# 中国語を対象とした意味訳型翻字手法

**黄　海湘**

図書館情報メディア研究科

筑波大学

2012年4月

# A Method of Semantic Transliteration for Chinese

**Huang HaiXiang**

Graduate School of Library,

Information and Media Studies

University of Tsukuba

April, 2012

# 概要

科学技術や文化の発展に伴って，新しい技術や概念を表す専門用語や固有名詞が次々に作られている．そこで，外国語の文化を取り入れるために，外国語の新語を迅速に母国語へ翻訳する必要性が高まっている．

外国語を翻訳する方法には「意味訳」と「翻字」がある.「意味訳」は原言語の意味を翻訳先の言語で表記する方法である.「翻字」は原言語の発音を翻訳先の言語における音韻体系で表記する方法である．企業名や商品名のような固有名詞や専門用語は翻字されることが多い．

外国語の専門用語や固有名詞を翻字するときに，日本語や韓国語ではカタカナやハングルなどの表音文字を用いる．中国語では漢字を用いて翻字する．しかし，漢字は表意文字であり，同じ発音に複数の漢字が対応している．また，異なる漢字が異なった意味と印象を持っているので，中国語へ翻字する際に，使用する漢字の選択に注意を払う必要がある．この点は外国の企業が自社の企業名や商品名を中国に輸入する際に特に重要である．さらに，人名や地名のように，漢字の選択は原言語が属する種別に関連している場合が多い．従って，中国への翻字には単に原言語の発音だけではなく，漢字の意味や原言語が属する種別も考慮する必要がある．

中国語への翻字に関する従来の研究は，主に発音と言語モデルの組み合わせである．本研究は，翻字対象の種類を制限せずに，発音，意味，および種別をモデル化する確率的な意味訳型翻字手法を提案する．

翻字対象が与えられた場合，本研究の発音モデルはその翻字対象を発音に似ている複数の翻字候補に変換する．ただし，ローマ字表記に変換できれば，ほかの言語を入力することも原理的に可能である．翻字対象の関連語が与えられた場合，本研究の意味モデルはこれらの関連語を関連する漢字のセットに変換する．ただし，関連語が中国語ではない場合，機械翻訳を利用して中国語に翻訳する．翻字対象の種別が与えられた場合，本研究のカテゴリモデルは対象種別に良く出現する漢字のセットを選択する．この 3 つのモデルを統合するために，本研究は確率的な枠組みを提案する．この確率的な枠組みより，確率が高い翻字候補は出力されるリストの上位に出現し，提示する．さらに，人手で翻字対象のために 1 つ以上の関連語を与えるのは手間がかかり高価である．本研究は Web を利用して，自動的に翻字対象の関連語を抽出する手法も提案する．翻字対象の情

報源として，Wikipedia の日本語版を使用する．

翻字評価実験で本研究が提案した自動翻字手法の有効性を評価する．本研究は意味モデル，カテゴリモデル，およびカテゴリモデルの適応について評価した．発音モデルのみの翻字手法は意味モデルとカテゴリモデルのそれぞれによって改良された．3 つのモデルを結合し，カテゴリモデルを適応させた場合，最も良い翻字結果が得られた．さらに，翻字における関連語自動抽出の有効性も評価した．自動抽出した関連語を利用した意味モデルの翻字結果は意味モデルを利用しない手法と比べると，より効果的だった．また，翻字精度を多少犠牲にして，人手で関連語を与えるコストを削減することができた．本研究は，自然言語処理における中国語への意味訳型翻字に関する重要な一歩である．

## Abstract

Reflecting the rapid growth of science, technology, and culture, words related to new concepts and proper names have progressively been created. These words have also been imported into different languages. These new words have also been imported into different languages.

There are two fundamental methods for translating foreign words into a language. In the first method–*translation*–the meaning of the source word in question is represented by an existing or new word in the target language. In the second method–*transliteration*–the pronunciation of the source word is represented by using the phonetic alphabet of the target language, such as Katakana in Japanese and Hangul in Korean. Technical terms and proper nouns are often transliterated.

To transliterate foreign words, in Japanese and Korean, phonograms such as Katakana and Hangul are used. In Chinese, the pronunciation of a source word is spelled out using Hanzi characters. Because Hanzi is an ideogrammatic script, a single pronunciation can be represented by more than one character. However, because different Hanzi characters convey different meanings and impressions, Hanzi characters must be selected carefully during transliteration into Chinese. This is especially important when foreign companies intend to introduce their names and products into Chinese-speaking communities. In addition, the selection of Hanzi characters often depends on the category of a source word, such as a person or place. Therefore, the process of transliterating to Chinese must consider not only the pronunciation but also the meaning of the Hanzi characters and word categories.

The existing methods for transliteration into Chinese are almost all combinations of the pronunciation and language. We propose a probabilistic semantic transliteration method that models pronunciation, lexical semantics, and word categories, without restricting the type of the source word.

Given a source word, our pronunciation model converts it into a set of transliteration candidates having pronunciations similar to that of the source word. In principle, any language that uses a phonetic script can be a source language for our method. Given

the related terms for the source word in Chinese, our lexical semantics model converts them into a set of Hanzi characters. If the related terms are not in Chinese, we machine translate them into Chinese. Given the category of a source word, our category model chooses a Hanzi character often used in the category to which the source word belongs. To combine the three models, we propose a probabilistic framework, which derives possible transliteration candidates and sorts them according to their probability. In addition, because providing one or more related terms for the source word manually is time-consuming and expensive, we automatically extract related terms for source words using the Web. We consult the Japanese Wikipedia as a resource for source words.

We show the effectiveness of our method experimentally. We evaluated the effectiveness of the lexical semantics and category models, and of the adaptation of the category model. A transliteration method that models only the pronunciation was improved by the lexical semantics model and category model independently, and the best result was obtained when we combined the three models and adapted the category model to the source word. In addition, we evaluated the effectiveness of the automatic extraction of related terms. The use of automatically extracted related terms combined with the lexical semantics model was more effective than the method that does not use the lexical semantics model. We also reduced the manual cost for providing related terms, sacrificing some transliteration accuracy. Our research is the first significant exploration of semantic transliteration in natural language processing for Chinese.

# 目 次

# 表 目 次

# 図 目 次

# 第1章　序論

## 1.1　研究の背景

近年，科学技術や経済の発展に伴い，新しい専門用語や固有名詞が次々に作られている．これらの新語はインターネットによって世界中に発信される．そこで，外国の文化を取り入れるために，外国語の新語を迅速に母国語へ導入する必要性が高まっている．

外国語を導入するには3つの方法がある．1つ目は意味訳である．意味訳は原言語の意味を翻訳先の言語で表記する方法である．例えば，英語の「address」は日本語では「住所」と訳され，中国語では「住址 /zhu–zhi/」と訳される．2つ目は翻字もしくは音訳であり，原言語の発音を翻訳先の言語における音韻体系で表記する方法である．例えば，英語の「adidas」は，日本語では「アディダス」と訳され，中国語では「阿迪达斯 /a–di–da–si/」と訳される．3つ目は外国語を原言語のままで使う方法である．しかし，この方法は翻訳先の言語における語の意味の理解性と可読性を損なう可能性がある．なお，本論文では中国語の発音をスラッシュ（/）で囲まれたピンインで表記する．

意味訳では，正しく外国語の意味を表すために，既存の語から訳語を選択するか，新しい語を生成する必要があり，高価である．そこで，外国語は翻字されることが増えてきた．特に，固有名詞や専門用語は翻字されることが多い．中国語においても同じ状況である．しかし，日本語や韓国語などの言語と比較すると，中国語への翻字は複雑である．

日本語や韓国語はカタカナやハングルなどの表音文字を用いて外国語を翻字する．それに対して，中国語には漢字しかないので，漢字を用いて翻字する．しかし，漢字は表意文字であるため，同じ発音に複数の文字が対応する．その結果，同音異義の問題が発生する．すなわち，翻字に使用する漢字によって，翻字された言葉に対する意味や印象が変わってしまう場合がある．

例えば，飲料水の名称である「コカコーラ（Coca-Cola）」に対して，様々な漢字列で発音を表記することができる．公式な表記は「可口可乐 /ke–ko–ke–le/」であり，原言語

と発音が近い．さらに,「可口」には「美味しい」,「可乐」には「楽しい」という意味があり，飲料水の名称として良い印象を与える．他方において,「Coca-Cola」の発音に近い別の漢字列として「口卡口拉 /ko–ka–ko–la/」もある．しかし,「口卡」には「喉に詰まる」という意味があり，飲料水の名称としては不適切である．

また,「人名」や「地名」といった翻字対象の種別によっても使用される漢字の傾向が異なる．例えば,「宝」と「堡」の発音はどちらも/bao/である.「宝」には「貴重」や「宝物」などの意味があり中国語で人名や商品名によく使われるのに対して,「堡」には「砦」や「小さい城」などの意味があり中国語で地名によく使われる．

以上より，外国語を中国語に翻字する場合は，発音だけではなく，漢字が持つ意味や印象，さらには，翻字対象の種別（人名や企業名など）も考慮して漢字を選択する必要がある．本論文では，このような翻字を発音だけを考慮する翻字と区別して,「意味訳型翻字（semantic transliteration）」と称する．意味訳型翻字は，漢字を使う中国語や日本語などにおいて重要である．漢字圏への進出を計画する企業にとっては，企業名や商品名のネーミングにおいて意味訳型翻字の果たす役割が大きい．

## 1.2　中国語への翻字に関する歴史的変遷

中国語への翻字において漢字の意味を考慮する必要性は古くから認識されていた．

中国語は表意文字である漢字を用いて表記する．現在，主に中華人民共和国（中国大陸），香港，マカオ，台湾，シンガポールなどの国と地域で使われている．しかし，同じ中国語でも，国と地域によって使う漢字の字体と発音の表記方法が異なる．中国大陸では，漢字の字体は読みや構成にも統一性を高めた簡体字が使われている．発音の表記は「拼音 /pin–yin/（ピンイン）」を使っている．一方，香港，マカオ，台湾などでは，主に簡体字以前の繁体字が使われている．さらに，台湾では，発音表記は「ㄅㄆㄇㄈ /b–p–m–f/」のような「注音符号」が使われている．簡体字は繁体字を元にして作られている．漢字の形だけが変化して，元の繁体字の意味は変わらない．以降は便宜上，簡体字とその発音表記であるピンインを使用する．

中国語への翻字活動の歴史は長く，後漢（25 年～ 220 年）の時代から始まった仏教経典の翻訳に遡る．当時，翻訳者らはサンスクリット語（梵語）で書かれている仏教経典を中国語に翻訳するため，発音に従って数多くの新しい漢字と語を生成した．ただし，こ

の時代の翻字は原言語の発音しか考慮していない.

明朝後期の科学者徐光启 /xu–guang–qi/（1562–1633）は翻訳対象の意味を考慮する翻字手法を進展させた [9]．彼はイタリア人の修道士 Matteo Ricci と一緒に「The Original Manuscript of Geometry」という本を翻訳する際に，単語「geometry」を「几何 /ji-he/（幾何)」と翻字した．漢字「几」には「どのぐらい大きい」の意味があり,「何」には「どんな形」の意味がある．この二文字の組合せは/geo-/の発音と類似しており，さらに,「地球を測ることに関心を持つ学問」の意味をうまく表現している．

同じ時代に「geometry」には「形学 /xing–xue/」という別の訳語も存在した．この訳語は翻字ではなく，意味訳である.「几何」と「形学」は暫くの間で併用されていた．しかし，20 世紀中期から「geometry」の訳語として,「几何」は完全に定着し,「形学」は淘汰された．この歴史的事実は，翻字対象の発音だけではなく，意味も考慮した訳語が意味訳より長く使われる一例である．

ここ数十年は，世界各国との交流が増えるにつれ，外来語が大量に入るようになり,「几何」のような翻字も生成された．例えば,「ベンツ（Benz）/奔驰 /ben–chi/ 奔：まっしぐらに向かっていく 驰：疾走する」,「SUNTORY（サントリー）/三得利 /san–de–li/ 顧客，社会，会社の三者とも利益を得るように」,「shampoo（シャンプー）/香波 /xiang–bo/ 薫る波のように」などがある．

一方，1 つの発音に複数の漢字が対応しているため，実際翻字を行う際に，どの漢字を使うかを決める規則がない．そのため，外来語が初めて中国語に翻字される際に，様々な組合せが使われている．これは利用者にとって，非常に不便なことである．従って，外来語が急増している今，中国語への翻字の標準化も重要である．中国語への翻字における標準化の作業は数十年前から行われている．特に，人名と地名に対して，新華通信社の訳名室が発音を考慮して，様々な翻字規則を制定した．その結果,「世界人名翻訳大辞典」[36] が出版された．

しかし,「世界人名翻訳大辞典」の翻字規則は人名と地名しか対象としていない.「奔驰」,「三得利」,「香波」などのような企業名や商品名や一般名詞などには対応できない．さらに，人名や地名でも実際の翻字作業の中でうまく機能していない場合がある．例えば，ミャンマーの女史 “Aung San Suu Kyi（アウンサンスーチー）” の名前は “昂山素季 /ang–shan–su–ji/”，“昂山素姬 /ang–shan–su–ji/”，“翁山蘇姬 /weng–shan–su–ji/” と “翁山淑枝 /weng–shan–shu–zhi/” に翻字されている．しかし，これらの翻字の

中に，新華通信社の “昂山素季” を含めて，翻字規則に従う名前は 1 つもない．

## 1.3　研究の目的

1.1 節で述べたように，漢字は表意文字であるため，同じ発音に複数の文字が対応する．その結果，同音異義の問題が発生する．すなわち，翻字に使用する漢字によって，翻字された言葉に対する意味や印象が変わってしまう場合がある．

さらに，1.2 節で述べたように，現在制定されている翻字規則では，翻字対象が限定されている上，不十分である．

本研究は，翻字対象に関する「発音」,「意味」,「種別」を考慮した漢字選択の手法を提案し，中国語への意味訳型自動翻字を目的とする．本研究は簡体字か繁体字の区別によらず，表意文字である全ての漢字への翻字に対応している．ただし，制限を加えても一般性は失わないため，便宜上簡体字とその発音表記であるピンインを使って説明する．

さらに，本研究は人名，地名，企業名などを問わず，翻字対象を制限しない．ただし，提案する翻字手法を現実世界へ応用する際に，特に企業名や商品名の場合，得られた翻字候補は既存の名前と商標の著作権や商標権を侵害しないように注意深く調べるべきである．しかし，この問題は本研究の範囲外である．

## 1.4　本論文の構成

2 章で翻字に関する関連研究について説明する．3 章で本研究で提案する手法について説明し，4 章で提案手法を評価する．最後に，5 章で本研究の貢献と残りの課題についてまとめる．

# 第2章　関連研究

## 2.1　翻字の分類

翻字の自動手法に関する関連研究は,「狭義の翻字」と「逆翻字」に大別することができる.「狭義の翻字」とは，外国語を母国語の文字で表記し，新しい語を生成する処理である [2, 12, 17, 18, 19, 21, 26, 28, 29]．例えば，前アメリカ大統領の名前「Bush」を中国語に翻訳するときに，発音が近い「布什 /bu–shi/」と翻字する．一方,「逆翻字」とは,既に翻字された語に対して原言語を特定する処理である [1, 5, 7, 11, 14, 15, 16, 22, 24]．例えば，既に訳された中国語の外来語「芒果 /mang–guo/」に対して原言語の「mango」を特定する．本研究の目的は狭義の翻字であるため，以降,「翻字」を「狭義の翻字」の意味で使用する．

多くの既存翻字手法 [2, 12, 17, 18, 21, 26, 28] は，目標言語が表音文字であるか，表意文字であるかを区別せずに，原言語と目標言語の発音及び目標言語だけをモデル化している．しかし，1 章で説明したように，目標言語が表意文字である漢字の場合は，翻字に使用する漢字の意味を考慮すべきである．漢字の意味を考慮する翻字手法 [19, 29] は少ない．

逆翻字は主に言語横断検索や機械翻訳に応用されている．定義により，逆翻字の場合,新しい語を生成するのではなく，対象としている語は既に翻字されていて，原言語の意味や印象を考慮する必要がない．従って，逆翻字は本研究とは目的が異なり，本研究の対象外である．しかし，原言語と目標言語の発音をモデル化して音訳を行う点では，本研究と関連する．

以下 2.2 と 2.3 節で，翻字と逆翻字に関する関連研究について説明する．

## 2.2 翻字に関する関連研究

### 2.2.1 発音と目標言語に基づく一般的な翻字手法

多くの既存翻字手法は，原言語と目標言語の発音及び目標言語をモデル化している．発音モデルを構築するために，通常，音素（phoneme）や書記素（grapheme）などの言語ユニットは原言語と目標言語を対応させることに使用される．また，目標言語のモデルを構築するために，通常，文字 N-gram モデルが使用される．そのため，翻字結果の文字列は元々目標言語にある単語に類似する．

英語の地名を中国語に翻字するため，Wan と Verspoor [28] は書記素から音素へのマッピングに基づく音訳手法を提案した．具体的には，まず，人手で発音規則を作成する．次に，英単語の書記素に対して，発音規則と実例の学習に基づき,「音節区切り」と「部分音節区切り」で音素に変換する．さらに，音素を「ピンインへのマッピング」で特定されたピイインに変換し,「漢字へのマッピング」で英語の発音と類似している中国語の漢字列に翻字する．

固有名詞を翻字するために，Virga と Khudanpur [26] は発音と言語の両方をモデル化する統計的手法を提案した．提案手法において，ピンインは英語の音素を漢字に変換するための仲介表現として使用された．

図 2.1 では彼らが提案した翻字手法の処理手順を示している．

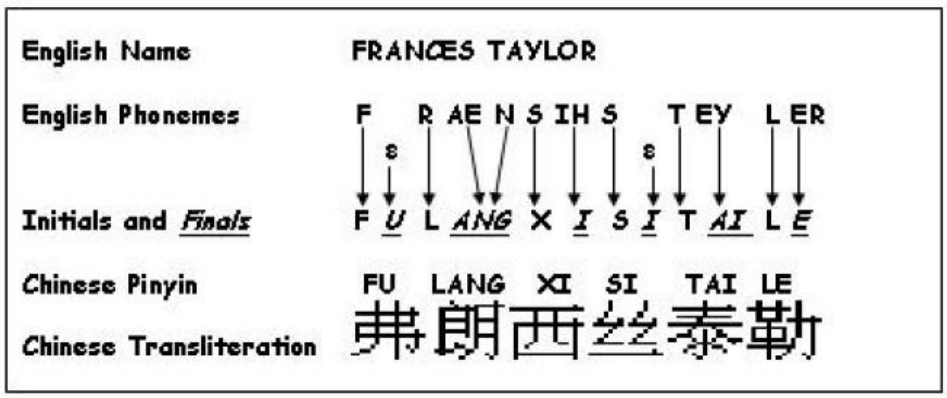


図 2.1: Virga と Khudanpur が提案した翻字手法の処理手順 [26]

以下，各ステップについて説明する．

1. 音声合成システムを用いて英語の固有名詞を音素に変換する．

2. 英語の音素列から GIFs (generalized initials and finals) の列に変換する．

3. GIF の列からピンイン列に変形する．

4. ピンイン列から漢字列に変換する．

システムの中心機能はノイズチャンネルモデル（Noisy Channel Model）である．英語単語を英語の音素に変換し，音素から一番近い中国語の発音表記ピンインに変換する．子音の挿入，変形，結合などの処理を行い，漢字に変換する．式（2.1）を用いて，英語 $\mathbf{f}$ が与えられた条件のもとで中国語 $\mathbf{e}$ が生成される条件付き確率を計算する．

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} P(\mathbf{f}|\mathbf{e}) \times P(\mathbf{e}) \tag{2.1}$$

$P(\mathbf{f}|\mathbf{e})$ は音訳モデルであり，$P(\mathbf{e})$ は言語モデルである．音訳モデル $P(\mathbf{f}|\mathbf{e})$ を計算するために，英中対訳辞書から音訳された固有名詞 3,875 対によって，英語の文字と漢字の対応頻度を計算した．言語モデル $P(\mathbf{e})$ は 3,875 の中国名から推定した．

Lee ら [17] はノイズチャンネルモデルに基づいて，英単語と対応する中国語の発音の類似度を求めた．まず，英語と中国語のテキストから二つの言語固有名詞リストを自動的に抽出した．次に，英語から中国語への変換において，中国語の発音表記である「Wade-Giles」を用いて中間言語として中継した．英語と中国語の発音との同時確率を求めるために，英語と中国語のテキストから抽出された対訳語の発音の出現頻度によって計算した．

Aramaki ら [2] は翻字作業を形態素解析の品詞特定手法と見なし，条件付確率場（CRFs: Conditional Random Fields）を利用する翻字手法を提案した．Jia ら [12] はフレーズベース統計機械翻訳（PBSMT: Phrase-based Statistical Machine Translation）の手法を翻字に適用する手法を提案した．Oh ら [21] は CRFs を含め，様々な統計機械翻訳の手法を取り入れた複合型翻字手法を提案した．これらの手法は英語と様々な言語間の翻字に対応できる．例えば，英語からロシア語，英語から中国語，英語から日本語などである．しかし，翻字するために，訓練コーパスを用意しなければならない．

