日本語学習者
誤用換言対コーパスに
見られる表記エラーについて

松田 真希子（金沢大学）

mts@staff.kanazawa-u.ac.jp

[付記]　この研究はJSPS基盤研究B「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」
Research Project Number:15H03216（研究代表者：山本和英）の助成を受けています。

Index

金沢大学 KANAZAWA UNIVERSITY

Introduction 1/6

金沢大学 KANAZAWA UNIVERSITY

日本語形態素解析技術の問題（山本他2015）

①分割単位の問題
慣用句や慣用表現が1語として出力されない

②表記統制の問題
明らかに同一であるべき語が別の語として出力される

③品詞の問題
形態による品詞付与を採用することで後の処理を複雑にしている（勉強する→勉強/する）

④意味の問題
本来形態素解析にさせるべきでない
問題を解いている（その結果解析を誤っている）

	Unidic-mecab	雪だるま
表記ゆれ	えり好み/選り好み	選り好み（一語）
サ変動詞	勉強（サ変）＋する（動詞）	勉強する（動詞）
複合名詞	日本＋料理	日本料理（複合名詞）

Introduction 2/6

日本語教育向け日本語単語解析器を開発中(2015.1～）
http://snowman.jnlp.org/

主な作業
〇表現統制を行う（表記レベル）
〇形態素を単語にまとめあげる

Project Leader
山本和英（長岡技術科学大学）

岩田　一成　（聖心女子大学）
内丸　裕佳子　（岡山大学）
建石　始　（神戸女学院大学）
中俣　尚己　（京都教育大学）
松田　真希子　（金沢大学）
茂木　俊伸　（熊本大学）
森　篤嗣　（帝塚山大学）

Introduction 3/6

金沢大学 KANAZAWA UNIVERSITY

日本語形態素解析技術の問題

⑤誤用への対応

日本語学習者の誤用は、現状の形態素解析器（単語解析器）で適切に解析されない。
母語話者がその誤用を見たときに「これのことだな」と正解情報に置き換えながら読むことができるものでも、置き換わらず誤解析となる。
主要な誤用だけでも、解析器に反映できると、日本語学習者データの分析に役立つのでは？

例

何もしんぽいしない。　→　何も/しんぽい（e:心配）/しない
成工の可能は低いだ。　→　成工（e：成功）/の/可能性/は
一兆懸命頑張るつもりだ。→　一兆懸命（e：一生懸命）/頑張る/つもりだ
誰も出来ことではない。　→　誰も出来（e：できる）/こと/ではない

Introduction 4/6

金沢大学 KANAZAWA UNIVERSITY

Unidic-mecabの場合
http://www4414uj.sakura.ne.jp/Yasanichi1/unicheck/

書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	書字形基本形	語種
そう	ソー	ソウ	そう	副詞			ソー	そう	和
らい	ライ	ライ	癩	名詞-普通名詞-一般			ライ	らい	漢
の	ノ	ノ	の	助詞-格助詞			ノ	の	和
ほしい	ホシー	ホシイ	欲しい	形容詞-非自立可能	形容詞	連体形-一般	ホシー	ほしい	和
士	シ	シ	士	名詞-普通名詞-一般			シ	士	漢
事	コト	コト	事	名詞-普通名詞-一般			コト	事	和
私	ワタクシ	ワタクシ	私-代名詞	代名詞			ワタクシ	私	和
の	ノ	ノ	の	助詞-格助詞			ノ	の	和
り	リ	リ	り	助動詞	文語助動詞-リ	終止形-一般	リ	り	和
ょ	ヨ	ヨ	よ	助詞-終助詞			ヨ	ょ	和
しん	シン	シン	芯	名詞-普通名詞-一般			シン	しん	漢

