社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－178 (6)
2007／3／28

シソーラスを用いた複層クラスn-gramモデル

市丸 夏樹

鳥取環境大学 環境情報学部 情報システム学科

ichimaru@kankyo-u.ac.jp

概要: 本稿では，携帯電話等における子音入力方式の正解率向上と打鍵数の削減のため，n-gramに名詞シソーラスを組み合わせた強力な優先付け機構を提案する．提案手法では，単語n-gramの単語間の条件付き連接確率が，シソーラス中の様々な抽象度の先行クラス間の連接を前提条件として計算されるものと考える．一般的には，より具体的なクラス列に後接し，かつ生起頻度が高い単語候補が優先される．その結果，文節中に派生語や複合語が含まれる場合にも正しい漢字候補を自動的に選び出すことができるようになるものと考えられる．派生語変換の実験では，マルチタップ入力と品詞bi-gramの場合と比較して，入力と変換を合わせた総合的な打鍵数を約56%削減できることを示した．

Multi-layer Class N-gram Model that Uses Noun Thesaurus

Natsuki Ichimaru

Department of Information Systems,
Tottori University of Environmental Studies

ichimaru@kankyo-u.ac.jp

Abstract: In this paper, we present a strong disambiguation method that combines class n-gram and a noun thesaurus, to improve accuracy and keystroke savings of consonant-kanji conversion method that is used on mobile phone. We presume that the preceding class sequence of a word can reside on any layers in a thesaurus. In general, a high frequency word that succeeds a specific class sequence has high priority. It is supposed that our method can automatically choose correct kanji candidates, even from a phrase that contains derivative word or compound word. As an experimental result of derivative conversion, we confirmed that the keystroke savings rate was about 56%.

1 はじめに

携帯電話上の小さな画面と少数のキーを用いた日本語入力方式として現在広く用いられているマルチタップ入力は，同一キーの連打を伴った手間の掛かるものである．ビジネス用途での本格的な利用を可能にするためには，日本語入力方式の効率化と打鍵数の削減が望まれる．

少数キーによる言語入力方式の開発は世界に共通する課題である．特に欧米では，携帯電話の1つのキーに3つ程度のアルファベットを割り当てて曖昧な形で入力し，単語辞書によって曖昧性を絞り込む入力方式T9 input method[8]が広く普及している．しかし，日本語は元来他の言語と比較して同音異義の曖昧性が高いため，さらにキー入力時の曖昧性を高めた場合，変換候補数が大幅に増加してしまう．この問題に対処するためには，従来の仮名漢字変換手法等よりも遥かに強力な優先付けの機構が必要である．

本稿では，シソーラス中の複数の層上でのクラス間の連接確率を学習し，より具体的なクラス列を優先するよう優先付けを施し，単語n-gramの1単語候補ごとに様々な抽象度の先行クラス列を動的に選択するモデルを提案する．これによって，派生語や複合語を含んだ文節変換の正解率の向上を目指す．

表 1: 派生語 1 語あたりの入力文字数の比較

入力デバイス	入力形式	平均文字数/語
フルキーボード	仮名	5.70
〃	ローマ字入力	9.34
10 キー	子音入力	6.41
〃	マルチタップ	17.09

2 日本語の子音入力

2.1 子音入力方式とは

携帯電話上の日本語入力方式としては，一つのキーに複数の文字を割り当て，同一キーの連打を行うマルチタップ方式が広く用いられている．その一方で，打鍵数を削減するために連打を除去し，シングルタップで入力する方式も開発されており，子音入力方式と呼ばれている．子音入力を用いた日本語入力システムとしては，米 Tegic Communications 社の T9(日本語版)[8] や，NTT の「あんないジョーズ」[12]，田中らによる TouchMeKey[14] 等が存在する．特に T9 は既に幾つかのメーカーの市販の携帯電話に搭載されている．

子音入力方式とは，日本語のローマ字表記の母音を省略し子音のみを打鍵する入力方式である．キー配列はマルチタップ方式と同様であるが，1 つのキーに最大 5 つ割り当てられた仮名文字のどれを入力する時も，1 文字につき 1 打のみ打鍵する．ただし，ア行の仮名は'a'，長音記号は'w'，濁点と半濁点は記号'*' または'#' で代用し，拗音には'y' を付加して入力する．

