概要

記述文を対象にした形態素解析や構文解析の技術が進歩してきた．機械翻訳はそれらの解析結果をベースに処理されている．会話文の機械翻訳を目指すとき，会話文と記述文とでは話し言葉特有の表現や敬語表現など構造の差異 [1,2] が問題になるのだが，形態素解析の段階で差異が吸収できるならば，機械翻訳の処理がそのまま利用できることになる．

そこで本研究では，形態素解析の段階で検出される構造の差異を定量的に分析し，その違いを吸収する方法について明らかにする．

調査のため，ATR 音声翻訳通信研究所が構築したバイリンガル旅行対話コーパスから収集した会話文 3,712 文を，記述文の解析精度が単語単位で 99.8%ある形態素解析システム ALT-JAWS を用いて解析した．解析に誤った箇所は記述文とは違う会話文特有の誤りとみなしてよいと考えられるため，解析に誤る文を収集した．

収集した文の誤り箇所を分析した結果，品詞別に分類すると，複数の品詞として解析すべき所が１つの形態素として解析してしまう箇所が最も多く 21.5%，続いて助詞が 13.6%，助動詞が 13.0%，接続詞が 9.9%の順であった．また，誤り箇所を単語が辞書に無い会話文特有の単語が原因の箇所 (未知語誤り箇所) と，単語が辞書にあるが会話文特有の使われ方をしているために誤った箇所 (既知語誤り箇所) に分けて分類した結果，未知語誤り箇所が 50.0%，既知語誤り箇所が 44.8%であった．

吸収方法として，既知語誤り箇所は，接続する単語によって決まった誤り方があり，それを if-then ルールによって正しく書き換える方法を考察した．これにより，既知語誤り箇所の 38.1%を正しく直すことが出来ると期待できる．

結論として，ALT-JAWS で解析に誤った箇所を会話文特有の表現のため誤ったとして分析したところ，50.0%が未知語誤りで，44.8%が既知語誤りであった．また，既知語誤り箇所のうち 38.1%に対して誤りを修正できる見込みを示した．

目 次

図 目 次

表 目 次

1　はじめに

記述文を対象にした形態素解析や構文解析の技術が進歩してきた．機械翻訳はそれらの解析結果をベースに処理されている．会話文の機械翻訳を目指すとき，会話文と記述文とでは話し言葉特有の表現や敬語表現など構造の差異 [1, 2] が問題になる．形態素解析の段階で差異が吸収できるならば，機械翻訳の処理がそのまま利用できることになる．

そこで本研究では，形態素解析の段階で検出される構造の差異を定量的に分析し，分析結果から違いを吸収する方法について明らかにする．

分析の方法を説明する．記述文用の形態素解析システム ALT-JAWS は単語単位で 99.8% の精度がある [7]．つまり，記述文ならば非常に高い精度で解析されるシステムに，会話文を入力して解析に成功したならば，その会話文は記述文と同じように扱え，解析に失敗したならばそれはその会話文に記述文には無い会話文特有の表現が入っており，それが記述文と会話文の差異とみなせると考えられるので，誤った箇所を分析する．分析の着目点として，まず誤った箇所は本来どうあるべきか，という観点から品詞に着目した分析を行う．次に差異を吸収する方法を考える際，その誤り箇所が既知語であるか未知語であるかによって大きく対策の仕方が異なるという観点から，誤り箇所が未知語によるものか既知語によるものかに着目した分析を行う．

差異の吸収について説明する．記述文と会話文の違いを吸収し，記述文用のシステムで会話文を翻訳するシステムを想定し，本研究で分析した記述文と会話文の差異が吸収できるかどうかについて考察する．

本論文の構成は以下の通りである．まず，第２章で記述文の形態素解析として，本研究で使用する記述文用翻訳システム ALT-JAWS について述べる．第３章で記述文との差異のある会話文を収集方法と収集結果について説明する．第４章で収集した特徴を品詞に着目して分類，および，既知語／未知語誤りかどうかという点に着目して分類をする．第５章で差異の吸収方法について考察し，机上シミュレーションを行う．第６章でまとめを述べ，今後の課題を提案する．