Introduction 5/6

雪だるまの場合

正答	分割結果			
心配	しんぽい	複合形容詞	しれる＋ぽい	igora,indar
自分	白	名詞	代	iekpz
	分	名詞|副詞	分	immtg
	ふたにく	数量詞	ぶた＋にく	imfmg,ikjsq
豚肉	てきしゅ	複合名詞	適＋しゅ	ijfhd,igfwt
摘出します	つ	助詞	つ	iixgl
	し	非自立動詞	する	ihjhf
	ます	助動詞	ます	iniol
将来会社	そうらい会社	複合名詞	そう＋らい＋会社	ihtyo,ioybh,icnit
面白いです	白面	名詞	白面	igobo
	です	助動詞	だ	iimqc

☆形態素及び単語に対してIDを付与

表記ゆれや誤用も同一IDにすることで統合する

Introduction 6/6

今めざしていること

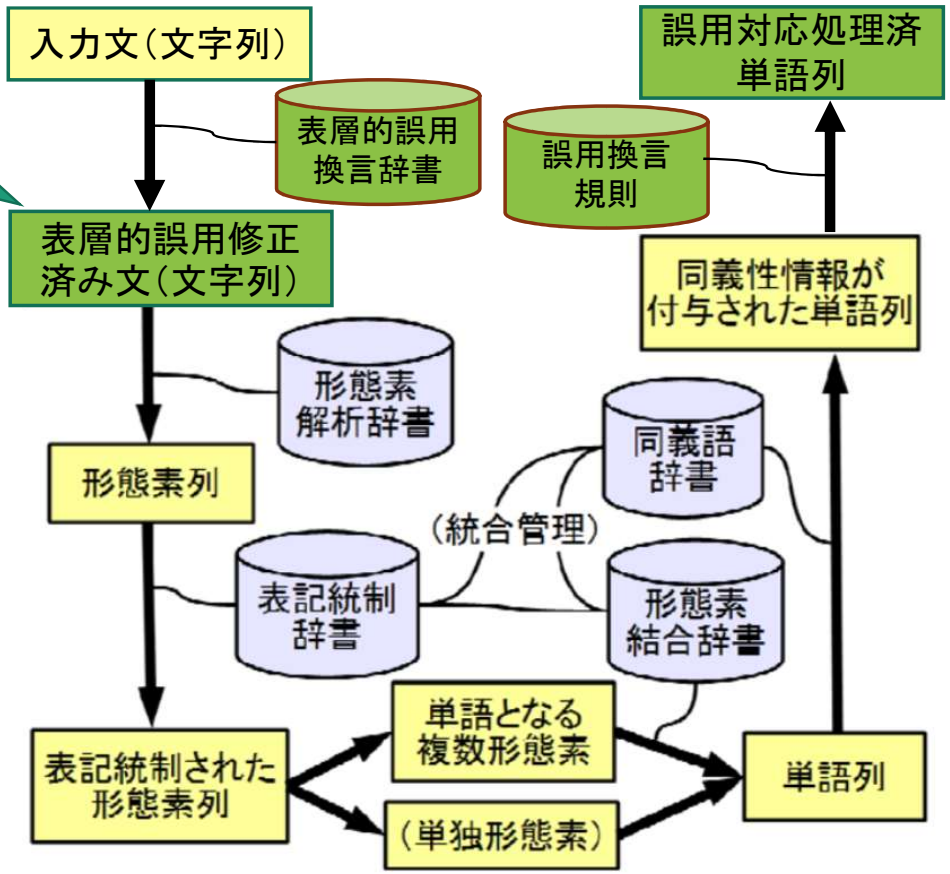


図 1：単語解析器の処理の流れと使用する言語資源

(山本ほか2016に緑部分を追加した図)

Related Works 1/4

金沢大学 KANAZAWA UNIVERSITY

日本語学習者誤用コーパス

	コーパス名	母語	レベル	表記 エラータグ
大山 (2012, 2016)	NAIST誤用コーパス (対訳作文DB(国語研)に情報付与)	中国、韓国、ベトナム、タイ、マレイシア、シンガポール他	中級	1,838
李(2012)	日本語学習者作文コーパス	中国 韓国	初級 中級 上級	3,387