ここで例を挙げる．携帯電話のメールの画面で「よろしく」と入力することを考えると，マルチタップ方式では，“yyyrrrrrsskkk” と 13 打タイプすることになる．一方，子音入力では連打が不要であるため “yrsk” の 4 打で済む．次に，入力方式ごとの平均入力文字数を派生語データ 15,000 語について調査した結果を表 1 に示す．このように子音入力方式を用いれば，マルチタップやローマ字入力と比べて仮名表記を入力する際の打鍵数を削減することが可能である．なお，ポケットベルの 2 タップ入力の場合にも，ローマ字入力とほぼ同様の打鍵数が必要である．

2.2 子音入力の問題点

子音入力では，キー割り当てに曖昧性があり，単語の仮名表記が確定しない曖昧な形で入力される．そのため従来の子音入力システムでは，単語辞書を用いて単語として成り立つ変換候補を選び出す方法がとられてきた．しかし，単語辞書による方法では次のような問題点が生じる場合がある．

- 未登録語の入力が難しい．
- ユーザの意図しない変換候補が大量に提示される．

前者について，曖昧性の解消に単語辞書を利用する関係上，そもそも単語辞書に登録されていない語は変換候補として提示されない．派生語や複合語といった膨大な数の語を全て予め辞書に登録しておくことは不可能であるから，何らかの形で新しい語を生成する機構が必要であると思われる．

後者については，特に派生語を品詞の bi-gram で解析すると，顕著に曖昧性が表れる．例えば「kkak」と打鍵した場合に名詞と接尾語の全ての組み合わせを考慮すると，「機械化」「機構化」「加工機」「下降機」など 5,500 通りを越える変換候補が得られる．子音表記に対応する多くの変換候補の中から尤もらしい単語候補を選び出すためには，従来よりも強力な優先付けを実現する機構が望まれる．

丸山ら [10] は様々な長さの単語 n-gram を組み合わせた PPM モデルを用いて子音入力の曖昧性を絞り込むことによって，打鍵数を半減できることを示している．しかし，派生語や複合語の変換の正解率をさらに向上するためには，構成要素となる名詞-接尾語間や名詞-名詞間の連接を捉えることが必要である．膨大な名詞間の連接傾向を効率的に捉えるためには，n を長くする方向での改良のみならず，名詞を様々なクラスに分け，学習サンプルが少なくても単語の意味的な区別が効くようにする方向での改良が必要であると思われる．

3 クラス n-gram モデルとその問題点

3.1 単層クラスモデル

似た意味を持つ単語は似た連接傾向を持つ可能性が高いものと考えられる．そこで単語を幾つかのクラスに分類しクラス間の連接確率を用いることによって，少数キーでの言語入力時の単語の曖昧性をある程度絞りこむことができると考えられる．

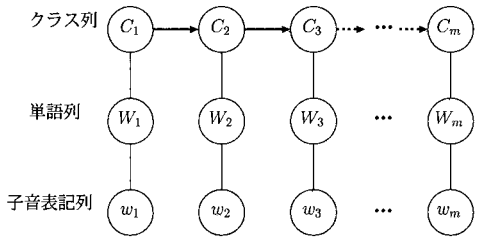

図 1: 単層クラスモデル

以降簡単のため，単語列あるいはクラス列 $A_i, A_{i+1}, A_{i+2}, \ldots, A_j$ を A_i^j と略記する．

クラス n-gram モデル [1] では，m 個の単語の列 W_1^m からなる文の生起確率 $P(W_1^m)$ は式 (1) のように表される．右辺は，先行する $n-1$ 個の部分クラス列が生起した場合に i 番目のクラス C_i が生起する条件付き確率と，クラス C_i が生起した場合に W_i が生起する条件付き確率との積となっている．従来のクラスモデルでは，1 文全体に渡る一連のクラス列の存在が仮定される (図 1)．本稿では，このようにある 1 層の意味分類上のクラス列のみを使用するクラスモデルを単層クラスモデルと呼ぶものとする．

$$P(W_1^m) \triangleq \prod_{i=1}^{m} P(C_i \mid C_{i-(n-1)}^{i-1})\, P(W_i \mid C_i) \quad (1)$$

3.2 単層クラスモデルの問題点

単層クラスモデルを仮名漢字変換や子音-漢字変換に用いた場合，次のような問題点が生じることがこれまでの我々のシソーラスを用いた派生語処理の研究過程において明らかになっている [5, 7]．