Li ら [18] は人名を英語から中国語（漢字）に翻字する際，中間言語を使わず，英語の音素と漢字間の直接マッピングが可能になる接合ソースチャンネルモデル（Joint Source-Channel Model）を提案した．

英語の音素から直接漢字に変換するため，有用な正字法文脈（Useful Orthographic Context）手法を利用した．有用な正字法文脈手法は，英語名前を適切な翻字ユニットに分割するために，訓練コーパスから得られた翻字規則に従い，翻字ユニット左右の文脈の組合せを利用する．

接合ソースチャンネルモデルでは，英語と中国語の同時確率 $P(E,C)$ を推定する．$P(E,C)$ を求めるために，式（2.2）を用いた．

$$\begin{aligned} P(E,C) &= P(e_1,e_2,...,e_k,c_1,c_2,...,c_k) \\ &= P(<e,c>_1,<e,c>_2,...,<e,c>_k) \\ &= \prod_{k=1}^{K} P(<e,c>_k \mid <e,c>_1^{k-1}) \end{aligned} \tag{2.2}$$

$E$ は英語の表記であり，$C$ は中国語の表記である．$e$ は英語の文字列であり，$c$ は中国語の漢字である．$<e,c>$ は英語と中国語の対応部分列である．

しかし，$e$ と $c$ の長さは必ずしも一致するわけではない．例えば，ある英語は $\alpha$ と仮定する．それに対応する中国語は $\beta$ と仮定する．英語 $\alpha = x_1x_2...x_m$，中国語 $\beta = y_1y_2...y_n$ とすると，英語と中国語の長さはそれぞれ $m$ と $n$ である．従って，以下のように，2 つや 3 つの英文字は 1 つの漢字に対応することもある．

$$<e,c>_1=<x_1,y_1> \quad <e,c>_2=<x_2x_3,y_2> \quad ...$$

すなわち，1 組みの音訳部分列は英語の文字数と中国語の漢字数が異なる場合もある．音訳部分列 $<e,c>$ の 1 組目から $K$ 組目までの同時確率は，対訳辞書から英語の発音と中国語漢字の出現頻度によって計算した．

### 2.2.2 翻字における意味情報の利用

Xu ら [29] は中国語への翻字に対して，翻字対象の発音と印象を考慮する統計的手法を提案した．図 2.2 は彼らが提案した翻字手法の概要を示している．

翻字手法の入力は翻字対象と翻字対象の印象を表す「印象キーワード」である．印象キーワードはユーザが中国語で与える必要がある．翻字対象は図 2.2 中左側の発音モデルによって，翻字対象と発音が近い翻字候補のリストが得られる．印象キーワードは図 2.2 中右側の印象モデルによって，印象キーワードに関連する漢字の集合が得られる．

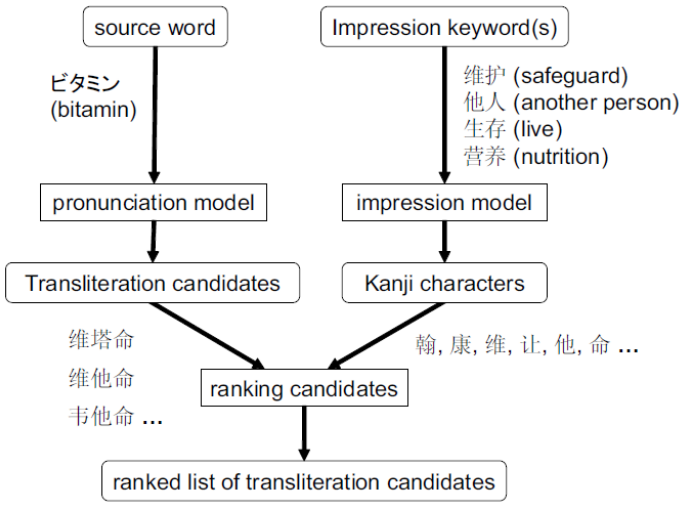


図 2.2: Xu らが提案した翻字手法の概要 [29]

最終的に，式（2.3）の確率翻字モデルによって，翻字候補リストに漢字の集合を加えて，ランキングされた翻字候補リストが出力される．

$$\begin{aligned} P(K|R,W) &= \frac{P(R,W|K) \times P(K)}{P(R,W)} \\ &\approx \frac{P(R|K) \times P(W|K) \times P(K)}{P(R,W)} \\ &\propto P(R|K) \times P(W|K) \times P(K) \end{aligned} \tag{2.3}$$

$R$ と $K$ はそれぞれ翻字対象と中国語の漢字列である．$W$ は印象キーワードである．$P(R|K)$ は発音モデルである．$P(W|K)$ は印象モデルである．$P(K)$ は言語モデルである．しかし，翻字は新しい語を生成する作業であるため，既存の語に関する言語モデルは効果がないと考え，$P(K)$ を定数とした．

Li ら [19] は人名のラテン語スクリプトから中国語へ翻字するために漢字の意味を考慮する統計的手法を提案した．彼らは人名を中国語へ翻字する際に，3 つの意味属性（semantic attribute）に依存すると考えている．

1. 語の起源：英単語は必ずしも英語から由来しているとは限らない．英語で書かれている人名は元々中国，日本，あるいは韓国から来た語かもしれない．それぞれの言語における音韻体系が異なるため，語の起源は翻字で使う発音規則と文字に影響を及ぼす．

2. 性別：名前は原言語と目標言語で同様な性別を連想すべきである．例えば,「Alice」は女性の名前である．中国語に「爱丽丝/ai–li–si/」と翻字され，漢字「丽」と「丝」から女性的な特徴が顕著に表れている．

3. 姓と名：中国の人名のパターンを見ると，姓に使う文字は限られている．名に対する文字の制限がない．

上記 3 つの意味属性を式（2.1）のノイズチャンネルモデルに取り入れ，式（2.4）を導入した．

$$\begin{aligned} P(T|S) &= \sum_{L\in\mathcal{L},G\in\mathcal{G}} P(T,L,G|S) \\ &= \sum_{L\in\mathcal{L},G\in\mathcal{G}} P(T|S,L,G)\times P(L,G|S) \end{aligned} \tag{2.4}$$

$S$ と $T$ はそれぞれ原言語と目標言語における名前である．$L$ は語の起源で，$G$ は性別である．$\mathcal{L}$ と $\mathcal{G}$ はそれぞれ $L$ と $G$ の集合である．$P(T|S,L,G)$ は Li ら [18] が提案した接合ソースチャンネルモデルを参考にして計算する．$P(L,G|S)$ は原言語の意味属性情報に依存して計算する．意味属性情報は訓練データから得られた各意味属性に関する事前知識を使用した．

## 2.3 逆翻字に関する関連研究

Chen ら [5] は，固有名詞を対象として，中国語の外来語を元の英語に逆翻字する手法を提案した．まず，中国語の発音表記である Wade-Giles(WG) とピンインの両方を用いて対象の中国語をローマ字に変換する．次に，中英人名の対訳 1,534 対に基づいて，音節単位で中国語と英語の照合を行う．音節単位で照合する際に，ローマ字化した音節中の先頭文字を重要視する．例えば，音節「chi」の中で,「c」は「h」と「i」よりも重要である．英語に関する発音規則も考慮した．例えば，英語の発音「ph」は通常中国語の発音「f」に対応する．しかし，この手法は特殊な人名を対象しているため，適用する範囲が狭い．

Knight ら [14] は，ベイズ定理に基づいて日本語のカタカナを元の英語に特定する．そのために，重み付き有限状態トランスデューサ（WFST: Weighted Finite-State Transducer）を用いた．WFST はオートマトンの一種であり，入出力シンボルを重みスコアを利用す

る情報変換の汎用計算モデルである．具体的には，式（2.5）を用いて，カタカナ文字列 $o$ が与えられた条件のもとで，最大の英単語 $w$ の条件付き確率 $P(w|o)$ を求める．この確率が最大になる $w$ を $o$ の訳語とする．

$$\arg\max_{w} P(w|o) = \arg\max_{w} \big\{ P(w) \times P(e|w) \times P(j|e) \times P(k|j) \times P(o|k) \big\} \tag{2.5}$$

$P(w|o)$ を計算するために，以下に示す 5 つの確率が必要である．

- $P(w)$ 英単語 $w$ の確率
- $P(e|w)$ 英単語 $w$ が与えられた条件のもとでの英語発音 $e$ の確率
- $P(j|e)$ 英語発音 $e$ が与えられた条件のもとでの日本語発音 $j$ の確率
- $P(k|j)$ 日本語発音 $j$ が与えられた条件のもとでのカタカナ表記 $k$ の確率
- $P(o|k)$ カタカナ表記 $k$ が与えられた条件のもとで OCR 入力 $o$ の確率

$P(w)$ は 262,000 単語の頻度を使用することによって推測した．$P(e|w)$ はオンライン CMU 発音辞書から音声記号なしで求めた．$P(j|e)$ は英日音の系列 8,000 組から EM アルゴリズムを使用することで計算した．$P(k|j)$ はカタカナ規則への日本の音に従って，人手で対応付けた．$P(o|k)$ はカタカナ単語と OCR が生成されたカタカナ間の対応確率を推測することによって求めた．しかし，この手法は大規模な音素辞書が必要である．

Stalls ら [24] は Knight ら [14] の手法をアラビア語への逆翻字に拡張した．Kwok ら [15] は Knight ら [14] の手法を英語から中国語への逆翻字に適応した．Al ら [1] はアラビア語へ逆翻字する際に，大規模な音素辞書が必要としない言語モデルの構築手法を提案した．

Jeong ら [11] は，韓国語の外来語から原言語の英語を特定する逆翻字手法提案した．韓国語から翻字候補を生成するために，隠れマルコフモデル（HMM: Hidden Markov Model）を用いた．翻字候補を絞るため，英語の辞書を利用した．

Fujii ら [7] は音素辞書を必要としない逆翻字手法を提案した．まず，文字列単位の翻字辞書を自動的に作成した．最短路アルゴリズムを用いて，ローマ字化されたカタカナ語と対応する英語の文字列での対応文字を自動的に抽出した．この処理によってカタカナ列 423 とアルファベット列 1,018 を含む翻字辞書を作成した．次に，翻字辞書のもとで，入力された単語を分割する．単語の分割は最小単位で分割を行い，式（2.6）に示す確率を用いて翻字の曖昧性を解消する．

式（2.6）は，原言語 $S$ が与えられた条件のもとで，目標言語 $T$ が生成される条件付き確率である．式（2.6）右辺の確率 $P(S|T)$ と $P(T)$ は式（2.7）に従ってそれぞれを計算する．$P(S|T)$ は音訳モデルである．$s_i$ と $t_i$ は翻字辞書で定義された対訳文字列であり，$P(s_i|t_i)$ は翻字辞書における対応する文字列の頻度で推定する．$P(T)$ は言語モデルである．

$$\arg\max_T P(T|S) = \arg\max_T P(S|T) \times P(T) \tag{2.6}$$

$$\begin{aligned} P(S|T) &\approx \prod_{i=1}^{n} P(s_i|t_i) \\ P(T) &\approx \prod_{i=1}^{n-1} P(t_{i+1}|t_i) \end{aligned} \tag{2.7}$$

Qu ら [22] の研究では，ラテン語で書かれた中国語や韓国語（CJK）人名を漢字に逆翻字する手法を提案した．提案した手法では，英日の逆翻字において，言語特有の規則を適応した．例えば，“koizumi” は日本人の名前として特定され，その結果，ローマ字と漢字の対応頻度を求めて，言語モデルに基づく翻字処理を行う．

まず，ローマ字から漢字に変換するために，Unihan データベースを使用した．Unihan データベースは中国語，日本語，韓国語を含め 54,728 の漢字情報がある．各漢字について日中韓 3 言語での発音表記が記載されている．日本語の場合は音読と訓読みの両方がある．このデータベースに定義された漢字とそれらの発音によって，ローマ字と漢字の対応頻度を求めた．

次に，ローマ字で表記された名前から漢字名前を以下の手順で逆翻字を行う．

1. ローマ字で表記された名前を分割する
2. 漢字で表記された名前を生成する
3. 漢字で表記された名前の候補からは日本語コーパスを用いて候補を絞る
4. ローマ字と対応する漢字候補のペアーを検索キーワードとして，Web 検索を行い，候補を絞る

ローマ字の名前から漢字の名前に変換するとき，ローマ字と漢字はすべての可能な対応を適用するため，生成された漢字列の組合せが膨大な数になる．例えば，日本人名「koizumi」

を「ko–i–zu–mi」と分割した場合，1 億 4900 万以上の可能な漢字列の組合せがある．そこで，日本語コーパスや Web 検索エンジンを用いて候補を絞る．

Kwok ら [16] は隠れマルコフモデルを用いて，中英と英中間の逆翻字手法を提案した．翻字対象を抽出するために，文書における人名前後の表現と品詞情報を利用した．

## 2.4 本研究の位置付け

既存の一般的な翻字手法は，主に言語の発音規則及び統計情報を用いて，原言語と目標言語の発音と目標言語をモデル化する手法であった．しかし，目標言語が中国語の場合，漢字は表意文字であるため，同音異義の漢字が存在する．「Coca-Cola」によって例示されたように，発音だけを考慮して，原言語と発音が似ているだけでは不十分であるということが分かった．従って，翻字対象の意味も翻字においてモデル化されるべきである．

さらに，意味の価値観が翻字対象の種別によって変わる．例えば，音楽家「ショパン（Chopin）」の名前は中国語で「肖邦 /xiao–bang/」と表記する．「肖」は中国の姓に良く使用される漢字である．「肖」と同じ発音の漢字には「消」がある．しかし，「消」は「消す」や「消滅する」などの意味があるため，人名の表記には不適切である．一方，もし，この名前が殺虫剤のような商品を指す場合では，「肖邦」より，「消邦 /xiao–bang/」の方が良いと考えられる．従って，翻字対象が属するカテゴリによって，良い意味と印象に関する定義が異なり，翻字対象のカテゴリも翻字においてモデル化されるべきである．

現在，各既存手法は，人名や地名などの種別の違いに対応せずに，全ての翻字対象に対して同じ言語モデルを適用した．音声処理で言語モデルを適合させる手法の効果について調べた研究 [8, 30] がある．しかし，中国語への翻字に適応する試みはなかった．

一般的な翻字手法と違い，Xu ら [29] の手法は翻字対象の印象を考慮している．しかし，翻字対象の印象を表す印象キーワードは人手で与えなければならない．さらに，翻字対象の種別を考慮していない．Li ら [19] の手法は翻字対象の意味属性を考慮している．しかし，人名だけに焦点を合わせたため，意味属性は人名に特有な属性である．従って，他の種別に属する翻字対象，例えば，企業名や，商品名などには使用できない．

以上をまとめると，中国語への翻字において，翻字対象の意味と種別のモデル化はまだ十分に検討されていない．本研究は，翻字対象の発音と同時に，意味と種別も考慮して，確率モデルに基づく中国語への意味訳型翻字手法を提案する．

# 第3章　意味訳型翻字手法

## 3.1　概要

本研究は中国語への翻字に対して，翻字対象の発音と同時に，意味と種別も考慮した意味訳型翻字手法を提案する．提案する翻字手法の概要を図 3.1 に示す．図 3.1 は,「発音モデル」,「意味モデル」,「カテゴリモデル」に大別される．提案する翻字手法を応用する状況の例として，中国に進出したい企業が企業名や商品名を中国語のネーミングを行う場合がある．ただし，利用者は中国語を知らなければならない．ネーミングでは，本研究の目的である翻字の他に，名前の「呼びやすさ」や「覚えやすさ」といった様々な観点を考慮する必要がある [32]．以下，図 3.1 を使用して本研究について説明する．

本研究の入力は 3 つある．1 つ目は，翻字対象となる外国語である．例えば,「エプソン /e–pu–so–n/（Epson)」である．2 つ目は，翻字対象が指す実体や概念に対して，その意味や印象を表す 1 つ以上の語である．例えば,「喜爱（好き)，普及（普及)，普通（普通)，生动（鮮明)」である．以後，本論文では，このような語を「関連語」と称する．現在，関連語は中国語で手動的または自動的に提供しなければならない．3 つ目は，翻字対象の種別を「人名」,「企業名」,「商品名」などのカテゴリで入力する．これらの入力に対して，本研究は 1 つ以上の漢字列を翻字の確率に従って，翻字の候補として出力する．

本研究において，翻字対象の関連語を得るために，2 つの手段を用意した．もし，利用者が中国語を良く知っているならば，翻字対象に関する中国語の関連語を 1 つ以上提供できると考えられる．しかし，これは本手法の適用性を制限してしまう．この問題を克服するために，本手法は World Wide Web を利用して，翻字対象の関連語を自動抽出する．Xu ら [29] の翻字手法は本研究と異なり，関連語の自動抽出を行っておらず，人手で与えなければならない．

発音モデルによって，翻字対象と発音が近い漢字列とそれぞれの確率が得られる．得られた漢字列が翻字候補となる．この処理は既存の音訳手法と基本的に同じである．現在，

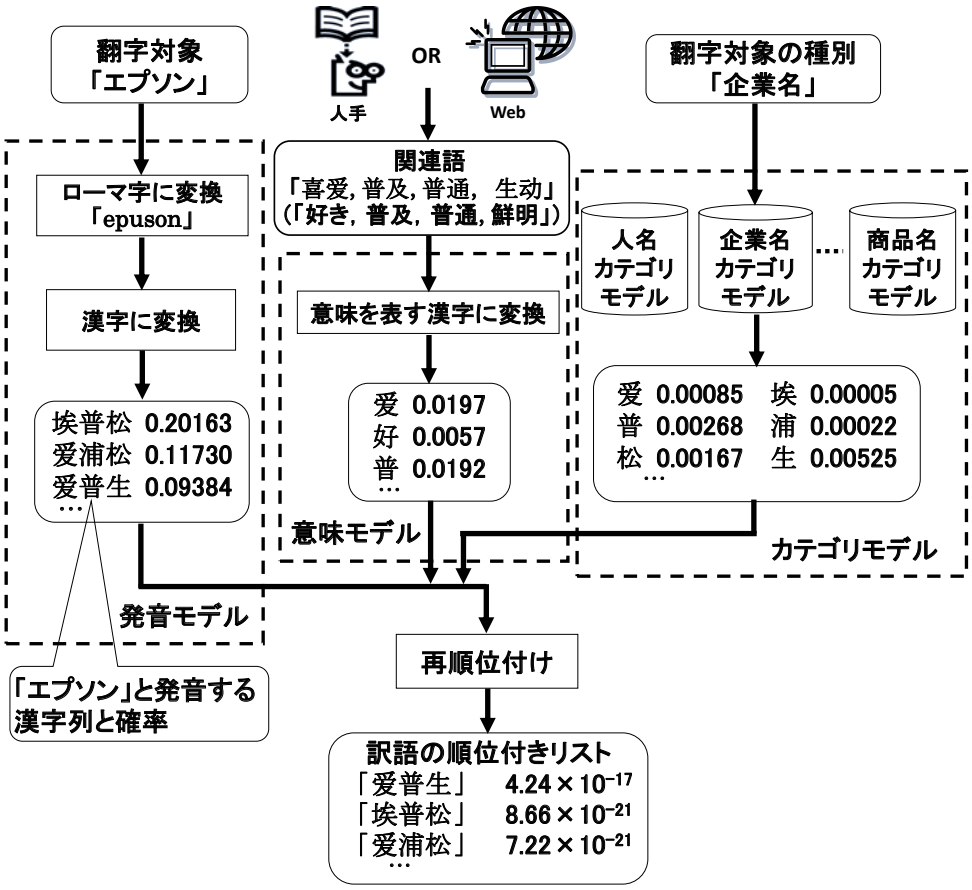


図 3.1: 提案する意味訳型翻字手法の概要

翻字対象となる外国語は日本語のカタカナ語を対象としている．カタカナ語は発音表記であるローマ字に変換することが容易だからである．図 3.1 では,「エプソン（Epson）」を翻字対象の例としている．ただし，ローマ字表記に変換できれば，他の言語を入力することも原理的に可能である．

発音モデルで得られた翻字候補は複数になる場合があるため，翻字候補を順位付けする必要がある．本研究では，翻字対象の発音，意味，カテゴリを 1 つの確率的な枠組みにモデル化した．翻字候補はそれぞれの確率スコアによって順位付けする．具体的には，発音モデルで得られた確率を意味モデルおよびカテゴリモデルで得られた漢字の確率と統合して，翻字候補の再順位付けに使用される．意味モデルとカテゴリモデルで得られた漢字をより多く含んでいる翻字候補がランキングの上位になる．

意味モデルによって，関連語に関連する漢字とそれぞれの確率が得られる．より簡単な方法としては，関連語をそのまま文字単位に分割することである．しかし，この方法では，関連語に関連する適切な文字のリストを得ることは難しい．本研究の意味モデルは，関連語に含まれていない関連する文字を引き出すことができる．

