

文庫本小説の自動点字翻訳／朗読音声出力 のための漢字かな変換の基礎実験

A basic experiment on Kanji-Kana translation for automatic Braille translation and reading voice output of novel books.

島田 恭宏 田中 良和 塩野 充
Yasuhiro SHIMADA Yoshikazu TANAKA Mitsuru SHIONO

岡山理科大学 工学部 電子工学科
Okayama University of Science

あらまし 小説書物の点字翻訳と朗読音声出力のための文節を単位とした、漢字かな変換の一手法について報告を行う。本実験は、パーズリスト作成を行う部分（パーズリスト作成部）と、どの単語によって文節が構成されているか検定を行う部分（文節構造検定部）から成る。パーズリスト作成部では、まず入力された漢字かな混じり文に対し、字種の変化する位置を抽出することによって文節へと分割する。次に、文節毎に単語辞書との照合を行い、候補単語を挙げパーズリストを作成する。文節構造検定部では、単語間の接続関係、文節末となる単語の検定を行うことによって文節を成すことのできる単語を選択する。これらの処理により選択された単語を辞書に登録している読みと置き換えることによって、漢字かな混じり文をかな文に変換する。また読みと置き換える際、点字の規則を用いて分かち書き処理を行う。

Abstract In this report, an experiment on Kanji-Kana translation for Braille translation and reading voice output using a Bunsetsu in Japanese sentence is described. This experiment consists of two units, that is, a parse list generator unit and a Bunsetsu structure examination unit. In the former unit, first, a Kanji-Kana sentence is segmented into Bunsetsu, and then, parse list is made using a word dictionary for each Bunsetsu. In the later unit, examinations of words connection are executed to extract correct words by consulting a Bunsetsu structure. To convert a Kanji-Kana sentence into a Kana sentence, KANJI is replaced with articulation using the dictionary. After that, spaces are inserted between words in a Kana sentence by the rule of Braille.

1. まえがき

視覚障害者の小説等の文字情報獲得は、点字本や朗読テープにより行われている。しかし、晴眼者が読むような活字本に比べると、

- ①種類が少なく読む本を自由に選べない
- ②速報性に乏しい

（新刊などすぐに入手できない）

などの問題点を抱えている。これらの問題点を解決するには、翻訳作業の機械化を目指し、かつ各個人の様々な読書要望にきめ細かく対処するためには、単に機械化を目指すのではなく、パーソナルコンピュータをベースとした個人でも所有できるようなローコストな点字翻訳システムが必要である。そこでまず入力装置として文庫本小説のための文字認識実験の報告を以前

行った⁽¹⁾。本報告は、その第2報として、自動点字訳／朗読音声出力のための漢字かな変換の基礎実験を行ったことに関して報告を行う。

変換対象となる一般の日本語文は漢字かな混じり文である。日本語文を機械により点字に変換する場合、あるいは音声出力しようとする場合、漢字の処理が問題となる。つまり漢字は表意文字であり、一般に使用される点字は表音文字である⁽²⁾。また、音声を出力する場合、音声合成装置自体がなんらかの変換システムを持たない限り⁽³⁾、表音文字であるかな文字に漢字を変換する必要がある。また、単に漢字かな変換を施しただけでは、かなのべた書き文となっ

↑点字には漢字を表す漢点字もある。しかしまだ一般的ではなく、かつその形式も統一されていないため本研究では、かな点字が一般的と考えた。

て読みにくく、又、点字翻訳しようとする場合、分かち書きが必要である。よって本報告において分かち書きは、点字の規則を用いて処理することとした。

図1に今回報告するシステムのブロック図を示す。システム内に漢字かな混じり文を一文単位で入力しているが、処理単位としては文節を単位としている。そこでまず、入力文に対して字種情報を用いた文節切りを行う。次に文節内における候補単語をすべて挙げる。ここで、辞書との照合は、最長一致の原則を用いている。これらの候補単語から文節構造をなすものを選択、処理対象となっている文節内を辞書の読みと置き換えて変換を行う。また置き換えを行う際、点字の分かち書きの規則を用いて分かち書きを加える。

以下の章より具体的な漢字かな変換手法について述べる。

2. 漢字かな変換用辞書

漢字かな変換は、基本的には単語を変換することで行っている。変換用の単語辞書は、漢字自立語、ひらがな自立語、付属語の各辞書を用意した⁽⁴⁾⁽⁵⁾。漢字かな変換処理なので漢字自立語辞書のみでもよいが、分かち書きを行う場合、各単語の品詞情報を用いて処理を行うため、ひらがな部分の単語情報も必要となる。従って、ひらがな自立語、付属語の各辞書も用意した。変換に用いる単語辞書の詳細について以下に述べる。

