に対応する符号に正しく対応付けすることができた．また，古ヘブライ語に同根語を持つウガリット語の単語の 60%について，その正しい古ヘブライ語の同根語を推定した．表 2.1 は Snyder による実験結果を示している．Baseline はウガリット語と古ヘブライ語の翻訳モデルを比較的シンプルな EM アルゴリズムによって実現した Knight らの手法[22]，Our Model は Snyder の手法，No Sparsity は Snyder の手法で構造的スパース性事前分布を用いないモデルを表す．Words と Morphemes は単語単位，形態素単位の同根語推定の正解率であり，Type と Token は単語の種類数と出現数(同じ単語がテキストに何回出現するか)を正解率の分母としていることを示す．なお，Baseline では形態素の境界を予測しないため Morphemes の正解率は算出できない．

表 2.1 Snyder らの実験結果[9]

	Words		Morphemes	
	Type	Token	Type	Token
Baseline	28.82%	46.00%	N/A	N/A
Our Model	**60.42%**	66.71%	75.07%	81.25%
No Sparsity	46.08%	54.01%	69.48%	76.10%

Tamburini は，ウガリット語から古ヘブライ語の同根語を推定する新たな手法を提案した[23]．組合せ最適化とシミュレーテッドアニーリング（高度な非凸最適化手続き）に基づく古代文字解読問題への新しいアプローチを提示している．解は，符号間の null，一対多，多対一のマッピングを可能にする k 乗法を用いて符号化される．提案システムは，標準的な評価による最先端のシステ

ムと比較した場合，同族識別においてより優れた結果を得ることができたと主張している．しかし，Luo ら[11]を state-of-the-art：SOTA として比較しており，より良い結果を導き出している Luo ら[12]との比較は行っていない．

2.2.2 ニューラルアプローチ

Luo らは，Snyder ら[9]と同じくウガリット語の単語を対象とした古ヘブライ語の同根語推定における新しい手法を提案した[11]．その手法は，言語間の高いマッチング性能を有し，失われた言語(ウガリット語)の自動解読のためのニューラルアプローチである．彼らのモデルは，ニューラルネットワークの学習に必要な強力な教師信号の欠如を補うために，歴史言語学によって文書化された言語変化の既知のパターンを含むように設計されている．まず文字レベルの対応付けのため，双方向 Recurrent Neural Network (RNN) を用いて失われた言語と既知の言語の単語を文字レベルでエンコードし，文字間の対応関係を学習する．次に単語レベルの対応付けのため，文字レベルの対応付け結果である埋め込みを編集距離に基づいてコスト計算した上で，単語間の対応付けを最小コストフロー問題として定式化し，解を探索する．

ここで最小コストフロー法は，単語レベルの一対一対応を効率的に制御し，適切な数の同源語(同根語)ペアを特定するために採用された．このアプローチにより，構造的疎性(Structural Sparsity)および十分な語彙カバレッジ(Significant Cognate Overlap)という言語学的制約を満たしつつ解を探索することが可能となった．

構造的疎性とは，失われた言語と既知の言語の単語レベルでの対応がほぼ 1 対 1 であることが歴史言語学で知られていることを意味する[24]．これは，同じ祖語から派生した語彙(同根語)が，基本的に 1 対 1 で対応する傾向があるためである．最小コストフローは，ネットワークにおけるフローの量を制限することで，この疎な関係を自然に表現できる．

十分な語彙カバレッジとは，関連言語間では対応する語彙（同族語）が一定程度存在することを意味する．最小コストフローでは，対応付けられるべき同根語のペアの総数をハイパーパラメータによって調整する．このパラメータを適切な値に設定することで，対応付けられる同族語の数が極端に少なくなることを防ぎ，失われた言語の語彙を幅広くカバーするようにモデルを学習させることができる．

Luo らが用いたウガリット語と古ヘブライ語の単語対である学習データセットについて説明する．図 2.2 は古ヘブライ文字とウガリット文字の例である．

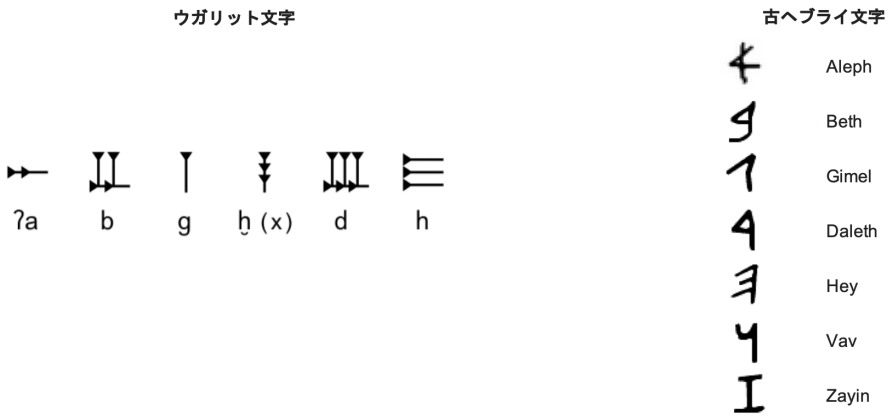

図 2.2 ウガリット文字と古ヘブライ文字[11]

また図 2.3 に示すように，ウガリット文字と古ヘブライ文字をそれぞれアルファベットに置き換えた同根語対を学習データとして使用している．

UGARITIC	Hebrew
msgr	msgr
win	wayn
brr	brr|brwr
nxl	nHl
kbkb	kwkb
agr$n	agr$nw
yrq	yrq
abm	abwt
n@ry	nSry
$n	$n

図 2.3 ウガリット語(UGARITIC)と古ヘブライ語(Hebrew)の同根語対の例[11]

Luo らが実験に用いたデータセットは Snyder らが作成したもの[9]であり，ウガリット語と古ヘブライ語の同根語ペアが 2,214 レコード収録されている．表 2.2 は Snyder らと Luo らの実験結果の比較であり，Luo らは Snyder らより良い正解率の結果を得た．

表 2.2 Luo らの実験結果 [11]

	Ugaritic - Hebrew 同根語変換正解率
Bayesian (Snyder ら[9])	0.604
Neurocipher (Luo ら[11])	0.659

さらに，ウガリット語とヘブライ語の組に加えて，Linear B 文字と古代ミケーネ語ギリシア語文字の組のデータセットに対して提案手法を適用したところ，良好な同根語のマッピング結果が得られたと報告している．

また，Luo らは，テキストが完全に単語に分かち書きされていない場合を新たに設定し，未知語(失われた言語の言葉)の分かち書きと未知語に対応する既知語の同根語推定を同時に行う手法も提案している[12]．分かち書きが十分にされていない言語であるゴシック語とイベリア語のデータセットを用いて同根語を推定する実験を行った．さらに，既存研究との公平な比較のため，分かち書きされてはいるがウガリット語を対象とした実験も行った．

単語の境界が不明なテキストに対応するというより困難な状況に取り組むためには，歴史的な音の変化の一貫したパターンを反映する豊富な言語的制約をモデル化する必要がある．そこで国際音声学アルファベット（International Phonetic Alphabet; IPA）に基づき，音韻的特徴を IPA 埋め込みとして学習する．この IPA 埋め込みを用いることで，文字の音韻的な特徴に基づいて単語の境界を推測する．例えば，ある音韻パターンが特定の単語の開始や終わりに現れやすいというような情報があれば，それを手がかりに単語の区切りを推測できる．

IPA シンボルで表される音素は，音声器官や発声方法に基づいて分類されている．例えば，図 2.4 に示すように， *b*， *p* という音素は VOICING(有声性)，MANNER(調音法)，PLACE(調音点)について異なる特徴を有する．

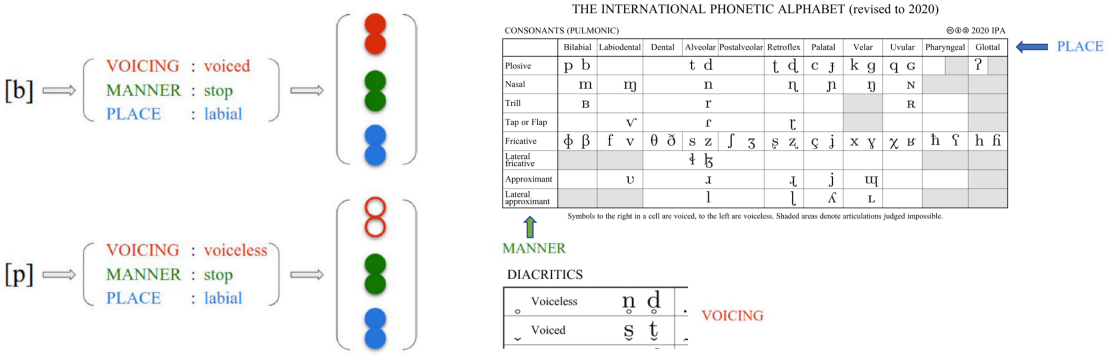

図 2.4 IPA シンボルが有する特徴情報の例

これらの音声学的特徴を定量的に扱うため，まず各特徴を二値ベクトルとして表現する．例えば，子音[b]は式(1)のように表される．

$$[\mathrm{b}] \Rightarrow \begin{bmatrix} VOICING: voiced & : & +1 \\ MANNER: stop & : & +1 \\ PLACE: labial & : & +1 \end{bmatrix} \quad (1)$$

この表現により，音素間の類似性を定量的に評価することが可能となる．例えば，[b]と[p]は調音法(stop)と調音点(labial)が共通しており，有声性のみが異なることが明確に表現される．ここで，この音素特徴の類似性を活用して，未知の文字と既知の文字との対応関係を確率的にモデル化する．例えば古ヘブライ語を既知言語（K），ウガリット語を未知言語（L）とすると，既知言語の文字 c^{K} と未知言語の文字 c^{L} の対応確率を式(2)のように定義できる．

$$Pr(c^{\mathrm{L}} \mid c^{\mathrm{K}}) \propto exp(E(c^{\mathrm{L}}) \cdot E(c^{\mathrm{K}})/T) \quad (2)$$

ここで式中の記号の意味は以下の通りである．

- $E(\cdot)$ は特徴ベクトルの埋め込み関数
- T は確率分布の鋭さを制御する温度パラメータ

未知言語の文字 c^{L} の埋め込み表現は，既知言語の文字の重み付き和として式(3)のように表される．

$$E(c^{\mathrm{L}}) = \Sigma_{\mathrm{i}} w_{\mathrm{i}} \cdot E(c^{\mathrm{K}}{}_{\mathrm{i}}) \quad (3)$$

ここで w_{i} は学習可能なパラメータであり，このパラメータの学習を通じて各文字の最適な表現が獲得される．具体的には，このパラメータの調整により音韻的に近い音素（例：[b]と[p]）は埋め込み空間で近い位置に配置され，特徴の類似性が自然に反映される．これにより音変化の妥当性を考慮した文字間のマッピングが可能になる．

まとめると，式(1)で表現された特徴ベクトルは，式(2)の確率計算を通じて対応関係が決定され，最終的に式(3)による埋め込み表現が得られる．これにより，IPA 埋め込みは音声発声の幾何学的構造を符号化した表現になる[25]．

但し，Luo らは，未知語であるウガリット語の単語音には多くノイズを含むとし，IPA 転写せずアルファベット表記による正書法のままとし，既知語である古ヘブライ語の単語に対してのみ IPA 転写を行った[12]．

Luo らは，この IPA 埋め込みに基づいて文字間の音韻的な類似度を計算し，その類似度を編集距離のコストとして，即ち，音韻的事前知識として利用した．コストは動的計画法で利用され，入力テキストの分かち書きと既知の単語

と対応付けを同時に最適化するため，累積コストを最小化するような経路を探索する．

ウガリット語と古ヘブライ語のデータセットを用いた実験の結果を表 2.3 に示す．Base が提案手法，その他は先行研究の手法を表す．IPA 埋め込みに基づく提案手法は論文発表の時点で SOTA を達成している．

表 2.3 Luo らの実験結果[12]

	Ugaritic - Hebrew 同根語変換正解率
Bayesian (Snyder ら[9])	0.604
Neurocipher (Luo ら[11])	0.659
Base (Luo ら[12])	**0.778**