カテゴリモデルによって，翻字対象が属するカテゴリに良く使用される漢字とそれぞれの確率が得られる．例えば，翻字対象が属するカテゴリが「企業名」であるならば，翻字に使用される漢字が中国の企業名に良く使われる漢字の方が望ましい．しかし，対応するカテゴリモデルを事前に用意しなければならない．モデル化されていないカテゴリに属する翻字対象は対応できない．このような場合には，本研究は特定のカテゴリに属さない一般的なカテゴリモデルを使用する．さらに，意味モデルの関連語より翻字対象が属するカテゴリを人手または自動的に判断できる．自動的な方法には，固有表現認識ための既存研究が利用可能である [3, 10, 23]．しかし，本研究ではこの問題について探究しない．Xu ら [29] の翻字手法はカテゴリモデルを構築していない．

以下，3.2 節で確率的な漢字選択手法の全体像について説明する．3.3〜3.5 節で「発音」,「意味」,「カテゴリ」のモデル化について個別に説明する．さらに，翻字対象の関連語を自動抽出する方法の詳細は 3.6 節で説明する．

## 3.2 確率的な漢字選択手法

本研究における翻字の目的は,「外国語のローマ字表記 $R$」,「関連語 $W$」,「カテゴリ $C$」が与えられた条件のもとで，$P(K|R,W,C)$ が最大になる漢字列 $K$ を選択することである．ベイズの定理を利用して $P(K|R,W,T)$ の計算を式（3.1）のように計算する．

$$\begin{aligned} P(K|R,W,C) &= \frac{P(R,W,C|K) \times P(K)}{P(R,W,C)} \\ &\approx \frac{P(R|K) \times P(W|K) \times P(C|K) \times P(K)}{P(R,W,C)} \\ &\propto P(R|K) \times P(W|K) \times P(C|K) \times P(K) \\ &= P(R|K) \times P(W|K) \times P(C,K) \end{aligned} \tag{3.1}$$

式（3.1）の 2 行目で，$R$，$W$，$C$ が互いに条件付き独立であると仮定する．3 行目で，$P(R,W,C)$ は $K$ に依存せず，$P(K|R,W,C)$ による漢字列 $K$ の順位付けに影響しないた

め，$P(R,W,C)$ を省略する．図 3.1 において，$R$，$W$，$C$ はそれぞれ「epuson（エプソン）」,「喜爱，普及，普通，生动」,「企業名」であり，$K$ は「埃普松」である．利用者が複数の漢字列を選ぶ場合は，確率が高い候補から順番に選ぶ．

式（3.1）は，最終的に，$P(K|R,W,C)$ は $P(R|K)$，$P(W|K)$，$P(C,K)$ の積として近似される．各要素をそれぞれ「発音モデル」,「意味モデル」,「カテゴリモデル」と呼ぶ．

## 3.3 発音モデル

発音モデルは，中国語の漢字列 $K$ が与えられた条件のもとで，ローマ字表記 $R$ が生成される条件付き確率 $P(R|K)$ である．中国語の発音表記ピンインを中間言語として，ローマ字を中国語の漢字に変換する．日本語をローマ字表記に変換するにはヘボン式と訓令式がある．本研究はピンインと似ているヘボン式を使用する．翻字対象カタカナ語のローマ字表記に含まれているローマ字音節の数（$R$ の長さ）は必ず中国語の漢字列に含まれている漢字の数（$K$ の長さ）と一致するとは限らない．本研究は，両者の長さの違いを補うために，長さ $L_K$ の $K$ が与えられた条件のもとで，長さ $L_R$ の $R$ が生成される条件付き確率を使用する．Xu ら [29] はこの長さの違いについて考慮していない．

$P(R|K)$ は式（3.2）を用いて計算する．

$$\begin{aligned} P(R|K) &\approx P(R|Y) \times P(Y|K) \times P(L_R|L_K) \\ &\approx \prod_{i=1}^{N} P(r_i|y_i) \times \prod_{i=1}^{N} P(y_i|k_i) \times P(L_R|L_K) \end{aligned} \tag{3.2}$$

$Y$ は漢字列 $K$ の発音表記ピンイン列である．$r_i$，$y_i$，$k_i$ はそれぞれローマ字の音節列，ピンインの音節列，漢字 1 文字である．$R$，$Y$，$K$ は同じ数（$N$）に分解される．式（3.2）の $P(r_i|y_i)$，$P(y_i|k_i)$，$P(L_R|L_K)$ は式（3.3）を用いて計算する．

$$\begin{aligned} P(r_i|y_i) &= \frac{F(r_i,y_i)}{\sum_j F(r_j,y_i)} \\ P(y_i|k_i) &= \frac{F(y_i,k_i)}{\sum_j F(y_j,k_i)} \\ P(L_R|L_K) &= \frac{F(L_R,L_K)}{\sum_j F(L_j,L_K)} \end{aligned} \tag{3.3}$$

$F(r_i, y_i)$ はローマ字の音節 $r_i$ とピンインの音節 $y_i$ が共起する頻度であり，$F(y_i, k_i)$ はピンインの音節 $y_i$ と漢字 $k_i$ が共起する頻度である．これらの共起頻度を計算するために，日中対訳辞書 [33] を利用する．辞書の中に人名，地名，企業名などの固有名詞約 6,500 語を収録し，原言語と発音が似ているピンイン付き中国語と対応するカタカナ語 1,346 対が含めている．本研究は 210 対をテストデータとして 4 章の評価実験で使用するため，残りの 1,136 対（異なり 590 字，延べ 3,891 字）を発音モデルの構築に使用した．なお，カタカナ語 1,136 の内訳は，商品名 117，企業名 106，地名 423，人名 286，一般名詞 204 である．

表 3.1 は使用した日中対訳辞書の一部を示している．表中のカラム「日本語」,「ローマ字」,「中国語」,「ピンイン」,「英語」はそれぞれ見出し語，見出し語のローマ字表記，見出し語の中国語訳語，中国語訳語のピンイン表記，見出し語が対応する英語である．カラム「ローマ字」は，元の対訳辞書になく，我々が追加した項目である．自動的に見出し語のローマ字表記を得るために，本研究はプログラミング言語 Perl のモジュール[1]を利用した．カラム「ピンイン」では，中国語の声調である四声に基づいて 1〜4 の識別子が付けられている．

$P(R|K)$ を計算する例として，式 （3.5）は，漢字列「爱普生」が与えられた条件のもとで，ピンインの音節「ai–pu–sheng」を中継して，ローマ字の音節「e–pu–son」が生成される確率の計算を示している．

$$
\begin{aligned}
& P(\text{e pu son}|\ \text{爱普生}) \qquad (3.4)\\
&= P(\text{e pu son}|\text{ai pu sheng}) \times P(\text{ai pu sheng}|\ \text{爱普生}) \times P(L_R = 3|L_K = 3)\\
&= P(\text{e}|\text{ai}) \times P(\text{pu}|\text{pu}) \times P(\text{son}|\text{sheng}) \times P(\text{ai}|\ \text{爱}) \times P(\text{pu}|\ \text{普}) \times P(\text{sheng}|\ \text{生})\\
&\quad \times P(L_R = 3|L_K = 3)
\end{aligned}
$$

$P(r_i, y_i)$ と $P(y_i, k_i)$ を計算するために，表 3.1 で示したローマ字，中国語，ピンインの組を利用して，$y$ と $r$，$y$ と $k$ をそれぞれ対応付ける必要がある．1 つの漢字は通常 1 つのピンインしか持っていないため，$k$ と $y$ の対応付けは難しくない．例えば，表 3.1 の 1 行目の場合では，$k$ と $y$ の対応である「爱，ai4」,「荷，he2」,「华，hua2」が簡単に得られる．一方，$r$ と $y$ の対応付けは容易ではない．例えば，表 3.1 7 行目の「エスプリ」の場合では,「e–su–pu–ri」にある 4 つのローマ字音節と「ai4–si1–pu3–rui4–te4」にある 5 つ

[1]http://www.srekcah.org/ũtashiro/perl/scripts/romkan_pl/

のピンイン音節との対応は明白ではない．この問題を解決するためには，Xu ら [29] は，手動で対応付けを行った．本研究は自動マッピングの手法を利用する．自動マッピングに利用できる方法として，EM アルゴリズム [14] と DP マッチング [7] がある．本研究は DP マッチングを利用する．

準備実験では，1,136 組の $R$–$Y$ に対して，$r$ と $y$ の対応付けを行い，精度は 91.7% だった．誤りの原因として，一致する部分が少ない $R$–$Y$ の対に対して，DP マッチングが有効に機能しなかった．例えば，ローマ字列/ko–pe–ru–ni–ku–su/（コペルニクス）とピンイン列/ge1–bai2–ni2/（哥白尼）の場合，共通するアルファベットは「e」,「n」,「i」の三つしかない．そのために，DP マッチングの対応付けは「ko–pe → ge1」,「ru–ni → bai2」,「ku–su → ni2」となっている．手動の対応付けは「ko → ge1」,「pe → bai2」,「ru–ni–ku–su → ni2」である．解決方法の 1 つは，一致する部分が少ない $R$–$Y$ の対を抽出して手動で対応を付けることである．しかし，本論文はこの問題を将来の課題として残す．

上記誤りによって，DP マッチングの対応付け結果は翻字作業への影響を把握する必要がある．我々は，手動で対応付けた $R$–$Y$ の結果を使い，4 章の評価実験と同じ実験を行った．DP マッチングの対応付けを使用する時と同じ翻字結果が得られた．

表 3.2 は対訳辞書における $r_i$，$y_i$，$r_i$ と $y_i$ の共起頻度，確率を示している．表 3.2 を見ると，複数の $y_i$ が 1 つの $r_i$ と対応している．例えば，$y_i$ の「a1」,「ai4」,「an1」が $r_i$ の「a」と対応している．

表 3.3 は対訳辞書における $y_i$，$k_i$，$y_i$ と $k_i$ の共起頻度，確率を示している．表 3.3 を見ると，確率 $P(y_i|k_i)$ は 1.00 になる場合が多い．従って，1 つの漢字は 1 つのピンインしか持っていないのは一般的である．例外として，漢字「佛」と「伽」が 2 つのピンインを持っている．

表 3.4 は対訳辞書における $L_R$，$L_K$，$L_R$ と $L_K$ の共起頻度，確率を示している．表 3.4 を見ると，$L_R$ の数が同じの場合，$L_R = L_K$ の時，確率 $P(L_R|L_K)$ が最大になる．例外として，$L_R$ の数が 6 の場合，$P(6|5) = 0.36$ が $P(6|6) = 0.34$ より大きい．

上記の計算は発音モデルの構築過程で行われる．実際，翻字を行う際に，翻字対象カタカナ語のローマ字表記 $R$ の分割が複数ある場合は，すべての可能な分割を考慮して式 （3.2）を計算する．例えば，ローマ字変換された翻字対象カタカナ語「*epuson* (エプソン)」は「e–pu–son」と「e–pu–so–n」に分割でき，それぞれピンイン列「ai–pu–sheng」と「ai–pu–sou–an」に対応している．

表 3.1: 発音モデル構築で使用した日中対訳辞書の一部

| 日本語 | ローマ字表記 | 中国語 | ピンイン表記 | 英語 |
|---|---|---|---|---|
| アーメン | aa–me–n | 阿门 | a1–men2 | amen |
| アイオワ | a–i–o–wa | 爱荷华 | ai4–he2–hua2 | Iowa |
| イスラム | i–su–ra–mu | 伊斯兰 | yi1–si1–lan2 | Islam |
| インド | i–n–do | 印度 | yin4–du4 | India |
| ウィリアム | wi–ri–a–mu | 威廉姆 | wei1–lian2–mu3 | William |
| ウェイコン | we–i–ko–n | 卫康 | wei4–kang1 | Weicon |
| ウェーバー | wee–baa | 韦伯 | wei3–bo2 | Weber |
| エスカーダ | e–su–kaa–da | 爱斯卡达 | ai4–si1–ka3–da2 | Escada |
| エスプリ | e–su–pu–ri | 爱斯普瑞特 | ai4–si1–pu3–rui4–te4 | Esprit |
| エニカ | e–ni–ka | 英纳格 | ying1–na4–ge2 | Enicar |
| エマソン | e–ma–so–n | 爱默森 | ai4–mo4–sen1 | Emerson |
| オキューラー | o–kyuu–raa | 奥克拉 | ao4–ke4–la1 | Ocular |
| オリンパス | o–ri–n–pa–su | 奥林巴斯 | ao4–lin2–ba1–si1 | Olympus |
| カーネギー | kaa–ne–gii | 卡内基 | ka3–nei4–ji1 | Carnegie |
| カロリー | ka–ro–rii | 卡路里 | ka3–lu4–li3 | calorie |
| キシニョフ | ki–shi–nyo–fu | 基希讷乌 | ji1–xi1–ne4–wu1 | Kishinev |
| ギター | gi–taa | 吉他 | ji2–ta1 | guitar |
| グリム | gu–ri–mu | 格林 | ge2–lin2 | Grimm |
| コーチゾン | koo–chi–zo–n | 可的松 | ke3–di4–song1 | cortisone |
| ザイール | za–ii–ru | 扎伊尔 | za1–yi1–er3 | Zaire |
| シアヌーク | shi–a–nuu–ku | 西哈努克 | xi1–ha1–nu3–ke4 | Sihanouk |
| スコピエ | su–ko–pi–e | 斯科普里 | si1–ke1–pu3–li3 | Skopje |
| セモア | se–mo–a | 施华 | shi1–hua2 | Cemoi |

表 3.2: 対訳辞書におけるローマ字音節とピンイン音節の対応例

| $r_i$ | $y_i$ | $F(r_i, y_i)$ | $P(r_i\|y_i)$ |
|---|---|---|---|
| a | a1 | 40 | 0.87 |
| a | ai4 | 5 | 0.23 |
| a | an1 | 4 | 0.22 |
| a | er3 | 5 | 0.04 |
| a | ha1 | 1 | 0.05 |
| a | he2 | 1 | 0.13 |
| a | le4 | 1 | 0.03 |
| a | wa3 | 1 | 0.04 |
| a | ya3 | 2 | 0.15 |
| a | ya4 | 73 | 0.87 |
| aa | a1 | 3 | 0.07 |
| abi | wei2 | 1 | 0.03 |
| agu | ai4 | 1 | 0.05 |
| ai | ai1 | 1 | 0.08 |
| ai | ai4 | 3 | 0.14 |
| an | an1 | 11 | 0.61 |
| an | qi2 | 1 | 0.05 |
| an | ya4 | 1 | 0.01 |
| ara | la1 | 1 | 0.01 |
| are | ya4 | 1 | 0.01 |
| aru | a1 | 1 | 0.02 |
| aru | ya3 | 1 | 0.08 |
| aru | ya4 | 1 | 0.01 |
| asu | zi1 | 1 | 0.08 |
| au | ao4 | 1 | 0.04 |
| ba | ba1 | 36 | 0.58 |
| ba | bo2 | 1 | 0.03 |
| ba | fa1 | 1 | 0.25 |
| ba | fa2 | 3 | 1.00 |
| ba | wa3 | 10 | 0.42 |

表 3.3: 対訳辞書におけるピンインと漢字の対応例

| $y_i$ | $k_i$ | $F(y_i, k_i)$ | $P(y_i \mid k_i)$ |
|---|---|---|---|
| a1 | 阿 | 48 | 1.00 |
| ai1 | 埃 | 11 | 1.00 |
| ai1 | 挨 | 2 | 1.00 |
| ai4 | 爱 | 18 | 1.00 |
| ai4 | 艾 | 4 | 1.00 |
| an1 | 安 | 19 | 1.00 |
| ao2 | 翱 | 1 | 1.00 |
| ao4 | 奥 | 26 | 1.00 |
| ao4 | 澳 | 1 | 1.00 |
| ba1 | 吧 | 1 | 1.00 |
| ba1 | 巴 | 54 | 1.00 |
| ba1 | 爸 | 1 | 1.00 |
| ba1 | 芭 | 6 | 1.00 |
| bai2 | 白 | 5 | 1.00 |
| bai3 | 百 | 2 | 1.00 |
| bai4 | 拜 | 3 | 1.00 |
| ban1 | 班 | 5 | 1.00 |
| bao1 | 褒 | 1 | 1.00 |
| bao3 | 保 | 4 | 1.00 |
| bao3 | 堡 | 2 | 1.00 |
| bao3 | 宝 | 9 | 1.00 |
| bao4 | 豹 | 1 | 1.00 |
| bao4 | 鲍 | 4 | 1.00 |
| bei4 | 卑 | 2 | 1.00 |
| bei4 | 蓓 | 1 | 1.00 |
| bei4 | 贝 | 18 | 1.00 |
| fo2 | 佛 | 8 | 0.67 |
| fu2 | 佛 | 4 | 0.33 |
| ga1 | 伽 | 1 | 0.50 |
| jia1 | 伽 | 1 | 0.50 |

表 3.4: 対訳辞書におけるローマ字音節の長さと漢字列の文字数の対応例

| $L_R$ | $L_K$ | $F(L_R, L_K)$ | $P(L_R\|L_K)$ |
|---|---|---|---|
| 1 | 1 | 2 | 0.50 |
| 1 | 2 | 1 | 0.25 |
| 1 | 3 | 1 | 0.25 |
| 2 | 1 | 1 | 0.01 |
| 2 | 2 | 176 | 0.95 |
| 2 | 3 | 9 | 0.05 |
| 3 | 2 | 79 | 0.20 |
| 3 | 3 | 299 | 0.77 |
| 3 | 4 | 8 | 0.02 |
| 4 | 2 | 8 | 0.03 |
| 4 | 3 | 111 | 0.40 |
| 4 | 4 | 155 | 0.55 |
| 4 | 5 | 7 | 0.02 |
| 5 | 2 | 3 | 0.02 |
| 5 | 3 | 28 | 0.19 |
| 5 | 4 | 49 | 0.33 |
| 5 | 5 | 65 | 0.44 |
| 5 | 6 | 3 | 0.02 |
| 6 | 3 | 4 | 0.05 |
| 6 | 4 | 18 | 0.24 |
| 6 | 5 | 27 | 0.36 |
| 6 | 6 | 26 | 0.34 |
| 6 | 7 | 1 | 0.01 |
| 7 | 4 | 2 | 0.06 |
| 7 | 5 | 9 | 0.28 |
| 7 | 6 | 10 | 0.31 |
| 7 | 7 | 11 | 0.34 |
| 8 | 5 | 3 | 0.20 |
| 8 | 6 | 3 | 0.20 |
| 8 | 8 | 3 | 0.20 |

## 3.4 意味モデル

意味モデルは，漢字列 $K$ が与えられた条件のもとで，関連語列 $W$ が生成される条件付き確率 $P(W|K)$ である．式（3.2）で計算される発音モデル $P(R|K)$ と同様，$W$ と $K$ を分解して $P(W|K)$ を計算する．具体的には，$W$ は単語 $w_i$ に，$K$ を漢字 1 文字 $k_j$ に分割する．$P(W|K)$ の計算は，漢字 $k_j$ が与えられた条件のもとで単語 $w_i$ が生成される条件付き確率 $P(w_i|k_j)$ の計算に置き換えられる．しかし，式（3.2）と違って，利用者が入力できる関連語の数に制限はないため，$W$ と $K$ より得られた $w_i$ の数と $k_j$ の数がいつも同じになるとは限らない．従って，各 $k_j$ について $P(w_i|k_j)$ が最大となる $w_i$ だけを考慮し，式（3.5)に示されているように，近似的に $P(W|K)$ を計算する．

$$P(W|K) \approx \prod_j \max_i P(w_i|k_j) \tag{3.5}$$

表 3.5 は 3 つの漢字「爰 普 生」と 4 つの関連語「喜爰 普及 普通 生动” and “爰普生」について，それぞれの $P(w_i|k_j)$ を示している.「—」は，対応する $w_i$ と $k_j$ に対して $P(w_i|k_j)$ が計算できないことを示す．表 3.5 において，太字の数字は各 $k_j$ の $\max_i P(w_i|k_j)$ 値である．$P(W|K)$ は式（3.6）で示したように計算される．

表 3.5: $P(w_i|k_j)$ の例

| $w_i$ \ $k_j$ | 爰 | 普 | 生 |
|---|---|---|---|
| 喜爰 | **0.02** | – | – |
| 普及 | – | **0.03** | – |
| 普通 | – | 0.02 | – |
| 生动 | 0.01 | – | **0.03** |

$$\begin{aligned} &P(\text{喜爰 普及 普通 生动} \mid \text{爰普生}) \\ &= P(\text{喜爰} \mid \text{爰}) \times P(\text{普及} \mid \text{普}) \times P(\text{生动} \mid \text{生}) = 0.02 \times 0.03 \times 0.03 = 0.000018 \end{aligned} \tag{3.6}$$

$P(w_i|k_j)$ は式（3.7）を用いて計算する．

$$P(w_i|k_j) = \frac{F(w_i, k_j)}{\sum_w F(w, k_j)} \tag{3.7}$$

式（3.7）において，$F(w_i, k_j)$ は $w_i$ と $k_j$ の共起頻度である．$P(w_i|k_j)$ を計算するために，中国語におけるすべての単語と漢字の共起頻度を計算する必要がある．中国語のコーパスからこれらの共起頻度を計算して得られる．しかし，意味モデルの意義から考えると，意味的に強い関連を持っている漢字と単語同士から共起頻度を計算するのは望ましいため，単にあるコーパスにおける共起の統計ではよくない．すなわち，それぞれの漢字の意味について説明する言語リソースが必要である．従って，本研究は中国語漢字字典 [35] を使用した．中国語漢字字典では，見出しの漢字が文で説明され，場合によって，その漢字を含んでいる 1 つ以上の単語によって例示されている．

