情報処理学会論文誌 Vol.54 No.12 2481–2491 (Dec. 2013)

テキストコーパスを用いた漢字詳細読みの自動生成

川崎 博章[1,a)] 笹野 遼平[1,b)] 高村 大也[1,c)] 奥村 学[1,d)]

受付日 2013年2月23日, 採録日 2013年9月13日

概要：スクリーンリーダは，コンピュータ上のテキスト情報を音声で読み上げるソフトウェアアプリケーションであり，視覚障害者がコンピュータを利用して情報にアクセスする際に重要な役割を果たす．スクリーンリーダに搭載されている重要な機能の 1 つに漢字詳細読みの出力がある．多くの漢字には同音異字が存在しており，漢字詳細読みには音声による説明のみでユーザに漢字を正しく想起させることが求められている．たとえば，一般的には “コウニュウ” という読みを持つ単語は “購入” しかないため，“購” という漢字は “コウニュウのコウ” という漢字詳細読みにより想起することが可能である．一方で，“コウバイ” という読みを持つ単語は “勾配” や “公売” などが存在するため，“コウバイのコウ” という漢字詳細読みから “購” という漢字を想起することは難しい．しかし，このような曖昧性を持つ漢字詳細読みは既存のスクリーンリーダの中にも存在しており，正しい漢字が想起できない要因の 1 つとなっている．また，漢字詳細読みで用いる単語はユーザに慣れ親しんだものであるべきだが，単語の親密度は時間の経過やユーザの背景により変化する．そこで，本論文では，同音異字の情報と単語の親密度を考慮に入れた，コーパスを用いた漢字詳細読みの自動生成法を提案する．さらに漢字想起実験により，提案手法はインタラクティブな要素を取り入れることで生成される漢字詳細読みの長さを既存のスクリーンリーダのものと同程度に抑えていること，および，提案手法により自動生成された漢字詳細読みの性能が既存のスクリーンリーダのものよりも高いことを示す．

キーワード：視覚障害者，スクリーンリーダ，漢字詳細読み

Automatic Generation of Distinctive Explanation for Kanji Using Text Corpus

Hiroaki Kawasaki[1,a)] Ryohei Sasano[1,b)] Hiroya Takamura[1,c)] Manabu Okumura[1,d)]

Received: February 23, 2013, Accepted: September 13, 2013

Abstract: Screen readers are software applications that produce voice output of text on a computer screen. They play an important role when visually impaired people access the information using a computer. One essential function attached to screen readers in the kanji-using area is to generate distinctive explanation for kanji. Most kanji characters have their homophones, and the role of distinctive explanation is to enable users to identify a unique kanji character only by listening to the explanation. For example, “購”(*kou*) can be identified by a distinctive explanation “*kou-nyū* (購入, purchase) *no kou*, because “購入 (purchase)” is the only word that has the reading “*kou-nyū*.” On the other hand, it is difficult to identify a unique kanji “購” by a distinctive explanation “*kou-bai no kou* (購買, purchase),” because there are multiple words with the reading “*kou-bai*,” such as “勾配 (gradient)” and “公売 (public auction).” Such confusing explanations for kanji are sometimes produced by the existing screen readers. In addition, although the words used in distinctive explanations have to be familiar to users, the familiarity of each word changes with the times and users’ backgrounds. In this paper, we propose a corpus-based method for automatically generating distinctive explanations for kanji, in which information about familiarity and homophones of kanji is taken into consideration. Through the kanji-identification experiments, we show that the quality of the explanations generated by the proposed method is higher than that of the manually-crafted distinctive explanations.

Keywords: visually impaired people, screen reader, distinctive explanation for Kanji

1. はじめに

日本語には平仮名，片仮名，漢字の 3 種類の文字がある．平仮名と片仮名は表音文字であり 1 つの音に 1 つの文字が割り当てられているのに対し，漢字は表意文字でありほとんどの漢字には同音異字が存在している．このため音声のみで漢字を特定することは簡単ではない．しかしながら，音声のみで漢字を特定することが必要な場合もある．たとえば，スクリーンリーダはユーザ，特に視覚障害者に音声のみで漢字を特定させる必要があるし，電話越しでの会話の際には音声のみで名前などの情報を交換する必要がある．単純に漢字の読みのみで漢字を説明しようとすると曖昧性が生じることから，このような場合，“コウニュウのコウ（購）” や “サンズイのカワ（河）” などのように，その漢字の読みだけでなくその漢字特有の性質や構成などを利用して漢字を説明することとなる．本論文では，このような音声による漢字の説明を先行研究 [2], [8] にならい “漢字詳細読み” と呼び，その漢字詳細読みの自動生成法を提案する[*1]．

漢字詳細読みによる漢字の説明では，曖昧性を減らすために，説明対象の漢字の音読みや訓読み，構成要素などの様々な特徴が利用される．漢字が独特な読みを持っているならば，その読みは漢字詳細読みの生成に利用することができ，たとえば “桜” のような漢字に対しては “サクラ” という独特な読みを利用できる．また，漢字には偏と旁に分けることができるものもあり，それらを用いて曖昧性なく漢字を説明することもできる．たとえば，“評” という漢字は “言” という偏と “平” という旁に分けることが可能であるので，偏と旁の情報を伝えることで漢字を説明することができる．“購” のような漢字を説明したいときに，“購入” のようにその漢字を含む単語を利用することもできる．このように，対象の漢字を含む単語もまた曖昧性を減らすことができるので，漢字を特定する際に有効である．

いくつかのスクリーンリーダにはすでに漢字詳細読みを出力する機能が搭載されている [2]．しかし，スクリーンリーダに搭載されている漢字詳細読みには，“アヤオリ（綾織）のアヤ” のように，対象の漢字の想起が難しいものが含まれている．“アヤオリ（綾織）のアヤ” により対象の漢字 “綾” を想起することが難しい要因の 1 つは，“綾織” という単語が普段使用されない，つまり親密度の低い単語であることだと考えられる．また，対象の漢字の想起が難しくなる他の要因としては，同音異字語の存在もいくつかの研究で指摘されている [8], [10]．しかしながら，これらの研究は定性的分析にとどまっており，これらの分析を基に自動で漢字詳細読みの生成に取り組んだ研究は行われていない．

本論文では大規模テキストコーパスを利用した漢字詳細読みの自動生成法の提案を行う．提案するシステムは，大規模テキストコーパスから得られる語の親密度と同音異字の出現数に関する情報を用いて漢字詳細読みを自動で生成する．また，自動音声案内をともなうチケット予約システムなどでの利用を想定しインタラクティブな要素を取り入れる．これらのシステムではユーザが音声案内の内容を理解した時点で音声案内を遮ってキー入力を行うといったことが一般的に行われていると考えられることから，提案するシステムは，漢字の想起のために重要な情報を早め出力し，ユーザの要求に応じて必要な情報を追加していくことを想定したシステムとなっている．