2.2.3 漢字系諸語の未解読文字の機械学習による解読

漢字系諸語における未解読漢字について機械学習を応用して解読するという研究は見当たらない．Luo は漢字の取り扱う上での困難な課題として次の 3 点を挙げている[26]．1)漢字はアルファベットに比べ文字種が圧倒的に多い．2)漢字は文字と音価の対応が一意ではない．3)漢字は表意文字であるので意味を考慮する必要がある．これらは漢字系の未解読文字の解読に機械学習の手法がこれまで適用されていなかった要因の一部であると考えられる．

第3章 提案手法

3.1 概要

万葉集の未解読歌を解読するために Vovin の手法を計算機処理により実現することを提案する．提案手法の概要を図 3.1 に示す．まず解読対象となる漢字列を入力する．本研究では万葉集のテキストとして西本願寺本を底本とした公開データ[27]を用いる．次に漢字音辞書を用いて漢字列を複数の音の系列に変換する．個々の音の系列に対して形態素解析を行い，単語列を得る．形態素解析用の辞書として古い朝鮮語の辞書を用いる．すなわち，音の系列を朝鮮語の単語列に変換する．得られた複数の単語列の候補の中から適切な候補を選択する．以上の処理は自動的に行う．その後，人手によりそれぞれの単語列の候補を日本語へ翻訳し，未解読歌の解読を試みる．

なお，本研究において，「古い朝鮮語」は「中期朝鮮語」とする．従って中期朝鮮語としての単語列の生成までを本研究の対象範囲とし，Vovin が行ったような古朝鮮語の再構築は本研究の対象範囲外とする．中期朝鮮語より古い朝鮮語は辞書として使える情報が少ないためである．また，Vovin は西本願寺本のテキストに対して校正を施していたが，このような漢字の置換は本研究の範囲を超えているため，公開データのテキストをそのまま使用することとする．

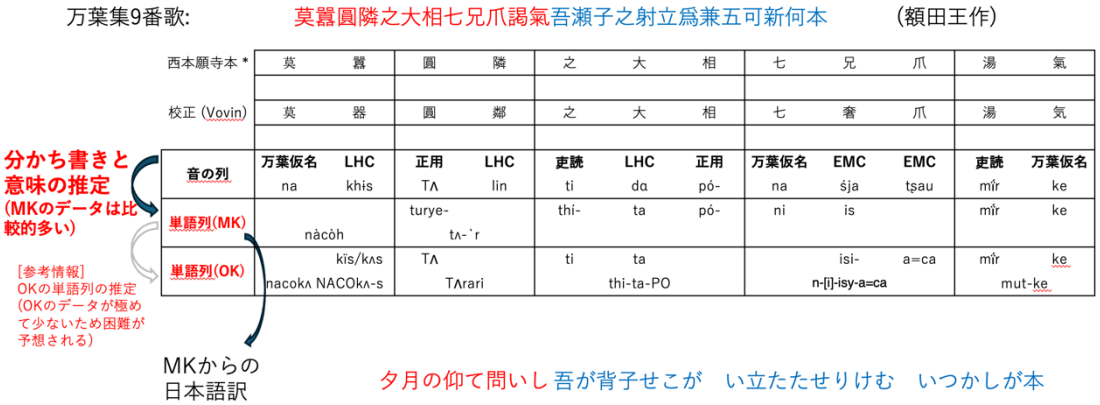

図 3.1 提案手法の概要

図 3.2 に上記の一連の処理を流れ図として示す．

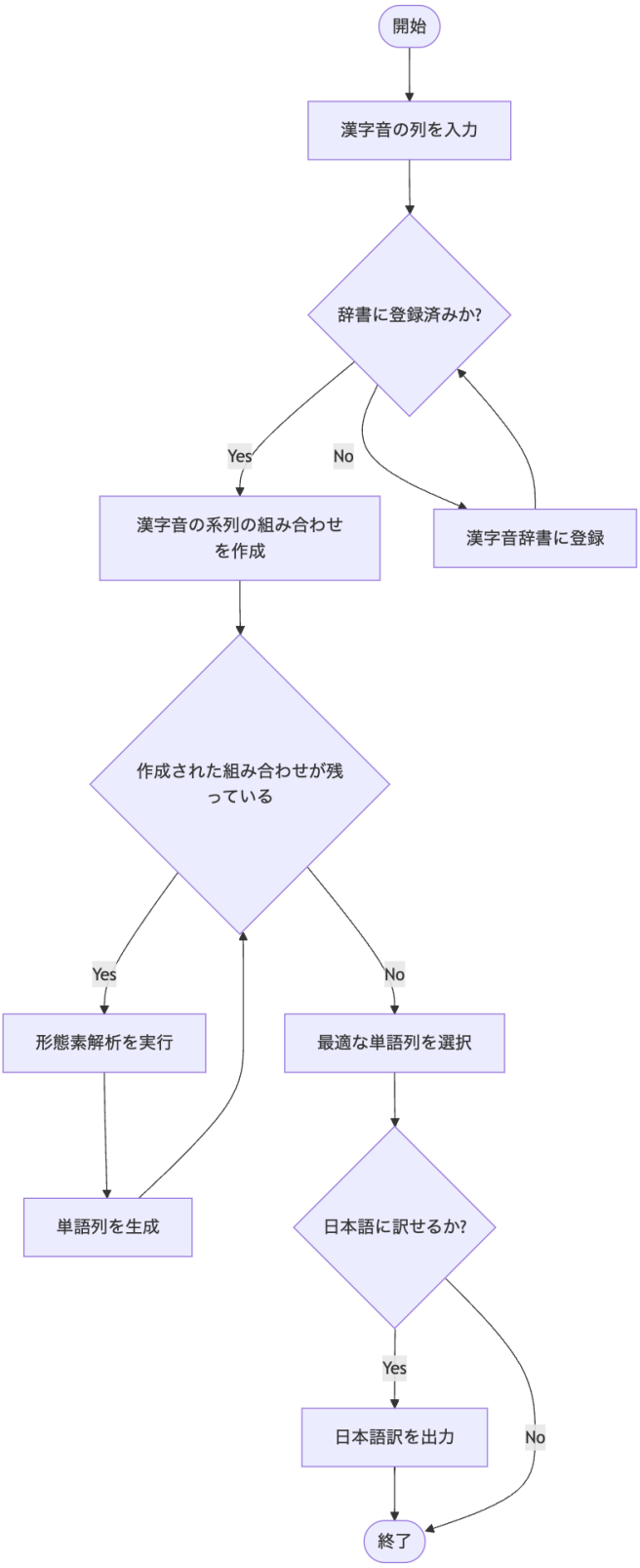

図 3.2 解読処理の流れ図

この流れ図をもとに処理の流れを改めて説明する．解読処理の第一段階として，システムは漢字列を入力として受け取る．この際，西本願寺本を底本とした万葉集テキストを使用する．第二段階では，入力された漢字列における個々の漢字の辞書登録状態を確認し，未登録の場合には漢字音辞書への登録を行う．漢字音辞書には漢字音５種類(詳細は後述する)を資料に基づいて登録する．その際，資料にはない中国の後漢・六朝時代の漢字音(Late Han Chinese: LHC)から転じた中期朝鮮語音(Middle Koerean: MK)，及び中国の隋唐時代の漢字音(Early Middle Chinese: EMC)から転じた中期朝鮮語音(Middle Korean: MK)については，「中期朝鮮語音の推定」(詳細は3.2節で述べる)を行い漢字音辞書に登録する．第三段階では，漢字音辞書を参照し，可能な漢字音の系列の組み合わせを生成する．第四段階では，生成されたそれぞれの音の系列に対して形態素解析を実行し，単語列を得る．正確には，前段階で生成される音の系列には重複があるため，音の系列に対する既存の処理結果の有無を確認し，処理結果がない場合には形態素解析を実行し，その音の系列に対する単語列を得る．第五段階では，得られた多数の形態素解析の結果(単語列)の中から解釈可能性のある単語列の候補を選択する．第六段階では，得られた単語列の日本語への翻訳可能性を判定する．不可能な場合は処理を終了する．第七段階として，翻訳可能と判定された場合は日本語訳を生成する．

以降の説明で使用する用語を表 3.1 にまとめる．また，「形態素」という単語は本来は単語よりも細かい言語単位を指すが，本研究では単語と同じものを指す．「形態素解析」や後述する「形態素数最小法」など自然言語処理の分野で長年使われてきた専門用語を使うときのみ「形態素」という語を用いる．

表 3.1 本研究で使用する用語

用語	説明
万葉仮名	中国語の漢字音を借りて上代日本語の音節を表す表記法であり「借用」とも言われる
吏読	中国語の漢字音を借りて朝鮮語の音節を表す表記法
LHC (Late Han Chinese)	1〜2 世紀後漢時代中国語
EMC (Early Middle Chinese)	7〜9 世紀隋唐時代の中国語
正用	漢字の意味である字義を表す表記法 本研究では日本語ではなく朝鮮語の字義
OK (Old Korean)	14 世紀以前の古朝鮮語
MK (Middle Korean)	15〜16 世紀の中期朝鮮語

3.2 中期朝鮮語音の推定

前節で述べたように，既存の漢字音辞書に中国語音(LHC または EMC)は登録されているが中期朝鮮語(MK)の音が登録されていないとき，LHC または EMC の音から MK の音を推定する手法を提案する．LHC，EMC，MK はいずれも IPA シンボルによって表記可能であることに着目し，IPA シンボル列の各 IPA シンボルを音素に展開して，音を音素の系列として表す．その上で，LHC または EMC の音素の系列を MK の音素の系列に変換する系列変換モデルを学習する．モデルを学習する際には，IPA シンボルによって表記された同一漢字に対応する中国語音と朝鮮語音のペアデータを訓練データとして使用する．

各音素は 2 種類のベクトルで表現する．ひとつは，声性，調音位置，調音方法などの音韻素性を特徴量ベクトルとして表現した IPA 特徴量ベクトルである．もうひとつは IPA シンボルの one-hot ベクトルである．

音素の系列変換モデルとして Long Short-Term Memory (LSTM)と Attention メカニズムを基盤としたエンコーダ/デコーダモデルを用いる．具体的には以下の 4 種類のモデルを学習する．

(1) 双方向 LSTM と Attention 機構を用い，エンコーダ・デコーダ共に IPA 特徴量のベクトルを扱うモデル (BiLSTM-IPA-IPA)
(2) 双方向 LSTM と Attention 機構を用い，エンコーダは IPA 特徴量，デコーダは one-hot ベクトルを扱うモデル (BiLSTM-IPA-OH)
(3) 双方向 LSTM と Attention 機構を用い，エンコーダ・デコーダ共に one-hot ベクトルを扱うモデル (BiLSTM-OH-OH)
(4) 単方向 LSTM を用い，エンコーダは IPA 特徴量，デコーダは one-hot ベクトルを扱うモデル (LSTM)

これらのモデルはいずれも，IPA 特徴量のベクトルもしくは one-hot ベクトルを用いて，IPA シンボルで表記された中国語音を一旦音素列に展開し，変換後の音素列を IPA シンボルで表記された朝鮮語音として復元するモデルの学習を目指すものである．

音素のベクトル(IPA 特徴量ベクトル，one-hot ベクトル)の詳細について述べる．IPA 特徴量ベクトルは，IPA 素性特徴が定義されている UNICODE_TO_IPA(ipapy ライブラリー[46])を参照し，式(4)のようなバイナリベクトルと定義する．

$$\varphi(x) = [v1,\ldots,vV;\ p1,\ldots,pP;\ m1,\ldots,mM;\ h1,\ldots,hH;\ b1,\ldots,bB;\ r1,\ldots,rR;\ d1,\ldots,dD;\ s1,\ldots,sS;\ l1,\ldots,lL;\ br1,\ldots,brBR;\ t_l1,\ldots,t_lL;\ tc1,\ldots,tcC;\ tg1,\ldots,tgG] \in Rd \quad (4)$$

ここで各記号の意味は以下の通りである.