表記エラーは細分化されていない

Related Works 2/4

金沢大学 KANAZAWA UNIVERSITY

韓国語母語話者（中上級）の表記エラーの分析（姜 2006）

表２　日本語作文に見られる誤用分析結果

音声項目誤用実態		誤用数	誤用割合
清音と濁音	清音の濁音化	76	17%
	濁音の清音化	125	28%
長音	添加	69	15%
	脱落	103	23%
促音	添加	37	8％
	脱落	4	1％
撥音	添加	1	0％
	脱落	1	0％
拗音と直音	拗音の直音化	14	3％
	直音の拗音化	8	2％
その他	子音添加	3	1％
	子音交替	10	2％
合　計		451	100%

清濁に関する
エラーが多い

長音の脱落・添加
に関する
エラーが多い

Related Works 3/4

韓国語、中国語母語話者の作文誤用の分析（李ほか 2013）

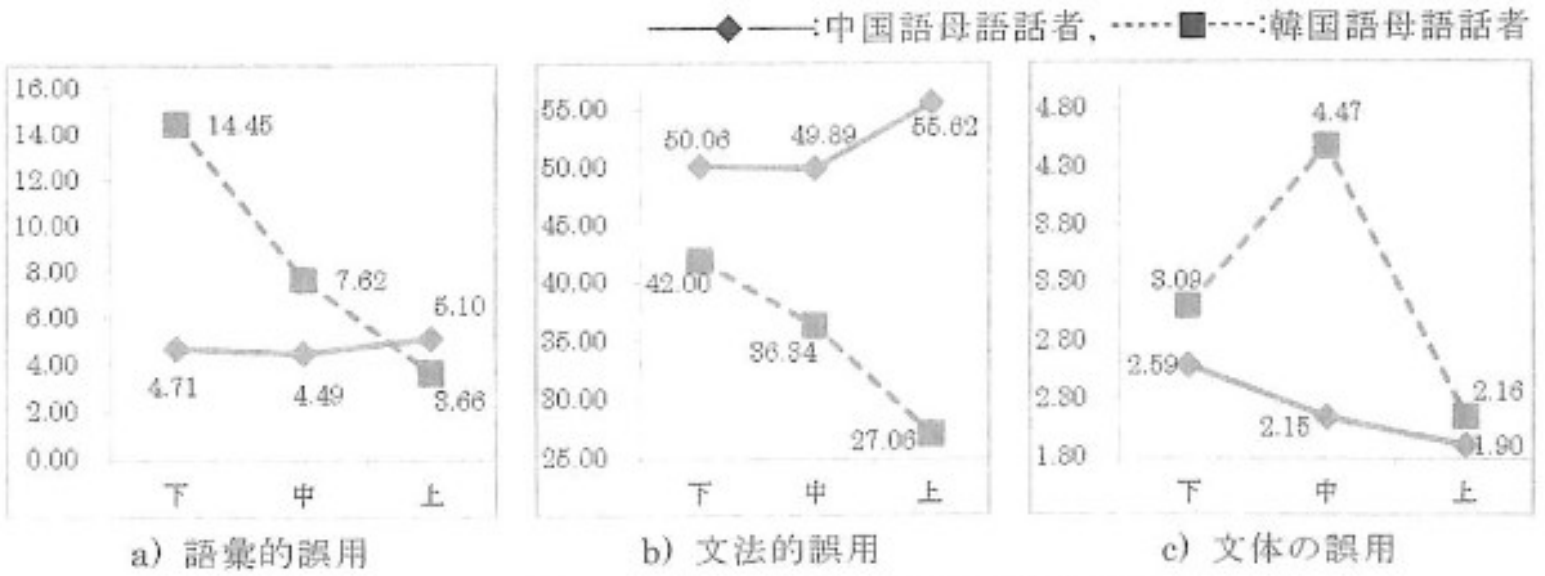


図 4. 母語別の一作文あたりの誤用の平均出現値

エラーと
レベルの
相関あり

文字？

Related Works 4/4

金沢大学 KANAZAWA UNIVERSITY

中国語母語話者の日本語の表記に関する誤用分析（李2009）

誤用の種類	誤用数	比率
簡体字をそのまま日本語に持ち込んだ誤用	39	26.9%
同音異字語の誤選択	22	15.2%
類義異字語の誤選択	15	10.3%
発音類似の誤用	34	23.4%
形類似の誤用	17	11.7%
その他の誤用	18	12.4%
合計	145	100%