- 粗い意味分類を用いた場合には最尤解の正解率があまり上がらない．
- 意味分類を過度に細分化した場合には解候補数が減少し，細分化前には提示できていた正解を消失する場合がある．
- 正解率を最大にする最適な意味カテゴリ数は学習サンプル数の増加に伴って変動するが，1, 3, 10 位解など順位によって最適な意味カテゴリ数が異なるため，全ての順位について同時に最適な正解率を得ることができない．

以上のことから，単層クラスモデルは子音-漢字変換への応用には適さないものと思われる．

4 提案手法

4.1 シソーラスを用いたクラスの複層化

そこで本研究では，人手によって作成されたシソーラス (EDR 概念辞書) を利用してクラスを複層化する．学習時には，部分単語列を各階層上に汎化した様々な抽象度を持つクラス列を作成しておく．このクラス列を連接ルールと呼ぶ．そして解析時には，抽象的なルールよりも具体的なルールの方が優先されるようにし，各単語候補の生起確率を予測するために最も適した抽象度を持つ階層上の連接ルールを動的に選択して用いる．

本手法では名詞のみを汎化するものとする．この理由は次のようなものである．第一に，名詞は品詞の中で単語数が最も多く，学習されにくい．特に派生語や複合語は日々新しく生まれているため，学習サンプルから生起頻度を予め得ておくことが難しい．第二に，その他の品詞，特に付属語は 1 単語 1 品詞と考えた方がよいほど用法がまちまちであるため，汎化することが適当でない場合が多いものと考えられる．第三に，もし複数の品詞を汎化した場合はシソーラスを品詞ごとに複数用意する必要がある．以上のことから，今回は名詞のみを汎化し，その他の品詞は汎化せず単語の形のままで用いるものとした．

また，本手法では n-gram の連接ルール数を削減するために，次のような方法で汎化を行う．まず，平川らの均等確率法 [4] を用いて，シソーラス中のノードを 3〜7 層程度の複数の層状に分ける．次に，藤井らの複合語翻訳 [3] のためのモデルと同様に，学習したい単語列に含まれる全ての名詞の汎化のレベルを同期させることによって，組み合わせ爆発を回避する．以上の工夫により，汎化された連接ルールの数を取り扱い可能な大きさに抑え込む．

本手法による解析時には，単語の n-gram の条件付き確率を求める際に個々の単語ごとに個別に先行クラス列を選択する．ここには従来の単層クラスモデルにあるような文全体に渡って一貫したクラス列は存在しない．単語の条件付き生起確率は，条件部の先行クラス列の抽象度や，汎化の源となった単語列の生起確率に依存して決まる．連接ルールは先行クラス列の抽象度が低いほど，また，汎化の源となった単語列の生起確率が高いほど，優先されることにな

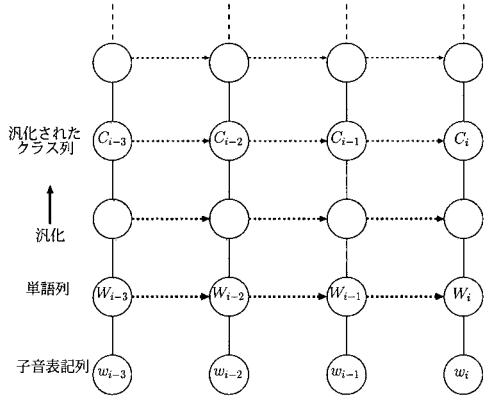

図 2: 複層クラスモデルにおける単語列の汎化

る．このように様々な抽象度を持つクラス列を単語ごとに個別に使用することによって，単語の n-gram の条件付き確率がきめ細かく求められるものと考えられる．

以上のようにクラスを複層化することによって，第 3.1 節で述べた単層クラスモデルの問題点はほぼ解決されるものと予想される．

4.2 複層クラス n-gram モデル

単語の n-gram モデルでは，m 個の単語の列 W_1^m からなる文の生起確率 $P(W_1^m)$ を式 (2) のように表す．i 番目の単語 W_i の生起確率は先行する $n-1$ 語の部分単語列が生起した場合の条件付き確率として式 (3) のように求められる．

$$P(W_1^m) \triangleq \prod_{i=1}^{m} P(W_i \mid W_{i-(n-1)}^{i-1}) \tag{2}$$

$$P(W_i \mid W_{i-(n-1)}^{i-1}) = \frac{P(W_{i-(n-1)}^{i})}{P(W_{i-(n-1)}^{i-1})} \tag{3}$$

