なんで日本語はこんなに難しいの？

—リアルな日本語学習者コーパスの分析と言語処理の課題—

水本 智也　小町 守
奈良先端科学技術大学院大学情報科学研究科

日本語学習者の増加と多様化

戦後の日本経済の発展とアニメやゲームといったソフトパワーの拡大による世界的な日本語ブームのため，日本語学習者は増加傾向にある．日本語学習者は 2009 年時点で海外 133 の国・地域でおよそ 365 万人となっており，ここ 30 年で 30 倍にまで増加している（国際交流基金調べ）．国際交流基金による地域別の学習者の割合を図 -1 に示す．図 -1 を見て分かるとおり，日本語学習者の 80% 以上がアジア太平洋圏の国である．またアジア圏の学習者に比べると数は少ないが，オーストラリアやアメリカといった英語を主な公用語とする地域でも学習者がいることが分かる．

タイトルにある「なんで日本語はこんなに難しいなの？」は日本語学習者が実際に書いた文である．この文中の“難しいなの”は中国語を母語（第一言語）とする日本語学習者に典型的に表れる誤りである．このような誤りは形容詞には“な”をつけると学習者が覚えてしまっていることから起こる活用の仕方に関する誤りである．日本語は膠着語と呼ばれ，助詞や活用語尾が文法的な意味を担っているが，中国語のように孤立語と呼ばれる言語は語順によって文法的な意味表現をするため，活用の習得が難しいのである．

日本語学習者の誤りにはこのような誤りのほかに，コロケーション誤り，格助詞誤り，スペル誤りなどがある．第二言語学習者の誤りは学習段階によってさまざまである一方，母語の干渉による誤り傾向の違いもあり，第一言語獲得とは異なる問題が存在する．

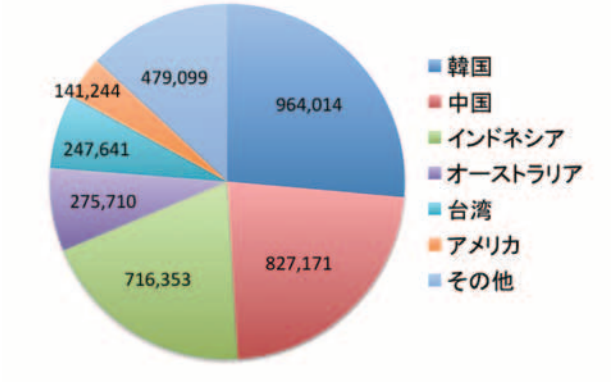

図 -1　国際交流基金が公開している地域別日本語学習者数

本稿では近年急速に発展してきた Web と自然言語処理を組み合わせることで，日本語学習者の語学学習支援を行う取り組みについて解説する．まず自然言語処理技術を用いた日本語学習者支援システムの紹介を行い，次にこれまで日本語学習者支援に用いられてきたテキストデータ（コーパス）について概観し，Web マイニングによる集合知を活用した新しい日本語学習者コーパスについて紹介する．そして現実の日本語学習者コーパスに自然言語処理技術を適用する際の問題点について考察し，最後にこの問題を解決する新しい大規模データを用いたアプローチと今後の課題について述べる．

自然言語処理を用いた日本語学習者支援システム

日本語学習者にとって，日本語を教えてくれる日本語教師の存在は大きい．しかし，特に日本に在住しない日本語学習者にとって，日本語を母語とする日本語教師に教わることは難しい．また，日本語教師

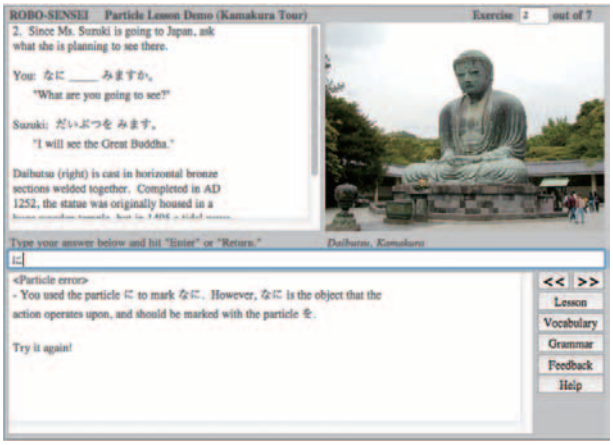

図 -2　Robo-Sensei の鎌倉観光スキットで格助詞のレッスンを受ける

に教わることが可能であっても，常に質問をしたりマンツーマンでインタラクティブに教わったりできる環境が得られることは稀である．そこで，日本語教師がいなくても，自然言語処理の技術を用いて，自学自習の補助を行う支援システムが開発されている．