図 3.2 は中国語漢字字典における漢字「普」の例を示している．図 3.2 において，カラム「POS」は見出し漢字，あるいは，見出し漢字を含んでいる単語の品詞を示している．「形」,「名」,「外」はそれぞれ形容詞，名詞，外来語の意味である．カラム「意味記述」は見出し漢字の意味，あるいは，見出し漢字を含む単語を示している．「[ ]」中の英語は前にある単語の英語訳である．「( )」中の単語は直前にある単語の意味を示している．

| 見出し漢字 | ピンイン | POS | 意味記述 |
|---|---|---|---|
| 普 | pu3 | 形 | 遍 普遍; 全面 [general; universal; wide spread] |
| 普 | pu3 | 形 | 普席 (全席); 普及本 (即普及版);<br>普加 (普遍 赐与; 普遍 施与);<br>普存 (普遍 富足); 普讯 (遍及; 普遍);<br>普施 (普遍 施与); 普恩 (普施的恩泽) |
| 普 | pu3 | 形 | 广大 [universal] |
| 普 | pu3 | 名 | 吐蕃俗称丈夫 (成年男子) 为普 [husband] |
| 普 | pu3 | 外 | 用于外国语音译. 如:普朗克 |

図 3.2: 中国語漢字字典における漢字「普」の例

本研究では，中国語漢字字典から，外国語の表記によく使われ，字典の品詞部分に「外」という印が付いている見出し漢字 599 文字（異なり）を選択した．このうち，発音モデルの構築に使用した 590 漢字との重複は 464 字あった．選択した各見出し漢字の説明に出現した単語の頻度を数えた．図 3.2 のカラム「意味記述」で示したように，中国語は分かち書きしない．単語を抽出するために，本研究は SuperMorpho[2]を利用して各見出し漢字の意味記述を形態素解析を行った．その結果，16,943 単語（異なり）を抽出した．本研究は抽出した単語の品詞を考慮せず，すべての単語を出現頻度の計算に使用した．

表 3.6 は中国語漢字字典における漢字，単語，漢字と単語の共起頻度，確率を示している．表 3.6 を見ると，強い関連が持っている対「$k_j$，$w_i$」の確率 $P(w_i|k_j)$ も高い．例えば,「高（高い），加高（高さを増やす）」,「好（良い），好吃（美味しい）」,「乐（楽しい），乐于（喜ばせる）」の確率 $P(w_i|k_j)$ はそれぞれ 1.00，1.00，0.83 である．しかし，意味モデルを構築するための 599 文字に含まれていない漢字については，確率 $P(w_i|k_j)$ の計算ができない．従って，確率 $P(w_i|k_j)$ の計算できない漢字に対して，正数の定数を与えて平滑化する．現在，この定数は経験的に 0.001 としている．

[2]http://www.omronsoft.com/

表 3.6: 中国語漢字辞典における見出し漢字と意味記述の単語との対応例

| $k_j$ | $w_i$ | $F(w_i, k_j)$ | $P(w_i\|k_j)$ |
|---|---|---|---|
| 高 | 加高 | 3 | 1.00 |
| 高 | 增高 | 1 | 0.50 |
| 高 | 崇高 | 2 | 0.40 |
| 高 | 高大 | 8 | 0.28 |
| 高 | 高 | 39 | 0.26 |
| 高 | 距离 | 2 | 0.14 |
| 高 | 高尚 | 2 | 0.13 |
| 高 | 远 | 4 | 0.08 |
| 高 | 俗 | 2 | 0.06 |
| 高 | 下 | 4 | 0.03 |
| 好 | 好吃 | 2 | 1.00 |
| 好 | 好看 | 2 | 0.50 |
| 好 | 好不 | 2 | 0.50 |
| 好 | 貌美 | 2 | 0.40 |
| 好 | 好 | 43 | 0.38 |
| 好 | 喜爱 | 2 | 0.17 |
| 好 | 同意 | 2 | 0.15 |
| 好 | 美丽 | 2 | 0.08 |
| 好 | 美 | 3 | 0.03 |
| 好 | 表示 | 4 | 0.01 |
| 乐 | 乐于 | 5 | 0.83 |
| 乐 | 乐 | 51 | 0.53 |
| 乐 | 快乐 | 5 | 0.50 |
| 乐 | 音乐 | 11 | 0.46 |
| 乐 | 安乐 | 7 | 0.44 |
| 乐 | 乐意 | 2 | 0.40 |
| 乐 | 喜悦 | 2 | 0.22 |
| 乐 | 笑 | 5 | 0.21 |
| 乐 | 幸福 | 2 | 0.20 |
| 乐 | 喜 | 3 | 0.14 |

## 3.5 カテゴリモデル

カテゴリモデル $P(C,K)$ はカテゴリ $C$ に関するコーパスを用いてモデル化する．具体的には，式（3.8）を用いて計算する．

$$P(C,K) = P(C) \times P(K|C) \propto P(K|C) \tag{3.8}$$

$P(C)$ は $K$ に依存しないので省略する．すなわち，原理的には，カテゴリ $C$ のコーパスが与えられた条件のもとで，漢字列 $K$ が生成される条件付き確率 $P(K|C)$ を計算する．実際には，カテゴリ $C$ に関するコーパスを用いて漢字の $N-gram$ 確率を計算する．準備実験では，unigram と trigram より bigram の方が翻字により効果的なので，実際の翻字には bigram を使用した．

本研究では，入手可能な公開コーパスより以下に示す 3 つのカテゴリモデルを構築した．

- 標準カテゴリモデル：中国北京大学「计算语言学研究所（Institute of Computational Linguistics）」[3]が富士通[4]と共同で作成した「PFR 人民日报注语料库（人民日報タグ付きコーパス）」1998 年 1 月の新聞記事一ヶ月分から構築したモデルであり，異なり漢字 4,540（延べ 12,229,563 漢字）を含む．

- 企業名カテゴリモデル：「中文自然语言处理开放平台（中国語自然言語処理オープンソース）」[5]が提供している 22,569 社を含む「公司名录库（企業名リスト）」から構築したモデルであり，異なり漢字 2,167（延べ 78,432 漢字）を含む．

- 人名カテゴリモデル：上記「中文自然语言处理开放平台」が提供している「带词性词频的扩展词典（品詞および出現頻度付き拡張辞典）」から 38,406 件の人名を抽出し，構築したモデルであり，異なり漢字 2,318（延べ 104,443 漢字）を含む．

「標準カテゴリモデル」は特定なカテゴリに適応していないのに対して,「企業名カテゴリモデル」と「人名カテゴリモデル」はそれぞれ「企業名」と「人名」カテゴリに適応している．標準カテゴリモデルで得られた結果と他のカテゴリモデルで得られた結果を比較することで，カテゴリモデルの適応が翻字に及ぼす影響について考察することができる．

[3]http://icl.pku.edu.cn/
[4]http://www.frdc-fujitsu.com.cn/
[5]http://www.nlp.org.cn/

表 3.7–3.9 は上記 3 つのカテゴリモデルにおいて，それぞれの漢字 bigram，出現頻度，出現確率を示している．表 3.7 において，数多くの漢字 bigram は一般名詞が含まれている．例えば，「中国（中国）」，「经济（経済）」，「人民（人民）」などである．単語にならない bigram もある．例えば，「日电」，「了一」，「业的」などである．表 3.8 において，企業名によく用いられる bigram の漢字列が含まれている．例えば，「兴业（企業は繁栄する）」，「长荣（繁栄は続ける）」，「富邦（国を豊かにする）」などである．これらの漢字は企業理念や哲学を反映している．表 3.9 において，「斯基」，「里亚」，「克里」などのような人名によく用いられる bigram の漢字列が含まれている．例えば，これらの漢字列は人名の「乔姆斯基（Chomsky）」，「阿德里亚诺（ Adriano）」，「克里（Kerry）」に含まれている．

上記の各カテゴリモデルを構築する際に，漢字列を抽出するため，SuperMorpho を用いてコーパスの形態素解析を行った．下記に示す 11 種類の機能語と 30 種類の記号を事前に除いた．各機能語の後には辞書 [34] における代表的な注釈を示している．

- 機能語
  - 啊：感動詞．感銘したり驚いたりした時に発する．
  - 唉：感動詞．返事する時発する．
  - 吧：助詞．文末に用い，同意・許可を表す．
  - 地：助詞．連用修飾語の後に用い，主として動詞・形容詞を修飾する．
  - 的：助詞．連体修飾語の後に用いる．主として名詞を修飾する．
  - 得：助詞．動詞の後にあって可能を表す．
  - 过：助詞．かつて，あるいはすでになされていることを表す．
  - 了：助詞．動詞や形容詞の後に置いて，動作や変化がすでに完了したことを表す．
  - 吗：助詞．疑問を表し，普通の平叙文の末尾に用いる．
  - 呢：助詞．疑問を表す．
  - 着：助詞．動作が進行していることを表す．
- 記号　〝 〞 “ ” ‘ ’ （ ） 《 》 〈 〉 『 』「 」〔 〕【 】 … 。 . ? ! ， 、 ； ： ※

表 3.7: 標準カテゴリモデルに含まれている漢字 bigram の例

| Bigram | 出現頻度 | 出現確率 |
|---|---|---|
| 中国 | 12 220 | 2.0E-03 |
| 经济 | 10 934 | 1.8E-03 |
| 工作 | 8 985 | 1.5E-03 |
| 国家 | 8 084 | 1.4E-03 |
| 人民 | 7 802 | 1.3E-03 |
| 企业 | 7 420 | 1.2E-03 |
| 我们 | 6 431 | 1.1E-03 |
| 政府 | 6 006 | 1.0E-03 |
| 问题 | 5 637 | 9.4E-04 |
| 全国 | 5 513 | 9.2E-04 |
| 北京 | 5 408 | 9.0E-04 |
| 本报 | 4 761 | 8.0E-04 |
| 公司 | 4 708 | 7.9E-04 |
| 进行 | 4 694 | 7.9E-04 |
| 领导 | 4 643 | 7.8E-04 |
| 中央 | 4 578 | 7.7E-04 |
| 他们 | 4 519 | 7.6E-04 |
| 抗洪 | 4 294 | 7.2E-04 |
| 我国 | 4 217 | 7.1E-04 |
| 地区 | 4 046 | 6.8E-04 |
| 世界 | 3 929 | 6.6E-04 |
| 国际 | 3 924 | 6.6E-04 |
| 的一 | 3 885 | 6.5E-04 |
| 日电 | 3 581 | 6.0E-04 |
| 了一 | 2 726 | 4.6.E-04 |
| 是一 | 2 420 | 4.0.E-04 |
| 这一 | 2 247 | 3.8.E-04 |
| 业的 | 2 130 | 3.6.E-04 |
| 的发 | 2 128 | 3.6.E-04 |
| 们的 | 2 030 | 3.4.E-04 |

表 3.8: 企業名カテゴリモデルに含まれている漢字 bigram の例

| Bigram | 出現頻度 | 出現確率 |
|---|---|---|
| 兴业 | 307 | 3.3E-03 |
| 统一 | 97 | 1.0E-03 |
| 解散 | 91 | 9.8E-04 |
| 农会 | 73 | 7.8E-04 |
| 租赁 | 70 | 7.5E-04 |
| 合作 | 58 | 6.2E-04 |
| 作社 | 54 | 5.8E-04 |
| 保全 | 51 | 5.5E-04 |
| 信用 | 51 | 5.5E-04 |
| 大同 | 50 | 5.4E-04 |
| 华新 | 49 | 5.3E-04 |
| 长荣 | 46 | 4.9E-04 |
| 第一 | 46 | 4.9E-04 |
| 法人 | 45 | 4.8E-04 |
| 财团 | 45 | 4.8E-04 |
| 团法 | 45 | 4.8E-04 |
| 用合 | 45 | 4.8E-04 |
| 中兴 | 44 | 4.7E-04 |
| 创新 | 44 | 4.7E-04 |
| 美商 | 43 | 4.6E-04 |
| 网际 | 43 | 4.6E-04 |
| 亚太 | 43 | 4.6E-04 |
| 合并 | 42 | 4.5E-04 |
| 育乐 | 42 | 4.5E-04 |
| 服务 | 42 | 4.5E-04 |
| 先进 | 42 | 4.5E-04 |
| 并解 | 41 | 4.4E-04 |
| 维京 | 41 | 4.4E-04 |
| 富邦 | 40 | 4.3E-04 |
| 群岛 | 39 | 4.2E-04 |

表 3.9: 人名カテゴリモデルに含まれている漢字 bigram の例

| Bigram | 出現頻度 | 出現確率 |
|---|---|---|
| 斯特 | 302 | 2.1E-03 |
| 斯基 | 143 | 1.0E-03 |
| 克斯 | 138 | 9.7E-04 |
| 尔斯 | 132 | 9.2E-04 |
| 阿尔 | 120 | 8.4E-04 |
| 德尔 | 116 | 8.1E-04 |
| 尔德 | 116 | 8.1E-04 |
| 里斯 | 110 | 7.7E-04 |
| 格尔 | 106 | 7.4E-04 |
| 斯塔 | 97 | 6.8E-04 |
| 利亚 | 95 | 6.6E-04 |
| 维奇 | 94 | 6.6E-04 |
| 贝尔 | 94 | 6.6E-04 |
| 瓦尔 | 91 | 6.4E-04 |
| 卡尔 | 91 | 6.4E-04 |
| 斯托 | 88 | 6.2E-04 |
| 里亚 | 88 | 6.2E-04 |
| 拉斯 | 88 | 6.2E-04 |
| 克拉 | 81 | 5.7E-04 |
| 克尔 | 81 | 5.7E-04 |
| 夫斯 | 81 | 5.7E-04 |
| 尼亚 | 81 | 5.7E-04 |
| 耶夫 | 79 | 5.5E-04 |
| 尔曼 | 79 | 5.5E-04 |
| 马尔 | 78 | 5.5E-04 |
| 克里 | 76 | 5.3E-04 |
| 诺夫 | 75 | 5.2E-04 |
| 尔特 | 74 | 5.2E-04 |
| 罗斯 | 74 | 5.2E-04 |
| 塞尔 | 72 | 5.0E-04 |

## 3.6 関連語の自動抽出

本研究では，翻字対象が指す実体や概念に対して，その意味や印象を中国語で表記した関連語を Web から自動的に抽出し，翻字に利用する．

図 3.3 に,「エプソン」の関連語を自動抽出する過程を示す．図 3.3 の上部では翻字対象に関連する関連語候補を抽出し，下部では抽出する関連語を選択している．以下，それぞれについて説明する．

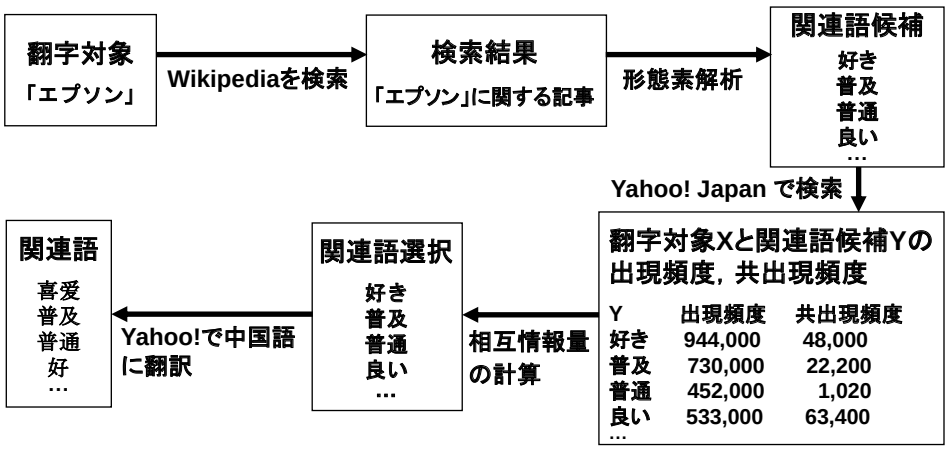


図 3.3: 関連語自動抽出の概要

翻字対象の関連語を抽出するためには，翻字対象に関する文書が必要である．例えば，翻字対象が商品名であれば，その商品を紹介する文書であり，翻字対象が企業名であれば，企業の理念などに関する文書である．このような文書として，フリー百科事典「ウィキペディア（Wikipedia)」日本語版[6]の記事を利用した．2012 年 1 月 1 日の時点では約 80 万の項目があり，一般名詞，人名，地名，企業名，商品名などが登録されている．

図 3.4 は地名「カラチ」を Wikipedia で検索して得られた記事ページの抜粋である．図 3.4 において，最上部の「カラチ」は記事の名称（記事名）であり，その下は本文である．図 3.4 の「目次」に示されているように，本文は「1 歴史」，「2 気候」，「3 人口統計」などの「セクション（節)」によって構造化されることがある．

[6]http://ja.wikipedia.org/wiki/

カラチ

カラチ(英: **Karachi**、ウルドゥー語表記:کراچی)は、パキスタン南部、アラビア海沿岸にあるパキスタン最大の都市。インダス川河口の西に位置する。シンド州の州都であり、世界有数のメガシティである。

2010年の近郊を含む都市的地域の人口は1,308万人であり、世界第20位である[1]。また、パキスタンにおける商業・金融の中心地でもある。位置は、北緯24度48分、東経66度59分。

目次 [非表示]

1 歴史
- 1.1 概要
- 1.2 カラチの発展史

2 気候
3 人口統計
4 交通
- 4.1 航空便
- 4.2 海運
- 4.3 鉄道
- 4.4 キンキラバス

5 姉妹都市
6 脚注
7 ギャラリー
8 外部リンク

歴史 [編集]

概要 [編集]

バローチスタン州やen:Makranに住んでいたバローチ人が漁村を作ったのがカラチの始まりである。バローチ人の多くが今もなお、カラチに居住しており、バローチ語ではこの都市のことをKolachiと呼ぶ[2]。しかしながら、カラチが現在の姿に発展するようになったのは19世紀から始まるイギリス植民地時代に起因する。1947年、パキスタンが独立を達成すると、カラチはパキスタンの首都となり、インドからムスリムが多く移住した。独立直後の人口移動により、カラチは、急速に人口が拡大するとともに、パキスタンにおける政治・経済の中心として機能するようになった。カラチはインフラストラクチャーが貧弱だったこともあり、社会経済的に大きな問題に直面したが、現在では、パキスタンにおける産業・ビジネスの中心地として発展を遂げた。

図 3.4: Wikipedia における「カラチ」の記事ページの抜粋

関連語候補の抽出は以下の手順に従って行う.

1. 翻字対象を Wikipedia で検索して記事ページを取得する．現在の手法では，記事ページがない用語に対しては関連語を抽出することができない.

2. 取得した記事ページから HTML タグを削除し，各文中の単語と品詞を特定する．しかし，日本語の文は分ち書きをしないため，日本語の形態素解析ソフト「茶筌」[7]で形態素解析を行う.