提案モデルでは式 (3) の右辺の分子に表れる長さ n の単語列 $W_{i-(n-1)}^{i}$ の生起確率 $P(W_{i-(n-1)}^{i})$ が，式 (4) のように表されるものと仮定する．式 (4) の右辺は，汎化されたクラス列 $C_{i-(n-1)}^{i}$ の生起確率 $P(C_{i-(n-1)}^{i})$ と各クラス C_k から各単語 W_k を導出する条件付き確率 $P(W_k \mid C_k)$ の積となっている．$gen(W_{i-(n-1)}^{i})$ は，単語列 $W_{i-(n-1)}^{i}$ を汎化して得られるクラス列の集合を表す．式 (3) の右辺の分母についても，単語列の長さを $n-1$ とすれば同様に求められる．

$$P(W_{i-(n-1)}^{i}) \triangleq \sum_{C_{i-(n-1)}^{i} \in gen(W_{i-(n-1)}^{i})} P(C_{i-(n-1)}^{i}) \prod_{k=i-(n-1)}^{i} P(W_k \mid C_k) \tag{4}$$

EDR 概念体系のように単語が所属する階層が一定でなく，抽象度の高い単語が中間ノードにぶら下がる形のシソーラスでは，クラス列を用いて抽象度の高い単語間の連接を表すことを考えると，クラス列から単語列が得られる確率は，シソーラスを辿るパス長が長くなるほど減衰することが望ましい．そこで，クラス列 C_k から W_k を導出する確率 $P(W_k \mid C_k)$ を導出パス上の親子ノード間の遷移確率 $P(C' \mid C)$ の積で表す (式 (5))．このとき単語に多重継承が存在する場合は複数のパスが対応する場合がある．$path(C_k \overset{*}{\Rightarrow} W_k)$ はクラス C_k から単語 W_k までの導出パスの集合を表す．

$$P(W_k \mid C_k) \triangleq \sum_{t \in path(C_k \overset{*}{\Rightarrow} W_k)} \prod_{C \to C' \in t} P(C' \mid C) \tag{5}$$

4.3 テキストコーパスからの学習

提案モデルの各確率パラメータは，テキストコーパスから収集した大量の学習データを用いて最尤推定法により学習する．学習サンプル集合 S 中の単語の頻度総和を $|S|$ とする．

クラス列 $C_{i-(n-1)}^{i}$ の生起確率 $P(C_{i-(n-1)}^{i})$ は，$C_{i-(n-1)}^{i}$ の生起頻度 $f(C_{i-(n-1)}^{i})$ から式 (6) のように求められる．学習サンプル単語列 $W_{i-(n-1)}^{i}$ を汎化して得られるクラス列 $C_{i-(n-1)}^{i}$ は抽象度が高くなるほど信頼性が下がって行くため，より具体的なものを優先するように重み付けすることが重要である．そこで，クラス列 $C_{i-(n-1)}^{i}$ の生起頻度 $f(C_{i-(n-1)}^{i})$ を求める際には，元となる学習サンプル単語列 $W_{i-(n-1)}^{i}$ の生起頻度 $f(W_{i-(n-1)}^{i})$ に重み $weight(C_{i-(n-1)}^{i})$ をかけて正規化する (式 (7))．また，クラス列は抽象度が高いほど多くの用例から学習された頻度が集積しやすい．1 つのクラスに対する頻度の集積の度合は概ね子孫単語数に比例するものと考えられる．その

ため，これを丁度打ち消すように，各クラス C_j の子孫単語数 $|C_j\downarrow|$ の逆数の積をクラス列 $C^i_{i-(n-1)}$ に対する重み $weight(C^i_{i-(n-1)})$ として用いる (式 (8))．

$$P(C^i_{i-(n-1)}) \triangleq \frac{f(C^i_{i-(n-1)})}{|S|} \tag{6}$$

$$f(C^i_{i-(n-1)}) \triangleq \sum_{W^i_{i-(n-1)}\in S} f(W^i_{i-(n-1)}) \cdot \frac{weight(C^i_{i-(n-1)})}{\sum_{C'^i_{i-(n-1)}\in gen(W^i_{i-(n-1)})} weight(C'^i_{i-(n-1)})} \tag{7}$$

$$weight(C^i_{i-(n-1)}) \triangleq \prod_{j=i-(n-1)}^{i} \frac{1}{|C_j\downarrow|} \tag{8}$$