2.1 漢字自立語辞書

漢字で始まる自立語を登録した辞書であり、その形式は見出し、読み、文法情報である。体言など活用が無い単語に関しては単語全てを見出し語とし、用言は、語幹のみを見出し語とした。ここで、上一段、および下一段活用については単に一段活用とすることで登録している。これによって語幹と語尾の一体となった動詞についても見出し語をたてることができる。用言以外の単語については固有名詞、接辞なども一括してこの漢字自立語辞書に登録している。図2(a)に漢字自立語辞書の例を示す。

2.2 ひらがな自立語辞書

ひらがなで始まる自立語を登録した辞書で、形式および登録語については漢字自立語辞書と

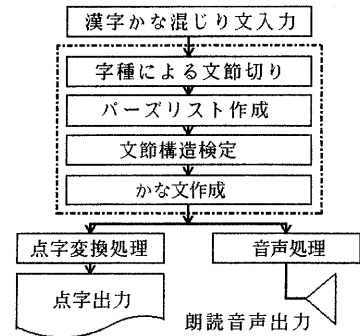


図1. システムブロック図
Fig.1 Block diagram of the proposed system.

同様である。この辞書はおもにひらがな表記される単語を中心に登録している。漢字表記する単語でも場合によっては（幼児用図書など）ひらがな表記をする場合があり、ある程度、漢字自立語辞書に登録した単語もひらがな自立語辞書に登録すべきであるが、今回作成した辞書にはこれらの単語は登録していない。ひらがな自立語辞書には漢字との混ぜ書き語も登録している。頻繁に見られる例としては女性の名前で「子」のみが漢字の場合などである。これら二つの自立語辞書は、パソコンのOSに付属している日本語入力フロントエンドプロセッサの辞書等を基に文献⁽⁵⁾を参考に作成した。図2(b)に漢字自立語辞書の例を示す。

2.3 付属語辞書

付属語（助詞、助動詞）に関して登録した辞書である。辞書形式は、前述の2つの辞書とは異なり見出し語、前接続条件、品詞活用形情報としている。見出し語は、助動詞に関しては各活用形に展開し、各々の活用形を異なる単語として別々に登録した。なお、今回作成した辞書は付属語単独のもののみを登録している。前接続条件は、その付属語の前に来ることができる品詞活用型、活用形情報は付属語毎に限定できるため、それらを示す情報を記述している。品詞活用形情報は、付属語の品詞、活用形を示している。付属語辞書は、文献⁽⁶⁾⁽⁷⁾を参考に作成した。図2(c)に漢字自立語辞書の例を示す。

表1に以上3つの辞書の登録語数を示す。

哀愁 アイウ 20	かき集め かアツメ 67
哀惜 アイキ 23	かき分け かワク 66
哀悼 アイウ 23	かき揚げ かアク 67
哀別 アイヅ 20	かき乱 かミダ 56
愛 アイ 41	かくして かシテ 11
愛 アイ 72	かくして かシテ 1C
愛くるし アイクルシ 02	かぐわし かワシ 02
愛し合 アイシア 64	かけ かけ 66
愛で め 66	かけ かけ 67
愛らし アイラシ 02	かけつけ かけツケ 67

(a)漢字自立語辞書 (b)ひらがな自立語辞書

う 0 @1_3
う 0 @1_4
か 1_2_3 #4
か 1_2_3 #7
かしら 1_2_3 #7
から 1_2 #5
から 3 #3
が 1_2 #5
が 3 #3
きり 1_4 #1

(c)付属語辞書

図 2 . 各辞書の例

Fig.2 Examples of each dictionary.

- (a)Kanji independent word dictionary.
- (b)Kana independent word dictionary.
- (c)An auxiliary word dictionary.