3. 形態素解析の結果から，名詞と形容詞を翻字対象の関連語候補として抽出する．ただし,「名詞」のうち「名詞-数」,「名詞-接尾-助数詞」,「名詞-副詞可能」,「名詞-非

[7]http://chasen.naist.jp/hiki/ChaSen/

自立」,「名詞-代名詞」は抽出しない．図 3.3 では,「普及」や「普通」などの名詞と「好き」や「良い」などの形容詞が関連語の候補として抽出されている．

図 3.4 に示した記事ページの本文は構造化されているため，上記の手順（2）において，関連語抽出に有効な特定のセクション内だけを解析対象とする手法が考えられる．しかし，Wikipedia のガイドブックによる記事ページの編集方針[8]では，記事は「記事名（項目名）」と「本文」で構成され，本文の基本構成は概要から次第に詳細内容になり，段落の数が多くなるようなら,「見出し」を付けて「セクション（節）」に分けるとしか規定していない．見出しの付け方やセクションの分け方は記事の著者によって方針が異なる．

例えば，図 3.4 で示した「カラチ」の記事ページは,「歴史」,「気候」,「人口統計」,「交通」,「姉妹都市」,「脚注」,「ギャラリー」,「外部リンク」の 8 セクションで構成されている．一方,「ハワイ」の記事ページは,「歴史」,「地理」,「人口動勢」,「政治と法律」,「経済」,「教育」,「芸術・文化」,「日本との関わり」,「その他」,「注」,「関連項目」,「外部リンク」の 12 セクションで構成されている．同じ地名に関する記述であるにも拘らず,「カラチ」と「ハワイ」の記事に共通するセクションは「歴史」だけである．さらに，同じ見出しのセクションでも，著者によって記述の方針が異なる可能性がある．そのため，関連語抽出に有効なセクションを事前に定義することが困難である．現在では本文全体を対象として関連語の候補を抽出した．

Wikipedia から抽出した名詞と形容詞の中には，翻字対象との関連が低い語も含まれているため，翻字に使用する関連語を選択する必要がある．単語間の関連度を計算する手法 [4, 31] が複数提案されている．本研究では，翻字対象と関連語候補間の相互情報量 [6, 25] を計算して，その値が高い語を関連語として抽出する．ここでいう相互情報量とは，正確には pointwise mutual information であり，式 (3.9) を用いて計算する．

$$I(X,Y) = \log \frac{P(X,Y)}{P(X) \times P(Y)} \tag{3.9}$$

$P(X)$ と $P(Y)$ は単語 $X$ と $Y$ それぞれの出現確率であり，$P(X,Y)$ は $X$ と $Y$ が同時に出現する確率である．ここでは便宜上，$X$ を翻字対象，$Y$ を 1 つの関連語候補とする．図 3.3 の例では，$X$ は「エプソン」であり，$Y$ は「好き，普及，普遍，良い」のいずれかである．

[8] http://ja.wikipedia.org/wiki/Wikipedia:ガイドブック_編集方針

関連語の選択は以下の手順に従って行う．

1. $P(X)$，$P(Y)$，$P(X,Y)$ を計算する．しかし，我々は $X$ と $Y$ をカーバできる十分大きなコーパスを持っていないため,「$X$」,「$Y$」,「$X$ and $Y$」を検索キーワードとして Yahoo! JAPAN[9]で検索し，検索結果の総数でそれぞれの確率を近似する．検索エンジンによって検索されたページの数は一定ではなく，完全に信頼できない [13]．我々は将来の仕事としてこの問題を残す．

2. 式 (3.9) の値が高い候補を関連語として選択する．図 3.3 では,「エプソン」の関連語として「好き」,「普及」,「普通」,「良い」が選ばれている．選択する関連語の件数は実験的に決めるパラメタである．4.4 節の評価実験では，関連語の件数を段階的に変化させて翻字への影響について考察する．

3. 上記 2. で選択した関連語を中国語に翻訳する．原理的には，この作業は機械翻訳システムを利用することで自動化することができる．しかし，現在は Yahoo! JAPAN[10]を利用して人手で翻訳している．ただし，Yahoo! JAPAN で翻訳できずに原言語がそのまま返される関連語は削除する．図 3.3 では,「喜爱」,「普及」,「普通」,「好」はそれぞれ「好き」,「普及」,「普通」,「良い」に対する訳語であり，翻字対象の関連語として使用される．

表 3.10 は実際関連語自動抽出で得られた翻字対象「ミサ（mass）」の抽出結果を示している.「ミサ」との相互情報量 $I(X,Y)$ が関連語候補の上位にある語「典礼」や「聖歌」などが関連語として選択された．一方,「ミサ」との関連が重要ではないと思われ，相互情報量が関連語候補の下位にある語「義務」や「とき」などが正しく除外された．

[9]http://www.yahoo.co.jp/
[10]http://honyaku.yahoo.co.jp/

表 3.10: 翻字対象「ミサ (mass)」の関連語自動抽出の結果

| 抽出した語 | | $I(X,Y)$ | 除外された語 | | $I(X,Y)$ |
|---|---|---|---|---|---|
| 日本語 | 英語 | | 日本語 | 英語 | |
| 典礼 | ceremony | 4.0E-08 | 義務 | obligation | 5.3E-10 |
| 聖歌 | hymn | 2.5E-08 | とき | time | 5.3E-10 |
| 司祭 | chaplain | 2.3E-08 | 多い | many | 5.3E-10 |
| 司教 | bishop | 1.6E-08 | 地方 | locality | 4.9E-10 |
| 聖餐 | communion | 1.4E-08 | 毎日 | every day | 4.5E-10 |
| 奉献 | dedication | 1.4E-08 | 派遣 | dispatch | 4.1E-10 |
| 信徒 | believer | 1.2E-08 | 参加 | participation | 4.0E-10 |
| 祭壇 | altar | 9.3E-09 | 時間 | time | 3.6E-10 |
| 礼拝 | worship | 8.6E-09 | すべて | all | 3.4E-10 |
| 教会 | church | 7.0E-09 | こと | thing | 2.8E-10 |

# 第4章　評価実験

## 4.1　実験方法

本研究で提案していた意味訳型翻字手法について，以下の 3 つの方面から評価した．

- 意味モデルとカテゴリモデル（4.2 節）
- カテゴリモデルの適応（4.3 節）
- 関連語の自動抽出（4.4 節）

本研究の成果を評価するために，2 つのテストコレクションを用意した．まず，意味モデル，カテゴリモデル，カテゴリモデルの適応を評価するため，発音モデル（3.3 節）を構築する際に利用した日中対訳辞書からランダムで 210 のテスト単語を選択した．210 語は発音モデルを構築する際に利用した 1,136 語の中に含まれていない．また，カテゴリモデルを構築する際に使用した 3 つのコーパスにも含まれていない．人手の判断コストを削減するために，本研究はプーリング法 [27] を使用した．すなわち，評価対象のすべての手法より生成された一定数の候補が蓄積され，人間の判定者はこの蓄積された候補だけを判断する．上記 210 語の中に，82 語が Wikipedia の記事を利用することができなかったため，関連語の自動抽出を評価するには別のテストコレクションを用意しなければならない．また，評価対象の手法が前のテストコレクションとも異なる．

両方のテストコレクションの生成手順が同じであるため，この節では，最初のテストコレクションのみについて説明する．4.4 節では，最初のテストコレクションとの違いだけに焦点を合わせて 2 番目のテストコレクションについて説明する．各基本モデルを以下に示すシンボルを使って表記する：

- P: 発音モデル
- Mm: 人手で与えた関連語を利用する意味モデル

- Ma: 自動抽出した関連語を利用する意味モデル
- Cg: 標準カテゴリモデル
- Cc: 企業名カテゴリモデル
- Cp: 人名カテゴリモデル
- Ca: すべての翻字対象に適応したカテゴリモデル

評価対象の手法は,「P」と「P+Mm+Cg」のように，発音モデルのもとで,「+」符号で上記 1 つ以上のシンボルを結合した組み合わせで指定する．

自動翻字における意味モデルとカテゴリモデルの有効性を評価するために，手法「P」,「P+Mm」,「P+Cg」,「P+Mm+Cg」の翻字結果を比較する．カテゴリモデル適応の有効性を評価するために，手法「P+Mm+Cg」,「P+Mm+Cc」,「P+Mm+Cp」,「P+Mm+Ca」の翻字結果を比較する．合計は 7 つの手法の結果が蓄積される．

評価実験で使用する翻字対象 210 語は表 4.1 に示されたように，5 つのカテゴリに分類することができる．評価手法 P+Mm+Ca では，翻字対象が人名と企業名の場合は，翻字する際にそれぞれ Cc と Cp で適応させ，それ以外は Cg を使用した．

日本語を理解できる 2 名の中国人留学生を判定者になってもらい，関連データを作成してもらった．判定者の中に，著者が含まれていない．判定結果の客観性を高めるために，2 名の判定者は同じ翻字対象語に対して独立して同じ作業を行った．具体的には, 以下の手順で行った．

まず，各翻字対象語に対して，判定者それぞれは中国語で 1 つ以上の関連語を与えた．翻字対象語ごとに与える関連語の語数について制限していないため，判定者によって決められた．また，判定者がそれぞれの翻字対象語の意味を理解させるために，日中対訳辞書に記載された注釈を判定者に示した．表 4.2 と 4.3 はそれぞれ判定者 A と B によって与えられた関連語の例を示している．

次に，各翻字対象語に対して，6 通りの手法（P，P+Mm，P+Cg，P+Mm+Cg，P+Mm+Cc，P+Mm+Cp）を適応し，各手法ごとによって，合計 6 つの翻字候補の順位付きリストが生成される．手法 P+Mm+Ca の翻字候補は P+Mm+Cg，P+Mm+Cc，と P+Mm+Cp の翻字候補に含まれているため，P+Mm+Ca で生成した順位付きリストを省略した．

最後に，各翻字対象語に対して，判定者がそれぞれ翻字対象語に与えた関連語を考慮して，1つ以上の正解訳語を選択した．判定者には，どの手法で得られた翻字候補なのかを分からないようにすることが重要である．上記6つの手法で得られた6つの順位付きリストから，それぞれ上位100位までの翻字候補を抽出し，重複を取り除き，文字コードで並べ替えた．結果として，判定者はすべての翻字対象語に対して，最大600翻字候補を含むリストについてそれぞれの正解訳語を判定をした．しかし，いくつかの翻字対象語に関して，判定者は翻字候補リストから正しい翻字を見つけられなかった．

各翻字手法における翻字の精度を評価する尺度として,「正解訳語の平均順位」を用いた．１つの翻字対象に対して複数の正解訳語がある場合は，それらの順位を平均してから，さらに全翻字対象語を横断して順位を平均した．評価の段階では，上位100件に限定せず，各手法が出力した順位付きリストを全て使用した．従って，正解訳語の平均順位は100より大きくなる場合がある．翻字対象語210語に対して，生成された翻字候補の平均は31,779であった．すべての手法のリストが同じ翻字候補を含んでいるので，正解訳語の平均順位で各手法を比較するのは妥当である．

各翻字対象語の「正解訳語」として，以下に示す3種類の解釈がある．

(a) 判定者のうち最低1名が適切と判定した訳語

(b) 判定者2名が適切と判定した訳語

(c) 日中対訳辞書 [33] に定義された訳語

（a）は正解訳語の網羅性が最も高い．しかし，判定者の主観に依存するため，評価の客観性は最も低い.（c）は評価の客観性は最も高い．しかし，辞書に定義されていない言葉でも訳語として適切な場合があるため，評価の網羅性は最も低い．例えば，辞書において，カタカナ語「ディスコ（disco)」と「アンモニア（ammonia)」の訳語はそれぞれ「迪斯科」と「阿摩尼亚」である．しかし，辞書以外で使われている訳語として，それぞれ「的士高」と「亚摩尼亚」がある.（b）は正解訳語の網羅性と評価の客観性ともに（a）と（c）の中間である．評価実験では，それぞれの解釈について評価を行った

表 4.1: 翻字対象語 210 語の内訳

| カテゴリ | 語数 | 例 | | |
|---|---|---|---|---|
| | | 日本語 | 中国語 | 英語 |
| 企業名 | 48 | アグファ | 爱克发 | Agfa |
| | | インテル | 英特尔 | Intel |
| | | カネボウ | 嘉娜宝 | Kanebo |
| | | ライカ | 莱卡 | Leica |
| | | ナイキ | 耐克 | Nike |
| 商品名 | 64 | アウディ | 奥迪 | Audi |
| | | コルゲート | 高露洁 | Colgate |
| | | コンタック | 康泰克 | Contac |
| | | フェンディ | 芬迪 | Fendi |
| | | エルメス | 爱马士 | Hermes |
| 人名 | 21 | アナン | 安南 | Annan |
| | | ビゼー | 比才 | Bizet |
| | | ショパン | 肖邦 | Chopin |
| | | リスト | 李斯特 | Liszt |
| | | ニュードン | 牛顿 | Newton |
| 地名 | 36 | アンマン | 安曼 | Amman |
| | | カラカス | 加拉加斯 | Caracas |
| | | ドーハ | 多哈 | Doha |
| | | オハイオ | 俄亥俄 | Ohio |
| | | ヤンゴン | 仰光 | Yangon |
| その他 | 41 | エンジェル | 安琪儿 | angel |
| | | コーヒー | 咖啡 | coffee |
| | | マンゴー | 芒果 | mango |
| | | オンス | 盎司 | ounce |
| | | ワルツ | 华尔兹 | waltz |

表 4.2: 判定者 A が与えた関連語の例

| カテゴリ | 翻字対象語 (英語) | 関連語 |
|---|---|---|
| 企業名 | ナイキ (Nike) | 成功 (成功), 名牌 (有名ブラント), 运动 (運動), 对号 (フィートする), 鞋 (靴), 坚韧 (強くて粘り強い). |
| 商品名 | アウディ (Audi) | 轿车 (乗用車), 酷 (格好良い), 高级 (高級感), 高品质 (高品質), 德国 (ドイツ), 四轮驱动车 (四輪駆動車), 历史悠久 (歴史が長い). |
| 人名 | ショパン (Chopin) | 钢琴家 (ピアニスト), 作曲家 (作曲家), 神童 (神童), 波兰 (ポーランド), 爱国 (愛国者), 浪漫主义 (ロマンチスト). |
| 地名 | アンマン (Amman) | 约旦首都 (ヨルダンの首都), 政治经济中心 (政治と経済の中心), 人口多 (人口が多い), 宗教 (宗教), 历史悠久 (歴史が長い), 中东 (中東). |
| その他 | エンジェル (angel) | 僮 (召し使われている未成年者), 孩 (子供), 长发 (長髪), 可爱 (可愛い), 金发 (金髪), 善良 (善良), 会飞 (飛べる), 白衣 (白衣). |

表 4.3: 判定者 B が与えた関連語の例

| カテゴリ | 翻字対象語 (英語) | 関連語 |
|---|---|---|
| 企業名 | ナイキ (Nike) | 名牌 (有名ブランド), 运动 (運動), 健康 (健康), 锻炼 (訓練), 忍耐力 (忍耐力), 坚强 (強靭), 拼搏 (奮闘する), 持久 (持久力). |
| 商品名 | アウディ (Audi) | 典范 (模範), 男孩 (男の子), 坚强 (強靭), 德国 (ドイツ), 速度 (スピード). |
| 人名 | ショパン (Chopin) | 钢琴 (ピアノ), 音乐 (音楽), 波兰 (ポーランド), 感情细腻 (感性が豊富かつ繊細), 维也纳 (ウィーン), 作曲家 (作曲家), 演奏家 (音楽家), 顽强 (頑固かつ強い), 干脆 (シンプル), 利落 (素早い). |
| 地名 | アンマン (Amman) | 温柔 (優しい), 女孩 (女の子), 中东 (中東), 石油 (石油), 贫穷 (貧困), 战争 (戦争). |
| その他 | エンジェル (angel) | 安详 (落ち着いた様子), 花 (花), 孩子 (子供), 女孩 (女の子), 可爱 (可愛い), 伶俐 (賢い), 优雅 (エレガンス), 幸运 (幸運), 翅膀 (翼). |

## 4.2 意味モデルとカテゴリモデルの評価

意味モデルとカテゴリモデルの有効性を評価するために，手法 P，P+Mm，P+Cg，P+Mm+Cg の正解訳語の平均順位を比較した．

正解訳語の種類（a）〜（c）に対する結果は表 4.4〜表 4.6 に示している．各表において，判定者「A」と「B」の結果を個別に示す．

表 4.4〜表 4.6 において，2 列目の「翻字対象語数」は，正解訳語が少なくとも 1 つ存在する翻字対象語の総数である．3 列目の「関連語数の平均」と 4 列目の「正解訳語数の平均」は，各判定者が翻字対象語 1 つにつき与えた関連語数の平均と正解と判定した正解訳語数の平均である．表 4.4 の 4 列目は，表 4.5，表 4.6 の場合と違い，判定者によって異なる．

表 4.5，表 4.6 では，正解訳語は判定者によらず共通である．さらに，P と P+Cg の結果は関連語に依存しない．表 4.5 と表 4.6 における P と P+Cg の値は判定者によらず同一になる．しかし，表 4.4 では，判定者 2 名の正解訳語が同一であると限らないので，表 4.4 における判定者 2 名の P と P+Cg の値は異なる．

正解訳語の種類と関係なく，表 4.4〜表 4.6 における P+Mm と P+Mm+Cg の値は判定者の関連語に依存するため，判定者 2 名の結果が異なる．

表 4.4〜表 4.6 より，判定者や正解訳語の種類に関係なく，各表における正解訳語の平均順位は，P と比べると，P+Mm，P+Cg，P+Mm+Cg がそれぞれ高い．すなわち，発音モデルに，意味モデルとカテゴリモデルをそれぞれ加え場合，翻字の精度が良くなる．さらに，3 つのモデルを一緒に使った場合（P+Mm+Cg）は翻字の精度がもっと良くなる．

表 4.4: 正解訳語の種類（a）に対する実験結果

| 判定者 | 翻字対象語数 | 関連語数の平均 | 正解訳語数の平均 | 正解訳語の平均順位 | | | |
|---|---|---|---|---|---|---|---|
| | | | | P | P+Mm | P+Cg | P+Mm+Cg |
| A | 210 | 7.1 | 1 | 1130 | 168 | 234 | **120** |
| B | 210 | 5.4 | 1 | 930 | 176 | 376 | **144** |
| 平均 | 210 | 6.2 | 1 | 1030 | 172 | 305 | **132** |

表 4.5: 正解訳語の種類（b）に対する実験結果

| 判定者 | 翻字対象語数 | 関連語数の平均 | 正解訳語数の平均 | 正解訳語の平均順位 | | | |
|---|---|---|---|---|---|---|---|
| | | | | P | P+Mm | P+Cg | P+Mm+Cg |
| A | 125 | 7.2 | 1.1 | 432 | 59 | 57 | **26** |
| B | 125 | 5.5 | 1.1 | 432 | 69 | 57 | **29** |
| 平均 | 125 | 6.4 | 1.1 | 432 | 64 | 57 | **28** |

表 4.6: 正解訳語の種類（c）に対する実験結果

| 判定者 | 翻字対象語語数 | 関連語数の平均 | 正解訳語数の平均 | 正解訳語の平均順位 | | | |
|---|---|---|---|---|---|---|---|
| | | | | P | P+Mm | P+Cg | P+Mm+Cg |
| A | 210 | 7.1 | 1 | 1874 | 102 | 130 | **53** |
| B | 210 | 5.4 | 1 | 1874 | 124 | 130 | **59** |
| 平均 | 210 | 6.2 | 1 | 1874 | 113 | 130 | **56** |

表 4.7〜表 4.9 は表 4.4〜表 4.6 までの結果をそれぞれカテゴリごとに集計した結果である．表 4.7〜表 4.9 では，正解訳語の種類およびカテゴリと関係なく，P の結果と比べて，P+Mm，P+Cg，P+Mm+Cg の方は高い．表 4.7〜表 4.9 の「人名」を見ると，商品名および地名と比べ，P+Mm と P+Mm+Cg は人名の翻字により効果的である．1 つの理由としては，中国語漢字字典 [35] には「この漢字がよく人名に使われる」という趣旨の意味記述があるためである．例えば,「娜」という漢字の意味記述に,「女子人名用字（女性の人名に使う漢字）」と書かれている．判定者が「人名」を関連語に入れると，意味モデルによって,「娜」の確率が大きくなる．その結果，P+Mm と P+Mm+Cg における人名の順位が向上した．この例では，意味モデルとカテゴリモデルの間には依存関係があると考えられる．現在，本研究の翻字手法では考慮していない．この問題は今後さらに検討する必要がある．

図 4.1〜図 4.3 は表 4.4〜表 4.6 の結果に対して，正解訳語の順位に関する分布を示して

表 4.7: 正解訳語の種類（a）におけるカテゴリごとの実験結果

| カテゴリ | 正解訳語の平均順位 | | | |
|---|---|---|---|---|
| | P | P+Mm | P+Cg | P+Mm+Cg |
| 企業名 | 1884 | 157 | 305 | **139** |
| 商品名 | 941 | 226 | 247 | **142** |
| 人名 | 1247 | **66** | 187 | 71 |
| 地名 | 231 | 183 | 559 | **118** |
| その他 | 759 | 141 | 188 | **81** |
| 平均 | 1012 | 155 | 297 | **110** |

表 4.8: 正解訳語の種類（b）におけるカテゴリごとの実験結果

| カテゴリ | 正解訳語の平均順位 | | | |
|---|---|---|---|---|
| | P | P+Mm | P+Cg | P+Mm+Cg |
| 企業名 | 748 | 34 | 57 | **18** |
| 商品名 | 436 | 117 | 68 | **33** |
| 人名 | 879 | **27** | 73 | 43 |
| 地名 | 108 | 40 | 9 | **8** |
| その他 | 106 | 62 | 62 | **33** |
| 平均 | 455 | 56 | 54 | **27** |

いる．図 4.1〜図 4.3 より，正解訳語の種類（a）〜（c）によらず，正解訳語が上位 10 まで入った語数は P より P+Mm と P+Cg の方がそれぞれ多かった．さらに，P+Mm+Cg の場合は一番多く，P の約 2 倍だった．正解訳語の種類（a）〜（c）において，P+Mm+Cg で上位 10 以内に入った語数はそれぞれ翻字対象語全体の約 41%，61%，56% だった．従って，P+Mm+Cg で翻字を行う場合は，上位 10 に適切な翻字候補が入っている確率は高い．

表 4.4〜表 4.6 の結果に対して，P と P+Mm+Cg を比べた場合に正解訳語の順位がどのように変化するか考察した．結果を表 4.10 に示す．

表 4.9: 正解訳語の種類（c）におけるカテゴリごとの実験結果

| カテゴリ | 正解訳語の平均順位 | | | |
|---|---|---|---|---|
| | P | P+Mm | P+Cg | P+Mm+Cg |
| 企業名 | 3457 | 204 | 274 | **127** |
| 商品名 | 2482 | 130 | 116 | **40** |
| 人名 | 882 | **18** | 51 | 30 |
| 地名 | 241 | 101 | 48 | **29** |
| その他 | 1012 | 39 | 98 | **36** |
| 平均 | 1615 | 98 | 117 | **52** |

表 4.10 の「正解訳語数」は各正解訳語の種類において，判定者「A」と「B」を横断した正解訳語の総数である.「向上」,「等しい」,「低下」は，P と比べた場合に，P+Mm+Cg において順位が向上，等しい，または低下した正解訳語の件数を示している.