2　記述文の形態素解析

本研究では，会話文と記述文の表現の差異を抽出するため，日本語形態素解析システム ALT-JAWS を使用する．

ALT-JAWS は日英機械翻訳システム ALT-J/E の日本語形態素解析部分をパッケージ化したものであり，以下の特徴がある．

- 記述文を対象としている
- 単語区切りの多義は出力しない
- 「ひらがな」だけから成る語の解析は，漢字かな混じり文にくらべて正解率が下がることがある

ALT-J/E の翻訳における日本語の形態素解析技術は [6] の方式を基に解析精度向上を行い，現在は単語単位の正解率 99.8%，1 文が 20 形態素とすると文単位の正解率は 96%である [7].

形態素解析結果の出力は単語毎に ’’/ ’’で分割され，１文節１行で標準出力される．さらに，各単語の直後には単語種別，品詞活用形コードが表示される．特に単語種別コードのうち未知語の単語に関しては [*] というコードが出力される．

以下に出力の例を示す．

入力文 :

漢字を入力する。

出力結果 :

1. 漢字 (1100)/を (7430)
2. 入力する (2636)/。([P]0110)

この出力結果から，名詞「漢字」，格助詞「を」，動詞「入力する」，記号「。」と正しく解析されていることが分かる．

3　特徴的な会話文の収集

3.1　収集の目的と調査対象

会話特有の表現を持った文を収集する．ATR 音声翻訳通信研究所では音声翻訳研究のためにバイリンガル旅行会話コーパス [3, 4] を構築している．バイリンガル旅行会話コーパス内の会話は多くの人に利用可能なホテル予約を中心とした旅行会話が選ばれている．

本研究ではそのコーパス中から 3,206 事例を対象に分析を行う．事例はいくつかの文で構成されているので文単位で調査する．対象となる日本語文は 3,712 文である．以下に事例の１つを示す．

事例

かしこまりました。では、ご昼食で、三千円以下で、禁煙席をご希望ですね。

また，データの特性を表１に記す．

表 1: データ特性

事例数	3,206
１事例あたりの平均文数	31.5
最長事例文字数	115
最短事例文字数	5
日本語文数	3712
１事例あたりの平均文数	1.2
１文あたりの平均文字数	27.1
日本語文の最長文字数	82
日本語文の最短文字数	3

3.2　記述文と差異のある会話文の抽出方法

記述文との差異を持つ会話文は，記述文のための解析ツールでは正しく解析されないことが予想される．2 章で述べた記述文用の形態素解析システム ALT-JAWS を用いて会話文を解析し，解析に誤る会話文を特徴的な会話文として収集する．

解析誤りの判定は，形態素境界・文節境界の誤り，語義の誤りに注目して行う．ただし，漢字表記に直せば正しく解析される箇所や，固有名詞のために誤る箇所は，会話文特有の誤りではないため収集はしない．

3.3 収集結果

3.1 節で述べた文を ALT-JAWS で解析したところ，表 2 のようになった．

表 2: 形態素解析の結果

	成功	失敗	計
文数 (%)	2632(70.9%)	1080(29.1%)	3712

1 文中に複数の誤り箇所が含まれており，失敗した箇所の数を調べた結果を表 3 に示す．

表 3: 失敗した文数と失敗した箇所の数

失敗した文数	1080
失敗した箇所数	1195
1 文辺りの平均失敗数	1.11

以後の分析はこの 1195 箇所を対象に行う．以下に誤り箇所の例を示す．

入力文 :

本当はそのくらいがいいのですが、もしキャンセルがあれば取り替えていただけますか。

出力結果 :

1. 本当 (1240)/は (7530)
2. その (4200, 其の)/くらい (7520)/が (7410)
3. いい (3107, 良い)
4. の (1800)/です (7246)/が (7610)/、([P]0210)/も (7530)
5. し (2433, する)
6. キャンセル (1220)/が (7410)
7. あれ (2188, ある)/ば (7660)
8. 取り替え (2413, 取り替える)/ていただけ (2719, て頂く)
9. ます ([*]1100)/か (7700)/。([P]0110)