かそか　過疎化――过疎化
かんたん　簡単――简单
ふうすいがい　風水害――风水害
ざっし　雑誌――杂志
りゅうちょう　流暢――流畅
しっぱい　失敗――失败
せんしん　先進――先进
らいほう　来訪――来访
せいかく　正確――正确
がっき　楽器――乐器
きつえん　喫煙――喫烟
えいよう　栄養――栄养
せいぼ　歳暮――岁暮
かね　鐘――钟
とうけい　統計――统计
さんせい　賛成――赞成
みたす　満たす――满たす

Research Questions

（１）表記エラーのうちどれが「よくあるエラー」なのか
（母語別、レベル別）

（２）表記エラーの分類を細かくし、
　　レベルと母語で比べたらどうなるか

（３）形態素解析器で拾えるエラーと拾えないエラーの区別は？

Data

「日本語学習者誤用換言対コーパス」（2016年末完成予定）

現在収集済のデータ

○金沢大学留学生センター日本語プログラムの作文課題
（初級～上級）

○ホーチミン市工科大学・ホーチミン市師範大学の
日本語クラスの作文課題（井上徹先生提供）
（初級・中級）

○インドネシア教育大学日本語クラスの作文課題
（初級～上級）

☆デジタル入力と紙で筆記したものと両方あり

☆できるだけ多くの国・母語（現在３２カ国）

☆初級～上級

☆目標10,000対（現在約3,000)

Data

金沢大学 KANAZAWA UNIVERSITY

構築状況（誤用対の数）

	上級 JLPT（N2-N1）	中級 JLPT（N4-N3）	初級 JLPT（N5）	合計
インドネシア	39	3	382	424
ベトナム	75	515	165	755
中国	253	145	120	518
英語（アメリカ、イギリス、オーストラリア）	1	240	103	330
タイ	64	24	125	213
その他	73	160	302	**549**
	505	**1087**	**1197**	2789

Data

アノテーションルール

① 表記エラー（C）

単語として切り出された際、単語の内部で誤りがみられるもの

C1	文字の異なり・文字順の異なり	ABC→ADC/CBA
C2	異表記（通常は漢字なのにカタカナやひらがな）	ABC→abc/aB c
C3	要素の脱落	ABC→AB
C4	要素の余剰	ABC→ABCD

二人の
日本語教師による人手判定
C1, C3, C4 の一致99.5%
C2の一致率は低い
（漢字を拾わない）

② 文法エラー(G)

単語として分割された際、それぞれは正しいが、
共起で誤りが見られるもののうち文法ルールにかかわるもの
あるいは単語内部の誤りのうち、活用の誤り

二人の
日本語教師による人手判定
一致率は未測定
（8割程度か）

③語彙エラー（V)