自学自習を目的とした日本語学習支援システムとしては，サンフランシスコ大学で開発された「BANZAI」そしてその後継である「Robo-Sensei」☆1 システムがある．図 -2 に Robo-Sensei の実行画面を示す．ここでは格助詞「を」の使い方を学ぶため，「なに __ みますか.」という穴埋め形式の作文を日本語学習者に促し，間違った答えを入力した場合，なぜ間違いで，正しくはどうするべきか，といった情報が表示されている．Robo-Sensei は 24 課の教材に従って学習者が日本語の作文をこなし，入力された文の作文間違いに関する詳細なフィードバックを行うことで，学習者支援を行う．Robo-Sensei は辞書・形態素解析・構文解析・誤り検出・フィードバック生成といったモジュールから構成されている．

国内で開発された日本語作文支援システムとしては東工大の「なつめ」☆2 がある．なつめは学習者が指定した名詞について，その名詞と共起頻度が高い動詞を格助詞ごとに表示するシステムである．図 -3 はなつめで「カレー」を検索している例である．共起の強さは棒グラフの長さで示され，「カレーを作る」あるいは「カレーを食べる」という表現がよく使われていることが分かる．また，日本語の母語話者が書いたテキストから例文を取得して表示することができる．

同じグループで日本語の読解支援システム「あすなろ」☆3 も開発されている．あすなろは入力文に対し形態素解析・語義曖昧性解消・構文解析を行い，さまざまな言語情報を表示するシステムである．図 -4 にあすなろで「六本木でカレーを食べたいです」と入力し，形態素解析結果と「食べる」の例文を表示させた例を示す．学習者は読んでいて分からない文があれば，このシステムを用いて読みや品詞を知ることができ，動詞の用例も検索することができる．単語の意味を表示する辞書は日本語のほかに英語・中国語など 6 言語用意されている．

日本語の読解学習支援システムはほかにも「リー

☆1 http://usf.usfca.edu/japanese/RSdemo/preRSfiles/Robo-Sensei.htm
☆2 http://hinoki.ryu.titech.ac.jp/natsume/
☆3 http://hinoki.ryu.titech.ac.jp/asunaro/index-j.php

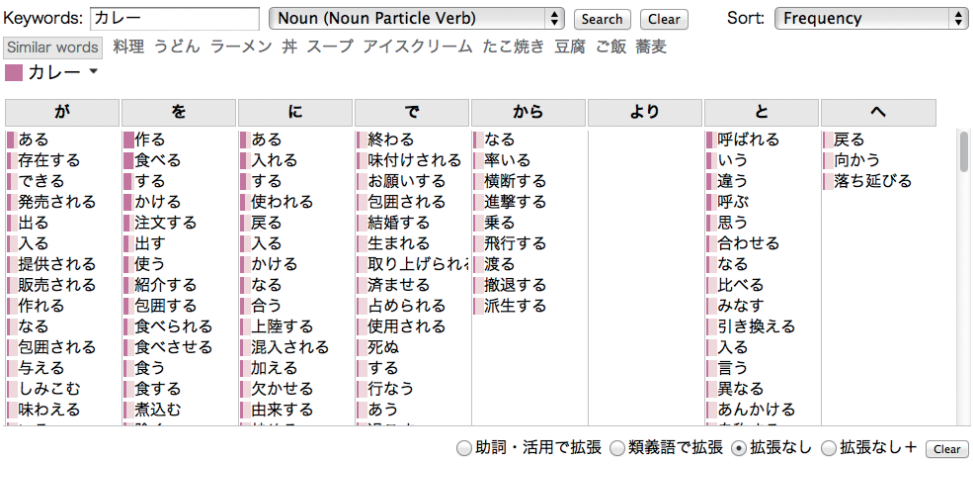

図 -3　なつめで「カレー」を検索し，格助詞ごとに共起する動詞を一覧する

ディングチュウ太」[☆4]が著名である．チュウ太は入力された文章に対し自動で辞書引きを行うことができるシステムで，日本語能力試験の級の情報を利用して文章中の単語のレベルを判定することもできる．現在，英語・ドイツ語・オランダ語がサポートされている．入力文にふりがなをつける「ひらがなめがね」[☆5]とともに，日本語教師が読解教材を作る際にも広く用いられている．