親クラス C の汎化サンプル中の生起頻度を $f(C)$，親クラス C から子クラス C' への遷移の頻度を $f(C\rightarrow C')$ とすると，親クラス C から子クラス C' への遷移確率 $P(C'\mid C)$ は式 (9) のように表されると考えられる．しかし，式 (9) をこのまま用いた場合学習サンプルの汎化時に使用されなかった親子クラス間の遷移確率が 0 になってしまう．そのため，辞書中の全ての単語が微小な頻度で出現しているものと仮定して補間を行う必要がある．そこでここでは，シソーラスの上界から各単語までのパス上のクラス遷移を微小頻度 α で学習させるものとする．これを考慮すると，親子クラスの遷移確率 $P(C'\mid C)$ は，式 (10) のように加算法 [9] に似た式で表される．ただし，$|C\uparrow|$，$|C\downarrow|$ は，それぞれシソーラス中のクラス C の先祖方向のパスの数と，子孫である単語の数を表す．

$$P(C'\mid C) \triangleq \frac{f(C\rightarrow C')}{f(C)} \tag{9}$$

$$P(C'\mid C) \simeq \frac{f(C\rightarrow C')+\alpha|C\uparrow||C'\downarrow|}{f(C)+\alpha|C\uparrow||C\downarrow|} \tag{10}$$

この学習法には，計算が簡単で，学習データ量を増加させることが容易であるという利点がある．

5 派生語変換の実験結果

現在，文節単位の変換実験の準備を進めている段階であるため，ここでは，これまでに行った子音入力派生語変換の実験結果について述べる．

表 2: 変換回数の比較

入力形式	変換手法	平均変換回数/語
マルチタップ	品詞 bi-gram	1.75
ローマ字	4,067 分類 1 層	1.20
仮名	68,048 分類 1 層	1.19
	提案手法	1.13
子音 (T9)	品詞 bi-gram	5.52
	4,067 分類 1 層	2.39
	68,048 分類 1 層	2.10
	提案手法	1.83

表 3: 入力と変換を合算した総合的な打鍵数の比較

入力形式	変換手法	平均打鍵数/語	KSR
仮名	提案手法	6.83	63.8%
子音 (T9)	〃	8.24	56.2%
ローマ字	〃	10.46	44.4%
マルチタップ	〃	18.21	3.3%
ローマ字	品詞 bi-gram	11.08	41.1%
子音 (T9)	〃	11.93	36.3%
マルチタップ	〃	18.83	-

学習データとしては，EDR[2] 日本語単語辞書 (JWD)，日本語コーパス (JCO)，新聞記事 6 年分から抽出したのべ約 360 万語の派生語サンプルを使用した．シソーラスとしては，EDR 概念体系辞書 (CP) の全ノードを使用した．学習サンプルはシソーラス上の全ての階層を用いて無段階に汎化した．試験データとしては，人手チェック済みの形態素データ [13] より抽出した派生語正例のべ 14,823 語を使用した．これらのデータを用いて，試験データのローマ字表記から求めた子音文字列を入力とした子音-漢字変換実験を行い，単語列の生起確率の降順に提示される変換候補の中に正解が表れる順位，すなわち正解が表れるまでの変換回数を計測した．計算量の増大を防ぐため，解析時には子音表記から仮名表記への変換は行わず，見出し語を子音表記とする辞書を用いて子音表記から漢字表記へ直接変換した．

各入力方式で表現した派生語を様々な手法で変換した場合の変換回数を表 2 に示す．変換手法としては，品詞 bi-gram，名詞の分類に 4,067 分類または 68,048 分類の 1 層のみを用いた単層クラスモデル，提案手法の 4 つの手法を比較した．この結果による

と，子音入力を変換した場合は，仮名表記が確定するマルチタップ入力，ローマ字入力，仮名入力の場合に比べ確かに変換回数が増加した．しかし，提案手法を用いて変換した場合には，平均的には 1.83 回程度までに抑えられた．提案手法は変換候補の絞りこみに有効に機能しているものと思われる．