本論文の主な貢献は，統計的手法に基づく漢字詳細読みの自動生成法を提案したこと，および，インタラクティブな要素を取り入れることにより既存の人手により生成された詳細読みと同程度の長さで高い漢字想起率を実現したことの 2 つである．語彙や親密度は，日常的に新聞を多く読むのか，Web 上のテキストを多く読むのかなどといったユーザの特性により異なっていると考えられるが，コーパスから自動で詳細読みを自動構築できるようになれば，ユーザの特性に応じたコーパスを用いて漢字詳細読みの自動生成を行うことにより，容易にユーザに適応した漢字詳細読みを生成することが可能となると考えられる．

2. 漢字詳細読み

本章では，まず漢字詳細読みに関する既存研究を紹介し，次に漢字詳細読みの分類と，既存の漢字詳細読みの問題について記述する．

2.1 関連研究

1990 年代のコンピュータの爆発的な普及にともない，漢字詳細読みの音声出力を機能として持つスクリーンリーダは視覚障害者の間で広まった．それに従い，スクリーンリーダにより出力される漢字詳細読みの問題について議論されるようになった．

渡辺ら [10] はスクリーンリーダで使用される漢字詳細読みにおいてユーザに対する親密度の低い表現や同音異字を持つフレーズを利用することの問題を報告したうえで，児童を対象とした単語親密度調査を行い，その結果に基づき漢字詳細読みを人手で作成している．渡辺らは小学 5 年生

[1] 東京工業大学
Tokyo Institute of Technology, Yokohama, Kanagawa 226–8503, Japan
[a)] kawa@lr.pi.titech.ac.jp
[b)] sasano@pi.titech.ac.jp
[c)] takamura@pi.titech.ac.jp
[d)] oku@pi.titech.ac.jp

[*1] 音声のみで漢字を特定させる方法として，“買うの意味のコウバイ（購買）” などのように単語単位で説明する方法も考えられる．しかし，たとえば電話口で自分の名前に含まれる漢字を説明する場合や，ユーザにとって初めて聞く単語の構成漢字を説明する場合などは単語単位で説明するのは難しいと考えられることから，本研究では漢字単位の説明の生成に焦点を当てる．

に対し，小学5年生が習う教育漢字206個を聞かせ，以下の選択肢から1つを選ばせることにより語彙親密度を計測し，すでに習っている単語の場合語彙親密度が高くなる傾向にあることを報告している：

a よく知っている．
b だいたい分かる．
c 知らない．

さらに渡辺らはこの結果に基づき，漢字詳細読みの作成を実際に行っている．この際，他の単語よりも親密度が高く，同音異字を持たない単語を優先し，また小学生に適した漢字詳細読みを生成するために，ネガティブな表現と，“Goldのキン”のような英単語の利用を避けている．小学生による漢字書き取り調査の結果，渡辺らが作成した漢字詳細読みを利用した漢字書取り調査の正解率は既存のものよりも14.1%高いことが示されている．

西田ら [4] は漢字の意味に基づく漢字詳細読み作成法を提案している．たとえば，一般的な漢字詳細読みでは，“情報”という単語は“ジョウネツ（情熱）のジョウ”，“ホウコク（報告）のホウ”のように説明されるが，意味に基づく漢字詳細読みでは，“インフォメーション”，“チョウホウ(諜報)”，“ヒミツジョウホウ（秘密情報)”のように説明される．さらに実験を行い，一般的な漢字詳細読みと意味に基づく漢字詳細読みとの間には正解率に差がないことを示している．しかし，西田らの研究の主な対象単語は類語辞典に現れるものであり，使用可能な単語に制限がある．

大山ら [12] は，人名で使用されている漢字に焦点を当て，日本人の名前の漢字表記を自動的に説明し合成音声で出力する説明文生成の対話システムEXPLANETを提案している．EXPLANETは多くの同音の他の漢字から目的の漢字を明確に区別することが可能で，説明の際には目的の漢字の構成や他の単語を用いている．説明の生成には姓名コーパスを利用し，コーパス中の度数などにより説明順位を決定している．しかしながら，EXPLANETは同音異字語の存在は考慮していない．

2.2 漢字詳細読みの分類

渡辺ら [9] はスクリーンリーダに搭載されている従来の漢字詳細読みの構造の詳細な分類を行っている．“アツリョク（圧力）のアツ”などの，“対象漢字を含む熟語 + 対象の漢字の読み”という構造の漢字詳細読みが最も使用されていて，次に続くのは“サクラ（桜)”などの“対象漢字の読み”という構造である．この分類の情報を基に，漢字詳細読みはその構成から3タイプに分類できる．

タイプ1 対象の漢字を含む単語とその読み．
“コウバイ（購買）のコウ”（“購”）
“ヒョウカ（評価）のカ”（“価”）
“カダイ（課題）のカ”（“課”）

タイプ2 対象の漢字の独特な読み．このタイプは，独特な訓読みを持つ漢字の説明によく用いられる．
“サクラ（桜)”（“桜”）
“フタタビ（再び)，サイ（再)”（“再”）

タイプ3 対象漢字の特徴とその読み．
“サンズイのカワ”（“河”）
“カンスウジのイチ”（“一”）
“タベルアメ”（“飴”）

このうちタイプ1は3章で示すように統計情報を利用することにより想起率の高い漢字詳細読みの生成が可能であり，また，一般で使用されるほとんどの漢字に適用可能であることから，本論文ではタイプ1の漢字詳細読みの自動生成を試みる．

2.3 既存の漢字詳細読みの問題点

すでに指摘したように，既存の漢字詳細読みには対象の漢字を特定することが困難なものがある．ここからはその主な要因について，例を用いて説明する．まず，既存研究 [9] で報告されている主な要因を列挙する．

要因1 “チヨガミのヨ”という漢字詳細読みで用いられている“千代紙”のような低い親密度の単語の存在
要因2 “購買”と“勾配”のような同音異字の存在
要因3 “瑠”のような難しい漢字の存在

要因1と要因2は，漢字詳細読みの作成の際に最適な単語を用いることにより対応できると考えられる．しかし，ユーザが対象の漢字を知らない場合はその漢字の音声による説明だけで未知の漢字を想起することは非常に困難であると考えられる．本論文では，漢字詳細読みによる対象漢字の想起率の向上を目的とし，要因1と要因2に焦点を当てる．

3. 漢字詳細読みの自動生成

3.1 概要

本論文では，2段階で構成される漢字詳細読みの自動生成法を提案する．提案システムの第1段階では“対象の漢字を含む単語とその読み”で構成される漢字詳細読み（2.2節のタイプ1）を1つ出力する．そのうえでユーザが1つの漢字を想起できない場合には，提案システムの第2段階に移行し，インタラクティブに2つ目の別の漢字詳細読みを出力する．