図-4　あすなろで「六本木でカレーを食べたいです」と入力して例文を見る

これまでの日本語学習者コーパス

これらの自然言語処理を用いた日本語学習者支援システムを高度化するためには，日本語学習者がどのような文を書くか，あるいは日本語学習者がどのように誤るか，といった情報が必要になる．そのためには，実際に日本語学習者が書いたテキストデータ，つまりコーパスを用いた処理を行う方法が自然言語処理では広く使われている．

最も有名な日本語学習者コーパスの 1 つに，「寺村誤用データ」（1990）[☆6]がある．このコーパスの大部分は 1986 年に収集され，主にアジアの学生からデータが取得されている．自由作文・穴埋め問題・パターン作文などいろいろなスタイルの作文からなっている．寺村誤用データはエラーの種類がタグ付けされているので，誤り検出に用いることができる．

一方，大曽（1998）による「日本語学習者の作文コーパス」も広く用いられている．こちらはエラーの種類だけでなく，訂正後の文字列も含めてタグ付けされているので，誤り検出だけではなく，誤り訂正にも用いることができる．

また，最近は国立国語研究所で「作文対訳 DB コーパス」（2009）という日本語学習者コーパスが作成されている．このコーパスの特徴は，学習者が自分の書きたかった意図を自分が使いやすい言語で説明する対訳文になっていることである．学習者の意図が分からないためにどのように訂正すればよいか分からない，といったことがしばしば起こるが，学習者の意図と実際に書く文のズレについての分析が行えるようになっている．

語学学習 SNS から作る新しい日本語学習者コーパス

前章で述べた日本語学習者コーパスは，日本語教師や研究者などの専門家によるもので，信頼性は高いが，大規模に収集することができないという欠点があった．また，近年の日本語学習では会話文や Twitter，インスタントメッセンジャーなどで日本人とコミュニケーションをとるためのくだけた文体の学習の需要が高いが，従来の日本語学習者コーパスは教室内での課題作文や穴埋め問題といった特定の状況における作文コーパスであり，分野やトピックが限られるといった問題がある．

一方，Web の急速な発展によって多くの人がソーシャルネットワークサービス（SNS）を使うようになり，最近では語学学習 SNS も誕生した．代表的な語学学習 SNS としては，iKnow![☆7]，Livemocha[☆8]や Lang-8[☆9]がある．本稿では日本

☆4 http://language.tiu.ac.jp/
☆5 http://www.hiragana.jp/
☆6 http://teramuradb.ninjal.ac.jp/
☆7 http://iknow.jp/
☆8 http://www.livemocha.com/
☆9 http://lang-8.com/

語学習者の利用が多い Lang-8 およびそこから作成されたコーパスについて詳しく説明を行う．

Lang-8 は相互添削型 SNS とも言われている．学習者が学習している言語で日記を書くとその言語を母語とするユーザが添削してくれ，また学習者も自分の母語で書かれた日記を添削することができる．Lang-8 では 2010 年 12 月（2011 年 10 月）の時点で 77 の言語をサポートしており，214,170 人（同 317,307 人）のユーザが登録している．

Lang-8 は語学学習 SNS であるため日本語以外の学習者のデータも存在している．図 -5 に Lang-8 登録ユーザの学習言語の分布を示す（注：複数の言語を学習している人もいるため，合計すると 214,170 人を超えている）．英語学習者が 142,311 人と最も多く，日本語学習者がそれに続き 75,162 人となっている．3 番目以降の言語学習者数は 2 万人以下である．また，表 -1 の学習言語別の投稿文数を見ると，英語が 1,069,549 文と最も多いが，日本語も 925,588 文と学習者の割合に対して比較的多い投稿文数となっている．