- $\mathrm{v_c} \in \{0, 1\}^{|\mathrm{V}|}$ は子音の声性（voicing）
- $p_c \in \{0,1\}^{|P|}$ は子音の調音位置（place）
- $\mathrm{m_c} \in \{0,1\}^{|\mathrm{M}|}$ は子音の調音方法（manner）
- $h_v \in \{0,1\}^{|H|}$ は母音の高さ（height）
- $b_v \in \{0,1\}^{|B|}$ は母音の舌の前後位置（backness）
- $r_v \in \{0,1\}^{|R|}$ は母音の円唇性（roundness）
- $d \in \{0,1\}\text{^}|D|$ は補助記号（diacritics）
- s, l, br は超分節音素（suprasegmentals）
  - ✧ s: 強勢（stress）
  - ✧ l: 長さ（length）
  - ✧ br: 休止（break）
- t_l, t_c, t_g は声調（tones）
  - ✧ t_l: 音高レベル（tone level）
  - ✧ t_c: 音高の輪郭（tone contour）
  - ✧ t_g: グローバルな音高変化（global tone）

V, P, M, H, B, R, D はそれぞれの特徴量の集合である.

一方，IPA 音素の one-hot ベクトルは式(5)のように定義する.

$$onehot(x) = e_i \in \{0,1\}\text{^}|V| \tag{5}$$

ここで，$|V|$ は語彙サイズ（全 IPA シンボル数 + 特殊トークン）で，e_i は i 番目の要素のみが 1 で他が 0 のベクトルである.

入力系列 X と出力系列 Y は以下のように表される.

$$X = \left(x_1, \ \ldots, \ x_T\right), \quad x_t \in \phi(x) \ or onehot(x) \tag{6}$$

$$Y = \left(y_1, \ \ldots, \ y_T\right), \quad y_t \in \phi(y) \ or onehot(y) \tag{7}$$

先に述べた 4 つのエンコーダ/デコーダモデルの詳細を表 3.2 にまとめる.

表 3.2 中期朝鮮語音推定モデル

設定＼モデル名	BiLSTM-IPA-IPA	BiLSTM-IPA-OH	BiLSTM-OH-OH	LSTM
エンコーダ LSTM	**双方向**	**双方向**	**双方向**	**単方向**
デコーダ LSTM	Attention 付き	Attention 付き	Attention 付き	Attention なし
エンコーダのベクトル	**IPA 特徴量ベクトル**	**IPA 特徴量ベクトル**	**one-hot ベクトル**	**IPA 特徴量ベクトル**
デコーダのベクトル	**IPA 特徴量ベクトル**	**one-hot ベクトル**	**one-hot ベクトル**	**one-hot ベクトル**
パディング	0 でパディング	0 でパディング	0 でパディング	0 でパディング
損失関数	CrossEntropyLoss	BCEWithLogitsLoss（カスタム)	BCEWithLogitsLoss（カスタム)	BCEWithLogitsLoss（カスタム)

ここで one-hot エンコードを用いた３つのモデルでは，損失関数として BCEWithLogitsLoss を用いる．BCEWithLogitsLoss は one-hot ベクトル形式の正解ラベルと組み合わせることを想定した損失関数である．ただし，本研究ではこれを多クラス(IPA 文字の全タイプ)に適用するためのカスタマイズを行っている．

モデルの学習にあたり，ハイパーパラメータの最適化をベイズ最適化手法により行った．探索空間 Θ は，連続的なドメインと離散的なドメインの直積空間とする．

$$\Theta = \Theta_{hidden} \times \Theta_{embed} \times \Theta_{layers} \times \Theta_{dropout} \times \Theta_{batch} \times \Theta_{lr} \tag{8}$$

ここで，各部分空間は以下のように定義される．

$$\Theta_{hidden} = [64,\ 512] \quad \text{(continuous)} \tag{9}$$
$$\Theta_{embed} = [32,\ 256] \quad \text{(continuous)} \tag{10}$$
$$\Theta_{layers} = \{1,\ 2,\ 3\} \quad \text{(discrete)} \tag{11}$$
$$\Theta_{dropout} = [0.\ 1,\ 0.\ 5] \quad \text{(continuous)} \tag{12}$$
$$\Theta_{batch} = \{16,\ 32,\ 64\} \quad \text{(discrete)} \tag{13}$$
$$\Theta_{lr} = [10^{-4},\ 10^{-2}] \quad \text{(log-scale)} \tag{14}$$

これにより最適化問題は次のように定式化される．

$$\theta * = argmin[\theta \in \Theta] f(\theta) \tag{15}$$

$$f(\theta) \sim GP\left(\mu(\theta),\ k\left(\theta,\ \theta'\right)\right) \tag{16}$$

ここで，$f(\theta)$ はハイパーパラメータθに対する検証損失(validation loss)である．

3.3 漢字音辞書の作成

漢字音辞書には，それぞれの漢字に対し，万葉仮名，吏読，LHC，EMC，正用の５種類の音を登録する．万葉集で使われている漢字の数(タイプ数)は 1,964 である．表 3.3 に漢字音の辞書作成に用いた資料の出典を示す．但し中期朝鮮語時代のハングルは古ハングルを使用している[28]．

表 3.3 漢字音の出典

漢字音辞書に登録する漢字音の種類	出典の著者名とその漢字音表記法
1．万葉仮名	Vovin[5]にある Vovin 式ローマ字
2．吏読	吏読と万葉仮名の研究[3]にある Yale 式ローマ字, 朝鮮吏読辞典[29]にある古ハングル, 吏讀辭典[30]にある古ハングル 古語辞典[31]にある古ハングル, 記紀万葉の朝鮮語[32]にある古ハングル
3．LHC から推定する MK	漢字音：簡約上古音和東漢音[33]にある IPA 中期朝鮮語音：訓蒙字会[34]にある古ハングル
4．EMC から推定する MK	漢字音：周法高上古音韻表[35]にある IPA 中期朝鮮語音：訓蒙字会[34]にある古ハングル
5．正用	古語辞典[36]にある古ハングル

漢字に対して正用の音を与えるには，まずその漢字の意味を把握し，その意味に該当する中期朝鮮語音を南の辞書[36]に見出す必要がある．そして辞書の第一項目にある「意味・意義」に相当する音を採用する，という手続きを取る．複数の音の中から代表的なものを選んで辞書に登録していることになるが，これによって特段の恣意性が働くことはないと考える．ただし，機械的な手続きには成り難いことから，他の４種類の漢字音と同様に必要最小限の音を登録するに留めた．

計算機処理に用いる漢字辞書における漢字音の表記方法は，3.4 節で述べる形態素解析器の仕様に従って，福井式ローマ字表記[37]の改訂版である須賀井

式ローマ字表記を採用する[38]．よって，出典における文字表記から須賀井式ローマ字へ変換が必要となる．変換のために用いた対応表は表 3.4 に示す資料から作成し，自動処理と人手処理の併用で変換作業を行う．

表 3.4 須賀井式ローマ字への変換に用いた資料

出典にある 漢字音表記の種類	須賀井式ローマ字変換のための資料
Vovin 式ローマ字	mecab-k2alpha syllables.csv [39]
古ハングル	mecab-k2alpha syllables.csv [39]
Yale 式ローマ字	Wikipedia Revised Romanization of Korea [40] mecab-k2alpha syllables.csv [39]
IPA	Wikipedia IPA/Korean [41] mecab-k2alpha syllables.csv [39]

後続の処理では漢字音辞書を用いて漢字列を音の系列に変換するが，組み合わせ爆発により音の系列の候補数が増大し，処理の効率が悪くなる可能性がある．例えば，9 番歌の未解読部分である 12 文字の場合，最大 5^{12}=244,140,625 個の音の列が得られる．そのため，漢字音辞書作成の際に 5 種類全ての漢字音を登録するのではなく，その一部のみを登録することによって組み合わせ爆発を抑えることも検討した．しかし，未解読歌の解読のためには解釈可能な音の列をできるだけ残す方が適切であると考え，全ての漢字について 5 種類全ての漢字音を登録することにした．ただし，使用した漢字音の出展は万葉集で使われている全ての漢字を網羅しているわけではないため，実際には全ての漢字に 5 種類の漢字音が登録できたわけではなかった．

漢字音辞書は tsv ファイルとして作成した．図 3.3 は作成した漢字音辞書の抜粋である．左端の列の漢字の右列にタブ区切りで漢字音である IPA シンボルを配置している．

丸 'oa
丹 da ni in an
主 ju
乃 no na nei noi nu 'ai n in
久 gu ku ggugo
之 si no da ji jiq ti i il ili
乍 tutu
乎 'ue 'o 'on 'oi ho hoi 'o'on'ol

図 3.3 漢字音辞書ファイルの抜粋

漢字音辞書の統計情報として登録された音の数毎の漢字のタイプ数を表 3.5 に示す．登録音数が少ない漢字タイプが多くあるのは，4 章で後述する実験での解析対象となる漢字についてのみ辞書を構築しているためである．また，6 以上の登録音数がある漢字が存在するのは，万葉仮名の音のように 1 つの種類につき複数の音があり得るためである．

表 3.5 漢字音辞書の統計情報

登録音数	0	1	2	3	4	5	6	7	8	9	合計
漢字数	1,151	414	173	123	59	24	7	8	3	2	1,964

3.4 音の系列の形態素解析

本節では，漢字音辞書を用いて漢字列を音の系列に変換した後，音の系列に対して形態素解析を行い，朝鮮語の単語列を得る処理について説明する．形態素解析器には MeCab[42]を使用する．中期朝鮮語辞書を有する唯一の形態素解析器であるからである．MeCab 用の中期朝鮮語辞書である MkHanDic[43]は 9,653 語を収容しており，11,315 語を収録した中期朝鮮語辞書である南の辞書[36]に匹敵する規模の辞書である．しかしながら，MkHanDic はその収録語を『釈譜詳節(1447)』『阿弥陀経諺解(1464)』『金剛経諺解』『三綱行実図諺解』『月印釈譜』から採っており，内容が仏典に偏る．MkHanDic の作者である須賀井も将来の登録数拡大の必要性を認めている．よって中期朝鮮語の形態素解析器として必要十分な単語が登録されているとも言い難いが，現時点で他に利用可能な計算機用辞書は存在しない．また MkHanDic の辞書はバイナリファイルだけが公開されており，ソースファイルは非公開である．したがって，自身で新しい単語を追加登録してから MeCab 用辞書をコンパイルすることはできない．

MeCab は辞書と学習用コーパスから学習した連接の強度などから形態素解析を行う．また辞書に登録されていない項目については，その品詞を「未知語」として出力するか，MeCab による品詞の推定結果を出力するかを選択できる．本研究では後続の処理に品詞の情報は使わないため，辞書に未登録の語の品詞を「未知語」として出力させる．

3.5 最適な単語列の選択

基本的に MeCab は最適解を選び出力するだけであり，正解を保証するもので

はない．また MeCab に投入する音の系列が膨大にあるということは，即ち MeCab からの出力である形態素解析結果も大量な数に上ることを意味する．よって MeCab からの出力である形態素解析結果から最適な単語列を選択する必要がある．

研究の初期の段階では，解析結果から得た古ハングルの文字列を中期朝鮮語から現代朝鮮語に翻訳する翻訳器[44]にかけ，そこから得られた現代語のハングル文字列の「自然さ」を評価し，そのスコアが最大の単語列を選択する手法を検討した．しかし，指定されたサイトから翻訳器をダウンロードできなかった．そのため，現代朝鮮語に翻訳してその流暢性を測ることで最適な単語列を選択するアプローチは断念した．

本研究では，MeCab による形態素解析の後処理として次の二つを行うことで最適な単語列を選択する．

最初の処理は文法的制約を考慮したフィルタリングである．ここでは，未知語を含む単語列や，明らかに文法的な誤りを含むと思われる単語列を除外する．後者について，具体的には表 3.6 に示す条件を満たす単語列は文法的に正しくないとみなして除外する．これらの条件は中期朝鮮語の文法規則[46]を参考に設計した．ただし，表 3.6 で ＊ をつけた除外条件は，文法的に正しい単語列を誤って除外する可能性がある．ただし，多くの場合には文法的に誤った単語列が除外され，最適な単語列の候補の絞り込みに有効であるため，採用している．