図1 に提案システムの概要を示す．“購”という漢字を提案システムに入力した場合，第1段階が入力漢字“購”に対し“コウバイ（購買）のコウ”という漢字詳細読みを出力し，ユーザに提示する．ユーザは提示された漢字詳細読みから正しい漢字を想起できると考えられる．一方，“科”という漢字を入力した場合，第1段階が入力漢字“科”に対し“カガク（科学）のカ”という漢字詳細読みを出力し，ユーザに提示したとしても，“科学”には“化学”などの同音異字が存在するので，ユーザは1つの漢字を想起するこ

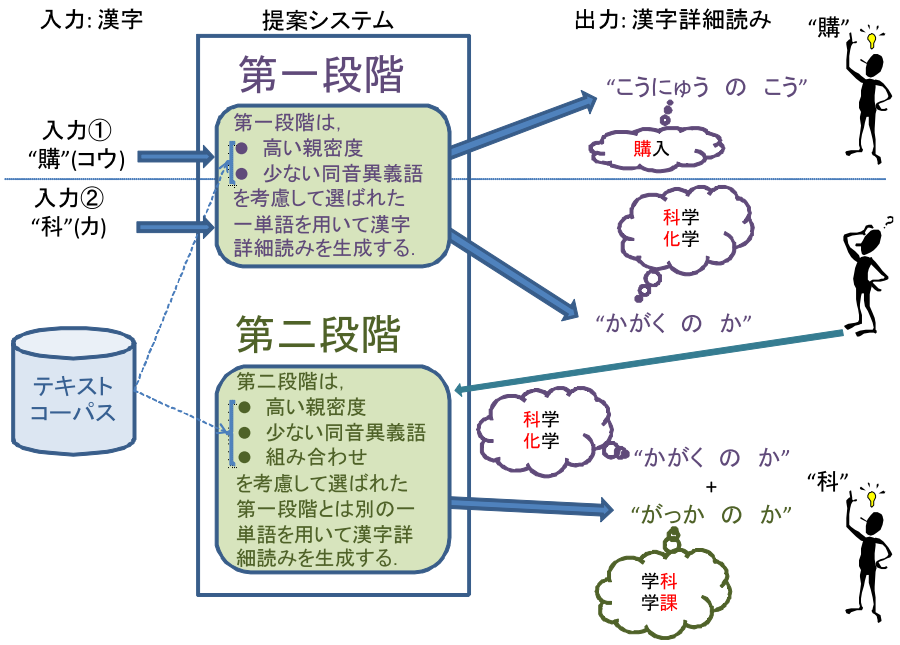

図 1 提案システムの概要
Fig. 1 Overview of our system.

とができない．そこでシステムはユーザからの要求を受け取り，追加の漢字詳細読みを出力する．システムの第 2 段階は “ガッカ（学科）のカ” という漢字詳細読みを出力し，ユーザは 2 つの漢字詳細読みから正しい漢字を想起する．

3.2 第 1 段階の漢字詳細読み生成法

提案手法はまず以下の手順により各漢字に対しタイプ 1 の漢字詳細読みを 1 つ生成する．

i. コーパスから，2 文字以上を含み，かつ対象の漢字を含む単語を抽出する．その際，“高校” などの，対象の漢字の読みを複数持つ単語は，それらを用いて漢字詳細読みを生成しても漢字 1 字だけを想起することができないので，除外する[*2]．

ii. 各単語に対しスコアを以下の式により計算する：

$$score_1(w) \triangleq p(w)^{\alpha} \cdot u_1(w), \qquad (1)$$

ただし，

w ： 候補単語
$p(w)$ ： コーパス中の単語 w の出現確率
$u_1(w)$ ： 単語 w と同じ読みを持つコーパス中の全単語の合計出現頻度に対する，単語 w の出現数の割合
$\alpha > 0$ ： $p(w)$ と $u_1(w)$ の重みを制御するパラメータ

とする．確率 $p(w)$ は

$$p(w) \triangleq \frac{c(w)}{\sum_{w' \in W} c(w')}$$

$c(w)$ ： コーパス中の単語 w の出現頻度
W ： コーパスに出現する単語の集合

とする．割合 $u_1(w)$ は

$$u_1(w) \triangleq \frac{c(w)}{\sum_{w' \in H(w)} c(w')}$$

$H(w)$ ： 単語 w と同じ読みを持つ単語の集合

とする．

iii. 最も高いスコアとなる単語を用いて，漢字詳細読みを生成する．生成の際に，提案システムは選択した単語の読みと，その単語中の対象漢字の読みを利用する．

式 (1) では，$p(w)$ が親密度度合いを，$u_1(w)$ が同音異字の少なさを表し，α はこれらの重みを制御するパラメータである．$\alpha > 1$ であれば親密度度合いを，$\alpha < 1$ であれば同音異字の少なさを重視していることになる．ここで，単語の親密度と出現頻度は厳密には異なる概念であるが，本研究ではユーザが普段，目にするドメインのコーパスから詳細読みを生成することを想定し，出現頻度が高い単語は親密度が高いものとの仮定をおいている．

第 1 段階により “購” という漢字に対し漢字詳細読みを作る手順は次のとおりである．

i. コーパスから “購読” や “購入”，“購買” などの，“購” という漢字を含む単語を抽出する．

ii. これらの単語のスコアを計算する．たとえば，$p(x)$ と $u_1(x)$ がそれぞれ

$$p(\text{“購読”}) \simeq 1.2 \times 10^{-5}\ ,\ u_1(\text{“購読”}) \simeq 0.193$$

[*2] “コウコウ（高校）のコウ” という漢字詳細読みからは，少なくとも “高” と “校” という 2 つの漢字が想起されるため．

p(“購入”) $\simeq 1.1 \times 10^{-5}$, u_1(“購入”) $\simeq 0.320$

p(“購買”) $\simeq 3.3 \times 10^{-7}$, u_1(“購買”) $\simeq 0.228$

と計算されたとすると，パラメータ α を 0.1 に設定した場合，上記 3 単語のスコアはそれぞれ以下のように計算される．

$$score_1(\text{“購読”}) = p(\text{“購読”})^{\alpha} \cdot u_1(\text{“購読”}) \simeq 0.062$$
$$score_1(\text{“購入”}) = p(\text{“購入”})^{\alpha} \cdot u_1(\text{“購入”}) \simeq 0.103$$
$$score_1(\text{“購買”}) = p(\text{“購買”})^{\alpha} \cdot u_1(\text{“購買”}) \simeq 0.051$$

iii. “購入” という単語が最も高いスコアを得ている．ここから，提案システムは “コウニュウのコウ” という漢字詳細読みを生成する[*3]．

上記の例では，“購読” という単語が 3 単語のうちで最も高い出現確率を付与されているが，“購読” には “鉱毒” などの同音異字が存在する．したがって，“購読” という単語よりも出現確率は低いが，同音異字の少ない “購入” という単語が使用される．