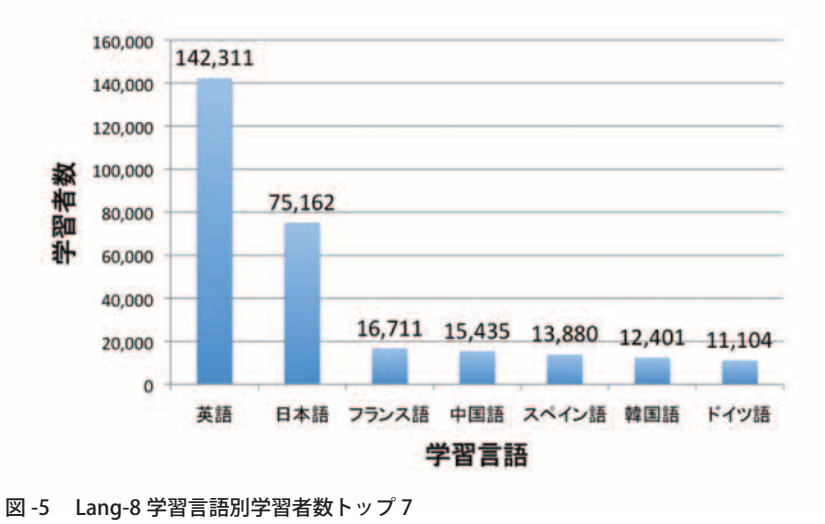

図 -5　Lang-8 学習言語別学習者数トップ 7

学習言語	英語	日本語	中国語	韓国語	フランス語	スペイン語	ドイツ語
文数	1,069,549	925,588	136,203	93,955	58,918	51,829	37,886

表 -1　Lang-8 学習言語別投稿文数トップ 7

母語	英語	中国語 (繁体字)	中国語	韓国語	ロシア語	スペイン語
文数	509,924	269,536	265,340	48,406	19,499	17,133

表 -2　Lang-8 日本語学習者の文を母語別に分類したときの母語ごとの投稿文数

日本語学習者の書いた 925,588 文のうち，実際に添削のついている文は 763,971 文あり，93.4% の文が添削されている．また，実際の添削は 1 文に対して 2 つ以上の添削がつくこともあり，添削後の文数は 1,288,934 文となっている．従来の日本語学習者コーパスが数千文から数万文であることに比べると，比較的大規模なコーパスであるといえる．

Lang-8 のデータでは学習者および添削者の母語も知ることができる．本稿冒頭で述べたように，学習者は母語によって誤り方が違うことが知られており，母語の違いを考慮した研究も行われている．国立国語研究所の作文対訳 DB コーパスにおいても学習者の母語と添削者の母語は属性情報として収集されており，誤り・添削傾向を知るための貴重な情報となっている．Lang-8 の日本語学習者の添削後の文を母語ごとに分類した結果を表 -2 に示す．英語を母語とする学習者の文が最も多く，中国語（繁体字），中国語，韓国語と続いている．図 -1 の地域別学習者数の場合と同様に，アジア圏において日本語学習の需要が高いことが分かる．

図 -6 に Lang-8 で実際にあった添削の例を 3 つ示す．1 つ目の例は学習者が書いた文で“う”と“あ”が抜けてしまっていて，それに対して文字を挿入することで添削を行っている．典型的な添削はこのように挿入・削除・置換によって，学習者の書いた文を訂正しているものである． 2 つ目の例は学習者がローマ字表記で日本語を書いているものである．

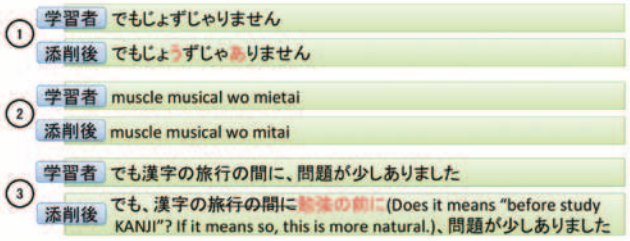

図 -6　Lang-8 から実際に取ってきた添削例

これは学習者がかなを入力できない環境にいる，もしくは，ローマ字表記を用いた日本語学習を始めたばかりの学習者がローマ字を用いて入力するためだと考えられる．Lang-8 を使う学習者のレベルやバックグラウンドは多様であり，学習者の習熟度に応じた学習支援を提供する必要があることが分かる．3 つ目の例は学習者の文に " 漢字の旅行の前に " と書いてあり，何が言いたいのか分からない文である．そこで，添削者は英語でコメントを加えてどういった意味で書いているのか確認，アドバイスしている．確認以外にも代替表現をコメントで追記したり，文法事項を説明したりするなど，添削者から多様な学習支援情報が学習者に提示されることがある．

日本語学習者の文を扱う上での従来の自然言語処理の問題点

図 -6 で 3 つの Lang-8 の学習者の文とその添削例を紹介したが，この日本語学習者の文を扱う上で従来の自然言語処理では問題が生じる．1 つ目の問題は形態素解析の問題である．2 つ目の問題は従来の日本語処理ではローマ字表記で書かれた文を扱っていないことである．3 つ目が作文意図推測の問題である．