単語として分割された際、それぞれは正しいが、
共起で誤りが見られるもののうちの語彙選択に関わるもの

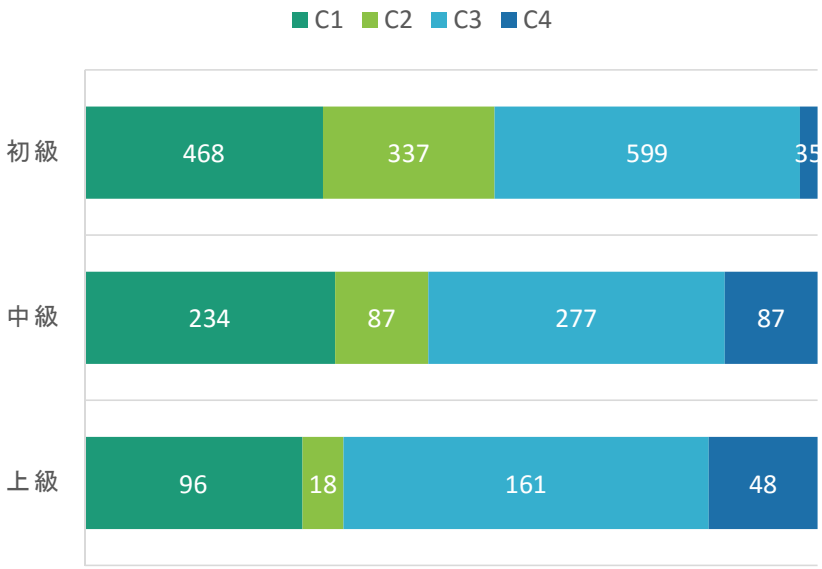
Data

誤用還元対コーパス例

レベル	国	誤用	正解	元の文	タグ1	タグ2（複合型の場合）
211	I	VN	初める	始めた	日本話の勉強を初めるのは9か月前です。	c1	
218	I	TH	ザイス	サイズ	ちょうどザイスがあって、お持ち運びやすいですよ。	c1	
222	I	KR	グレン	グレー	ピンク、黒、グレンなどがあるからご自由に選べます。	c1	
224	I	KR	だった2800円	たった2800円	こんなに性能がいいものがだった2800円しかしません	c1	
231	I	VN	選んだち	選んだり	ファーストフードを選んだち、店で食事したりする傾向があった。	c1	g
232	I	TW	もって	もっと	もって大切だと感じた。	c1	
234	I	KR	工産品	工業製品	工産品に関するトラブルは私たちの周辺に多く存在している。	c1	v

Results

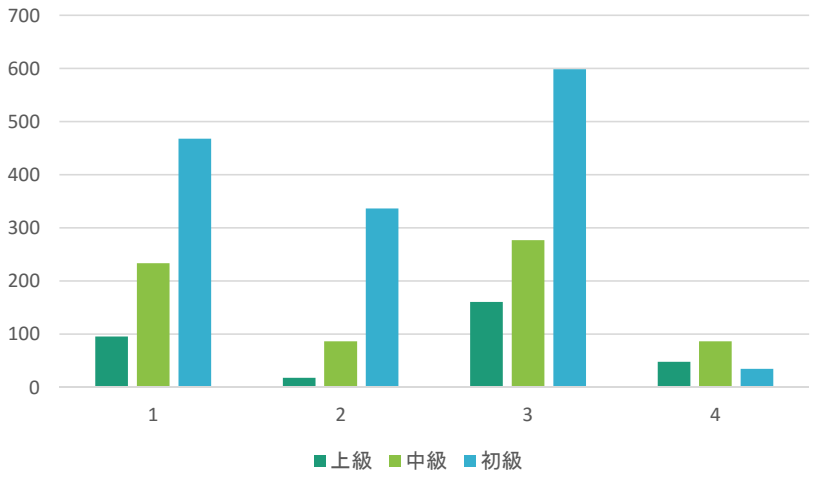
RQ1　表記エラーのうち、どれがよくあるエラーなのか　（レベル別）



			上級	中級	初級	合計
C1	文字の異なり・文字順の異なり	ABC→ADC/CBA	96	234	468	798
C2	表記ゆれ	ABC→abc/aBc	18	87	337	442
C3	文字の脱落	ABC→AB	161	277	599	1037
C4	文字の過剰	ABC→ABCD	48	**87**	35	170
			323	685	1439	2447