表 1 . 各辞書の登録語数

単語辞書	登録語数
漢字自立語辞書	6 6 4 7 2
ひらがな自立語辞書	6 7 9 0
付属語辞書	2 1 5

2. 4 単語辞書の検索

図 3 に単語辞書の構造を示す。各単語辞書はインデックスと辞書本体からなる。各辞書本体は、見出し語の先頭文字のシフト JIS コード順に並べている。管理単位は、辞書内容 5 1 2 バイトとし、見出し語の先頭文字が同じ単語を一つのブロックに格納する。もちろん 1 ブロックに納まらない場合は複数のブロックに渡ることとなる。そしてこの一つのブロックに対して 1 つのインデックスを与えており、インデックスは見出し語の先頭文字を持つ。インデックス自体は図のように 1 次元配列をなしており、その要素番号が辞書ファイルにおける 1 ブロック

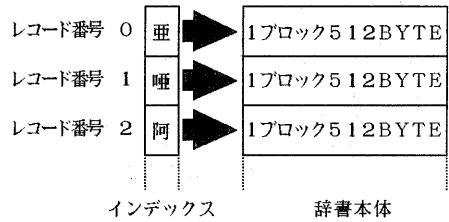


図 3 . 辞書構造

Fig.3 Structure of the dictionary.

のレコード番号を示す。単語の辞書照合を行う場合、このインデックスを用いて辞書ファイルにおいてその単語と同じ先頭文字を持つブロックの最初の位置を検出し、そのブロック内の単語と文節内単語の検索を行う。その位置から以後のブロックについても同様に単語検索を行う。

3. 変換アルゴリズム

前述の変換辞書より単語の変換を行うための諸情報を得る。変換は単語を単位としているが、一度に漢字かな変換する処理単位としては文節を単位としており、以下より変換アルゴリズムについて示す。

3. 1 文節切り^{(4) (8)}

漢字かな混じり分を文字種の移り変わる位置(ひらがな→漢字など)によって文節単位に切る。記号「。」、「、」などは別の処理を施すためそれ単体で文節と見なすようにしている。この処理で完全に文節に区切ることができるわけではなく、また、文法上のそれとは異なるものであり、処理をする上での単位である。つまり長い文を一度に処理するのではなく、この文節を単位として処理を行う。単語照合を行う場合、最長一致の原則を用いているためその時間的効果は大きいと考える。

3. 2 パーザリスト作成

ここでは、前述の処理により決定された文節内において、その文節を構成していると考えられる候補単語をすべて挙げる。図 4 にこの処理部での流れを示す。

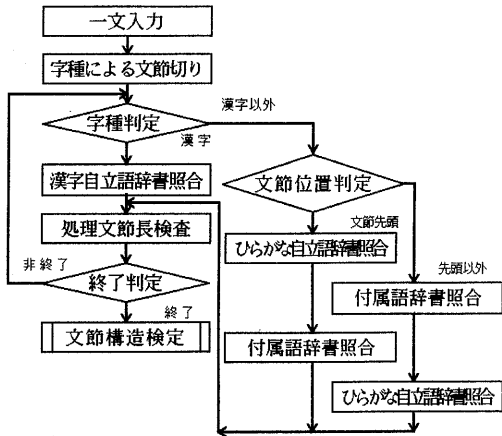


図4. パーズリスト作成部のジェネラルフロー
Fig.4 General flowchart of the parse list generator unit.

3. 2. 1 処理対象文節

処理対象は1つの文節であるが、場合によっては先の文節切りで1つの単語が文節に分けられてしまう場合がある。すなわち「成り立つ」という単語は、「成り」と「立つ」に分割されてしまい、正確な単語照合ができない。よって、単語照合を行う場合、接続する2つの文節を処理対象とし、最初の文節内において照合した単語の最長の単語が次の文節にかかるならば、その2つの文節を1つの文節と見なし、そうでなければ独立した2つの文節として以後の処理を行う。

3. 2. 2 単語照合

変換処理を行う上でまず必要なことは、照合すべき単語辞書の選択である。これは、文節ないし対象単語の先頭文字の字種によって行う。すなわち、先頭文字が漢字であれば漢字自立語辞書を、またひらがなであればひらがな自立語辞書を選択する。付属語辞書は、文節の構造が単純にはく自立部+付属部>となるので、文節内において最初の単語を照合した後、残りの文字列がひらがなから始まっている場合にのみ付属語辞書の照合を行う。但しこの場合、文節が自立語のみでも構成できるため、ひらがな自立語辞書の照合も行っておく。文節先頭の単語を照合した後、次の単語の照合は、直前に行った照合単語の最短のものの次の文字から同じように照合を行う。このような照合処理により、不