P と比べると，正解訳語の種類と関係なく，P+Mm+Cg では約 74% の正解訳語は順位が向上した．しかし，同時に，約 20% の正解訳語は順位が低下した．原因として，意味モデルで生成されない漢字に対する平滑化が有効に機能しなかった.

例えば,「コニカ」に対して，発音モデルによって正解訳語「柯尼卡」と不正解訳語「柯尼伽」が生成された．正解訳語中の「卡」は意味モデルに含まれていなかったために，確率として定数の 0.001 が与えられた（3.4 節）．それに対して，不正解訳語中の「伽」は意味モデルに含まれており，確率が 0.053 だった．その結果，意味モデルを考慮すると，正解訳語の順位が不正解訳語よりも低下してしまった.

この問題を解決するためには，平滑化の手法を改善することや，意味モデルに含まれる漢字や単語を増やすことが必要である.

以上をまとめると，発音のみをモデル化する翻字手法には，意味モデルとカテゴリモデルはそれぞれ加えると，翻字精度が良くなる．さらに，発音，意味，およびカテゴリモデルを結合すると，最も良い結果が得られた．同時に，意味モデルには改善の余地がある．例えば，意味モデルとカテゴリモデル間の依存関係を考慮する必要がある．さらに，意味モデルに含まれている漢字の適用範囲を高める必要がある.

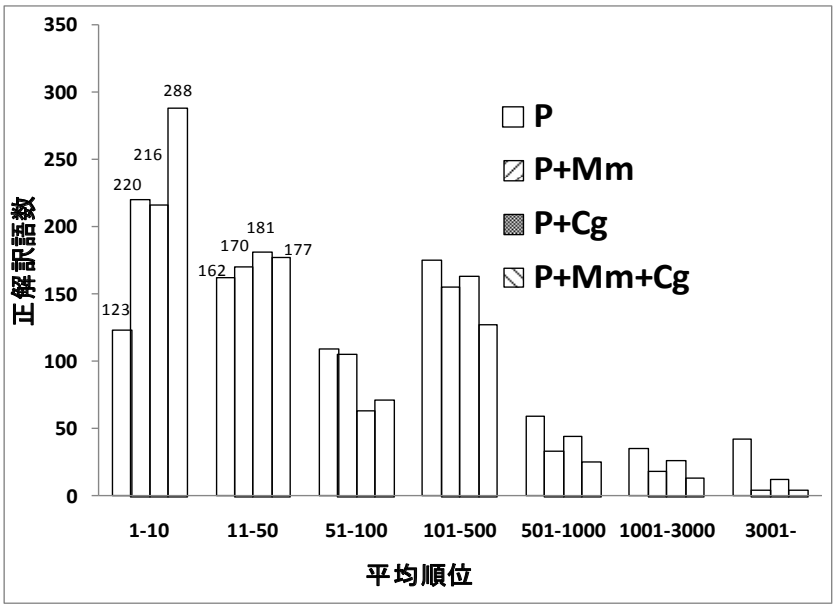


図 4.1: 正解訳語の種類（a）における正解訳語の順位分布図

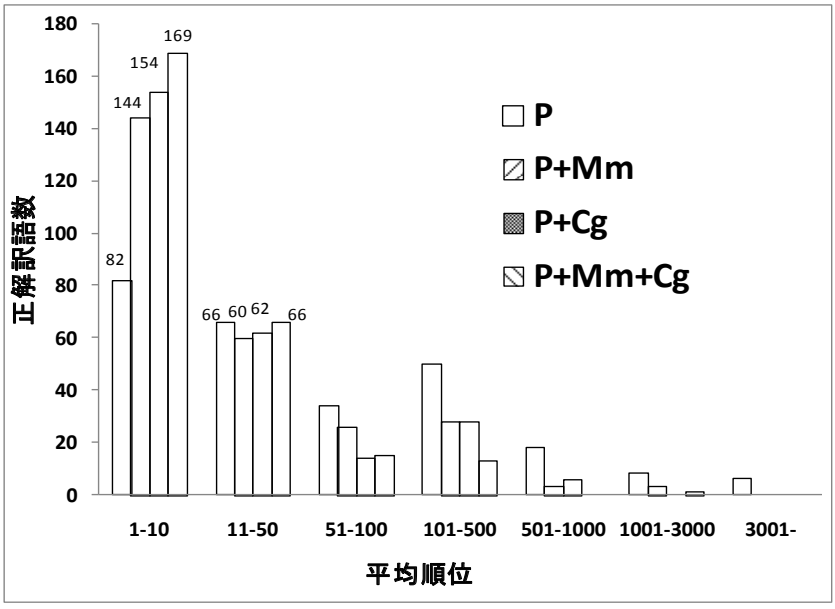


図 4.2: 正解訳語の種類（b）における正解訳語の順位分布図

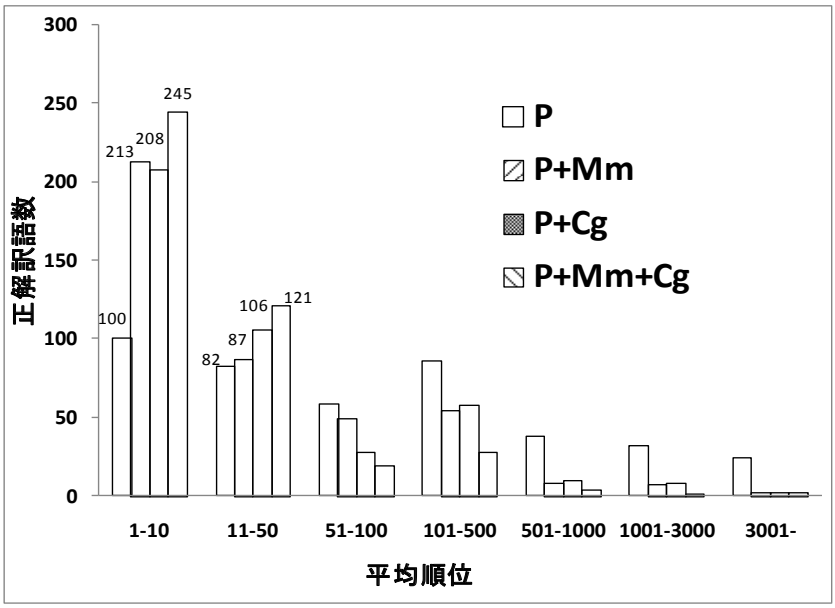


図 4.3: 正解訳語の種類（c）における正解訳語の順位分布図

表 4.10: 正解訳語の順位変化

| 正解訳語の種類 | 正解訳語数 | 正解訳語の順位 | | |
|---|---|---|---|---|
| | | 向上 | 等しい | 低下 |
| (a) | 705 | 522 | 39 | 144 |
| (b) | 264 | 194 | 14 | 56 |
| (c) | 420 | 343 | 20 | 57 |

## 4.3 カテゴリモデル適応の評価

翻字におけるカテゴリモデル適応の有効性を評価するために，手法P+Mm+Cg，P+Mm+Cc，P+Mm+Cp，P+Mm+Caの翻字結果を比較した．表4.11〜表4.13は翻字対象語のカテゴリごとに，正解訳語の種類（a）〜（b）における実験結果をそれぞれ示している．表記方法は表4.7〜表4.9と同じである．

表4.11〜表4.13より，翻字対象語のカテゴリが「企業名」である場合は，正解訳語の種類に関係なく，企業名カテゴリモデルを用いたP+Mm+Ccの結果が最も良かった．同様に，翻字対象語のカテゴリが「人名」である場合は、正解訳語の種類に関係なく，人名カテゴリモデルを用いたP+Mm+Cpの結果が最も良かった．

表4.11〜表4.13の正解訳語の平均順位を見ると，正解訳語の種類に関係なく，P+Mm+Caの結果が最も良かった．従って，翻字にはカテゴリモデル適応が有効である．

図4.4〜図4.6は，表4.11〜表4.13の結果に対して，正解訳語の種類（a）〜（c）における正解訳語の順位に関する分布をそれぞれ示している．図4.4〜図4.6を見ると，正解訳語の種類に関係なく，正解訳語が上位10まで入った語数は，P+Mm+Cg，P+Mm+Cc，およびP+Mm+Cpと比べると，P+Mm+Caが一番多かった．

表4.11: 正解訳語の種類（a）におけるカテゴリモデル適応に関する実験結果

| カテゴリ | 正解訳語の平均順位 | | | |
|---|---|---|---|---|
| | P+Mm+Cg | P+Mm+Cc | P+Mm+Cp | P+Mm+Ca |
| 企業名 | 139 | **120** | 124 | **120** |
| 商品名 | **142** | 164 | 163 | **142** |
| 人名 | 71 | 90 | **57** | **57** |
| 地名 | 118 | 112 | **83** | 118 |
| その他 | **81** | 115 | 113 | **81** |
| 平均 | 110 | 120 | 108 | **104** |

表 4.12: 正解訳語の種類（b）におけるカテゴリモデル適応に関する実験結果

| カテゴリ | 正解訳語の平均順位 | | | |
|---|---|---|---|---|
| | P+Mm+Cg | P+Mm+Cc | P+Mm+Cp | P+Mm+Ca |
| 企業名 | 18 | **14** | 27 | **14** |
| 商品名 | 33 | 57 | **29** | 33 |
| 人名 | 43 | 44 | **9** | **9** |
| 地名 | **8** | 21 | 52 | **8** |
| その他 | 33 | 51 | **20** | 33 |
| 平均 | 27 | 37 | 28 | **20** |

表 4.13: 正解訳語の種類（c）におけるカテゴリモデル適応に関する実験結果

| カテゴリ | 正解訳語の平均順位 | | | |
|---|---|---|---|---|
| | P+Mm+Cg | P+Mm+Cc | P+Mm+Cp | P+Mm+Ca |
| 企業名 | 127 | **76** | 251 | **76** |
| 商品名 | **40** | 107 | 66 | **40** |
| 人名 | 30 | 38 | **5** | **5** |
| 地名 | **29** | 72 | 70 | **29** |
| その他 | 36 | 59 | **18** | 36 |
| 平均 | 52 | 70 | 82 | **37** |

表 4.14 は正解訳語の種類（c）に対して，カテゴリモデル適応が有効だった例を示している．表 4.14 では,「カネボウ（Kanebo）」は企業名であり，P+Mm+Cg，P+Mm+Cc，P+Mm+Cp における正解訳語「嘉娜宝 /jia–na–bao/」の平均順位がそれぞれ 51，2 と 41 だった．企業名カテゴリモデルを使用するとき（P+Mm+Cc）の結果が最も良かった．正解訳語に含まれている漢字「嘉」と「宝」はそれぞれ「喜ばしい」と「宝物」の意味で，中国の企業名によく使われるためである．さらに,「ダビンチ（Da Vinci）」は人名であり，P+Mm+Cg，P+Mm+Cc，P+Mm+Cp における正解訳語「达芬奇 /da–fen–qi/」の

平均順位がそれぞれ 129，13 と 2 だった．中国語において，正解訳語に含まれている漢字「达」と「芬」は，国籍を問わず人の姓名によく使われるため，人名カテゴリモデルを使用するとき（P+Mm+Cp）の結果が最も良かった．

対照的に，表 4.15 は正解訳語の種類（c）に対して，カテゴリモデル適応が有効でなかった例を示している．表 4.15 では,「ヒルトン（Hilton)」は企業名であり，P+Mm+Cg, P+Mm+Cc，P+Mm+Cp における正解訳語「希尔顿 /xi–er–dun/」の平均順位がそれぞれ 2，28 と 1 だった.「ヒルトン」は企業名であるにも拘らず，人名カテゴリモデルを使用するとき（P+Mm+Cp）の結果が最も良かった．原因の 1 つとしては，正解訳語に含まれている漢字「希」と「尔」は中国で企業の名前に使うのは稀である．さらに,「ビゼー（Bizel)」は人名であり，P+Mm+Cg，P+Mm+Cc，P+Mm+Cp における正解訳語「比才 /bi–cai/」の平均順位はそれぞれ 1，8 と 3 だった.「ビゼー」は人名であるにも拘わらず，標準カテゴリモデルを使用するとき（P+Mm+Cg）の結果が最も良かった．原因の 1 つとしては正解訳語に含まれている漢字「比」は，中国で人名に使うのは稀である．

以上をまとめると，外国語を中国語に翻字するのに，カテゴリモデル適応が有効であった．同時に，カテゴリモデルには改善の余地がある．表 4.15 で挙げた例を見ると，例外的な翻字には，提案した統計的手法はいつも効果的であるとは限らない．

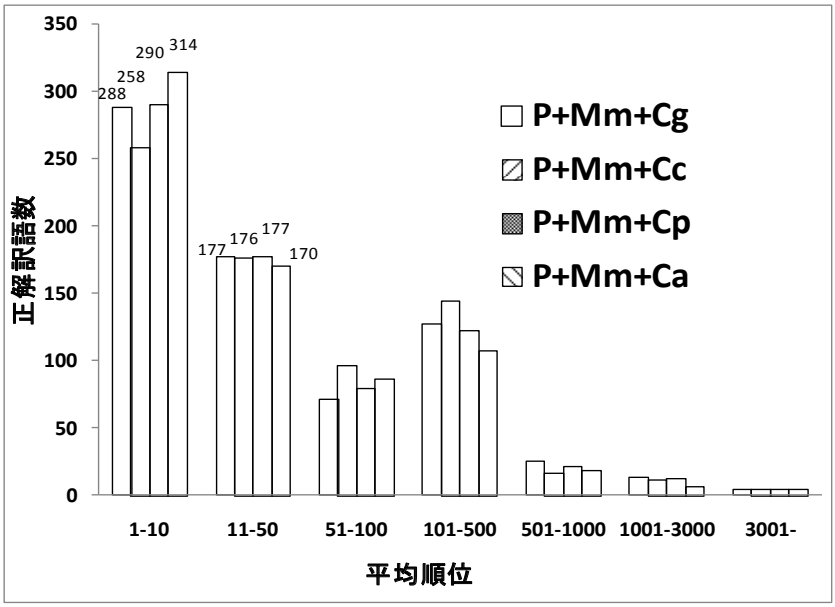


図 4.4: 正解訳語の種類（a）における正解訳語の順位分布図

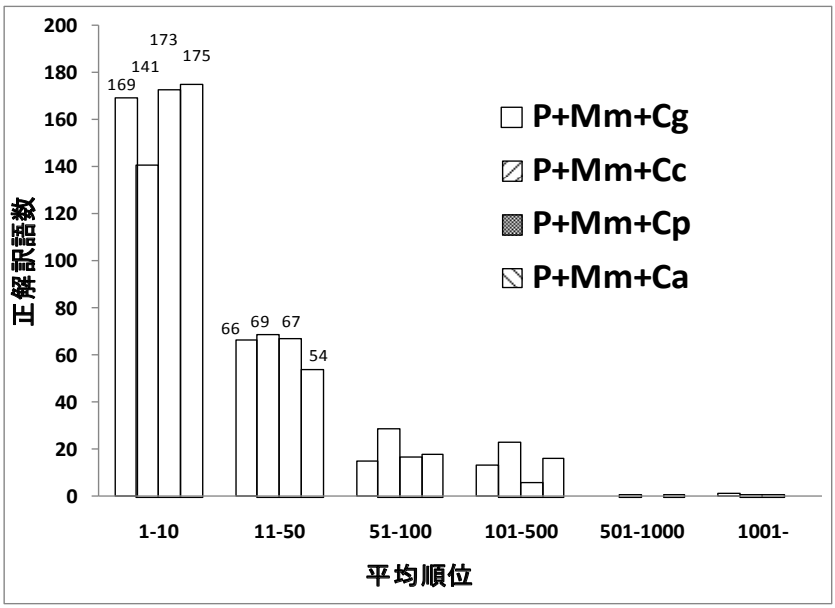


図 4.5: 正解訳語の種類（b）における正解訳語の順位分布図

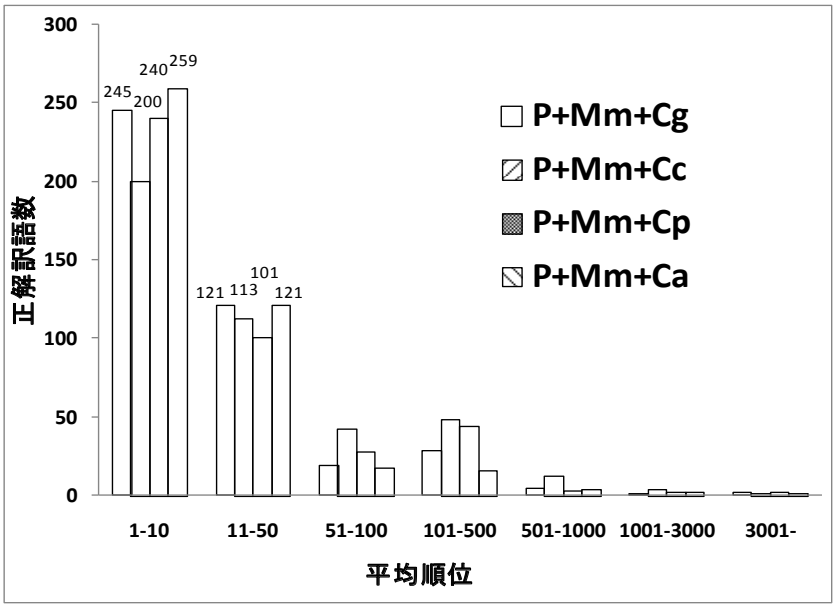


図 4.6: 正解訳語の種類（c）における正解訳語の順位分布図

表 4.14: カテゴリモデル適応が有効だった例

| 翻字対象語 | 正解訳語 | カテゴリ | 正解訳語の平均順位 | | |
|---|---|---|---|---|---|
| | | | P+Mm+Cg | P+Mm+Cc | P+Mm+Cp |
| カネボウ | 嘉娜宝 | 企業名 | 51 | **2** | 41 |
| ダビンチ | 达芬奇 | 人名 | 129 | 13 | **2** |

表 4.15: カテゴリモデル適応が有効でなかった例

| 翻字対象語 | 正解訳語 | カテゴリ | 正解訳語の平均順位 | | |
|---|---|---|---|---|---|
| | | | P+Mm+Cg | P+Mm+Cc | P+Mm+Cp |
| ヒルトン | 希尔顿 | 企業名 | 2 | 28 | **1** |
| ビゼー | 比才 | 人名 | **1** | 8 | 3 |

## 4.4 関連語自動抽出の評価

翻字における関連語自動抽出の効果を評価するために，手法 P+Cg，P+Ma+Cg と P+Mm+Cg を比較する．P+Ma+Cg 評価対象で，は本研究の提案手法である．P+Ma+Cg にとって，P+Cg と P+Mm+Cg はそれぞれ期待される翻字精度の下限と上限を推定するための手法である．また，手法 P+Ca，P+Ma+Ca と P+Mm+Ca についても比較し，関連語自動抽出の効果はカテゴリモデル適応によって変化するかどうかを考察する．

翻字対象語は 4.1 節で使用された 210 語を使う．2006 年 8 月 20 日に 210 語を Wikipedia で調べ，Wikipedia の見出し語として存在する 128 語を選んだ．その理由は，Wikipedia で記事ページが検索されなかった用語は関連語を抽出できないため，翻字対象語から削除した．また，Wikipedia で「曖昧さ回避のためのページ」が検索された用語も翻字対象語から削除した．例えば,「アポロ」で検索すると「アポロ（小惑星)」や「アポロ（曲)」といった異なる語義について書かれた記事へのリンクが表示される．本研究を実際に運用する場合，現状では複数のリンクから対象の語義に関する記事ページを自動的に特定することができない．今回の実験では多義語を翻字対象語から削除した．最終的に翻字対象として残った 128 語の内訳を表 4.16 に示す．P+Cg，P+Ma+Cg と P+Mm+Cg の翻字結果は 4.1 節でプーリングされたデータを使用した．

4.1 節と同じ 2 人の中国人留学生が判定者として，意味モデルの関連語を与えて，翻字結果の正解判定を行った．各翻字対象語の「正解訳語」も 4.1 節と同じ，正解訳語の種類（a）〜（c）を使用した．

正解訳語の種類（a）〜（c）に対する結果は表 4.17〜表 4.19 に示している．各表において，判定者「A」と「B」の結果を個別に示す．

表 4.17〜表 4.19 において，2 列目の「翻字対象語数」は，正解訳語が少なくとも 1 つ存在する翻字対象語の総数である．3 列目の「関連語数の平均」と 4 列目の「正解訳語数の平均」は，各判定者が翻字対象語 1 つにつき与えた関連語数の平均と正解と判定した正解訳語数の平均である．翻字結果を公平に比較するために，翻字対象ごとに Ma と Mm で使用する関連語の数を揃えた．具体的には，Ma では式 (3.9) の値が高い関連語候補のうち，Mm で使用された関連語と同じ件数だけを使用した．従って，表 4.17〜表 4.19 において，判定者ごとに，P+Ma+Cg，P+Mm+Cg，P+Ma+Ca，P+Mm+Ca の「関連語数の平均」が同じである．表 4.17 の 4 列目は，表 4.18，表 4.19 の場合と違い，判定者によって異な