レベルを通じて高めなのは
要素の脱落

Results



		上級	中級	初級	合計
C1	頻度 残差	96 −1.189	234 1.019	468 −0.112	798
C2	頻度 残差	18 −6.263**	87 −4.299**	337 8.229**	442
C3	頻度 残差	161 2.915**	337 −1.211	599 −0.900	1037
C4	頻度 残さ	48 6.004**	87 6.979**	35 −10.496**	170
		323	685	1439	2447

初級は表記ゆれが有意に多い

中級は過剰が有意に多い

上級は脱落や過剰が有意に多い

Results

C1　の例

E	CH	初めで	初めて
E	CH	友たち	友だち
E	CH	おかげて	おかげで
E	CH	自家	実家
I	VN	好きななった	好きになった
I	VN	ヒューモア	ユーモア
I	CH	チュロ	チェロ
I	FR	だぁろうか	だろうか
I	FR	手にんさん	店員さん
I	CH	直しなきゃ	直さなきゃ
A	TUR	プラグラム	プログラム
A	POL	学黒人	外国人
A	TW	時間とおり	時間どおり
A	CH	クゴソ	ワゴン

トヨタが生産している車種はセダン、スポーツ、SUV、ステーションクゴソ 、ミニバン、コンパクトカー、商用車、軽自動車に分類される。

Results

C2　表記ゆれ誤用の例
（赤字は形態素解析エラーになるもの）

あめりか（アメリカ）（ベトナム・初級）
こんびに（コンビニ）（タイ・初級）
だい好です
もスク（モスク）（インドネシア・初級）
カメら（カメラ）（ベトナム・初級）
ほっかいどう（北海道）（タイ・初級）
おいしkて（おいしくて）（ベトナム・初級）

→混在するものはエラー。かな→漢字は問題なし

Results

C3　要素脱落の例
（赤字は形態素解析エラーになるもの）

げつよび（げつようび・月曜日）(インドネシア・初級）
ときょう（とうきょう・東京）（タイ・初級）

→脱落は形態素解析器は拾わない？

Results

C4　要素余剰の例
（赤字は形態素解析エラーになるもの）

上級の余剰の多くは
カタカナ表記エラー

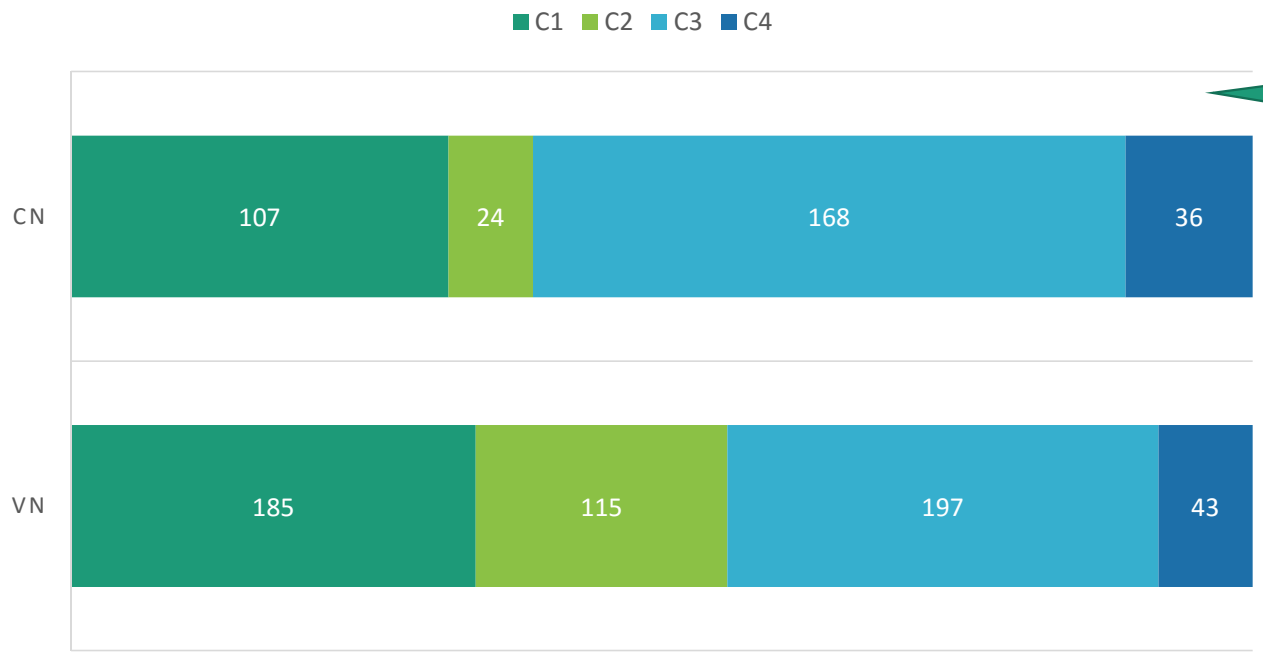
初級
みんなさん（みなさん）(インドネシア・ロシア他・初級）
きょうとう（きょうと・京都）（フィンランド・初級）
ほうかの（ほかの・他の）（マレイシア・初級）
いっしょうに（いっしょに・一緒に）（インドネシア・初級）