表 3.6 文法的制約を考慮した単語列のフィルタリング

除外する条件	理由
語尾で始まる	語尾は文節先頭に出現しない
指定詞で始まる ＊	特定の物や人を指し示す言葉だから
依存名詞で始まる	依存名詞は主に助詞を伴って文中で役割を果たすとされるから
数詞で終わる ＊	数詞は通常語尾に来ない
連続する語尾を含む	語尾は連続しない
未知語を含む単語列 ＊	未知語を含む事により解釈が困難なため

2 番目の処理は形態素数最小法による単語列の絞り込みである．形態素数最小法とは．複数の単語列の候補の中から，その形態素数(ここでは単語数)が最も小さい解を選択する手法である．形態素数最小の解が複数存在する場合は．その全てを解の候補として残す．

ただし，この一連の処理は根本的なジレンマを抱えている．理想的には一意に解釈可能な単語列を自動的に求めたい．すなわち多数の単語列の候補の中から

正しいものを 1 つ選択したい．しかし，詩歌の色彩が濃い万葉集の一節が分析の対象であり，その適切な解釈を表す単語列が必ずしも形態素数最小の解であるとも限らない．すなわち，形態素数最小の解のみを残すことによって，正しい解釈を表す単語列を除外する危険性がある．そのため，本研究では形態素数が最小の候補だけでなく，それが閾値 T より小さい全ての単語列の候補を選択する．例えば，T=5 と設定した場合，形態素数が 5 以下の単語列を全て解釈の可能性がある単語列の候補として残し，これらについて人手による解釈を試みる．

3.6 解読処理の擬似コード

図 3.4 は，漢字音辞書が用意できている前提のもとに，入力された漢字列から可能な読みの組み合わせを生成し，形態素数が少ない単語列を生成(提示)するアルゴリズムの擬似コードである．アルゴリズムの処理は大きく 3 つのフェーズから構成される．

まず初期化フェーズ(2-7 行目)では，提示する単語列の最大形態素数(単語数) MAX_MORPHEMES を設定する(この擬似コードでは例として 5 を設定している)．読みの集合 P と結果セット results を空集合として初期化する．その後，入力された漢字列の各文字について，辞書を参照して可能な読みを取得する．

次の分析フェーズ(8-13 行目)では，取得した読みの直積を計算し，得られた各組み合わせに対して形態素解析を実行する．この過程で有効な分析結果 (単語数が MAX_MORPHEMES 以下の単語列)が得られた場合，その結果を結果セットに追加する．これにより，入力された漢字列に対する可能な全ての形態素解析結果を収集する．

最後の可視化フェーズ(14-17 行目)では，解析結果を単語数の少ない順にソートし，各分析結果に対してパス図を生成する．形態素情報を含むノードを作成し，ノード間にリンクを張って接続関係を明示することで，形態素解析の結果を視覚的に表現する．これにより，分析結果の理解が容易になり，また異なる分析結果の比較も直感的に行うことが可能となる．

Algorithm 1 Minimal Morpheme Analysis

Require: Kanji String K, Dictionary D
Ensure: Morphme analysis results and path diagram

```
function MinimalMorphemeAnalysis(K, D)
    const MAX_MORPHEMES ← 5                    ▷ Maximum morpheme count
    P ← ∅                                      ▷ Set of readings
    results ← ∅                                ▷ Set of results
    for Each Kanji k in K do
        P ← P ∪ GetReadings(k, D)
    end for
    for Each reading p in CartesianProduct(P) do
        (nodes, count) ← MorphemeAnalysis(p)
        if nodes ≠ ∅ and count ≤ MAX_MORPHEMES then
            results ← results ∪ {(p, nodes, count)}
        end if
    end for
    sortedResults ← SortByMorphemeCount(results)
    for Each result (p, nodes, count) in sortedResults do
        GeneratePath(K, p, nodes, count)
    end for
end function
function GeneratePath(K, p, nodes, count)
    Create path graph G with label "K/p (Number of Morphemes: count) "
    for i ← 1 to Length(nodes) do
        nodes[i] Add nodes with morphological information
        if i > 1 then
            Connect to previous node
        end if
    end for
    Output path diagram
end function
function SortByMorphemeCount(results)
    sorted ← results sorted by (count) ascending
    return sorted
end function
```

図 3.4 単語列生成アルゴリズム

第4章 実験・評価

4.1 中期朝鮮語音の推定の評価

本節では, 中国語音(LHC または EMC)から中期朝鮮語音を推測するモデルの評価実験について述べる.

4.1.1 実験設定

データセットに関して述べる. まず朝鮮語音に関して, 中期朝鮮語音は訓蒙字会[34]を用いる. 訓蒙字会は, 崔世珍が嘉靖 6 年（1527 年）に著した漢字学習書である. 約 3,600 の漢字にハングルで音と義を示したもので, 発音辞書である東国正韻(1447 年)と並んで中期朝鮮漢字音の重要な資料でとされる. 但し訓蒙字会に記載されている漢字音表記はハングルによるものであるため, これを IPA シンボルに変換する必要がある. 本実験では資料[41]に基づいた変換テーブルを用い機械的に変換した.

なお万葉集で使われている漢字 1,964 タイプの内, 958 文字は訓蒙字会に含まれる. これは訓蒙字会を訓練データとして使用するのみならず, 漢字音辞書のデータとして直接使うことができることを意味する. 逆に言えば, 1,006 タイプの万葉集にあるが訓蒙字会には無い漢字は, 中期朝鮮語音の推定によって獲得することになる.

次に中国語音に関しては, まず LHC の音は簡約上古音和東漢音[33]を用いる. 簡約上古音和東漢音は許思莱が作成した東漢（日本では後漢と言う）時代の漢字音, 即ち LHC の音を 3,767 漢字について収録し, 公開データとして使用できる資料である. 次に EMC の音は周法高上古音韻表[35]を用いる. 周法高上古音韻表は周法高が作成した隋唐時代の漢字音, 即ち EMC の音を 25,528 字について収録し, 公開データとして使用できる資料である. また, 簡約上古音和東漢音と周法高上古音韻表の漢字音は IPA シンボルによるデータとして公開されている. これを手動でデータセットに加工する.

以上のようにして得られた朝鮮語音と中国語音を使って LHC-MK, EMC-MK のペアのデータセットを作成する. 図 4.1 は作成したデータセットの抜粋である.

MK	LHC
ɦil	ʔit
tjəŋ	teŋ
tsʰil	tsʰit
tjaŋ	ʥiɑŋᴮ
sam	sɑm
sjaŋ	dźɑŋᴮ
…	…

MK	EMC
soan	ʂiuæn
kai	kɛi
tʰu	dəu
koa	kua
zjən	ɳiuæn
tsʰil	tsʰiɪt
…	…

図 4.1 データセット抜粋

このデータセットを訓練用・検証用・テスト用に 8:1:1 の割合で分割する．LHC-MK データセット，EMC-MK のデータセットの統計を表 4.1 に示す．

表 4.1 データセットの統計

	訓練データ	検証データ	テストデータ	合計
LHC-MK	1,852	232	232	2,316
EMC-MK	2,395	299	299	2,993

構築したデータセットの訓練データを用いて中期朝鮮語音を推測するモデルを学習する．実験の再現性を確保するため，seed は固定する．検証用データを用いてハイパーパラメタの調整を行い，PyTorch から出力される最良モデル best_model.pt を保存する．得られたモデルをテストデータに適用し，推測された中期朝鮮語音がデータセットの正解の音とどれだけ一致するかを評価する．実験は Google Colab(T4)環境で実施した．

4.1.2 結果と考察

モデルの学習にあたり，ベイズ最適化手法によってハイパーパラメータの最適化を実施した．最適化の結果得られたハイパーパラメータの値を表 4.2 及び表 4.3 に示す．前者は LHC を MK に変換するモデル，後者は EMC を MK に変換するモデルの最適化の結果である．

表 4.2 LHC から MK への変換モデルのハイパーパラメータ

ハイパーパラメータ	BiLSTM-IPA-IPA	BiLSTM-IPA-OH	BiLSTM-OH-OH	LSTM
HIDDEN_SIZE	209	470	411	505
EMBEDDING_SIZE	232	232	34	167
NUM_LAYERS	1	1	1	1
DROPOUT	0.31	0.49	0.18	0.36
BATCH_SIZE	16	32	64	16
LEARNING_RATE	0.0006	0.0003	0.0044	0.0013
NUM_EPOCHS	21	22	25	19
Encoder_Output_SIZE	HIDDEN_SIZE * 2	HIDDEN_SIZE * 2	HIDDEN_SIZE * 2	HIDDEN_SIZE * 2

表 4.3 EMC から MK への変換モデルのハイパーパラメータ

ハイパーパラメータ	BiLSTM-IPA-IPA	BiLSTM-IPA-OH	BiLSTM-OH-OH	LSTM
HIDDEN_SIZE	314	146	458	273
EMBEDDING_SIZE	228	94	60	130
NUM_LAYERS	1	1	1	1
DROPOUT	0.3	0.32	0.31	0.41
BATCH_SIZE	32	16	16	16
LEARNING_RATE	0.0009	0.0024	0.0009	0.0019
NUM_EPOCHS	22	22	20	28
ENCODER_OUTPUT_SIZE	HIDDEN_SIZE * 2	HIDDEN_SIZE * 2	HIDDEN_SIZE * 2	HIDDEN_SIZE * 2

これらのハイパーパラメータを用い LHC から MK への変換(推測), EMC から MK への変換を行った. テストデータに対する損失(Test Loss), 正解率(Test Accuracy), 実行時間を表 4.4, 表 4.5 に示す.

表 4.4 LHC から MK への変換モデルの学習・評価結果

	BiLSTM-IPA-IPA	BiLSTM-IPA-OH	BiLSTM-OH-OH	LSTM
Test Loss 最良値	**0.818**	1.041	1.108	1.130
Test Accuracy 最良値	0.820	**0.839**	0.831	0.830
T4 実行時間 (min)	**1:06**	2:09	1:45	2:06

表 4.5 EMC から MK への変換モデルの学習・評価結果

	BiLSTM-IPA-IPA	BiLSTM-IPA-OH	BiLSTM-OH-OH	LSTM
Test Loss 最良値	**0.947**	1.224	1.303	1.425
Test Accuracy 最良値	0.795	0.799	**0.800**	0.779
T4 実行時間 (min)	**2:10**	4:08	3:45	3:53

各モデルの訓練データ，検証データに対する損失(Loss)と正解率(Accuracy)の学習曲線を図 4.2-図 4.9 に示す.