る．表 4.18，表 4.19 では，正解訳語は判定者によらず共通である．表 4.17〜表 4.19 では，手法 P+Cg と P+Ca は関連語に依存しないため，正解訳語の種類ごとに判定者 2 人の値が同じである．しかし，P+Ma+Cg，P+Mm+Cg，P+Ma+Ca，P+Mm+Ca は使用する関連語に依存するため，値が判定者ごとに異なる．

表 4.17〜表 4.19 を見ると，正解訳語の種類に関係なく，正解訳語の平均順位は P+Ma+Cg と P+Mm+Cg が P+Cg より良かった．そして，P+Mm+Cg の翻字結果が最も良かった．P+Ma+Cg は正解訳語の種類に関係なく，正解訳語の平均順位は下限の P+Cg より良かった．P+Ca，P+Ma+Ca，P+Mm+C でも同じ傾向だった．

図 4.7〜図 4.9 は，表 4.17〜表 4.19 の P+Cg，P+Ma+Cg と P+Mm+Cg の翻字結果に対して，正解訳語の種類ごとに正解訳語の順位に関する分布を示している．図 4.7〜図 4.9 を見ると，正解訳語が上位 10 まで入った語数では，P+Ma+Cg は P+Mm+Cg より少なかった．しかし，P+Cg より多かった．

図 4.10〜図 4.12 は，表 4.17〜表 4.19 の P+Ca，P+Ma+Ca と P+Mm+Ca の翻字結果に対して，正解訳語の種類ごとに正解訳語の順位に関する分布を示している．図 4.10 〜図 4.12 を見ると，正解訳語が上位 10 まで入った語数では，図 4.7〜図 4.9 の場合と同じ傾向だった．

以上をまとめると，翻字における関連語自動抽出の効果は，カテゴリモデルの選択と無関係で有効だった．そして，正解の種類と関係なく，自動抽出した関連語を利用して翻字を行う手法は，意味モデルを利用しない手法よりも有効であった．さらに，翻字精度を多少犠牲にして，人手で関連語を与えるコストを削減することができた．

正解訳語の種類（b）と（c）に関して，自動抽出した関連語を上位から 1 つずつ増やして，P+Ma+Cg と P+Ma+Ca における正解訳語の平均順位が変化する様子を調べた結果をそれぞれ図 4.13 と図 4.14 に示す．

図 4.13 より，関連語を 1 つしか使用しない場合，正解訳語の種類 (b) と (c) における P+Ma+Cg の平均順位はそれぞれ 46 と 140 であり，表 4.18 と表 4.19 にそれぞれ示した「P+Cg」の平均順位（54 と 155）よりも高かった．また，正解訳語の平均順位は関連語数が増えるにつれ高くなり，関連語数が 7 を超えたところでほぼ一定になった．この傾向は図 4.14 でも同様であった．

表 4.16: 翻字対象語 128 語の内訳

| カテゴリ | 語数 | 例 | | |
|---|---|---|---|---|
| | | 日本語 | 中国語 | 英語 |
| 企業名 | 35 | ボーイング | 波音 | Boeing |
| | | エンロン | 恩隆 | Enron |
| | | ゲラン | 盖兰 | Guerlain |
| | | オメガ | 欧米茄 | Omega |
| | | オペル | 欧宝 | Opel |
| 商品名 | 27 | キャデラック | 凯迪拉克 | Cadillac |
| | | シャネル | 夏奈尔 | Chanel |
| | | フェラーリ | 法拉利 | Ferrari |
| | | ペンティアム | 奔腾 | Pentium |
| | | プラダ | 普拉达 | Prada |
| 人名 | 13 | エンヤ | 恩雅 | Enya |
| | | ガウス | 高斯 | Gauss |
| | | カラヤン | 卡拉扬 | Karajan |
| | | プーチン | 普京 | Puchin |
| | | ラビン | 拉宾 | Rabin |
| 地名 | 29 | チリ | 智利 | Chile |
| | | カラチ | 卡拉奇 | Karachi |
| | | ロメ | 洛美 | Lome |
| | | キト | 基多 | Quito |
| | | リヤド | 利雅得 | Riyad |
| その他 | 24 | バレエ | 芭蕾 | ballet |
| | | ミサ | 弥撒 | missa |
| | | ナイロン | 尼龙 | nylon |
| | | オーム | 欧姆 | ohm |
| | | サウナ | 桑拿 | sauna |

表 4.17: 正解訳語の種類（a）に対する実験結果

| 判定者 | 翻字対象語数 | 関連語数の平均 | 正解訳語数の平均 | 正解訳語の平均順位 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | +Cg | | | +Ca | | |
| | | | | P | P+Ma | P+Mm | P | P+Ma | P+Mm |
| A | 128 | 7.3 | 1 | 257 | 151 | **134** | 248 | 147 | **130** |
| B | 128 | 5.4 | 1 | 510 | 191 | **184** | 513 | 186 | **181** |
| 平均 | 128 | 6.3 | 1 | 384 | 171 | **159** | 381 | 166 | **156** |

表 4.18: 正解訳語の種類（b）に対する実験結果

| 判定者 | 翻字対象語数 | 関連語数の平均 | 正解訳語数の平均 | 正解訳語の平均順位 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | +Cg | | | +Ca | | |
| | | | | P | P+Ma | P+Mm | P | P+Ma | P+Mm |
| A | 76 | 7.4 | 1.04 | 54 | 29 | **22** | 58 | 30 | **22** |
| B | 76 | 5.6 | 1.04 | 54 | 30 | **29** | 58 | 30 | **26** |
| 平均 | 76 | 6.5 | 1.04 | 54 | 30 | **25** | 58 | 30 | **24** |

表 4.19: 正解訳語の種類（c）に対する実験結果

| 判定者 | 翻字対象語数 | 関連語数の平均 | 正解訳語数の平均 | 正解訳語の平均順位 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | +Cg | | | +Ca | | |
| | | | | P | P+Ma | P+Mm | P | P+Ma | P+Mm |
| A | 128 | 7.3 | 1 | 155 | 119 | **59** | 147 | 101 | **46** |
| B | 128 | 5.4 | 1 | 155 | 128 | **68** | 147 | 110 | **53** |
| 平均 | 128 | 6.3 | 1 | 155 | 124 | **64** | 147 | 106 | **49** |

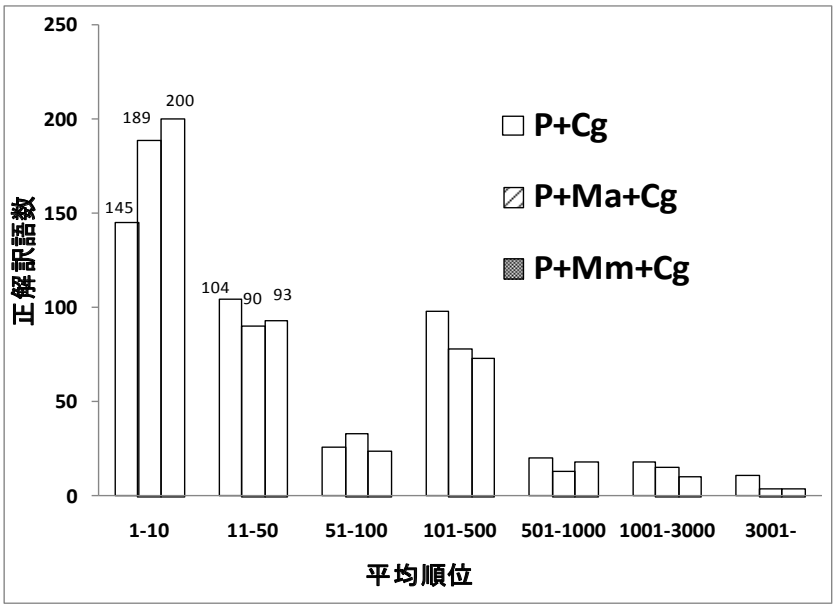


図 4.7: 標準カテゴリモデルを利用して正解訳語の種類（a）における順位分布図

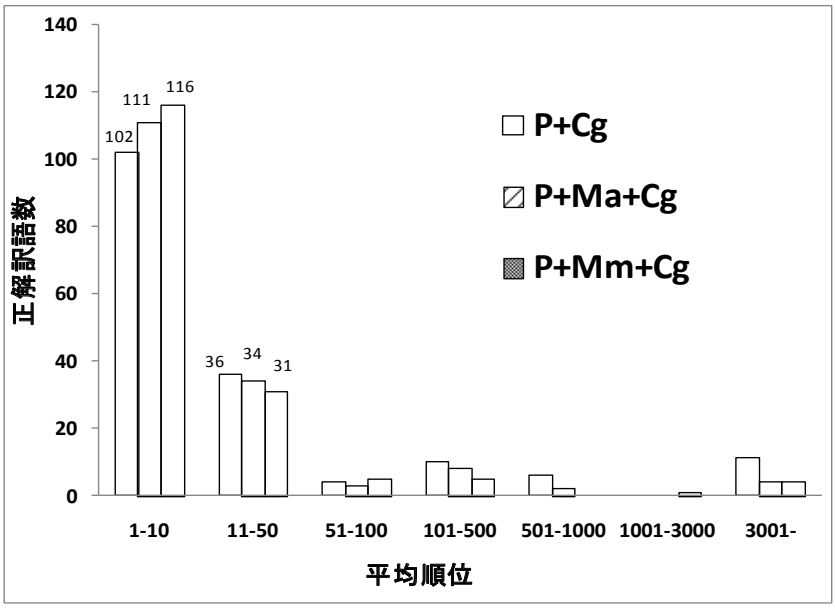


図 4.8: 標準カテゴリモデルを利用して正解訳語の種類（b）における順位分布図

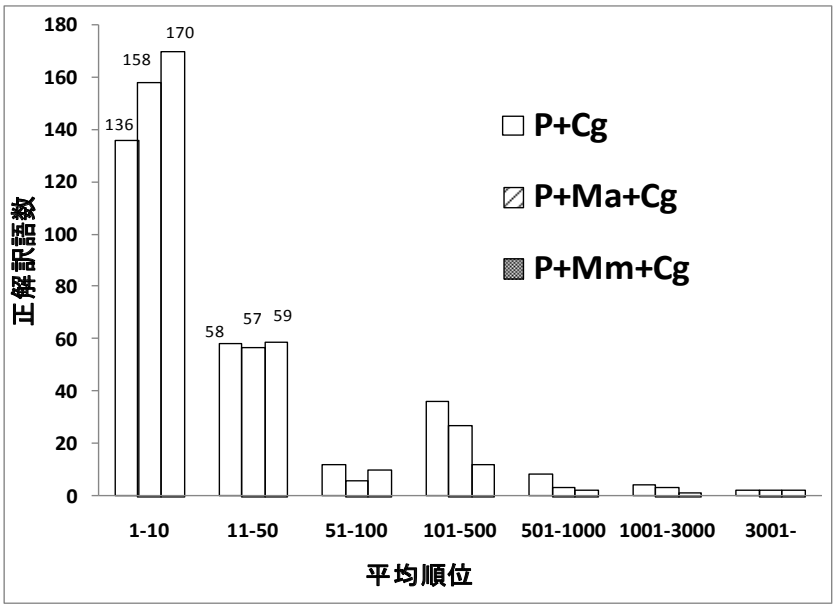


図 4.9: 標準カテゴリモデルを利用して正解訳語の種類（c）における順位分布図

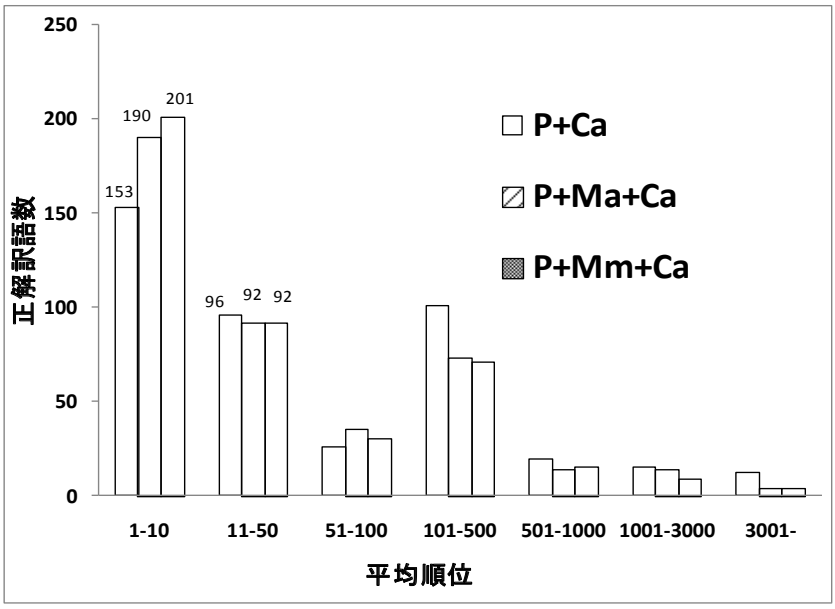


図 4.10: カテゴリモデル適応を利用して正解訳語の種類（a）における順位分布図

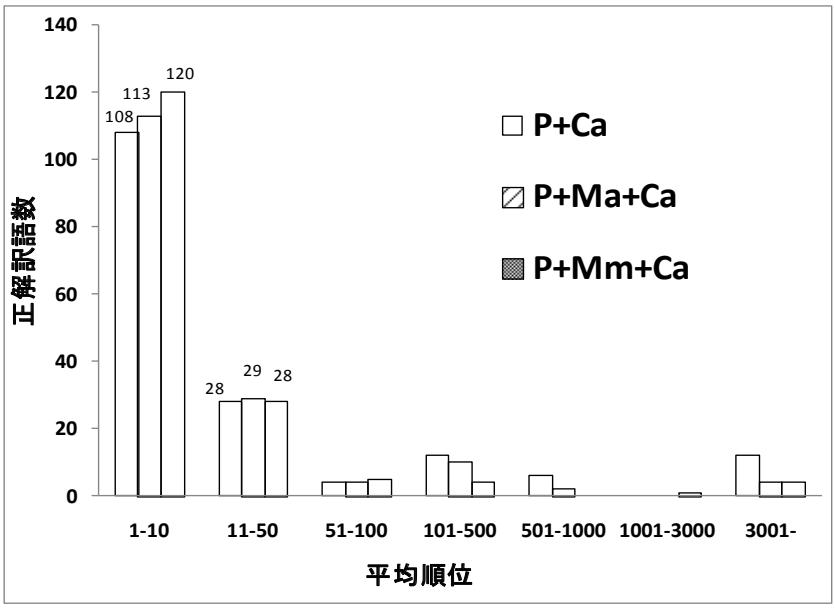


図 4.11: カテゴリモデル適応を利用して正解訳語の種類（b）における順位分布図

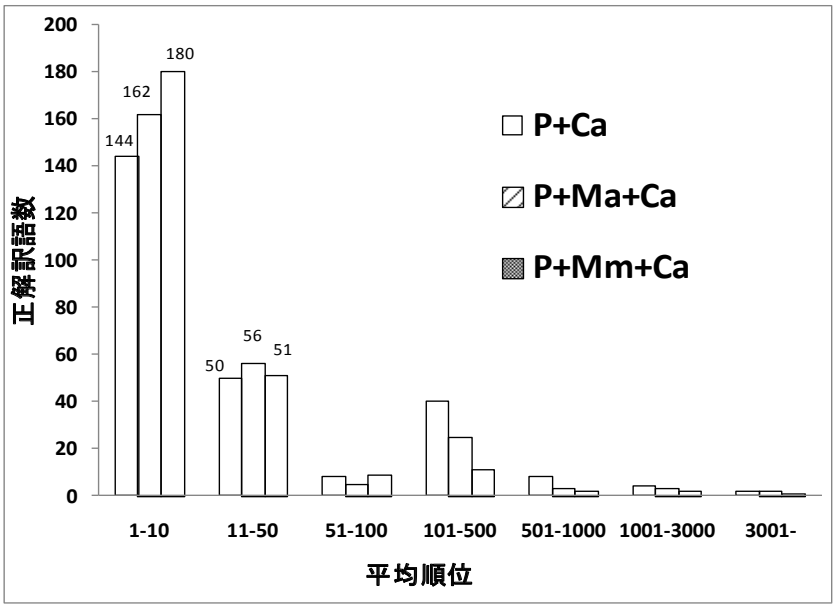


図 4.12: カテゴリモデル適応を利用して正解訳語の種類（c）における順位分布図

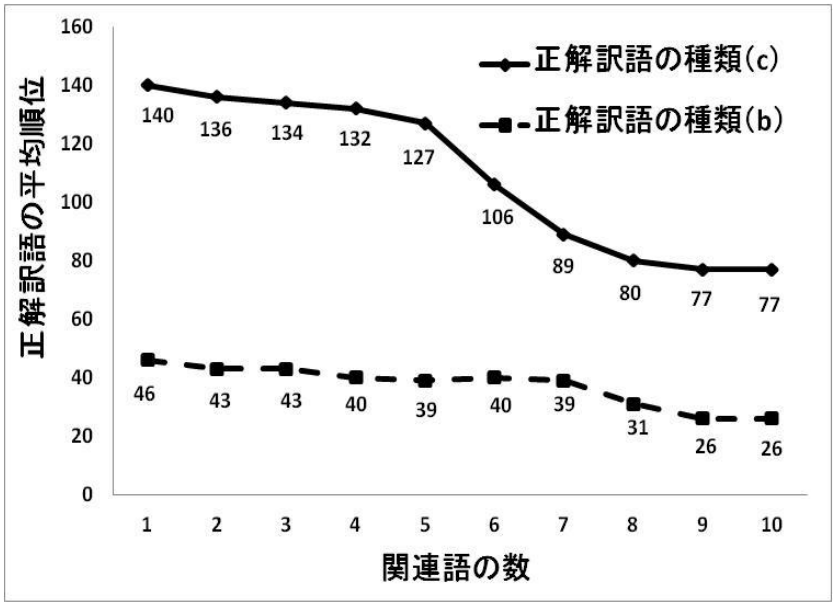


図 4.13: P+Ma+Cg における関連語の数と正解訳語の平均順位

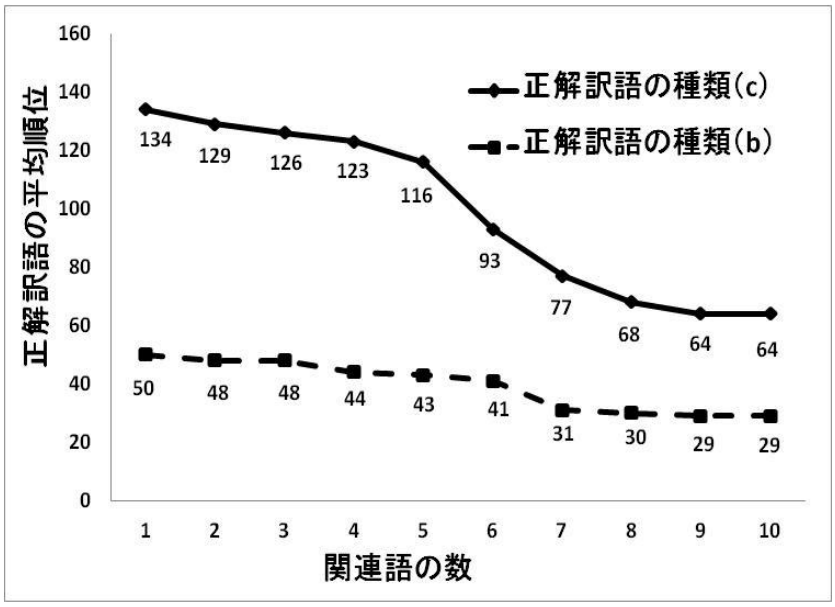


図 4.14: P+Ma+Ca における関連語の数と正解訳語の平均順位

表 4.20 と 4.21 は，用語のカテゴリごとに翻字対象を 1 つずつ選んで，翻字に使用した関連語と正解の種類 (a) における正解訳語の平均順位を示している．表 4.20 は自動抽出した関連語が有効だった翻字対象の例を示し，表 4.21 は自動抽出した関連語が有効でなかった翻字対象の例を示している．表 4.20 の「順位」では「自動」における正解訳語の平均順位は「人手」より高く，逆に表 4.21 では「自動」の順位は「人手」より低い．表 4.20 と 4.21 の「中国語の関連語」において,「自動」では中国語に翻訳する前の日本語を括弧内に示す．ただし,「人手」の関連語は判定者が直接中国語で入力したため，筆者が日本語訳を与えた.「中国語の関連語」を見ると，自動的に抽出した関連語は人手で与えた関連語とあまり一致していない．以下，この点について具体例を挙げながら考察する．

表 4.20 を見ると,「カネボウ」では，企業名を付ける際には使用されないであろう「破产（破産）」と「傲慢（傲慢）」が関連語として使用されていた．判定者が不適切な関連語を与えたことで評価実験の妥当性が損なわれていないか調べるために，全ての翻字対象について著者が関連語を吟味した．その結果,「カネボウ」の「破产（破産）」と「傲慢（傲慢）」以外の関連語には問題がなかった．さらに,「カネボウ」の関連語から「破产（破産）」と「傲慢（傲慢）」を削除し，自動抽出した関連語数を人手の関連語数に揃えて再度実験を行った．その結果，正解の種類や「人手」と「自動」といった手法の違いによらず，上記 2 つの関連語を削除する前と比べて実験結果は変わらなかった．以上より，人手による不適切な関連語によって評価実験の妥当性が損なわれていないことを確認した．

別の例として,「カラチ」では，自動抽出した関連語の中に，Wikipedia の記事ページにある「カラチ」と関連が強いと考えられるいくつかの語が含まれていない．例えば,「ムスリム」や「パキスタン」である.「ムスリム」と「パキスタン」は関連語候補として抽出されたものの「カラチ」との相互情報量は関連語候補中それぞれ 11 位と 12 位だった．他方において，判定者 A と B が「カラチ」に与えた関連語の数はそれぞれ 8 語と 10 語だった．人手と自動と関連語の数を揃えたため,「ムスリム」と「パキスタン」は最終的に関連語として選択されなかった．本来関連が強い語を自動的に関連語として選択するためには，関連語候補を抽出する段階と抽出した候補を一定の基準で順位付ける段階のそれぞれにおいて改善の余地がある．

まず，関連語候補を抽出する段階では，記事ページ本文全体が抽出対象となっている点に問題がある．表 4.20 において「カラチ」の関連語を見ると,「外部リンク」というセクションから抽出された「リンク」が関連語として選択されている．しかし，Wikipedia で

は「外部リンク」に見出し語に関する説明が書かれることは稀である.「カラチ」の例に関して言えば,「外部リンク」のセクションを関連語抽出の対象から削除すれば,「リンク」は関連語候補として抽出されず，その結果「ムスリム」や「パキスタン」の順位が相対的に上がる．しかし,「ハワイ」の記事ページにおける「外部リンク」のセクションには,「カウアイ観光局」や「オアフ島観光局」などのアンカーテキスト（リンクをはるためのテキスト）が記述されており,「カウアイ」や「オアフ島」などの「ハワイ」に関連する語を含んでいる．すなわち，関連語抽出の対象から「外部リンク」を一律削除すればよいとは限らない．また，3.6 節で議論したように，セクションの分け方，見出しの付け方，セクション内の記述内容に関する方針は記事の著者によって異なる．以上より，関連語抽出において対象にすべきセクションとそれ以外を正確に区別することは難しい．これは今後も検討して解決すべき課題である.