A	TH	レーポート
A	IN	シェーア
A	KR	マーケッティング
A	POL	メーディア
A	TW	ウェッブサイト

I	BR	高かい	高い
I	VN	これられの	これらの
I	SR	守もり	守り
I	USA	じゅうぎょう	じゅぎょう
I	USA	かっこよっく	かっこよく

Results

国別比較 (ベトナムー中国のみ)



中国は脱落の比率が高い？

Conclusion

金沢大学 KANAZAWA UNIVERSITY

（1）表記エラーのうちどれが「よくあるエラー」なのか

→レベルを問わず文字の脱落が多い

→初級は漢字をかなで書く「エラー」が多い

→カタカナ表記エラーはレベルを問わず多い

（2）表記エラーの分類を細かくし、
レベルと母語で比べたらどうなるか

→中国語母語話者は脱落が多い？

（3）形態素解析器で拾えるエラーと拾えないエラーの区別は？

→表記がまざったもの、文字の入れ替わりは拾えない？

Future Work

（1） データを増やし、母語別・レベル別の分析を進める

（2） 形態素解析器に実装する

（3） 換言対を機械学習させる

引用文献

- 山本和英, 宮西由貴, 高橋寛治, 猪俣慶樹, 須戸悠太, 三上侑城(2015)「日本語解析システム「雪だるま」 ～ 単語解析部の設計思想 ～」『電子情報通信学会 テキストマイニングシンポジウム 信学技報』Vol.115, No.222, pp.13-18.
- 大山浩美、小町守、松本裕治(2016)「日本語学習者の作文における誤用タイプの階層的アノテーションに基づく機械学習による分類」『自然言語処理』V23-2 .195-225.
- 大山浩美、小町守、松本裕治(2012)「日本語学習者の作文における誤用タグつきコーパスの構築について―NAIST誤用コーパスの開発―」『第一回テキストアノテーションワークショップ』
- 姜枝廷「韓国人学習者の日本語の文字表記に見られる音声項目の誤用―長母音を中心に―」杏林大学大学院国際協力研究科『大学院論文集』No３， 2006.3
- 李在鎬・林炫情・宮岡弥生・柴崎秀子(2012)「言語処理の技術を利用したタグ付き日本語学習者コーパスの構築」（日本語教育学会2012年度秋季大会
- 李在鎬・宮岡弥生・林炫情(2013)「学習者コーパスと言語テスト― 言語テストの得点と作文のテキスト情報量の関連性」『言語教育評価研究（AELE）』p.22-31
- 李被(2009)「日本語の表記に関する誤用分析ー東華大学日本語学習者の場合ー」『日本言語文化研究』第13号』

謝辞

本発表資料作成にあたり、長岡技術科学大学山本和英先生と、
長岡技術科学大学大学院高橋寛治氏、
金沢大学松田佳子先生に大変お世話になりました。
心より感謝申し上げます。