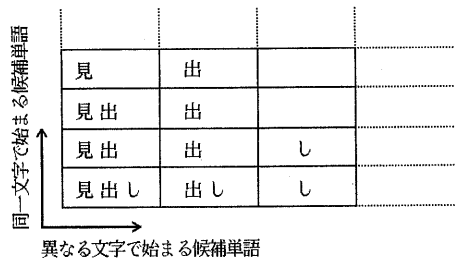
必要な候補単語も挙がってしまうが、文節内における候補単語の照合漏れをなくす目的でこのような手法をとった。各辞書における照合手順を次に示す。

①漢字自立語辞書照合

対象文字列の先頭が漢字のとき、漢字自立語辞書と照合する。一致するものがあれば候補語スタック(後述)に積む。また、品詞情報より用言であれば活用語尾の検定を行い、活用形および活用語尾も候補語スタックに積む。活用語尾検定は、各活用形に対応する活用語尾だけを納めたテーブルを用意し、品詞情報が示す活用形のテーブルを選択し、語幹の次の文字列と照合を行う。ただこのとき、例えば五段活用の終止形と連体形、仮定形と命令形の活用語尾は同一のものとなり、一意に決めることができない。よって、複数の活用形が存在する場合には、全ての活用形、及び活用語尾を候補語スタックに積むこととした。ここで活用語尾検定がうまく行かない場合も語幹のみの照合として候補語スタックに積む。この候補語スタックイメージを図5に示す。

②ひらがな自立語辞書照合

対象文字列の先頭がひらがなのとき、あるいは次の照合位置の文字列の先頭文字がひらがなの場合、ひらがな自立語辞書と照合し、一致するものがあれば候補語スタック(後述)に積む。



スタック1セルの要素

見出し語	1文節において見出し語の始まる位置
読み(前後読条件)	活用形
品詞情報	品詞
(品詞活用型活用形条件)	活用語尾
見出し語長	活用語尾長

図5. 候補語スタックの構成
Fig.5 Structure of a stack of candidate words.

また漢字自立語辞書照合と同様に、品詞情報より用言であれば活用語尾の検定を行い、活用形および活用語尾も候補語スタックに積む。また、活用語尾検定がうまく行かない場合は、語幹のみの照合として候補語スタックに積む。

③付属語辞書照合

文節の前段の辞書照合で検定された文字列を除いた残りの文字列の先頭がひらがなである場合、まず付属語辞書の照合を行う。一致するものがあれば候補語スタックに積む。付属語（助動詞）の場合、活用形が存在するが、活用形を別の単語として各々登録しているため活用語尾検定は行わない。

以上の処理で挙げられた候補語は、2次元配列状の候補語スタックに積む。

3.3 文節構造検定

作成されたパーズリストより、この処理で文節構造の検定を行い、それにより選択される単語を読みと置き換え、漢字かな変換を行う。またその際に、分かち書き処理も実行する。

3.3.1 文節構造⁽⁸⁾

筆者らは、ある程度の文法情報を用いて、より正確な処理が実現できないかと考え、日本語文における文節に注目した。つまり、漢字かな混じりのべた書き文をはじめになんらかの方法で文節に分離し、その文節において含まれるであろう候補単語をすべて挙げた後、文節条件を満たす単語を選択し変換しようというものである。日本語文と文節との関連を以下に示す。

日本語文:=〈文節〉〈文節〉…〈文節〉

文節:=〈自立部〉|〈自立部〉〈付属部〉

自立部:=〈自立語〉|〈接頭語〉〈自立部〉|

〈自立部〉〈接尾語〉|

〈数詞〉〈助数詞〉

付属部:=〈付属語〉|〈付属部〉〈付属語〉

よって文節は

〈文節〉:=〈自立部〉〈付属語〉…〈付属語〉

と表される。これらの関係を基に、各部、語の接続を検定し、文節構造をなすかどうか検定してもよいが、本研究では、もう少し簡単な構造検定手法をとった。そのための文節構造に関する定義を以下に示す。ただし、日本語文 $s = s(1)s(2)\dots s(n)$, ($s(i)(i=1,2,\dots,n)$ は文字) が与えられているものとする。

[定義1] 単語構造

単語 w の綴り W (活用語の場合は終止形の綴り)、品詞 H 、活用情報 K からなる3項系列 (W, H, K) を単語 w の単語構造と呼ぶ。

[定義2] 述語 WS, J, E, C

I) 入力記号列 s の部分列 $s(i+1)s(i+2)\dots s(j)$ なる単語 w が存在し、その単語構造が α であることを $WS(i, j, \alpha)$ で表す。