3.3 第 2 段階の漢字詳細読み生成法

第 1 段階では，対象漢字を含む単語がすべて同音異字を持つ場合，その漢字を特定できる漢字詳細読みを生成することができない．たとえば “科” という漢字の場合，その漢字を含む最も一般的な単語は “科学” や “教科”，“単科” などであるが，“科学” には “化学”，“教科” には “強化”，“単科” には “炭化” や “単価” などの，同音異字が存在するため第 1 段階により生成される漢字詳細読みだけでは不十分である．そこで追加の漢字詳細読みを生成することを考える．

第 1 段階の漢字詳細読みが曖昧であるとユーザが返答した場合，第 2 段階では第 1 段階でユーザに提示した漢字詳細読みと組み合わせることによりユーザに漢字を想起させることが最も期待できる，第 1 段階とは別の漢字詳細読みを生成する．

i. コーパスから，2 文字以上を含み，かつ説明したい漢字を含む単語を抽出する．

ii. 抽出した単語の組すべてにスコアを付ける．スコアは

$$score_2(w_1, w_2) \triangleq score_1(w_1) \cdot score_1(w_2) \cdot u_2(w_1, w_2)^{\beta} \quad (2)$$

として計算する．ただし，

w_1 ： 第 1 段階により選択された単語

w_2 ： 第 1 段階とは別の漢字詳細読みを生成する候補単語

$u_2(w_1, w_2)$ ： 単語の組 w_1, w_2から想起可能な漢字の曖昧性の少なさ

$\beta > 0$ ： 組合せの曖昧性に関するパラメータ

とする．u_2 は

$$u_2(w_1, w_2) \triangleq \frac{\min(c(w_1), c(w_2))}{\sum_{(w_1', w_2') \in C} \min(c(w_1'), c(w_2'))}$$

により計算する．ここで C は，w_1 と w_2 の 2 単語からそれぞれ生成された漢字詳細読みにより想起可能な漢字すべてに対し，それらの漢字の想起に必要な 2 単語の組合せからなる集合である．たとえば，“カガクのカ” という漢字詳細読みに対する想起可能漢字は “科学” から “科” や “化学” から “化” などであり，同様に，“タンカのカ” に対する想起可能漢字は，“炭化” から “化” や “単価” から “価” などである．したがって，“科” という漢字に対し w_1，w_2 としてそれぞれ “科学”，“単科” が選択された場合，$C = \{$(科学，単科), (化学，炭化)$\}$ となる．ユーザが C 中の漢字の各組合せをどのくらい想起するかの度合いは，その組合せに含まれる 2 単語のうち親密度の低い単語の出現頻度 $\min(c(w_1), c(w_2))$ に比例すると考え，対象漢字に対するその頻度を，全候補のその頻度の和で割った値を想起可能な漢字の曖昧性の少なさとして使用している．

iii. 最も高いスコアの単語を選択した後，単語 w_1 と w_2 を用いて漢字詳細読みを生成する．

式 (2) は $score_1(w_1)$ と $score_1(w_2)$ と $u_2(w_1, w_2)^{\beta}$ の積からなり，$u_2(w_1, w_2)^{\beta}$ は 2 単語を用いたときの曖昧性の少なさを表している．たとえば，“コウドウのコウ” と “コウギのコウ” は，“公道” と “公議” から “公” という漢字と，“講堂” と “講義” から “講” という漢字の少なくとも 2 つの漢字を想起させる可能性があるが，項 $u_2(x, y)^{\beta}$ はこのような曖昧な漢字詳細読みのスコアを下げる．

3.4 第 2 段階の漢字詳細読み生成例

“科” という漢字を用いて，提案システムの第 2 段階による漢字詳細読みの生成例を示す．

第 1 段階

まず第 1 段階では，“科” という漢字を説明するのに最も適した 1 単語を用いて漢字詳細読みを生成する．コーパスから，2 文字以上を含み，かつ “科” という漢字を含む単語を抽出する．例として，“科学” と “単科”，“学科” の 3 単語が抽出され，$score_1(x)$ は以下のように計算されたとする：

$score_1$(“科学”) $\simeq 0.100$,

$score_1$(“単科”) $\simeq 0.050$,

$score_1$(“学科”) $\simeq 0.030$.

この場合は，“科学” という単語が最も高いスコアを得ているので，第 1 段階では “科” という漢字に対し “カガクの

[*3] 単語 w の読みは MeCab で与えられる読みを使用した．

カ” という漢字詳細読みが生成される．もしユーザが漢字を想起できない場合，ユーザはシステムに対し追加の漢字詳細読みの出力を要求し，システムは第 2 段階で 2 つ目の漢字詳細読みを生成する．

第 2 段階

各々の 2 単語の組合せ “科学–単科” と “科学–学科” に対し，$score_2$ を計算する．たとえば，まず $u_2(x)$ を

$$u_2(\text{“科学”}, \text{“単科”}) \simeq 0.50$$
$$u_2(\text{“科学”}, \text{“学科”}) \simeq 1.00$$

と計算する．パラメータ β を 1.0 に設定すると，$score_2$ は

$$\begin{aligned} &score_2(\text{“科学”}, \text{“単科”}) \\ &= score_1(\text{“科学”}) \times score_1(\text{“単科”}) \times u_2(\text{“科学”}, \text{“単科”}) \\ &\simeq 0.0025 \\ &score_2(\text{“科学”}, \text{“学科”}) \\ &= score_1(\text{“科学”}) \times score_1(\text{“学科”}) \times u_2(\text{“科学”}, \text{“学科”}) \\ &\simeq 0.0030 \end{aligned}$$

と計算できる．これらの組合せのうち，“科学–学科” という単語の組合せが最も高いスコアを得ているので，システムは第 2 段階で “科” という漢字に対し，“ガッカのカ” という漢字詳細読みを生成する．

システムの第 2 段階までを用いることにより，“科” という漢字に対し “カガクのカ” と “ガッカのカ” という 2 つの漢字詳細読みを得ることになる．もし第 1 段階におけるスコア上位 2 件の単語を用いて 2 つの漢字詳細読みを生成する場合，用いる単語の組合せは “科学–単科” である．しかし第 2 段階では，最も高いスコアの単語と 3 番目に高いスコアの単語の組合せ，つまり “科学–学科” の 2 単語が選ばれている．これは，“科学–単科” の 2 単語よりそれぞれ生成される漢字詳細読みからは少なくとも “科” と “化” という 2 つの漢字を想起しうるが，“科学–学科” の 2 単語からは “科” だけ想起することができることを反映した結果である．

4. 実験

4.1 実験設定

実験では以下の 3 つのコーパスを用いる．

- Google 日本語 N グラムコーパス（Google コーパス）[6]
- 読売新聞コーパス [11]
- 現代日本語書き言葉均衡コーパス（BCCWJ）[3]