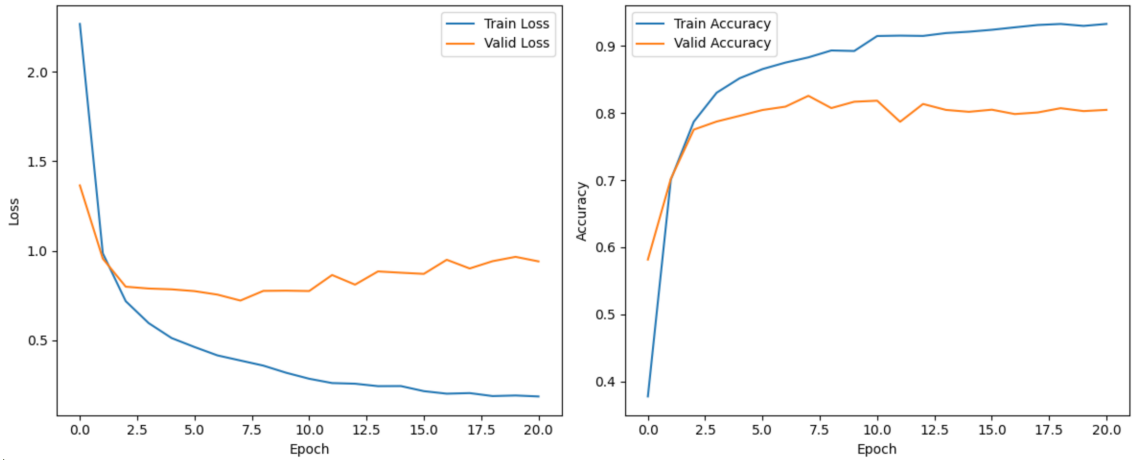

図 4.2 LHC→MK BiLSTM-IPA-IPA モデルの損失(左)と正解率(右)の学習曲線

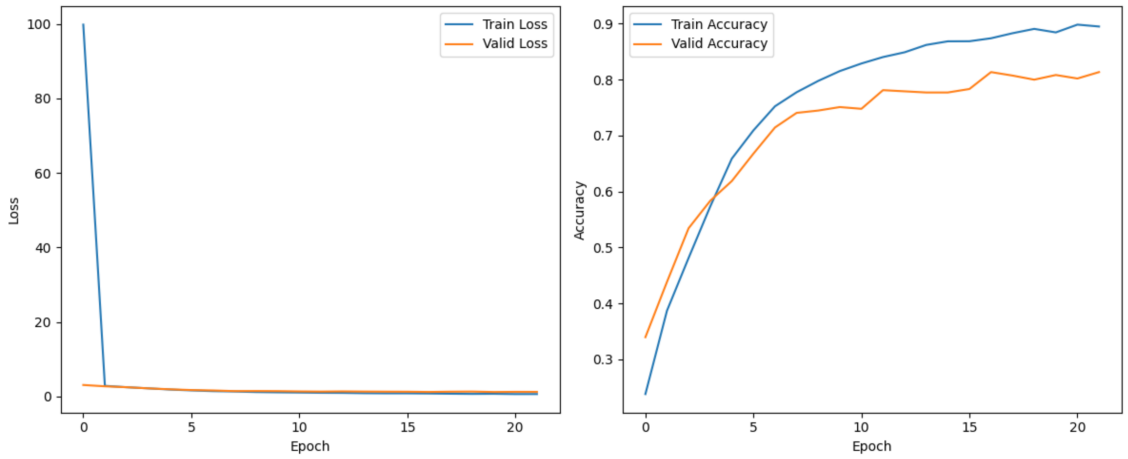

図 4.3 LHC→MK BiLSTM-IPA-OH モデルの損失(左)と正解率(右)の学習曲線

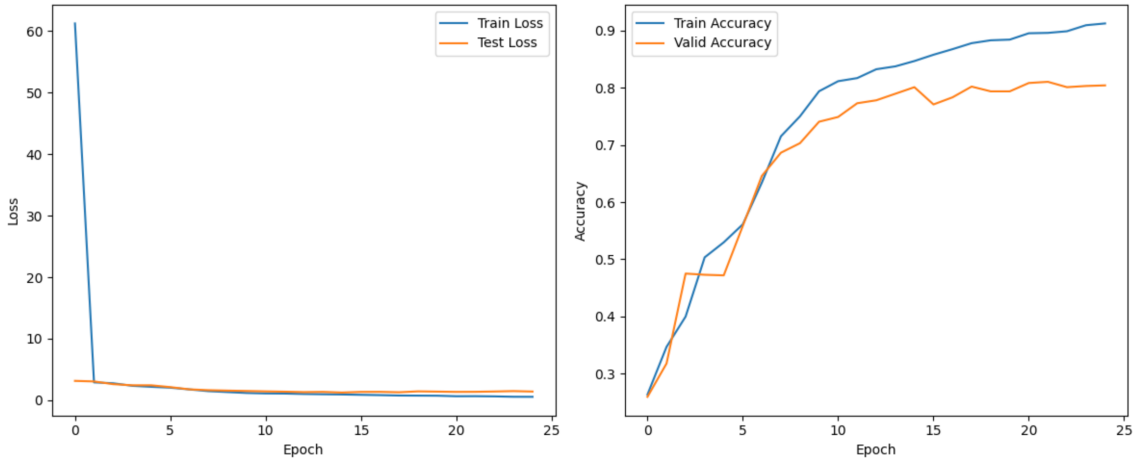

図 4.4 LHC→MK BiLSTM-OH-OH モデルの損失(左)と正解率(右)の学習曲線

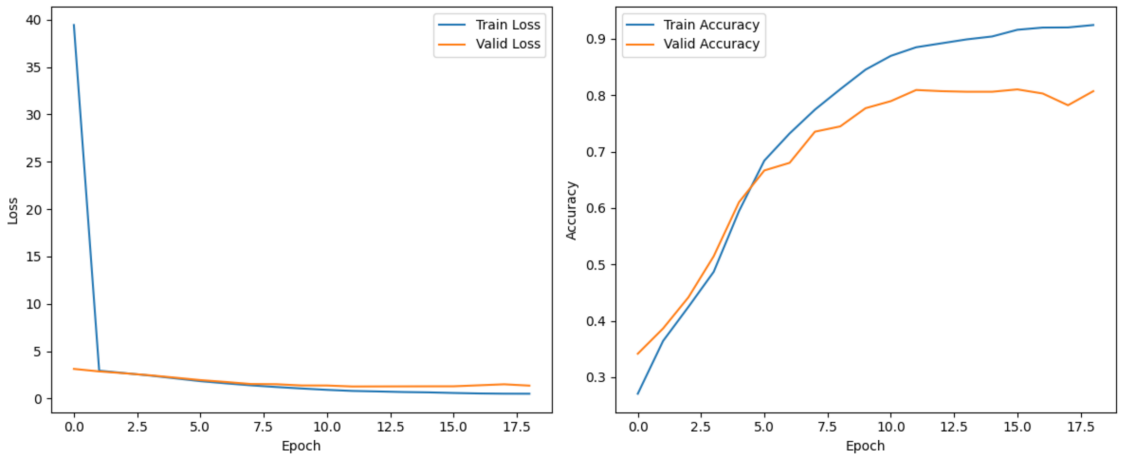

図 4.5 LHC→MK LSTM モデルの損失(左)と正解率(右)の学習曲線

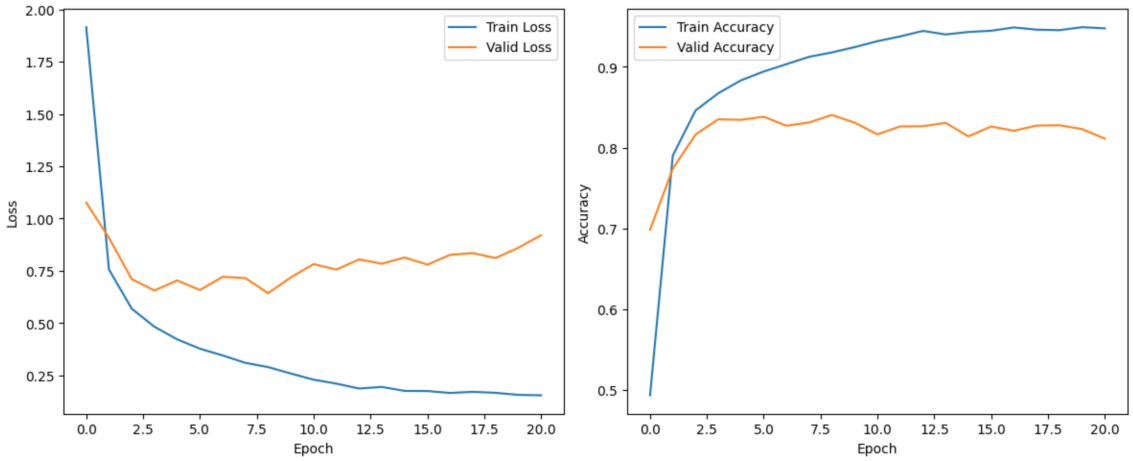

図 4.6 EMC→MK BiLSTM-IPA-IPA モデルの損失(左)と正解率(右)の学習曲線

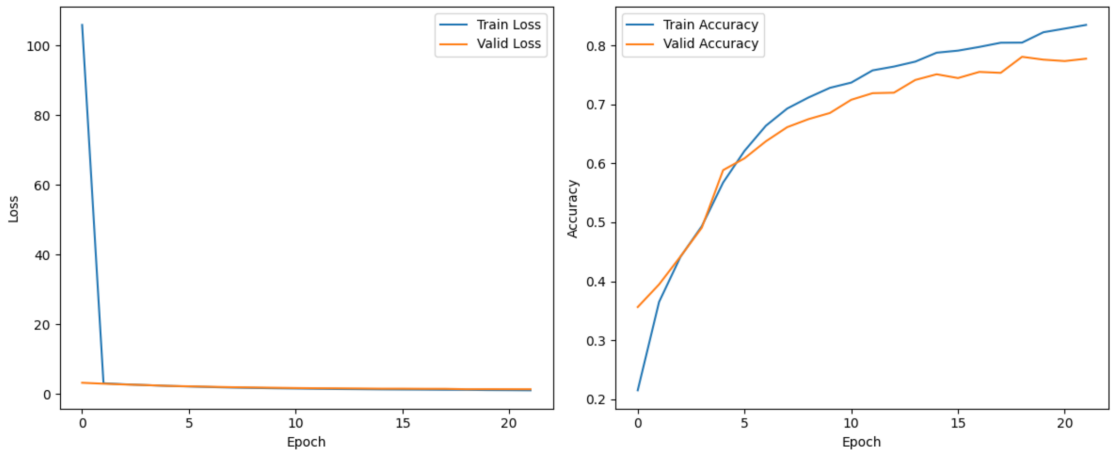

図 4.7 EMC→MK BiLSTM-IPA-OH モデルの損失(左)と正解率(右)の学習曲線

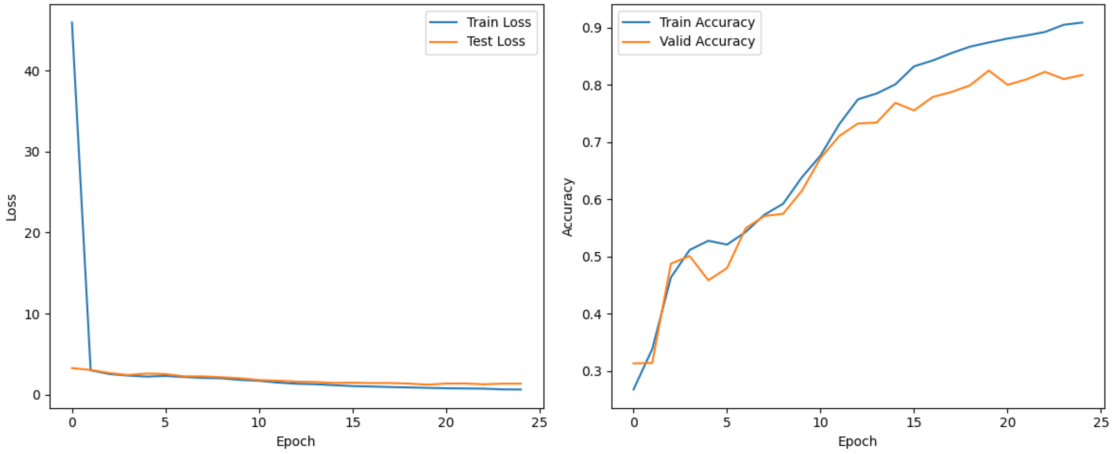

図 4.8 EMC→MK BiLSTM-OH-OH モデルの損失(左)と正解率(右)の学習曲線

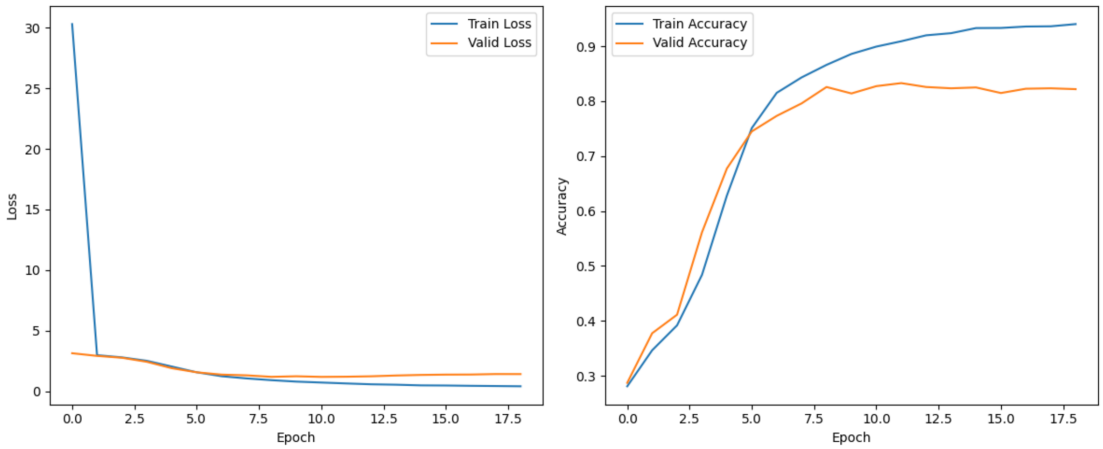

図 4.9 EMC→MK LSTM モデルの損失(左)と正解率(右)の学習曲線

LHC から MK への変換タスクについて，BiLSTM-IPA-IPA モデルは，Test Loss の値が最も低く，学習曲線も安定しており，最も短時間で学習を終了した．BiLSTM-IPA-OH モデルは，Test Accuracy の値が最も高いが，初期の段階で訓練データに対する損失が急激に低下し，その後の損失は大きく変化しない．BiLSTM-OH-OH と LSTM モデルも同様に初期の急激な損失低下を示し，またその後の損失の変化がほとんどない．これらのモデルは局所的な解に早期に収束している可能性がある．