II) 単語構造 α が自立語の単語構造であることを $J(\alpha)$ で表す。

III) 単語構造 α_1 の単語 w_1 と単語構造 α_2 の単語 w_2 の接続 w_1w_2 が文節の部分列になり得ることを $C(\alpha_1, \alpha_2)$ で表す。

IV) 単語構造 α の単語が文節末の語になり得ることを $E(\alpha)$ で表す。

これらの述語を用いて文節構造規則を次に示す。

[定義3] 文節構造規則

長さ n の入力記号列 s に対して、 $i_0 (=0) < i_1 < \dots < i_m (=n)$ なる整数列 i_0, i_1, \dots, i_m と単語構造 $\alpha_0, \alpha_1, \dots, \alpha_m$ が存在し、以下を満たすとき s は文節をなすという。

I) $WS(i_{k-1}, i_k, \alpha_k)$ ($k=1,2,\dots,m$)

II) $J(\alpha_1)$

III) $C(\alpha_{k-1}, \alpha_k)$ ($k=2,3,\dots,m$)

IV) $E(\alpha_m)$

上記の規則に従って、文節構造規則を満たす候補単語を選択し、変換を行ってゆく。但し、規則 I に関しては、パーズリストを作成した時点で単語の存在、単語構造をなしているかどうか判定できるので、ここでは特に処理は行っていない。

3.3.2 単語間接続条件

前述の定義における III の述語 C に関する各品詞における接続条件を記す。

①自立語と自立語の接続⁽⁸⁾

自立語同士の接続の場合、おもに複合語、あるいは派生語と考えられ、次のような条件により接続検定を行う。

[複合語]

I. サ変動詞語幹 + 名詞 → 名詞

II. 名詞 + 名詞 → 名詞

III. 名詞 + 動詞連用形 → 名詞

IV. 動詞連用形 + 動詞 → 動詞

V. 動詞連用形 + 名詞 → 名詞

VI. 動詞連用形 + 形容詞 → 形容詞

VII. 動詞連用形 + 動詞連用形 → 名詞

但し、漢語による複合語の係り受け処理は含んでいない。また前の単語が人の姓の場合、次の単語は人の名を優先するようにしている。

[派生語]

今回作成した辞書の各単語（自立語）の品詞情報は、細分類しており、この情報を基に接続条件とする。例えば接尾語に関して、人の姓名について名詞をつくる（「君」、「さん」など）、地名について名詞をつくる（「駅」、「港」など）などである。

②自立語と付属語および付属語同士の接続

自立語とその直後に接続する付属語は、その接続する付属語毎に自立語の品詞、活用形が決まることが分かっている。また付属語同士の接続も、後続の付属語によって直前の付属語、およびその活用形が決まることも分かっている。よって、実際には口語において使用される自立語と付属語、付属語同士の接続関係を文献⁽⁶⁾、⁽⁷⁾を参考にビットテーブルで表現（ビットテーブルは、助動詞と助詞別々に作成した）し、その内容を調べることで接続検定を行う（図6参照）。

この他に、単独で文節をなす単語（付属語、他の自立語を伴わない単語、接続詞など）は接続検定は行わない。なお原則的には付属語への自立語の接続は禁止している。

3.3.3 文節末語条件

一つの文節内における単語の接続検定が終了したならば、最後に位置している単語の品詞・活用形情報より文節末になり得る単語かどうかを検定する。すなわち、条件としては活用語の未然形、仮定形でないことを必要とする⁽⁷⁾。

前述の定義では、 w_1w_2 と2つの単語しか扱わなかったが、実際にはそれ以上の候補が存在するのが一般的であり、かつどの単語とどの単語が接続でき文節を構成するかは不明である。そこで、この文節構造検定は再帰的な関数で構成し⁽⁹⁾、ある程度の情報を保存することで前段にかえって処理することを可能とする。図7にその手順を示しておく。

文節条件を満たすものを選択し、その単語を辞書項目の読みと単に置き換えると、かなのべた書き文となり読みにくい。また、点字は分かち書きを必要とする。しかし日本語文では、分かち書きの規則が定まっていないために点字の

$w_1 \backslash w_2$	形容詞	形容動詞	助動詞	形容詞型	助動詞型	形容動詞型
しめる	×	×	●●●●	×	×	●●●●
そうだ	1	1	●●●●	1	1	●●●●
そうです	1	1	●●●●	1	1	●●●●

w_1 : 付属語または自立語
 w_2 : 付属語

図6. 付属語接続テーブル
 Fig.6 The auxiliary word conjunctive table.