■形態素解析失敗の問題

通常自然言語処理で日本語の解析を行う場合は，まず形態素解析という処理を行って文を単語に分割する．しかしながら，図 -6 の例 1 にあるように学習者の文には誤りやひらがなが多く含まれているため，新聞記事を正しく形態素解析できるようにチューニングしている従来の形態素解析器では単語分割や品詞付与に失敗してしまう．図 -7 に図 -6 の例 1 の学習者の文を形態素解析器で単語分割した例を示す．" じ " と " ょずじゃりません " に分割されており，単語分割に失敗していることが分かる．従来の助詞の誤り検出・誤り訂正の研究は，形態素解析結果までは正しい前提で行われることが多いが，この前提は必ずしも正しくない．特に自由作文を誤り検出の対象にした場合，形態素解析は自動解析の結果を使わざるを得ないが，ひらがなを頻繁に使う学習者の作文には解析誤りが含まれることが多い．形態素解析に失敗してしまうと，従来行われていた助詞の誤り訂正の手法は適用できなくなる．

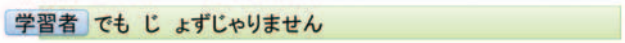

図 -7　日本語学習者の書いた文を形態素解析器で単語分割した例

■ローマ字表記による問題

日本語母語話者は通常ローマ字表記で文を書くことがないため，従来の自然言語処理ツールはそのまま適用することができない．ローマ字表記は基本的にはひらがなと 1 対 1 で対応しており（例：ka →か），変換ルールを使うことでローマ字表記からひらがな表記に変換することが可能である．しかしながら，日本語学習者の書いたローマ字には誤りも含まれている．図 -6 の例 2 の例で学習者の書いたローマ字表記を変換ルールに従ってひらがな表記に直すと " む sc ぇ むしか l を みえたい " となる．また，書き誤りや読みの学習誤りによって "hajimemashtei" と母音を抜かして子音だけで書いてしまうため，ひらがな変換に失敗する場合もある．

■作文意図推測の問題

図 -6 の例 3 の学習者の文は，機械はもちろんのこと日本語母語話者でも添削が難しい例である．このような文は現状の自然言語処理の技術では扱うことはできない．文を超えた談話構造を考慮した自然言語処理や，言語以外の情報も用いて書き手の意図

を推測する手法の研究は，現在の自然言語処理でも未解決な課題の 1 つである．

リアルな日本語学習者の文を扱う大規模データを用いた新しいアプローチ

ここでは前章で挙げた形態素解析失敗の問題，ローマ字表記の問題，そして作文の意図推測の問題について最近の大規模データを用いたアプローチの 3 つを紹介する．

■統計的機械翻訳の手法を使った文字ベースの誤り訂正

前章で挙げた形態素解析失敗の問題に対処した研究として，統計的機械翻訳を使った文字に基づく誤り訂正がある[1]．通常の英語から日本語への翻訳（例：I like English →私は英語が好き）を行う場合では，英語から日本語への翻訳ルールを用いて，単語に基づいて翻訳を行う．統計的機械翻訳をそのまま誤り訂正に応用した場合，学習者の書いた誤りを含む文から正しい日本語文への変換ルールを用いて訂正を行う．しかしながら，学習者の文は自動単語分割に失敗してしまうため，単語単位での変換を行うことは難しい（図 -8）．

そこで，単語よりも細かい文字単位に分割することで，文字から文字への変換ルールを用いて訂正を行う手法が提案されている．この手法は学習者の書いた文と正しい日本語文が対になったデータ（学習者コーパス）から，学習者の書いた文の誤り部分を抽出し，自動的に学習者の書いた文字列から正しい文字列への訂正の対応表を取得する．文字列がおおむね単語に相当するので，この手法は学習者コーパスの自動訂正に適した単語分割基準を自動で獲得し，同時に訂正も行っていることに相当する．単語単位の分割である図 -8 と文字単位の分割である図 -9 を比べると，図 -9 のほうが頑健な解析ができそうであるのが分かるだろう．実際，実験の結果，単語単位の分割よりも文字単位の分割で訂正を行うほうがよいことが分かっている．