EMC から MK への変換タスクについて， BiLSTM-IPA-IPA モデルは，4 つのモデルの中で最も良い Test Loss が得られている．学習曲線もエポック数が増えるにつれて学習データの損失が継続して低下しており，最も短時間で学習を終了した．BiLSTM-OH-OH モデルは，Test Accuracy の値が最も高いが，訓練データに対する損失は初期の段階で急激に低下し，その後では損失がほとんど低下しない．BiLSTM-IPA-OH と LSTM モデルも BiLSTM-OH-OH モデルと同様の傾向が見られる．

以上のことから，BiLSTM-IPA-OH，BiLSTM-IPA-OH，LSTM に見られる，学習の初期に損失が急激に低下しその後の損失の変動が小さい，という現象は，デコーダのベクトル表現として one-hot エンコーディングを使用しているためと考える．そのため，one-hot エンコーディングは音の特徴を表現する形式としては不適切であると言える．

一方で BiLSTM-IPA-IPA モデルは，滑らかな学習曲線を示し，学習過程が安定していることから，高い汎化性能を有していると考えられる．つまり，エンコーダおよびデコーダのベクトル表現として，IPA 音韻素性を特徴量として用いたエンコーディングが，IPA シンボルの one-hot エンコーディングに比べて MK の音の推測に適しており，モデル性能の向上に寄与したと言えよう．また，処理

速度も最も早いという点も他のモデルと比べて優れている．本実験の結果から，BiLSTM-IPA-IPA が MK 音の推測に最適なモデルであると言える．

実験では，3,000 件に満たない小規模なデータセットを用い，またモデル自体も標準的な双方向あるいは単方向 LSTM を採用したのにも関わらず，80%前後の高い正解率が得られた．これは，音素を IPA シンボルに基づくベクトルで表現したことで，中国語音から朝鮮語音への変換モデルを効率的に学習できたためと考えられる．

4.2 解読実験

4.2.1 予備実験

本項では，万葉集 9 番歌を提案手法によって解読する予備実験について述べる．Vovin は中期朝鮮語音によって万葉集 9 番歌の最初の 12 文字を解読しており[4]，提案手法で同様の解読ができるかを確認することで提案手法の妥当性を検証する．

4.2.1.1 実験設定

Vovin の結果を再現するため，入力として西本願寺本の漢字列ではなく，Vovin によって校正された漢字列を用いた．予備実験を実施する段階では，中期朝鮮語音を推定するモジュールは開発中であった．そのため，Vovin が解読に用いた中期朝鮮語音を人手で漢字音辞書に加えた．また，複数の単語列の候補から最適なものを選択する手法も検討中であった．そこで，提案手法によって得られた単語列の候補の中に Vovin による解釈と同じものが含まれているかを確認するに留めた．すなわち，本予備実験は万葉集の未解読歌を解読する手法を検討する段階で実施されたものであり，図 3.4 の擬似コードに示したアルゴリズムをそのまま適用した実験ではないことに注意されたい．

4.2.1.2 結果と考察

万葉集 9 番歌未解読部分「莫囂圓隣之大相七兄爪謁氣」の 12 文字について，漢字音を辞書に登録した．それぞれの漢字に 5 種類の音を登録できれば，最大 5^{12}=244,140,625 個の音の列が得られることになる．このような組み合せ爆発とも言うべき状況が生じることが危惧されたが，実際には 1 つの漢字に対して登録できた音の数は限られ，結果的に音の列の数は 27,648 となった．上記の結果を表 4.6 にまとめる．

表 4.6 9番歌の未解読部分に対する解析

	未解読の漢字の系列（文字数 n=12）	漢字音組み合わせ最大値 5^n	漢字音辞書による組み合わせ	Vovin [4]の解釈と同じ12形態素数を有する単語列の数
9番歌	莫器圓鄰之大相七兄爪湯氣	244,140,625	27,648	5,513

表 4.6 の右端の列にある「Vovin[4]の解釈と同じ12形態素数を有する単語列の数」とは，Vovin による解釈における形態素数は 12 であり，これと同じ数の単語を含む単語列の数を示す．本来，正しい解釈の形態素数は必ずしも 12 ではない．しかしながら，4.2.1.1 で述べたように，今回は Vovin の正解データがこの中に存在することを確認することを予備実験の目的としている．これらの単語列を人手で確認したところ，Vovin による解釈，すなわち正解に該当するものを見つけることができた．図 4.10 に MeCab による 9 番歌の解析結果，図 4.11 に 9 番歌の解析結果のパスを示す．

```
najoidaltitabodanaistahamudda
najoi    Noun,普通,*,*,*,나죄,나죄,*,*,NNG
dal      Noun,普通,*,*,*,달01,달,*,달/풀,NNG
ti       Verb,自立,*,語基1,*,티다,티,*,*,VV
ta       Ending,語尾,接続形,*,1接続,다,타,*,非終止,EC
bo       Verb,自立,*,語基2,*,보다,보,*,*,VV
da       Ending,語尾,接続形,*,1接続,다,다,*,非終止,EC
na       Verb,自立,*,語基3,*,나다01,나,*,*,VV
is       Verb,非自立,*,語基1,3接続,잇다02,?,*,*,VX
ta       Ending,語尾,接続形,*,1接続,다,타,*,非終止,EC
ha       Adverb,一般,*,*,*,하02,하,*,*,MAG
mud      Verb,自立,ㄷ不規則,語基1,*,묻다03,묻,*,묻다[問],VV
da       Prefinal,回想,*,語基4,*,더,다,*,*,EP
EOS
```

図 4.10 9番歌の MeCab による解析結果

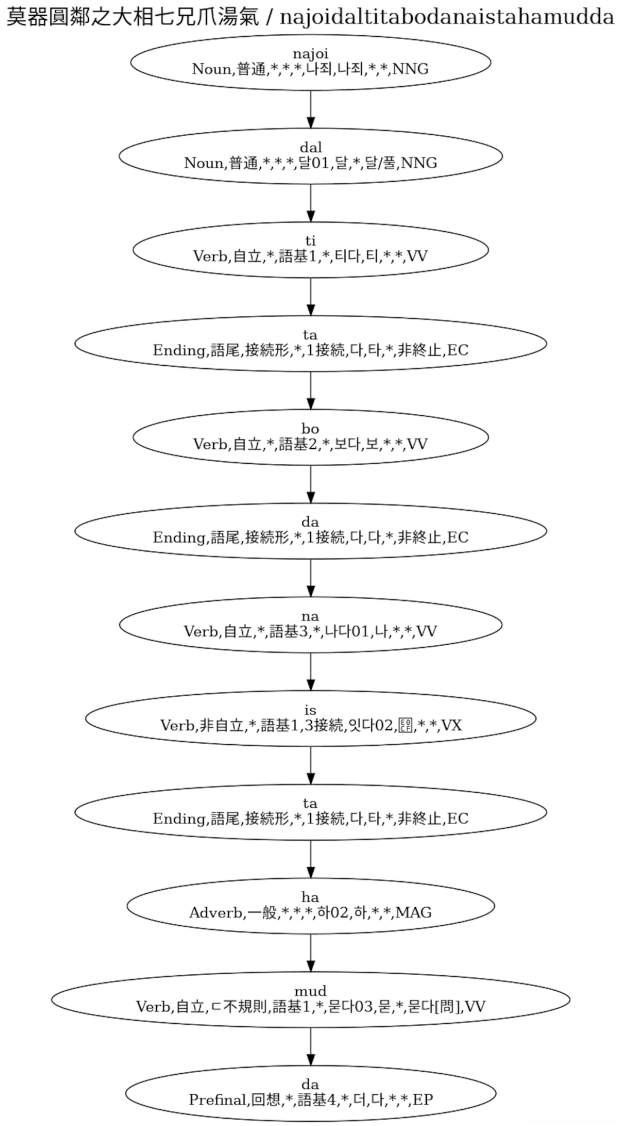

図 4.11 9 番歌の解読結果パス図

この結果を中期朝鮮語辞書[36]により現代韓国語に訳し，さらに韓国語辞書に[47]により現代韓国語から日本語へ翻訳すると，9 番歌の未解読部分の私訳は「夕月を仰いで問いし」になる．これは Vovin が正解として位置付けた仙覚の残した解釈「ユフツキノアフキテトヒシ」と一致する．このことは提案手法の有効性を示すものであり，提案手法によって 9 番歌以外の未解読歌を解読できる可能性があることを示唆する．すなわち，万葉集の他の未解読歌が古い朝鮮語を適当な漢字音で表したものであり，古い朝鮮語の単語列として解釈できる可能性がある．

提案手法は万葉集の未解読歌が古い朝鮮語で書かれていることを仮定してい

るが，その仮定が正しいとは限らない．また，提案手法には，漢字音として漢字音辞書に登録されたものしか考慮されないこと，中期朝鮮語の単語として MeCab が使用する中期朝鮮語辞書 MkHanDic に登録されたものしか考慮されないこと，MeCab から出力される形態素解析結果である単語列を一定程度絞り込んだ後で(表 3.6 文法的制約を考慮した単語列のフィルタリングを参照)中期朝鮮語辞書を引いて解釈できる単語が得られる必要があるなど，いくつかの制約がある．したがって，たとえ万葉集の他の未解読歌が古い朝鮮語を適当な漢字音で表したものだとしても，適切な解釈を導くことができることを保証するものではない．とはいえ，万葉集未解読歌の解読に向けた第一歩として，この方法により解読を試みる価値はあると考える．

4.2.2 未解読歌 6 つの解読

4.2.2.1 実験設定

万葉集における未解読歌は 40 首ほどある[2]とされるが，ここでは代表的なものを少数選び実験する．代表的な未解読歌の選択基準としては，岩波文庫が発行した万葉集[48]において，通常は上代日本語に訳されるべき所，原文の漢字のままで印刷されている歌を選んだ．具体的には表 4.7 に示す 6 首である．

表 4.7 未解読歌 6 首

	未解読の漢字の系列
9 番歌	莫器圓鄣之大相七兄爪湯氣
156 番歌	已具耳矣自得見監乍共
262 番歌	雪驪朝樂毛
537 番歌	哭者痛寸取物
655 番歌	邑礼左變
3889 番歌	葉非左思所念

本実験では歌全体ではなく漢字で印刷されている未解読の文字列のみを解析の対象とする．9 番歌は予備実験において解読対象としたが，提案手法の一部は未実装であったため，本実験で改めて解読を試みる．中期朝鮮語音の推定モデルを学習した上で漢字音辞書を作成した上で，図 3.2 に示した解読処理の流れ図にしたがい，漢字音辞書を用いて漢字列を音の系列に変換した後，音の系列を形態素解析し，最後に最適な単語列を選択する．

4.2.2.2 結果と考察

形態素数最小法の考え方に基づいて，MeCab が解析結果として出力する形態素数が少ない順に解釈可能な歌があるかどうかを検証する．表 4.8 は提案手法によって得られた単語列のうち，形態素数(単語数)が 1，2，3，4，5 である単語列の数を示している．その結果，形態素数が 4 個以下の単語列は 6 歌のいずれも得られなかった．形態素数が 5 個のとき，3889 番歌に関して 22 個の解釈できる可能性がある単語列が得られた．