Google コーパスは WEB 上の 200 億文から作られたコーパスであり 2,550 億単語からなる．オープンソースの形態素解析エンジンである MeCab [*4] を用いて単語分割した結果，20 回以上の頻度で出現した単語 1 グラムから 7 グラムとその頻度情報が含まれている．読売新聞コーパスは 1991 年から 2004 年までの読売新聞記事から作られており 4 億単語からなり，BCCWJ は現代の日本語で書かれた 1 億単語からなる．これらの 2 つのコーパスは MeCab により単語分割を行ったうえで使用した．この際，辞書としては IPA コーパスに基づき条件付き確率場でパラメータ推定された IPA 辞書を用いた [5]．

しかし，これらの単語分割基準は単語を細かく切り過ぎる傾向がある．たとえば提案手法では “炒め物” という単語は 1 単語として扱いたいところであるが，IPA 辞書では “炒め” と “物” という 2 つの単語として扱われる．そこで，このような IPA 辞書中では複数の単語として登録されている単語を 1 単語として扱うために，提案手法では補助的に SKK 辞書 [1] も用いた．SKK 辞書中に存在する項目をも 1 単語として扱い，漢字詳細読みで選択されうる単語を追補した．たとえば SKK 辞書には “炒め物” の項目があるので，“炒め物” という単語を漢字詳細読みの候補として扱うことができる．また，予備実験の結果に基づき，パラメータ (α, β) は $(0.1, 1.0)$ に設定した．具体的には，Google コーパス中に出現する単漢字からその出現頻度，上位・中位・下位であるものから 10 個ずつ，計 30 個の漢字を無作為に選出し，Google コーパスを用いそれらの漢字詳細読みを生成した結果を参考にパラメータを設定した．

漢字詳細読みそのものの性能に焦点を当てるため，実験では Google コーパス中に現れる出現頻度上位 2,000 個の漢字[*5]を用い，2.2 節の要因 3 により引き起こされる難しい漢字の存在に起因するエラーをなるべく無視できるようにした．上記 2,000 個の漢字の合計出現頻度は全出現漢字の 99%以上を占めており．実用上の観点から十分であると考えられる．

4.2 3 つのコーパスの比較

3 つのコーパスのうち，どのコーパスが提案手法に最も適しているかを調査するために，3 つのコーパスを用いて第 1 段階の漢字詳細読みを生成し，評価した．コーパスごとの傾向を調査するために各漢字に対し，Google コーパス，読売新聞コーパス，BCCWJ それぞれからの漢字詳細読みと，比較対象のスクリーンリーダ PC-Talker XP に搭載されているものとの，計 4 つの漢字詳細読みを用意し，評価を行った．比較対象のスクリーンリーダとして PC-Talker XP を使用した理由はこのソフトウェアがスクリーンリーダとして最も一般的に使われているソフトウェアの 1 つ [7] であるためである．

PC-Talker XP には様々なタイプの漢字詳細読みが存在するが，提案手法はタイプ 1，すなわち “コウバイ（購買）のコウ” のような漢字詳細読みを生成する．これらの漢字

[*4] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[*5] Google コーパス中の 1 グラムデータに含まれる各単語を文字単位で分割し，さらに各 1 グラムデータの頻度情報を考慮し出現頻度上位 2,000 個の漢字を求めた．

表 1 3 つのコーパスの比較の結果
Table 1 Comparison of identification rates for three corpora.

	Google コーパス	読売新聞コーパス	BCCWJ	PC-Talker XP
a	179	170	185	185
b	15	15	9	9
c	6	15	6	6
IR [%]	89.5	85.0	92.5	92.5

詳細読みを比較するために，出現頻度上位 2,000 個の漢字から PC-Talker XP による漢字詳細読みがタイプ 1 以外の漢字は除外し，残った 615 個の漢字から無作為に 100 個の漢字を評価のために選んだ．

評価は漢字詳細読みを紙に印字し，無作為に混ぜ，8 人の評価者に提示することにより行った．各漢字に対し 4 つの漢字詳細読みが存在するので，各漢字詳細読みを 2 人が評価するように調整し，評価者に提示した．漢字詳細読みの評価では書き取りを行う場合もあるが，実際に漢字詳細読みが使用される場面では漢字を書き取ることができる必要があることは稀で，漢字の想起ができれば十分である場合が多いことから，本実験では漢字の想起の可否により漢字詳細読みの評価を行った．評価者には提示された漢字詳細読みから最もふさわしい漢字 1 字の想起を試みた後に，次の選択肢から 1 つを選択するように伝えた：

a 漢字を想起し，正解だった．
b 漢字を想起したが，不正解だった．
c 漢字を想起しなかった．

各コーパスに対し，正しく漢字が想起された割合である想起率（Identification Rate, IR）を

$$\mathrm{IR} = \frac{n(a)}{n(a)+n(b)+n(c)} \times 100\%$$

により計算する．ここで $n(x)$ は選択肢 x が選ばれた回数である．

表 1 に結果を示す．PC-Talker XP と，BCCWJ を用いた提案手法が最も高い想起率を達成している．手法ごとに各漢字の正解率[*6]を算出し，その優劣を求め，優劣がつく場合，その分布が母比率 1/2 の場合と比べて有意に差があるかを検定した結果，読売新聞コーパスを用いた提案手法と PC-Talker XP との間と，読売新聞コーパスを用いた提案手法と BCCWJ を用いた提案手法との間には 0.05 水準で有意差があることが確認された．3 つのコーパス，つまり Google コーパスと読売新聞コーパスと BCCWJ との中では，BCCWJ が最も高い想起率を達成している．

表 2 に漢字詳細読みの例とその評価を示す．読売新聞コーパスにより生成されるいくつかの漢字詳細読みには，“貌” に対する “外貌” や，“儀” に対する “余儀無い” など，難しい単語が含まれている．“美貌” や “儀式” などのより簡単な単語が存在するにもかかわらず，読売新聞コーパスを用いた場合，これらの単語を含む漢字詳細読みは生成されなかった．この傾向は，読売新聞コーパスにより生成された漢字詳細読みの想起率が低いことの要因の 1 つだと考えられる．

[*6] 評価者が 2 人であるため 0%，50%，100%のいずれかとなる．

4.3 提案手法とスクリーンリーダの比較

続いて提案手法全体と PC-Talker XP の出力を比較する実験を行った．3 つのコーパスの比較の結果，最も想起率の高かった BCCWJ から生成した漢字詳細読みをこの実験では用いる．使用する漢字として出現頻度上位 2,000 個から無作為に 100 個の漢字を抽出した．提案手法ではタイプ 1 の漢字詳細読みのみを出力するが，ここでは PC-Talker XP 全体との比較をする目的で，4.2 節の比較の場合とは異なり PC-Talker XP の出力はタイプ 1 に限らなかった．**表 3** はこの実験で使用した 100 個の漢字のタイプ別の内訳である．タイプ 1 を含むものが 93 個あり，そのうち，タイプ 1 のみから構成される漢字は 32 個であった．タイプ 2 の漢字詳細読みのみから構成される漢字は含まれておらず，タイプ 3 のみから構成される漢字が 7 個含まれていた．