「もし」の形態素境界・文節境界が誤っており，さらに「ます」に [*] が出力されている．

4　会話文の特徴分析

この章では，3 章で収集した会話文特有の表現のため誤った箇所を分析する．

まず，4.1 節では品詞に着目し，本来どう解析されるべきかに分類する．次に 4.2 節で誤り箇所の単語が辞書に有るか無いかに分割して分類する．

4.1　品詞に着目した特徴の分類

形態素解析に失敗した品詞を表１にまとめる．今回，品詞の判定は [6],[7] および [8] の単語品詞分類を参考に行った.「複数の品詞」とは，本来複数の品詞として解析すべき所が一つの形態素として解析してしまった部分のことである．例えば「〜なんです」といった箇所では，助動詞「な」と形式名詞「ん」と解析すべき所を代名詞「何」と解析した．

表 4: 失敗箇所の品詞

品詞	数 (%)
複数の品詞	257(21.5)
助詞	163(13.6)
助動詞	155(13.0)
接続詞	118(9.9)
感動詞	108(9.0)
名詞	103(8.6)
補助動詞	91(7.6)
動詞	69(5.8)
形容詞	64(5.4)
副詞	40(3.3)
数詞	11(0.9)
連体詞	8(0.7)
接辞	6(0.5)
形容動詞	3(0.3)
計	1195

この結果，複数の品詞として解析されるべき箇所がもっとも多く 21.5%，続いて助詞が 1.6%，助動詞が 13.0%，接続詞が 9.9%，感動詞が 9.0%の順であった．表 5，表 6 に具体例を示す．

表 5: 複数の品詞

箇所	数	箇所	数
なん	73	風邪気味	2
、もし	23	夜着く	2
お〜くださいませ	22	約千円安い、	2
それでしたら	20	、くらい	1
ありがとうございます	17	、でしょうか	1
十分 (数詞)	15	、はい	1
ございまして	10	お電話かけていただく	1
なんでしょうか	8	こっから	1
食付き	8	この間会議場の	1
ご覧いただく	6	ご心配いりません	1
いいんじゃない	5	ご予約なさいます	1
のでしたら	5	じゃあ	1
いつから	2	ただやはり	1
お一つ	2	だいたい分かります	1
ご一名様	2	ちょっと見当たらない	1
ご覧ください	2	ていただけませんか	1
それともし	2	どうですか	1
ちょうどの	2	ねえ	1
のどのあたり	2	雨だし	1
もちろんなさいます	2	間に合うできるだけ	1
ようでしたらですね	2	左手の方すぐに	1
二分 (数詞)	2	食込み	1
風邪気味	2	**計**	257

表 6: その他の品詞

品詞	具体例 (数)	計
助詞	けど (99)，って (34)，かあ (6)，ね (6)，ねえ (5)，から (4)，かな (3)，て (1)，と (1)，として (1)，にて (1)，ので (1)，ばっかり (1)	163
助動詞	ます (126)，る (24) れる (2)，かもしれない (1)，で (1)	155
接続詞	では (116)，それじゃあ (2)	118
感動詞	ええ (58)，はい (34)，いえ (9)，いや (2)，うん (2)，ありがとう (1)，いえいえ (1)，まあ (1)	108
名詞	お一人 (32)，いくら (30)，日にち (7)，チェックイン (6)，満タン (5)，あたくし (4)，お二人 (4)，皆さん (4)，お出かけ (3)，フレンチ (2)，アメリカン (1)，イタリアン (1)，間欠泉 (1)，盗難届け (1)，紛失届け (1)，万一 (1)	103
補助動詞	てる (84)，でいただく (6)，とく (1)	91
動詞	お待たせ (20)，らっしゃる (11)，いただけ (4)，おあり (3)，お運びする (3)，チェックインする (3)，プラスする (3)，分かんない (3)，おっしゃい (2)，チェックアウトする (2)，プレーできる (2)，(薬) ください (2)，(時間) かかる (2)，(由緒) ある (1)，お見せできる (1)，ジョギングする (1)，チェックインできる (1) ツアーする (1)，(パーセント) 掛かる (1)，(円) 掛かる (1)，(分) 掛かる (1)	69
形容詞	申し訳ございません (25)，すいません (21)，すみません (7)，すみません (7)，申し訳ない (2)，とんでもございません (1)	64
副詞	実は (23)，いかが (7)，ごゆっくり (6)，とっても (2)，なんといっても (1)，パッと (1)	40
数詞	零 (11)	11
連体詞	どういった (8)	8
接辞	っかわ (4)，込 (2)	6
形容動詞	ほんとだ (3)	3