表 4.8 形態素数別の解釈対象となる候補数

	未解読の 漢字の系列	形態素数 1-4	形態素数 5
9 番歌	莫器圓鄣之大相七兄爪湯氣	0	0
156 番歌	已具耳矣自得見監乍共	0	0
262 番歌	雪驪朝樂毛	0	0
537 番歌	哭者痛寸取物	0	0
655 番歌	邑礼左變	0	0
3889 番歌	葉非左思所念	0	**22**

3.5 節で述べたように，単語列の候補を絞り込む際には，表 3.6 に示した条件を満たす単語列は文法的に不適切であるとして除外している．3889 番歌の形態素数 5 の単語列について，各条件に該当する単語列の数を表 4.9 に示す．ほとんどの単語列は文法的に不適切であるとして除外されている．

表 4.9 3889 番歌に対して得られた形態素数 5 の単語列の内訳

単語列の種類	単語列の数
総単語列数	32,415
適切単語列数	22
不適切単語列数(以下内訳)	32,393
- 語尾で始まる	0
- 指定詞で始まる	0
- 依存名詞で始まる	4
- 数詞で終わる	0
- 連続する語尾を含む	21,552
- 未知語を含む単語列	10,837

次に，3889 番歌の形態素数が 5 である 22 個の解釈対象となる単語列の候補について，MeCab の解析結果とそれを図示したパスを参照し，それぞれの単語を中期朝鮮語辞書[36]及び韓国語辞書[47]によって辞書引きし，解釈を試みた．図 4.12 と図 4.13 は 22 個の単語列のひとつについて MeCab の解析結果とそのパスを図示したものである．

```
babijoasikane
bab     Noun,普通,*,*,*,밥,밥,*,*,NNG
i       Ending,助詞,主格,*,*,이,ㅣ,*,*,JKS
joasi   Verb,自立,*,語基1,*,좌시다,좌시,*,*,VV
ka      Prefinal,強勢,*,語基4,*,거,카,*,*,EP
ne      Noun,代名詞,*,*,*,너,너,*,*,NP
EOS
```

図 4.12 3889 番歌の MeCab 解析結果

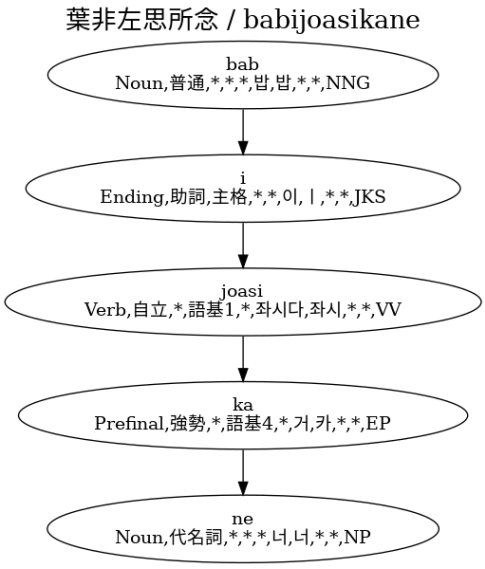

図 4.13 3889 番歌の解読結果パス図

この形態素解析結果では「i : 이」を主格の助詞としているが，人手による解釈では「i : 이」は対格の助詞として「을」(を)の意味を持つこととする．なぜなら，中期朝鮮語の「i : 이」には「을」(を)の意味があり用例も残っている[36]ことを踏まえると，この 3889 番歌の句の最後の単語「ne : 너」が倒置法の表現によって主語になっていると考えることが自然だからである．すなわち，ここでは「i : 이」は「joasi : 좌시다」(召し上がる)の目的語である「bab : 밥」(ご飯)にかかる対格助詞として機能すると解釈する．しかし一つ疑問が残る．「joasi : 좌시다」(召し上がる)は尊敬語であるが，主語の「ne : 너」(あなた，

お前)は通常は同等ないし目下の者に対して用いられる．ただし，ここで倒置法が使われていることを勘案するならば，強調表現の一種として尊敬語の主語に非日常的な語を当て嵌めた可能性も考えられる．

以上より，3889 番歌を私訳すると，「ご飯を召し上がるのか，あなたは」となる．3889 番歌のこの未解読部分は最後の句である．上の句の部分は，「人魂のさ青なる君がただひとりもあへりし雨夜の」とあり[48]，現代語に訳せば，「人魂である君がただひとりさまよっている．その君に私は雨の夜に出会い」のようになろう．ここで導き出された未解読部分の解釈はこの上の句と絶対に連接しないとも言い難く，その妥当性をどう評価するのかは難しい．また，詩歌の解釈は言語も文字もわかっていたとしても時に易しいものではない．したがって，本研究で得られた 3889 番歌の解釈が真に正しいと結論付けることはできない．一方，朝鮮語音を借りているという仮定の下，従来は未解読とされた歌を半自動的に解析し，日本語としての解釈が得られたことは，本研究の成果と言える．

第5章 枕詞の解読

5.1 枕詞解読の背景と目的

第 4 章の実験では万葉集の未解読歌における漢字が朝鮮語音を借りているという仮定の下で解釈できる可能性を示した．一方，これまで解釈が試みられていないものとして枕詞がある．枕詞における漢字も朝鮮語の音を借りていると仮定し，提案手法によって枕言葉を解釈することを試みる．最終的な目標は代表的な枕詞についてその意味を解読することである．

5.2 実験設定

多数ある枕詞から検証に用いる代表的な枕詞を選別する．選定基準としては万葉集における枕詞の出現頻度を考える．山口が調査した万葉集における枕詞の出現頻度[21]を参照し，出現頻度が高いものから 12 個の枕詞を選んだ．しかし実際には時間の制約から 6 個の枕詞の実験を行うに止まった．表 5.1 は出現頻度の高い枕言葉，それが係る単語(被枕詞)，その出現頻度，漢字列のタイプ数を示す．一番左の列に数字が記載されているものが実験の対象とした枕言葉である．

表 5.1 選出した 6 枕言葉

	枕詞	被枕詞	頻出数	漢字列のタイプ数
1	**あしひきの**	山（やま）	112	31
2	**ぬばたまの**	夜（よ）・闇（やみ）	79	24
	しろたへの	衣（ころも）・袖（そで）	57	28
3	**ひさかたの**	天（あめ）・光（ひかり）・空（そら）	51	11
	くさまくら	旅（たび）	49	4
	たまほこの	道（みち）	37	7
4	**あらたまの**	年（とし）・月（つき）	36	17
	しきたへの	枕（まくら）・衣（ころも）	31	14
	うつせみの	世（よ）・身（み）	30	10
5	**あをによし**	奈良（なら）	27	11
	やすみしし	大君（おおきみ）	27	3
6	**たらちねの**	母（はは）・親（おや）	24	10

表 5.1 では枕言葉をひらがなで表記しているが，万葉集は全て漢字で書かれており，枕詞も例外ではない．例えば「あしひきの」は以下に示す 31 通りの漢字表記(漢字列のタイプ)で万葉集中に出現している．これらの漢字列のタイプを全て分析対象とする．

> 葦引乃，足引乃，足引之，足曳之，足檜乃，足檜之，足疾乃，足病之，悪氷木乃，蘆檜木乃，蘆檜木笑，足日木乃，足日木之，足日木，足日木笑，足日木篁，足日木能，足檜木乃，足檜木之，足比奇乃，足比木乃，足氷木乃，安之比奇乃，安之比奇能，安之比紀乃，安之比紀能，安志比奇乃，安志比紀乃，安思必奇能，安思比奇能，阿之比奇能

「あしひきの」に限らず他の枕詞についても同様に全ての漢字列のタイプを分析対象とする．

枕詞は意味が不明であっても被枕詞を修飾するとされているので，枕詞の解釈が得られたとき，被修飾語との関係からある程度その解釈の妥当性が推しはかれると考えられる．

5.3 結果と考察

6 つの枕詞について提案手法による解析を行い，形態素数が 1，2，3，4 個の場合の解釈対象の候補となる単語列を調べた．その結果,「あしひきの」について形態素数 2 個の単語列の候補が得られた．このときの漢字記は「足檜乃」であった．しかし，他の枕詞については該当する単語列は得られなかった．表 5.2 は,「足檜乃」に対して得られた形態素数 2 の単語列に表 3.6 に示した条件によるフィルタリングを行った結果を示している．文法的に不適切な単語列を除いた結果 11 個の単語列が得られた．

表 5.2 「あしひきの」に対して得られた形態素数 2 の単語列の内訳

処理結果サマリー	単語列の数
総単語列数	1,177
適切単語列数	11
不適切単語列数(以下内訳)	1,165
- 語尾で始まる	164
- 指定詞で始まる	0
- 依存名詞で始まる	0
- 数詞で終わる	1
- 連続する語尾を含む	0
- 未知語を含む単語列	1，000

文法的に適切な 11 個の単語列を解釈対象の候補として，中期朝鮮語辞書[36]により現代韓国語に訳し，さらに韓国語辞書に[47]により現代韓国語から日本語に翻訳することを試みた．しかしながら中期朝鮮語の文として解釈可能な単語列はなかった．一方，漢字音の辞書を拡張すれば解釈が可能になりうる単語列がいくつか見つかった．例として，図 5.1 に漢字の系列が「足檜乃」で漢字音が「'abina」のときの単語列を示す．

```
'abina
'abi    Noun,普通,*,*,*,아비,아비,*,*,NNG
na      Ending,助詞,接続助詞,*,*,나,나,*,*,JC
EOS
```

図 5.1 'abina の単語列

辞書引きによる意味は父('abi：아비)だが乃の音 na が「인」(in)という指定詞の i と語尾の -n であるならば，音は'abiin となり，意味は「父である」「父なる」となるので，「山」(「あしひきの」の被枕詞)に掛かる言葉として不自然ではない．他にも同様に解釈の可能性がある単語列が見つかった．したがって，漢字音辞書に新たに音を登録すれば，「あしひきの」の解釈が可能になると考えた．

そこで，EMC の資料としてこれまで参照してきた周法高上古音韻表と同等の収容語数(25,528 語) を擁する漢語音韻學[49]と上古音研究[50]を参照し，新たに EMC の音を得ることとした．その結果「乃」について，漢語音韻學からは「nAi」を，上古音研究からは「nậi」という音を得た．その上で EMC から MK への変換の訓練で得た BiLSTM-IPA-IPA の最良モデルである best_model.pt を用い，中期朝鮮語音を推定したところ表 5.3 のような結果が得られた．

表 5.3 中期朝鮮語音推定器による 「乃」 の音の推定

Processing Input (EMC)	Predicted (MK)
nAi	in
nậi	in

「乃」に対して「in」という音を推定できたので，この「in」を須賀井式ローマ字に変換し「乃」に相当する部分を「in」に置き換えた音の系列を作成する．IPA シンボルの「in」は須賀井式ローマ字でも「in」であるので「'abiin」が生成される．これに対し MeCab による形態素解析を行ったところ図 5.2 に示す結果を得た．

```
'abiin
'abi    Noun,普通,*,*,*,아비,아비,*,*,NNG
i       Siteisi,非自立,*,語基2,*,이다,ㅣ,*,*,VCP
n       Ending,語尾,連体形,*,2接続,ㄴ,ㄴ,*,*,ETM
EOS
```

図 5.2「あしひきの('abiin)」の MeCab による解析結果

図 5.3 はこの解析結果を図示したものである．

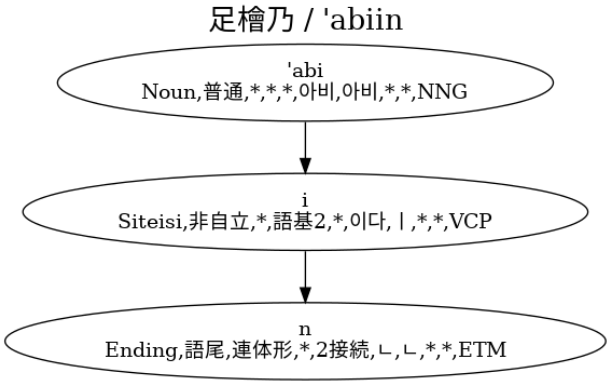

図 5.3「あしひきの('abiin)」の解読結果パス図

この結果より，中期朝鮮語辞書[36]により現代韓国語に訳し，さらに韓国語辞書[47]により現代韓国語から日本語へ翻訳すると，「父である」または「父なる」といった解釈が得られる．これは被枕詞である山に掛かる解釈として自然である．しかし「足檜乃」は 31 もの漢字表記がある「あしひきの」に関する漢字表記の一つに過ぎない．したがって，上記の分析から朝鮮語音を借りているという仮定の下で枕言葉「あしひきの」の解釈に成功したと断言することはできない．