関連語の候補に順位を付ける段階では，式 (3.9) で用いた相互情報量以外の計算方法を試し，本研究の目的にとって最適な手法について今後検討する必要がある.

表 4.21 を見ると,「シャネル」では，自動抽出した関連語のうちいくつかが人手で与えた関連語と一致した．例えば,「香水（香水)」や「名牌（ブランド)」などである．しかし，人手で与えられた「华丽（華やか)」や「耐久（耐久)」などのように翻字対象の印象を表す関連語がなく，翻字に有効でなかった.「インテル」では，自動抽出した関連語に「インテル」に関する印象を表す語がなかった.「カタール」では，自動抽出した関連語は全てカタール周辺国の国名であり,「カタール」自体を表す語として適切ではなかった.「モナリザ」と「ディスコ」では，自動抽出した関連語の中に，翻字対象と関係のない語がいくつかあった．例えば,「モナリザ」の「自己（自分)」,「手（手)」,「没有（無く)」や,「ディスコ」の「好（良い)」,「回来（帰り)」,「制（製)」である．また，Yahoo! JAPAN の翻訳システムによる誤訳もあった．例えば,「ディスコ」に対する日本語の関連語「非常」は「とても」という意味なので，中国語の「大」ではなく「非常」と訳されるべきであった.

表 4.20: 自動抽出した関連語が有効だった翻字対象の例

| カテゴリ | 翻字対象 | 正解訳語 | 手法 | 順位 | 中国語の関連語（日本語） |
|---|---|---|---|---|---|
| 企業名 | カネボウ | 嘉娜宝 | 自動 | 5 | 花王（花王），制（製），制药（製薬），化妆（化粧），公司（会社），食品（食品），肥皂（石鹸），粉饰（粉飾） |
| | | | 人手 | 28 | 破产（破産），化妆品（化粧品），美容（美容），美丽（綺麗），可爱（可愛い），傲慢（傲慢），气质（気質），女孩（女の子），才华（才気），食品（食品） |
| 商品名 | ボルボ | 沃尔沃 | 自動 | 141 | 标致汽车（プジョー），雷诺（ルノー），马自达（マツダ），轿车（セダン），小轿车（クーペ），福特（フォード），车型（車種），福特汽车（フォード・モーター），三菱汽车工业（三菱自動車），全部的型号（フルモデル） |
| | | | 人手 | 275 | 大型（大型），舒适（心地よい），身份（身分），豪华（豪華），快速（高速），汽车公司（自動車会社），安全（安全），重型（重機），卡车（トラック），北欧（北欧） |
| 人名 | カラヤン | 卡拉扬 | 自動 | 3 | 柏林菲尔（ベルリン・フィル），管弦乐（管弦楽），交响乐团（フィルハーモニー），伯恩斯坦（バーンスタイン），交响曲（交響曲），施特劳斯（シュトラウス），马勒（マーラー），瓦格纳（ワーグナー），菲尔（フィル） |
| | | | 人手 | 5 | 指挥家（指揮家），风范（風格），奥地利（オーストリア），音乐（音楽），洒脱（さっぱりしている），大方（気前が良い），激情（激情），豪迈（豪胆），细腻（繊細） |
| 地名 | カラチ | 卡拉奇 | 自動 | 130 | 美国（アメリカ），国际（国際），都市（都市），郡（州），机场（空港），交通（交通），滑冰场（リンク），西面（西），地域（地域），隐私（プライバシー） |
| | | | 人手 | 139 | 巴基斯坦（パキスタン），魔幻（幻），风韵（あでやかな姿），婀娜多姿（しなやかで美しい），高贵（高貴），清真寺（モスク），伊斯兰教（イスラム教），海港（港） |
| 一般名詞 | ミサ | 弥撒 | 自動 | 41 | 秘跡（秘跡），圣体（聖体），式司（司式），典礼（典礼），天主教（カトリック），奉献（奉献），主教（司教），约翰（ヨハネ），教会（教会），语言（ことば） |
| | | | 人手 | 52 | 神圣（神聖），仪式（儀式），端庄（荘重），富丽（華麗），天主教（カトリック），信仰（信仰），祭祀（祭祀），庄严（厳粛），圣餐（聖餐），教会（教会），典礼（典礼） |

表 4.21: 自動抽出した関連語が有効でなかった翻字対象の例

| カテゴリ | 翻字対象 | 正解訳語 | 手法 | 順位 | 中国語の関連語（日本語） |
|---|---|---|---|---|---|
| 企業名 | インテル | 英特尔 | 自動 | 16 | 制（製），各种各样（さまざま），最合适地（最適），大（非常），装置（デバイス），日语（日本語），闪光（フラッシュ），实现（実現），法人（法人），产品（製品） |
| | | | 人手 | 6 | 爽朗（さわやか），快乐（楽しみ），男孩（男の子），垄断（独占），电脑（パソコン），微软（マイクロソフト），普及（普及），网络（ネットワーク），世界（世界） |
| 商品名 | シャネル | 夏奈尔 | 自動 | 26 | 香水（香水），附件（アクセサリー），化妆（化粧），便宜（安く），名牌（ブランド），固定（固定），巴黎（パリ），收集（コレクション），黑色（ブラック），衣服（服） |
| | | | 人手 | 18 | 时尚（ファッション），性感（セクシー），野味（野性的），魅力（魅力），潮流（流行），香水（香水），服装（洋服），奢侈品（贅沢品），华丽（華やか），耐久（耐久），名牌（ブランド），法国（フランス），时髦（流行） |
| 人名 | モナリザ | 蒙娜丽莎 | 自動 | 64 | 列奥纳多（レオナルド），丽萨（リザ），大（非常），自己（自分），微笑（微笑），手（手），美术馆（美術館），没有（無く） |
| | | | 人手 | 2 | 美丽（綺麗），温柔（やさしい），贵妇（貴婦人），微笑（微笑），眼泪（涙），神秘（神秘），少妇（若い奥さん），达芬奇（ダビンチ），名画（名画） |
| 地名 | カタール | 卡塔尔 | 自動 | 452 | 科威特（クウェート），也门（イエメン），奥曼（オマーン），利比亚（リビア），约旦（ヨルダン），巴林（バーレーン），叙利亚（シリア），黎巴嫩（レバノン） |
| | | | 人手 | 239 | 阿拉伯（アラブ），酋长（酋長），石油（石油），狭小（狭い），沙漠（砂漠），干燥（乾燥），战乱（戦乱），混乱（混乱），独立（独立），伊斯兰（イスラム） |
| 一般名詞 | ディスコ | 迪斯科 | 自動 | 43 | 好（良い），音乐（音楽），舞厅（ダンスホール），大（非常），朱丽安娜（ジュリアナ），曲（曲），回来（帰り），制（製），跳舞（ダンス） |
| | | | 人手 | 34 | 消遣（気晴らし），娱乐（娯楽），跳舞（ダンス），节奏（リズム），快乐（楽しい），热闹（賑やか），喧闹（騒がしい），乱（乱れる），啤酒（ビール），美女（美女） |

# 第5章　結論

## 5.1　本研究の貢献

外国語を導入する際に，基本的な方法は，翻訳，翻字，原言語をそのまま出力するの3つである．企業名や商品名のような固有名詞は通常翻字を利用する．

中国語では漢字を用いて翻字する．しかし，漢字は表意文字であり，同じ発音に複数の漢字が対応している．また，異なる漢字が異なった意味と印象を持っているので，中国語へ翻字する際に，使用する漢字の選択に注意を払う必要がある．この点は外国の企業が自社の企業名や商品名を中国に輸入する際に特に重要である．さらに，人名や地名のように，漢字の選択は原言語が属する種別に関連している場合が多い．従って，中国への翻字には単に原言語の発音だけではなく，漢字の意味や原言語が属する種別も考慮する必要がある．

中国語への翻字に関する従来の研究は主に発音と言語モデルの組み合わせである．Xuら [29] の手法は翻字対象の印象を考慮している．しかし，翻字対象の印象を表す印象キーワードは人手で与えなければならない．さらに，翻字対象の種別を考慮していない．Liら [19] は意味翻訳の方法を提案した．ただし，彼らの方法は人名の翻字だけに適応できる．本研究は，翻字対象の種類を制限せずに，発音，意味，および種別をモデル化する確率的な意味訳型翻字手法を提案した．意味モデルを構築するために，漢字字典を使用した．漢字辞典において，見出し漢字は文で説明され，また，見出し漢字を含まれている1つ以上の単語によって例示されている．カテゴリモデルを構築するために，本研究は文字のN-gramモデルを利用した．そして，標準，企業名，人名の3つのカテゴリモデルを構築した．

翻字対象が与えられた場合，本研究の発音モデルはその翻字対象を発音に似ている複数の翻字候補に変換する．ただし，ローマ字表記に変換できれば，ほかの言語を入力することも原理的に可能である．翻字対象の関連語を与えられた場合，本研究の意味モデル

はこれらの関連語を関連する漢字のセットに変換する．ただし，関連語が中国語ではない場合，機械翻訳を利用して中国語に翻訳する．翻字対象のカテゴリが与えられた場合，本研究のカテゴリモデルは対象カテゴリに良く出現する漢字のセットを選択してくれる．この3つのモデルを統合するために，本研究は確率的な枠組みを提案した．そして，この確率的な枠組みより，確率が高い翻字候補は出力されるリストの上位に出現し，提出される．さらに，人手で翻字対象のために1つ以上の関連語を与えるのは手間がかかって高価である．本研究はWebを利用して，自動的に翻字対象の関連語を抽出する手法を提案した．翻字対象の情報源として，Wikipediaの日本語版を調査し，結果の記事ページに形態素解析を行った．そして，関連語の候補として，形態素解析の結果から名詞と形容詞を抽出し，関連語を選択するには翻字対象との相互情報量を使用した．

評価実験で本研究が提案した意味訳型翻字手法の有効性が示された．本研究は意味モデル，カテゴリモデル，およびカテゴリモデルの適応について評価した．発音モデルのみの翻字手法は意味モデルとカテゴリモデルのそれぞれによって改良された．3つのモデルを結合し，カテゴリモデルを適応させた場合，最も良い翻字結果が得られた．

さらに，翻字における関連語自動抽出の有効性も評価した．自動抽出した関連語を利用した意味モデルの翻字結果は意味モデルを利用しない手法と比べると，より効果的だった．また，翻字精度を多少犠牲にして，人手で関連語を与えるコストを削減することができた．

## 5.2 残された課題

まず，提案手法に含まれている各部分の改良である．

本研究が提案した意味訳型翻字手法は,「発音」,「意味」,「カテゴリ」モデルの三つから構成されている．発音モデルを構築する際に，ローマ字音節とピンイン音節の対応を得るため，DPマッチングを利用した．準備実験では，対応付けの精度は91.7%だった．この精度を高めなければならない．

意味モデルを構築する際に，中国語漢字字典 [35] を使用した．字典から外国語の表記によく使われ，字典の品詞部分に「外」という印が付いている見出し漢字599文字を選択した．しかし，599文字に含まれていない漢字については対応できなく，定数を与えて平滑化している．この問題も考えなければならない．

意味モデルにおけるもう 1 つ重要な課題は，関連語自動抽出のさらなる精緻化である．また，Wikipediaで説明を得ることができない用語や多義語への対応も今後の課題である．

カテゴリモデルも洗練しなければならない．例えば，Li ら [19] のように，姓，名，性別などの名前に関する意味属性を我々のカテゴリモデルに統合することができる．

# 参考文献


[1] Yaser Al-Onaizan and Kevin Knight. 2002. Translating Named Entities Using Monolingual and Bilingual Resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 400–408.

[2] Eiji Aramaki and Takeshi Abekawa. 2009. Fast Decoding and Easy Implementation: Transliteration as Sequential Labeling. In *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP*, pages 65–68.

[3] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An Algorithm that Learns What's in a Name. *Machine Learning*, 34(1–3), pages 211–231.

[4] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizaka. 2007. Measuring Semantic Similarity between Words Using Web Search Engines. In *Proceedings of the 16th International World Wide Web Conference*, pages 757–766.

[5] Hsin-Hsi Chen, Shen-Jie Hueng, Yung-Wei Ding, and Shih-Chung Tsai. 1998. Proper Name Translation in Cross-Language Information Retrieval. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 232–236.

[6] Kenneth Ward Church and Patrick Hanks. 1989. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83.

[7] Atsushi Fujii and Tetsuya Ishikawa. 2001. Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration. *Computers and the Humanities*, 35(4):389–420.

[8] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. 2006. LODEM: A System for On-Demand Video Lectures. *Speech Communication*, 48(5):516–531.

[9] Peter Hu. 2004. Adapting English to Chinese. *English today*, 20(2):34–39.

[10] Hideki Isozaki. 2001. Japanese Named Entity Recognition Based on a Simple Rule Generator and Decision Tree Learning. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 314–321.

[11] Kil Soon Jeong, Sung Hyon Myaeng, Jae Sung Lee, and Key-Sun Choi. 1999. Automatic Identification and Back-Transliteration of Foreign Words for Information Retrieval. *Information Processing & Management*, 35:523–540.

[12] Yuxiang Jia, Danqing Zhu, and Shiwen Yu. 2009. A Noisy Channel Model for Grapheme-based Machine Transliteration. In *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP*, pages 88–91.

[13] Adam Kilgarriff. 2007. Googleology is Bad Science. *Computational Linguistics*, 33(1):147–151.

[14] Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics*, 24(4):599–612.

[15] Kui-Lam Kwok and Peter Deng. 2002. Corpus-based Pinyin Name Resolution. In *Proceeding of the First SIGHAN Workshop on Chinese Language Processing*, pages 41–47.

[16] K.L. Kwok, P. Deng, N. Dinstl, H.L. Sun, W. Xu, P.Peng and J. Doyon. 2005. CHINET: a Chinese Name Finder System for Document Triage. In *Proceedings of 2005 International Conference on Intelligence Analysis.*

[17] Chun-Jen Lee and Jason S. Chang. 2003. Acquisition of English-Chinese Transliterated Word Pairs from Parallel-Aligned Texts Using a Statistical Machine Transliteration Model. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 96–103.

[18] Haizhou Li, Min Zhang, and Jian Su. 2004. A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 159–166.

[19] Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007. Semantic Transliteration of Personal Names. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 120–127.

[20] Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 19th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pages 768–774.

[21] Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Machine Transliteration using Target-Language Grapheme and Phoneme: Multi-engine Transliteration Approach. In *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP*, pages 36–39.

[22] Yan Qu and Gregory Grefenstette. 2004. Finding Ideographic Representations of Japanese Names Written in Latin Script via Identification and Corpus Validation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 183–190.

[23] Satoshi Sekine and Yoshio Eriguchi. 2000. Japanese Named Entity Extraction Evaluation: Analysis of Results In *Proceedings of the 18th conference on Computational linguistics*, pages 1106–1110.

[24] Bonnie Glover Stalls and Kevin Knight. 1998. Translating Names and Technical Terms in Arabic Text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, pages 34–41.

[25] Peter D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning*, pages 419–502.

[26] Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, pages 57–64.

[27] Ellen M. Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323.

[28] Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic English-Chinese Name Transliteration for Development of Multilingual Resources. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 1352–1356.

[29] LiLi Xu, Atsushi Fujii, and Tetsuya Ishikawa. 2006. Modeling Impression in Probabilistic Transliteration into Chinese. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 242–249.

[30] Xiaojin Zhu and Ronald Rosenfeld. 2001. Improving Trigram Language Modeling with the World Wide Web. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 533–536.

[31] 佐々木 靖弘, 佐藤 理史, 宇津呂武仁. 2006. 関連用語収集問題とその解法. 自然言語処理, 13(3), pages 151–175.

[32] 柴田 容子, 藤井 敦, 石川 徹也, 2006. “頭字語ネーミングの計算モデル”. 言語処理学会第 12 回年次大会発表論文集, pages 755–758.

[33] 鈴木 義昭, 王 文, 日本語から引ける中国語の外来語辞典, 東京堂出版, 2002.

[34] 宮田 一郎　編訳，新華字典，第 10 版，日本語版，光生館，2005.

[35] 新华字典電子版,v1.0.

[36] 新华通讯社译名室 编, 世界人名翻译大辞典, 中国对外翻译出版公司, 2007.

# 謝辞

本研究を遂行し学位論文をまとめるに当たり，研究の細部にわたり丁寧にご指導と激励をくださいました藤井敦准教授（東京工業大学）に心から感謝の意を表します．そして，日々の研究だけでなく多くのご指導をして頂いた石川徹也教授（東京大学），石塚英弘教授（筑波大学），田中和世教授（筑波大学）に深く感謝致します．また，論文の審査委員になって頂いた佐藤哲司教授（筑波大学），杉本重雄教授（筑波大学），山本幹雄教授（筑波大学）にも深くお礼申し上げます．

そして，博士課程後期に進学前から現在にわたり温かく見守って頂くとともに，多くのご支援ご指導を賜りました高柳敏子名誉教授（獨協大学）には深く感謝しております．

最後に，どのような状況においても辛抱強く応援してくれた父，母，妹家族に心から感謝します．そして，いつも心の支えになってくれた妻と息子にも深く感謝します．

# 著者の主要論文

### 査読付き論文誌

- 黄 海湘, 藤井 敦. 中国語への翻字における関連語抽出の応用. **自然言語処理**, Vol.17, No.2, pp.3-24, Apr. 2010.（本論文中 3.1，3.2，3.6 節と対応する）
- 黄 海湘, 藤井 敦, 石川 徹也. 中国語への翻字における確率的な漢字選択手法. **電子情報通信学会論文誌**, Vol.J90-D, No.10, pp.2914-2923, Oct. 2007.（本論文中 3.1〜3.5 節と対応する）

### 国際会議発表論文

- HaiXiang Huang and Atsushi Fujii. Effects of Related Term Extraction in Transliteration into Chinese. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pp.643-648, Jan. 2008. (Poster)

### 口頭発表

- 黄 海湘, 藤井 敦. 意味訳型翻字システムにおけるユーザフィードバックの応用. 言語処理学会第 16 回年次大会発表論文集, pp.629-632, Mar. 2010.
- 黄 海湘, 藤井 敦. 中国語への翻字における関連語抽出の効果. 情報処理学会研究報告, 2007-NL-177, pp.9-16, Jan. 2007.
- 黄 海湘, 藤井 敦, 石川 徹也. 中国語への翻字における漢字選択の手法. 電子情報通信学会技術研究報告, NLC2006, pp.7-12, Jul. 2006.