提案手法により生成された漢字詳細読みを 100 個，PC-Talker XP により生成された漢字詳細読みを 100 個，合計 200 個の漢字詳細読みを評価する．ここで評価者 60 人には提案手法により生成された漢字詳細読みと，PC-Talker XP で使用されている漢字詳細読みから 25 個ずつ計 50 個を，各漢字詳細読みの評価人数が等しくなるように無作為に提示した．200 個の漢字詳細読みに対し，60 人が 50 個ずつ評価したことから，各詳細読みは 15 人により評価されたことになる．また，4.2 節における実験と同様に各漢字詳細読みは紙に印字したうえで評価者に示した．

提案手法は 2 段階構成であり，第 1 段階の漢字詳細読みの出力で評価者が漢字を想起しない場合は第 2 段階の漢字詳細読みを出力する．したがって評価の際には，まず提案システムは第 1 段階の漢字詳細読みを評価者に提示し，評価者の要求に応じて第 2 段階の漢字詳細読みを追加で提示する．漢字が想起できた場合，それが正解だったか否かを評価者自身が答え合わせをする．評価者には以下の 5 つの選択肢から適切なものを 1 つ選ぶように伝えた．

- 第 1 段階の漢字詳細読みのみを見て，1 つの漢字を想起した
  - a 正解だった．
  - b 不正解だった．
- 第 2 段階の漢字詳細読みまで見て，1 つの漢字を想起した
  - a’ 正解だった．
  - b’ 不正解だった．
  - c 2 つの漢字詳細読みから漢字を想起することはなかった．

表 2 漢字詳細読みの例とその評価（“$(n/2)$” は，2 人のうち n 人が正解の選択肢を選んだことを意味する）

Table 2 Examples of distinctive explanations and their evaluation. “$(n/2)$” means n subjects out of two chose the positive answer.

漢字	Google コーパス	読売新聞コーパス	BCCWJ	PC-Talker XP
儀	“ソウギのギ” 葬儀（2/2）	“ヨギナイのギ” 余儀無い（0/2）	“ギシキのギ” 儀式（2/2）	“ギシキのギ” 儀式（2/2）
貌	“ビボウのボウ” 美貌（2/2）	“ガイボウのボウ” 外貌（0/2）	“ビボウのボウ” 美貌（2/2）	“ビボウのボウ” 美貌（2/2）
感	“カンジのカン” 感じ（0/2）	“カンジルのカン” 感じる（2/2）	“カンジルのカン” 感じる（2/2）	“カンシンスルのカン” 感心する（2/2）
遥	“ヨウハイのヨウ” 遥拝（0/2）	“ヨウハイのヨウ” 遥拝（0/2）	“ヨウハイのヨウ” 遥拝（0/2）	“ハルカカナタのハル” 遥か彼方（2/2）
餅	“センベイのヘイ” 煎餅（1/2）	“キリモチのモチ” 切餅（1/2）	“センベイのヘイ” 煎餅（2/2）	“モチツキのモチ” 餅つき（1/2）
欄	“クウランのラン” 空欄（2/2）	“ランカンのラン” 欄干（0/2）	“ランカンのラン” 欄干（1/2）	“ランガイのラン” 欄外（2/2）
点	“キョテンのテン” 拠点（1/2）	“キョテンのテン” 拠点（1/2）	“カンテンのテン” 観点（0/2）	“テンスウのテン” 点数（2/2）
輸	“ユニュウのユ” 輸入（2/2）	“ユニュウのユ” 輸入（2/2）	“ユニュウのユ” 輸入（2/2）	“ユシュツスルのユ” 輸出する（2/2）

表 3 PC-Talker XP で使用されている漢字詳細読みの数

Table 3 The number of distinctive explanations for each type used in PC-Talker XP.

タイプ	個数	例
タイプ 1 の詳細読みを含む漢字：93 個		
タイプ 1 のみから構成	32	“カゼイスル（課税する）のカ”（“課”） “ガッキマツ（学期末）のキ”（“期”） “ケンカスル（喧嘩する）のカ”（“嘩”）
タイプ 1 以外のものも含む	61	“コクゴ（国語）のゴ，カタル”（“語”） “キテキ（汽笛）のテキ，フエ”（“笛”）
タイプ 1 の詳細読みを含まない漢字：7 個		
タイプ 2	0	（含まれず）
タイプ 3	7	“ヨロコビヲイミスル カ”（“嘉”） “ツチヲフタツカサネタ ケイ”（“圭”）

表 4 提案システムとスクリーンリーダの比較結果

Table 4 Comparison between outputs of our system and distinctive explanations in screen reader.

	提案システム（BCCWJ）	PC-Talker XP
a	1,181	1,301
b	28	58
a’	163	-
b’	22	-
c	106	141
IR [%]	78.7（1 単語） 89.6（2 単語）	86.7

提案システムの第 1 段階または第 2 段階のどちらかで評価者は漢字を想起することができればよいので，選択肢 a と a’ を正解として想起率（Identification Rate, IR）を

$$\mathrm{IR}_{(2\text{ 単語})} = \frac{n(a)+n(a')}{n(a)+n(b)+n(a')+n(b')+n(c)} \times 100\%$$

として計算した．同様に，提案手法の第 1 段階の評価の際，選択肢 a を正解とし，想起率を

$$\mathrm{IR}_{(1\text{ 単語})} = \frac{n(a)}{n(a)+n(b)+n(a')+n(b')+n(c)} \times 100\%$$

として計算した．

PC-Talker XP の漢字詳細読みを評価する際には，評価者に選択肢

a 曖昧性なく 1 つの漢字を想起し，それは正解だった．

b 曖昧性なく 1 つの漢字を想起したが，それは不正解だった．

c 漢字 1 つを想起しなかった．

から最も適したものを選ぶように伝えた．

スクリーンリーダの評価のために，選択肢 a を正解とし，想起率を

$$\mathrm{IR}_{(\mathrm{SR})} = \frac{n(a)}{n(a)+n(b)+n(c)} \times 100\%$$

として計算した．提案手法の評価では評価者の要求を考慮したインタラクティブな設定を使用しているのに対し，PC-Talker XP の評価ではインタラクティブな設定を使用していないのは，PC-Talker XP で採用されている漢字詳細読みは一部を除き 2 つの詳細読みの組合せで構成されているわけではなく，提案手法のように 2 段階で表示することができないためである．

表 4 に実験結果を示す．4.2 節の評価（92.5%）よりも今回の評価（78.7%）のほうが提案手法の想起率が低くなっているがこの理由は 2 つある．1 つ目の理由は，評価方法

表 5 漢字詳細読みの平均文字数
Table 5 The average length of distinctive explanations.