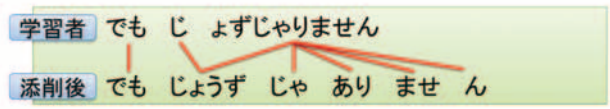

図 -8　単語単位分割時の学習者文と添削文の変換対応

図 -9　文字単位分割時の学習者文と添削文の変換対応

通常の統計的機械翻訳では，翻訳ルールを自動抽出するために，文単位で英語と日本語が対になった文が大量にある大規模な対訳コーパスが必要である．統計的機械翻訳を使って誤り訂正を行う場合も同じように，学習者の文とその添削文が対になった大規模な学習者コーパスが必要である．語学学習 SNS の登場により大規模な学習者コーパスが手に入るようになったことで，大規模データを用いた学習者の誤り訂正ができるようになってきたのである．

■誤り訂正ローマ字かな変換

日本語学習者，特にヨーロッパ言語を母語とする初心者は，ローマ字を用いた入力をすることも多い．たとえば Lang-8 からローマ字で書かれた文を抽出すると約 1 万文あり，日本語学習者の作文全体の 1% ほどを占める．

日本語母語話者がローマ字で表記されたものを理解するのは，漢字，ひらがな，カタカナで書かれた場合に比べて難しい作業であり，添削漏れが発生しやすいという問題点がある．そこで，学習者の書いた単語の言語判定を行い，単語の曖昧検索を行うことで誤りを含んだローマ字を正しいローマ字に自動的に修正し，ひらがなに変換することによって，添削効率の改善を行う手法が提案されている[2]．

■インタラクティブな作文支援

これまでに紹介した日本語学習者システムは，解析したい文を一括して入力し，結果を提示するシステムであった．しかし，先述したように，学習者の

図 -10　Chantokun で「あなたは夜何時で寝ますか。」という文の格助詞を自動訂正する

意図が分からないため解析できない場合がある．そこで学習者にインタラクティブに入力させることで，学習者が自ら誤りに気づくシステムが提案されている．

そういった日本語学習者の誤り検出・訂正システムとしては奈良先端大の「Chantokun」☆10 がある．Chantokun は学習者が入力した日本語文に対し，「が」「を」「に」といった格助詞の誤りを検出し，正しい格助詞候補を提示する．図 -10 に Chantokun で「あなたは夜何時で寝ますか。」という文を入力し，格助詞「で」は正しくは「に」であるという自動添削結果が表示されているところを表示している．Robo-Sensei では課題に沿った作文しかできないが，Chantokun は任意の文を入力・誤り訂正することができる．Chantokun は「あすなろ」同様単語の意味を表示するため英日辞書が用意されており，単語の読みと意味を表示することができる．Chantokun は Google N-gram という大規模な Web テキストから抽出した統計情報をもとに訂正候補を取得しており，このような大規模データの存在が学習者の支援システムの性能向上を後押ししている．

☆10 http://cl.naist.jp/chantokun/

自然言語処理を使った日本語学習支援のための今後の課題

最後に自然言語処理を用いて日本語学習者を支援するための課題を挙げる．1 つ目は日本語学習者の母語に応じたモデルを作ることである．母語によって誤りの傾向が異なるため，このモデルをうまく作ることができれば訂正の性能向上に繋がる．

2 つ目は学習者の文に対応した形態素解析器の作成である．学習者の文を単語単位に分割できるようになれば，従来の格助詞誤り訂正の手法を応用できるようになる．また，単語単位で訂正可能になれば単語の意味を使った訂正も可能になる．

3 つ目は学習者用の日本語入力システムの開発である．学習者が日本語を書く際に問題となる部分を検出し，読解や作文をサポートできるシステムを作ることで学習効率を向上させることが可能になる．

参考文献
1）Mizumoto, T., Komachi, M., Nagata, M. and Matsumoto, Y. : Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, Proceedings of 5th International Joint Conference on Natural Language Processing, pp.147-155（2011）.
2）Kasahara, S., Komachi, M., Nagata, M. and Matsumoto, Y. : Error Correcting Romaji-kana Conversion for Japanese Language Education, Proceedings of the Workshop on Advances in Text Input Methods（WTIM 2011）, pp.38-42（2011）.

（2011 年 11 月 20 日受付）

水本 智也 ▌tomoya-m@is.naist.jp
奈良先端科学技術大学院大学情報科学研究科博士前期課程．2010 年甲南大学理工学部情報システム工学科卒業．専門は自然言語処理．

小町 守（正会員）▌komachi@is.naist.jp
奈良先端科学技術大学院大学助教．博士（工学）．2005 年東京大学教養学部基礎科学科科学史・科学哲学分科卒業．2010 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．専門は自然言語処理．