分かち書きの規則を用いて処理を行う。以下におおよその点字の分かち書きの規則を示す⁽²⁾、⁽⁴⁾、⁽¹⁰⁾。

- (1)自立語は前を切って書く
- (2)付属語は自立語につけて書く
- (3)合成語は構成単語単位に切って書く
- (4)接頭語、および接尾語は自立語につけて書く

実際の分かち書きは、前述の文節構造検定において、文節における単語を辞書の読みと置き換える場合に、処理対象となる単語の品詞を調べ、自立語であるならば読みと置き換える前に分かち書きの区切りを挿入する。但し、(4)より自立語でも接辞の場合は、分かち書きの区切りは挿入しない。

その他必要な処理としては、助詞の「は」、 「へ」に関しては、点字の場合、「わ」、「え」とすること、また一般に発音する場合にも「わ」、「え」とすることから、変換処理を施す。

3.3.4 点字変換⁽²⁾、⁽¹⁰⁾

ここでは、目的の1つである点字変換について述べておく。点字は3行2列の6つの点の有無で表現される。よってその種類は（空白を含めて） $2^6 = 64$ 種と限られており、表現できる文字数は限られる。また表音文字であるが、かなとすべて一対一の対応にはなっていない。例えば「ぎゃ」、「ぎゅ」、「ぎょ」はそれぞれ

- 「ぎゃ」→「幼濁音記号」+「か」
- 「ぎゅ」→「幼濁音記号」+「く」
- 「ぎょ」→「幼濁音記号」+「こ」

示しておく。また点字変換例を図9にあげておく。但し、この点字は凸点を黒点で表示した形状であり、一般の日本語文同様に左から読む。

5. むすび

本稿では、文庫本小説の自動点字訳／朗読音声出力のための、文節を用いた漢字かな変換の一手法の提案を行った。処理対象をおもに文庫本小説としているが、一般の日本語文を扱っているので、入力データの入手が可能であれば、文庫本にとどまらず新聞などの他の文書への応用も可能と考える。また、処理途中で用いている単語、あるいは文節の情報は構文解析など他の分野への基礎段階としても活用できるものと考えている。今後、規則化などの作業を進めると共に、問題となるであろう同字異語、漢字列など⁽³⁾の問題を解決しつつ実験を進めて行きたいと考える。

[参考文献]

- (1) 島田, 塩野: "文庫本点字訳のためのパソコンによる印刷文字認識", 信学技法, PRU87-92.
- (2) 本間, 岩橋, 田中: "点字と朗読への招待", 福村出版, 1983.
- (3) 長倉, 箱田, 壁谷, 平原: "文・音声変換ユニット", 研実報, 37, 4/5, 1988.
- (4) 野村, 森: "漢字かな変換システムの試作", 信学論(D), Vol. J66-D, No. 7, 1983.
- (5) 長尾, 辻井, 山上, 建部: "国語辞書の記憶と日本語文の自動分割", 情報処理, 19, 6, 1978.
- (6) 久松, 佐藤編: "角川国語辞典(第164版)", 角川書店, 1977年発行.
- (7) 遠藤嘉基監修: "対照日本文法(第72版)", 中央図書, 1979年発行.
- (8) 長尾 真監修: "日本語情報処理", 電子情報通信学会, 1984.
- (9) 田中, 佐藤, 元吉: "自然言語処理のためのプログラミングシステム-拡張LINGOLについて-", 信学論(D), Vol. J66-D, No12, 1977.
- (10) 本間一夫: "標準点字表記辞典", 日本盲人福祉研究会, 1982.

吾輩は猫である
ウ ガ ハ イ ワ ネ コ テ ア ル。

名前はまだ無い。
ナ マ エ ワ マ タ ナ イ。

どこで生まれたか頼と見当がつかぬ。
ト コ テ ウ マ レ タ カ ト ト

ケン ト ウ ガ ッ カ ヌ。

何でも薄暗いじめじめした所で
ニャーニャー泣いていた事だけは記憶している。
ナ ン テ モ ウ ス ク ラ イ

シメ シメシタ トコロテ

ニャーニャー ナイテイタ

コト タ ケ ワ キ オ ク シ テ イ ル。

図9. 点字変換例

Fig.9 Examples of Braille-Kana translation.