	第 1 段階	第 2 段階	提案手法とスクリーンリーダの比較で評価者に実際に提示した文字数
BCCWJ	6.80	13.93	8.14
PC-Talker XP	-	-	8.96

の違いによるものであり，評価者が複数の漢字を想起した際に，4.2 節の評価の際には複数の漢字を 1 つに絞ってから答え合わせをするため正解の可能性があったが，今回の評価の際にはかならず不正解となるためである．2 つ目の理由は，使用する漢字が異なることによるものであり，前の評価の際には使用する漢字は PC-Talker XP による出力がタイプ 1 のものに限っていたが，今回の評価の際にはそのような制限はしなかったためである．つまり，タイプ 1 に適していない漢字も，提案手法によりタイプ 1 の形式で漢字詳細読みを生成されている可能性がある．

システム全体で見た場合，提案手法は PC-Talker XP よりも良い性能を示していることが確認できた．4.2 節の実験と同様に手法ごとに精度に差があった漢字の分布に対し 2 項検定を行った結果，BCCWJ を用いた第 2 段階までの提案手法の出力と，PC-Talker XP との間には，0.05 水準の有意差があることが確認できた．BCCWJ に基づく提案手法の漢字詳細読みは最多で 2 つの漢字詳細読みを出力するため，PC-Talker XP のものよりも長くなることが考えられる．しかし，80.6% [*7] の漢字詳細読みは第 1 段階の時点で漢字が想起されているので，評価者は漢字を想起する際にシステムにより出力されうる 2 つの漢字詳細読みを必ずしも見る必要はなく，また，第 1 段階の漢字詳細読みの平均文字数は PC-Talker XP のものよりも短いため，評価者が見た漢字詳細読みの平均文字数は**表 5** に示すように提案手法の方が短くなっている．このことから提案手法により実際にユーザに提示される漢字詳細読みは PC-Talker XP と比べ長くはなっていないことが確認できる [*8]．

システムの第 1 段階の出力の平均文字数が PC-Talker XP のものよりも少ない理由は，システムの第 1 段階の出力はタイプ 1 の漢字詳細読みに限定している一方で，PC-Talker XP にはその制限はなく，タイプ 1 よりも文字数が多いことが予測されるタイプ 3 [*9] の漢字詳細読みや，タイプ 1 を 2 つ組み合わせたような漢字詳細読みの出力をも許容しているからであると考えられる．また，すべての漢字に対し第 2 段階までの漢字詳細読みを出力したと仮定した場合の平均文字数は PC-Talker XP のものよりも多くなる．

4.4 漢字詳細読みの出力例と考察

表 6 に漢字詳細読みの例とその評価を示す．提案手法がスクリーンリーダよりも良い性能を示した漢字として，“課” と “灌” がある．“課” という漢字の場合，15 人の評価者のうち，13 人が提案手法の第 1 段階により生成された漢字詳細読みから漢字を想起しているが，スクリーンリーダの場合は 15 人中 6 人のみが想起に成功している．提案手法の第 1 段階から漢字を想起できなかった残りの 2 人は，第 2 段階の出力まで見ることにより正しい漢字を想起しているので，提案手法は 15 人全員に “課” という漢字を想起させることに成功している．スクリーンリーダのこの漢字に対する低い正解率の原因は，評価者が “課税する” の代わりに “加勢する” を想起したためであると考えられる [*10]．“灌” という漢字の場合，15 人の評価者のうち 1 人だけが提案手法の第 1 段階またはスクリーンリーダの出力から，正しい漢字を想起している．しかしながら，第 2 段階の出力までを見ることにより，7 人の評価者が正しい漢字を想起することに成功していることから，“灌” という漢字を想起することは難しいが，提案する 2 段階手法が効果的に働いていることが確認できたといえる．ただし，第 1 段階で出力された “湯灌” という語はほとんど漢字想起の手がかりとならなかったのに加えて，死という一般的にネガティブなイメージに関連する語であることから，詳細読みとして適当でない可能性がある．説明に使用する単語の極性の考慮は今後の課題の 1 つである．

提案手法がスクリーンリーダよりも悪い性能を示した漢字の例として，“藍” と “圭” がある．“藍” という漢字の場合，提案手法の第 1 段階で漢字を想起した評価者はおらず，第 2 段階まで見ることにより初めて 4 人の評価者に漢字を想起させることができている．これに対しスクリーンリーダ PC-Talker XP は 12 人の評価者に漢字を想起させていてる．これは色という重要な視覚情報である “藍” の特徴を提案手法が利用できなかったためであると考えられる．スクリーンリーダが色に関係する単語を利用している一方で，提案手法は “伽藍” や “藍原” などの単語を用いているが，それらが色に関する情報を持っているとは思えない．このような漢字特有の情報は重要だが，提案手法ではうまく扱えていない．

“圭” という漢字の場合，提案手法では曖昧性のある単語を用いて漢字詳細読みを生成しており，ほとんどの評価

*7 ((1,181 + 28)/1,500) = 80.6%
*8 提案手法の評価にのみインタラクティブな手法を用いているため，提示される詳細読みの長さの比較は必ずしもフェアでないと考えられる．しかし，インタラクティブな提示方法自体が提案手法の 1 要素であり，既存のシステムとの比較という観点からこのような設定を採用した．
*9 タイプ 3 とは “ツチヲフタツカサネタケイ”（圭）などの対象漢字の特徴を利用した説明のタイプであり，一般的に “対象漢字を含む熟語+ 対象の漢字の読み” というタイプ 1 のものよりも文字数が長くなる傾向があると考えられる．
*10 “加勢” は（カセイ）と読むが，“多勢” は（タゼイ）と読むことから，“勢” は（ゼイ）と読むこともあるので，評価者は “加勢する” を（カゼイスル）と読み間違えたと推測される．

表 6 BCCWJ を用いて提案システムが生成した漢字詳細読みと PC-Talker XP による出力の例とその評価（“(n/15)” は，15 人のうち n 人が正解の選択肢を選んだことを意味する）

Table 6 Examples of distinctive explanations generated by the our whole system and distinctive explanations in PC-Talker XP and their evaluation. “(n/15)” means n subjects out of 15 chose a positive answer.