第6章 おわりに

6.1 まとめ

本研究では，万葉集の未解読歌を朝鮮語により解読するアプローチを提案し，いくつかの重要な成果が得られた．

まず，漢字音辞書の作成と中期朝鮮語音の推定システムの開発を行った．後者について，IPA シンボルの音韻特徴を活用した双方向 LSTM ベースのモデルを構築し，約 80%の正解率で中期朝鮮語音を推定した．特に，エンコーダ・デコーダともに IPA シンボルによるベクトル表現を入力・出力する BiLSTM-IPA-IPA モデルは，本研究で提案した 4 つのモデルの中で最も正解率が高く，学習に要する時間も短かった．

次に，漢字音辞書を参照して万葉集歌の漢字列を音の系列に変換した後，それを単語に分割する形態素解析システムの実装に取り組んだ．MeCab を用いた中期朝鮮語の形態素解析システムを構築し，得られた複数の単語列の候補の中から解釈できる可能性のある単語列を選別するシステムも実装した．

実証実験においては，まず Vovin が解読した 9 番歌について，本研究でも同じ解釈が得られることを確認した．さらに，未解読歌 6 首への適用実験を実施し， 3889 番歌について解読の可能性がある候補を提示することができた．また，万葉集で頻出する枕詞についても解読を試み，「あしひきの」について新たな解釈の可能性を示唆した．

これらの成果は，漢字で書かれた未解読テキストに対する機械的解読の可能性を示すとともに，万葉集研究に新たな視点を提供するものである．

漢字で書かれた古文書の解読の困難性について Luo は以下の 3 点を指摘した[26]．1)漢字は alphabet に比べ文字種が圧倒的に多い，2)漢字は文字と音価の対応が一意ではない，3)漢字は表語文字であるので意味を考慮する必要がある．本研究は，1)について，中期朝鮮語音を推定するモデルを学習する際に漢字の音を比較的少数の IPA シンボルに縮約して表現することで，漢字の種類が多いという問題に対応した． 2)について，1 つの漢字音に対して最大で 5 種類の音(万葉仮名，吏読，LHC から推定する MK，EMC から推定する MK，正用)の可能性を考慮し，これらの組み合わせを網羅的に探索することで対応した．なお，3)意味の考慮は依然として課題として残された．

6.2 今後の課題

本研究で提案した手法には，データ，手法，評価の各側面において，いくつかの本質的な限界と課題が存在することが明らかになった．

まずデータに関して，入力データ，辞書データ，実験範囲の三点について課題が挙げられる．入力データについては，万葉集テキストとして西本願寺本を底本とするデータ[27]をそのまま使用したが，Vovin が自身による校正を施しているように，異本から採取された漢字を入力データとするという選択肢も存在した．また，漢字音辞書についても，さらなる充実の余地が大きい．「あしひきの」の解読実験で示唆されたように，新たな文献から得られる漢字音を適切に取り込むことで，未解読歌を解読する可能性が高まることが期待される．特に，Vovin の研究では言及されていない呉音の登録も検討に値する．これは，万葉集編纂当時，日本で流通していた漢字音の主流が呉音であった[51,52,53]という歴史的背景を考慮したものである．実験範囲については，未解読歌が 42 首ある[2]とされる中で，現時点では 6 首のみしか解読を試みていない．同様に，500 に上るとされる枕詞[21]についても，6 個のみの実験に留まっている．

手法面での課題としては，まず MeCab 用の中期朝鮮語辞書の制約が挙げられる．現在使用している MkHanDic は約 9,653 語を収録しているものの，仏典に偏った語彙構成となっており，より広範な解読を行うためには語彙の拡充が必要である．ただし，MkHanDic はバイナリファイルのみが公開されており，辞書の拡張が技術的に困難である状況にある．この制約を克服する方法を検討する必要がある．

また，形態素解析における単語列の選択方法にも改善の余地がある．未解読歌 6 つに関する実験では，形態素数最小法の考えに基づいて形態素数が少ない場合から実験を行ったが，形態素数が 5 つの場合までしか検討できていない．形態素数 6 以上の候補について人手による解釈を試みれば，適切な解釈が得られる可能性がある．さらに，現状の手法では未知語を含む単語列や文法的な誤りを含むと思われる単語列を機械的に除外しているが，この方法では詩歌としての微妙な表現を過度に切り捨てている可能性がある．より柔軟な選択方法の開発が求められる．

最後に，評価方法に関する重要な課題が存在する．解読結果の妥当性評価の方法論が確立されていないことである．特に歌集という性質上，機械的な評価基準だけでは不十分である可能性が高く，言語学的知見と計算機による解析を組み合わせた新たな評価手法の開発が求められる．これは，対象とする時代の

言語使用の実態が完全には解明されていないことに加え，和歌という文学形式特有の表現の多義性に起因する課題でもある．

このように本研究には多くの課題が残されているが，これらの課題を一つ一つ克服していくことで，より多くの未解読歌の解読可能性を広げることができると考えられる．特に，計算機による古文書解読という研究領域において，これらの課題は今後取り組むべき重要な研究テーマとなると考えられる．

謝辞

主指導教員の白井清昭准教授にご指導とご助言を数多く賜り，本研究の遂行に不可欠なものでありました．深く感謝いたします．中間審査会において的確なコメントとご助言を賜りました，池田心教授，井之上直也准教授に感謝の意を表します．また井之上直也准教授には副テーマのご指導を頂きましたことに感謝いたします．最終審査会において，今後の研究の指針を与えて下さった池田心教授，岡田将吾教授，井之上直也准教授にお礼申し上げます．

参考文献

[1] 大野晋，仮名文字・仮名文の創始 岩波講座 日本文学史 古代 2，岩波書店 (1958)

[2] 菊沢秀生，国語学論集，教育出版センター(1981)

[3] 姜斗興，吏読と万葉仮名の研究，和泉書店(1982)

[4] Alexander Vovin, Man’yōshū, Book 1, a new English translation containing the original text, kana transliteration, romanization, glossing and commentary, Brill(2017)

[5] Alexander Vovin, Man’yōshū, Book 2, a new English translation containing the original text, kana transliteration, romanization, glossing and commentary, Brill(2020)

[6] Alexander Vovin, Man’yōshū, Book 15, a new English translation containing the original text, kana transliteration, romanization, glossing and commentary, Brill(2009)

[7] Alexander Vovin, Man’yōshū, Book 17, a new English translation containing the original text, kana transliteration, romanization, glossing and commentary, Brill(2016)

[8] Alexander Vovin, Man’yōshū, Book 19, a new English translation containing the original text, kana transliteration, romanization, glossing and commentary, Brill(2018)

[9] Benjamin Snyder, Regina Barzilay, Kevin Knight, A Statistical Model for Lost Language Decipherment, ACL(2010)

[10] Taylor Berg-Kirkpatrick, Dan Klein, Simple Effective Decipherment via Combinatorial Optimization, ACL(2011)

[11] Jiaming Luo, Yuan Cao, Regina Barzilay, Neural Decipherment via Minimum-Cost Flow: form Ugaritic to Linear B, ACL(2019)

[12] Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, Yuan Cao, Deciphering Undersegmented Ancient Scripts Using Phonetic Prior, ACL(2021)

[13] 町田和彦，世界の文字とことば，河出書房新社(2009)

[14] 山口文彦，未解読言語の解析技術，人工知能，31 巻 6 号 p775-779(2015)

[15] 矢島文夫，解読古代文字，筑摩書房(1999)

[16] 沖森卓也編，陳力衛，肥爪周二他，日本語史概説，朝倉書店(2010)

[17] 今野真二，日本語と漢字 -正書法がないことばの歴史，岩波書店(2024)

[18] 仙覚，萬葉集註釈，国立国会図書館デジタルコレクション，https://dl.ndl.go.jp/ (2024-12 閲覧)

[19] 古橋信孝，古代和歌の発生，東京大学出版会 (1988)

[20] 大浦誠士，「枕詞は訳さない」でいいのか，松田治他編，古典文学の常識を疑う，勉誠出版(2017)

[21] 山口正，万葉修辞の研究，武蔵野書院 (1964)

[22] Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada, Unsupervised analysis for decipherment problems, ACL(2006)

[23] Fabio Tamburini, Decipherment of Lost Ancient Scripts as Combinational Optimization using Coupled Simulated Annealing, ACL(2023)

[24] Lyle Campbell, Histrical Linguistcs, Edinburgh University Press(2020)

[25] Kenneth N. Stevens,Acoustic phonetics, volume 30. MIT Press(2000)

[26] Jiaming Luo, Automatic Methods for Sound Change Discovery, MIT (2021)

[27] 吉村誠，万葉集テキストデータ，https://c-able.ne.jp/~y_mura/ (2024-12 閲覧)

[28] Wikipedia, Origin of Hangil, https://en.wikipedia.org/wiki/Origin_of_Hangul (2024-12 閲覧)

[29] 鄭光，北郷照夫，朝鮮吏読辞典，ペン・エンタープライズ(2006)

[30] 南豊鉉，吏讀辭典，檀国大学(2020)

[31] 劉昌惇，古語辞典，東国文化社 (1955)

[32] 金思燁，記紀万葉の朝鮮語，六興出版 (1989)

[33] 許思萊，簡約上古音和東漢音 (2008)，http://www.kaom.net/ny_word.php (2024-12 閲覧)

[34] 崔世珍，訓蒙字会 (1527)，https://dl.ndl.go.jp/pid/926032/1/19 (2024-12 閲覧)

[35] 周法高，周法高上古音韻表 (1973)，http://www.kaom.net/ny_word.php (2024-12 閲覧)

[36] 南廣祐，古語辞典，教學社（1997）

[37] 福井玲，中期朝鮮語文献の電子計算機による処理，明海大学外国語学部論集 2(1990)

[38] 須賀井義教，mecab-k2alpha のアルファベット転写表，https://ja.osdn.net/projects/handic/wiki/MkHanDicAlphabet（2024-12 閲覧）

[39] 須賀井義教，tools mecab-k2alpha，https://ja.osdn.net/projects/handic/releases/71709（2024-12 閲覧）

[40] Wikipedia，Revised Romanization of Korean，https://en.wikipedia.org/wiki/Revised_Romanization_of_Korean（2024-12 閲覧）

[41] Wikipedia，Help:IPA/Korean，https://en.wikipedia.org/wiki/Help:IPA/Korean（2024-12 閲覧）

[42] 工藤拓，MeCab: Yet Another Part-of-Speech and Morphological Analyzer，https://taku910.github.io/mecab/ （2024-12 閲覧）

[43] 須賀井義教，MkHanDicIntro，https://osdn.net/projects/handic/wiki/MkHanDicIntro（2024-12 閲覧）

[44] Chanjun Park， Chanhee Lee， Yeongwook Yang， Heuiseok Lim，Ancient Korean Neural Machine Translation，IEEE(2020)

[45] 河野六郎，河野六郎著作集 第１巻 朝鮮語学論文集，平凡社（1979）

[46] Alberto Pettarin，ipapy，https://pypi.org/project/ipapy/（2024-12 閲覧）

[47] 油谷幸利，門脇誠一，松尾勇，高島淑郎，韓日辞典，小学館（2018）

[48] 佐々木信綱，万葉集，岩波書店（1927）

[49] 董同龢，漢語音韻學（1993），http://www.kaom.net/ny_word. php（2024-12 閲覧）

[50] 李方桂，上古音研究（1980），http://www.kaom.net/ny_word. php（2024-12 閲覧）

[51] 犬飼隆，古代日本の文字世界，大修館書店（2000）

[52] 小倉肇，日本呉音の研究，新典社（2014）

[53] 小倉肇，続・日本呉音の研究，和泉書店（2014）