次に，各入力方式を用いて提案手法と品詞の bi-gram で変換した場合の入力と変換を合算した総合的な打鍵数を表 3 に示す．Keystroke Savings Rate(KSR)[11] はマルチタップ入力と品詞 bi-gram による変換をベースラインとした場合に，各入力方式と提案手法による変換によって打鍵数が削減された割合を示している．提案手法の KSR は，マルチタップ入力時には 3.3%，子音入力方式時には 56.2%であった．提案手法はマルチタップ入力にはあまり有効ではないが，子音入力方式には有効であると思われる．

また，品詞の bi-gram を用いて変換した場合は，子音入力方式よりローマ字入力の方が高い KSR が得られた．しかし提案手法を用いて変換した場合は，逆に子音入力方式の方がローマ字入力より高い KSR が得られた．子音入力方式は，フルキーボードを用いた仮名入力には及ばないものの，提案手法などの曖昧性の高い変換手法と併用すれば，フルキーボードを用いたローマ字入力の場合よりも打鍵数において優れていると言える．

6 おわりに

本稿では，シソーラスを用いて連続する単語列の名詞部分を多段階に汎化し，単語の bi-gram の条件付き確率の計算に用いるモデルを提案した．また，このモデルが派生語の子音入力時の打鍵数の削減に有効であることを実験によって示した．このモデルは，携帯電話や PDA などの携帯端末における少数キーを用いる日本語入力方式の効率化に役立つものと期待される．

今後の課題としては，大量データを用いて文節単位あるいは 1 文単位の変換実験を行うこと，提案モデルを拡張し予測変換に適用できるようにすること，ユーザ個人に適応した学習機能を備えること，携帯電話上で動作するクライアントを実装すること，などが挙げられる．

参考文献

[1] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[2] EDR. EDR 電子化辞書. 日本電子化辞書研究所, 第 2 版, 1999.

[3] 藤井敦, 石川徹也. 技術文書を対象とした言語横断情報検索のための複合語翻訳. 情報処理学会論文誌, Vol. 41, No. 4, pp. 1038–1045, 2000.

[4] 平川秀樹, 木村和広. 概念体系を用いた概念抽象化手法と語義推定におけるその有効性の評価. 情報処理学会論文誌, Vol. 44, No. 2, pp. 421–432, 2003.

[5] 市丸夏樹, 中村貞吾, 日高達. 汎化用例とシソーラスを用いた派生語の仮名漢字変換の特性. 自然言語処理, Vol. 12, No. 2, pp. 189–207, 2005.

[6] 市丸夏樹. シソーラスによる意味処理を用いた派生語の子音入力方式とその効果. 第 5 回情報科学技術フォーラム FIT2006 講演論文集, E-016, 2006.

[7] 市丸夏樹. 用例とシソーラスに基づく派生語処理に関する研究. 博士論文, 九州大学大学院システム情報科学研究府, 2006.

[8] Tegic Communications Inc. T9 home page. http://www.t9.com/, 2000.

[9] 北研二（編）. 確率的言語モデル. 計算と言語 4. 東京大学出版会, 11 月 1999.

[10] 丸山卓久, 田中 (石井) 久美子, 武市正人. PPM 法を用いたかな漢字変換の学習モデル. 情報処理学会研究報告 NL146-2, pp. 9–14, 2001.

[11] Johannes Matiasek and Marco Baroni. Exploiting long distance collocational relations in predictive typing. In *Proceedings of thes Workshop on Language Modeling for Text Entry Methods, Association for Computational Linguistics 10th Conference of The European Chaper EACL2003,*, pp. 1–8, 2003.

[12] NTT 東日本. あんないジョーズ. http://www.ntt-east.co.jp/anjozu/.

[13] RWC. RWC テキストデータベース. メディアドライブ, 第 2 版, 1998.

[14] 田中久美子, 犬塚祐介, 武市正人. 携帯電話における日本語入力–子音だけで日本語が入力できるか. 情報処理学会論文誌, Vol. 43, No. 10, pp. 3087–3096, 2002.

[15] Hirofumi Yamamoto, Shuntaro Isogai, and Yoshinori Sagisaka. Multi-class composite n-gram language model for spoken language processing using multiple word clusters. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, P01-1068*, 2001.