漢字	提案システムの出力		PC-Talker XP
	第 1 段階	第 2 段階	
悟	“カクゴのゴ” 覚悟（9/15）	“サトリのサト” 悟り（14/15）	“カクゴのゴ，サトル” 覚悟，悟る（13/15）
課	“カダイのカ” 課題（13/15）	“カゼイのカ” 課税（15/15）	“カゼイスルのカ” 課税する（6/15）
灌	“ユカンのカン” 湯灌（1/15）	“カンガイヨウスイのカン” 灌漑用水（7/15）	“カンガイスルのカン，ソソグ” （1/15）
藍	“ガランのラン” 伽藍（0/15）	“アイハラのアイ” 藍原（4/15）	“アイイロのアイ” 藍色（12/15）
圭	“ケイジロウのケイ” 圭二郎（2/15）	“ケイイチのケイ” 圭一（4/15）	“ツチヲフタツカサネタ ケイ” （11/15）
嘩	“フウフゲンカのカ” 夫婦喧嘩（4/15）	“オオゲンカのカ” 大喧嘩（5/15）	“ケンカスルのカ” 喧嘩する（1/15）
嘉	“カエイのカ” 嘉永（0/15）	“カデナのカ” 嘉手納（0/15）	“ヨロコビヲイミスル カ” （1/15）

者は漢字を想起できなかった．提案手法により評価者が漢字を想起できずにいる一方で，スクリーンリーダは “土の上に土が乗っている” 様子を評価者に伝えることにより，“圭” という漢字の形を説明していた．これにより提案手法の場合と比べ多くの評価者が正しい漢字の想起に成功しており，提案手法では扱っていないタイプ 3 の漢字詳細読みが効果的に働いていることが確認できた．

提案手法とスクリーンリーダのどちらを用いても想起が難しい漢字もいくつかあり，“嘩” という漢字の場合は提案手法とスクリーンリーダとは両方とも “喧嘩” という単語を用いている．通常は “喧嘩” ではなくカタカナ表記である “ケンカ” を用いることが多く，漢字表記である “喧嘩” を見ることは稀であるので，読みから “嘩” という漢字を想起することは難しい．“嘉” という漢字の場合，漢字そのものが難しいため，漢字詳細読みから想起することは難しいと考えられる．難しい漢字を除外する目的で実験対象の漢字を出現頻度上位 2,000 個のものに限定したが，“嘉” のように一般的に難しいと考えられる漢字もいくつか含まれていた．

5. おわりに

本論文では，漢字の親密度と同音異字の情報を考慮に入れた，テキストコーパスを用いた漢字詳細読みの自動生成法を提案した．評価者による評価の結果，システムにより生成された漢字詳細読みが，スクリーンリーダに搭載されているものよりも性能が良いことを確かめた．

今後の課題としては，発音情報の取り入れ，単語単位の説明読みの生成，中国語への適用，ユーザへの適用が考えられる．たとえば，“橋” と “箸” は同じ読みだが，イントネーションが異なるなどのように，単語のイントネーションは良い漢字詳細読みを作成するときの手がかりとなりうるので，システムにイントネーションの情報を取り入れることが考えられる．また，用途によっては本論文で行ったように漢字単位で説明読みを生成するのではなく，「買うの方のコウバイ」などのように単語単位で説明を行う方が適している場合もあると考えられる．さらに，漢字は日本だけでなく中国でも使用されているので，中国語の漢字詳細読みへも適用できると考えられる．また，法律を学んだ学生は法律用語に慣れ親しんでいるが，医学用語には不慣れであるなど，ユーザの属性は様々であることから，各ユーザに適したコーパスを選ぶことにより想起率を向上させられる可能性が考えられる．

参考文献

[1] SKK-JISYO.L: available from ⟨http://openlab.ring.gr.jp/skk/dic-ja.html⟩.

[2] 藤沼輝好，渡辺恵理子，鈴木沙耶：スクリーンリーダ使用者のための単漢字詳細説明読みガイドライン，第 27 回感覚代行シンポジウム，Vol.27, pp.67–71 (2001).

[3] 国立国語研究所：現代日本語書き言葉均衡コーパス，入手先 ⟨http://www.tokuteicorpus.jp/⟩.

[4] 西田昌史，堀内靖雄，市川 熹：視覚障害者のための意味情報に基づく仮名漢字変換（高齢者支援/肢体不自由者支援/一般），電子情報通信学会技術研究報告：WIT，福祉情報工学，Vol.105, No.186, pp.1–6 (2005).

[5] 工藤 拓，山本 薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告：自然言語処理研究会報告，Vol.2004, No.47, pp.89–96 (2004).

[6] 工藤 拓，賀沢秀人：Web 日本語 N グラム第 1 版，言語

資源協会 (2007).
[7] 渡辺哲也，長岡英司，宮城愛美，南谷和範：視覚障害者のパソコン・インターネット・携帯電話の利用状況調査2007 (2007).
[8] 渡辺哲也，藤沼輝好，渡辺文治，澤田真弓，鎌田一雄：視覚障害者用スクリーンリーダの「詳細読み」に関する検討，電子情報通信学会技術研究報告：HCS, ヒューマンコミュニケーション基礎，Vol.102, No.599, pp.25–28 (2003).
[9] 渡辺哲也，渡辺文治，藤沼輝好，大杉成喜，澤田真弓，鎌田一雄：スクリーンリーダの詳細読みの理解に影響する要因の検討：構成の分類と児童を対象とした漢字想起実験（福祉工学），電子情報通信学会論文誌 D-I，情報・システム，I-情報処理，Vol.88, No.4, pp.891–899 (2005).
[10] 渡辺哲也，渡辺文治，岡田伸一，山口俊光，大杉成喜，澤田真弓：スクリーンリーダの漢字詳細読みに関する研究：試作した詳細読みによる漢字書取り調査（福祉と音声処理，一般），電子情報通信学会技術研究報告：WIT，福祉情報工学，Vol.105, No.373, pp.7–12 (2005).
[11] 読売新聞社：読売新聞記事データ集，入手先 ⟨http://www.nichigai.co.jp/dcs/index2.html⟩.
[12] 大山芳史，浅野久子，高木伸一郎：姓名漢字表記を説明する対話システムの試作と評価，情報処理学会研究報告：SLP，音声言語情報処理，Vol.96, No.123, pp.53–58 (1996).

川崎 博章

1987 年生．2010 年東京工業大学工学部情報工学科卒業．2012 年同大学大学院総合理工学研究科物理情報システム専攻修了．修士（工学）．自然言語処理に興味を持つ．

笹野 遼平 （正会員）

2009 年東京大学大学院情報理工学系研究科博士課程修了．博士（情報理工学）．京都大学大学院情報学研究科特定研究員を経て，2010 年より東京工業大学精密工学研究所助教．自然言語処理，特に照応解析，述語項構造解析の研究に従事．言語処理学会，人工知能学会，ACL 各会員．

高村 大也 （正会員）

1997 年東京大学工学部計数工学科卒業．2000 年同大学大学院工学系研究科計数工学専攻修了（1999 年はオーストリアウィーン工科大学にて研究）．2003 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．博士（工学）．2003～2010 年まで東京工業大学精密工学研究所助教．2006 年にはイリノイ大学にて客員研究員．2010 年より同准教授．計算言語学，自然言語処理を専門とし，特に機械学習の応用に興味を持つ．言語処理学会，人工知能学会，ACL 各会員．

奥村 学 （正会員）

1962 年生．1984 年東京工業大学工学部情報工学科卒業．1989 年同大学大学院博士課程修了．同年東京工業大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，2000 年東京工業大学精密工学研究所助教授，2009 年同教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，テキスト評価分析，テキストマイニングに関する研究に従事．電子情報通信学会，人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．