4.2 既知語／未知語に着目した特徴の分類

形態素解析を誤る原因は会話文でしか使われない単語があることが挙げられる．そこで単語が辞書に無いがために失敗した (未知語誤り) 箇所と，辞書に単語があるにもかかわらず失敗した (既知語誤り) 箇所とに失敗箇所を分けて，分析を進める．1,195 箇所の誤りの中で，未知語誤り箇所は 597 箇所 (50.0%)，既知語誤り箇所は 535 箇所 (44.8%)，そしてどちらとも取れず保留としたものは 63 箇所 (5.2%) であった．表 7，表 8 にそれぞれの誤り箇所を，表 9 に保留とした箇所を示す．

表 7: 未知語誤り箇所

箇所	出現数	箇所	出現数
では (接続詞)	116	うん (感動詞)	2
けど (助詞)	99	チェックアウトする (動詞)	2
てる (補助動詞)	84	申し訳ない (形容詞)	2
ええ (感動詞)	58	プレーできる (動詞)	2
って (助詞)	34	フレンチ (名詞)	2
いくら (名詞)	30	盗難届け (名詞)	1
実は (副詞)	23	アメリカン (名詞)	1
らっしゃる (動詞)	11	ツアーする (動詞)	1
いえ (感動詞)	9	見当たる (動詞)	1
どういった (連体詞)	8	リムジンて	1
る (助動詞)	8	万一 (名詞)	1
日にち (名詞)	7	紛失届け (名詞)	1
ね (間投助詞)	8	間欠泉 (名詞)	1
チェックイン (名詞)	6	イタリアン (名詞)	1
かあ (助詞)	6	チェックインなさる (動詞)	1
でいただく (補助動詞)	6	いえいえ (感動詞)	1
いいんじゃない	5	お見せできる (動詞)	1
満タン (名詞)	5	こっから	1
ねえ (助詞)	5	なんといっても (副詞)	1
あたくし (名詞)	4	ばっかり (助詞)	1
皆さん (名詞)	4	なんで	1
お運びする (動詞)	3	ジョギングする (動詞)	1
分かんない	3	まあ (感動詞)	1
っかわ (接尾語)	4	にて (助詞)	1
ほんとだ (形容動詞)	3	チェックインできる (動詞)	1
プラスする (動詞)	3	とく (補助動詞)	1
チェックインする (動詞)	3	として (助詞)	1
込 (接尾語)	2	じゃあ	1
いや (感動詞)	2	パッと (副詞)	1
それじゃあ (接続詞)	2		
とっても (副詞)	2	**計**	597

表 8: 既知語誤り箇所

箇所	出現数	箇所	出現数
ていただけます	79	見れる	2
なん	73	夜着く	2
はい	34	いつから	2
お〜いただけます	33	風邪気味	2
お一人	32	もちろんなさいます	2
、もし	23	おっしゃいました	2
お〜くださいませ	22	約千円安い、	2
それでしたら	20	二分	2
お待たせ	20	お一つ	2
ありがとうございます	17	パーセント掛かる	1
お〜いただける	16	だいたい分かります	1
十分	15	食込み	1
願えますか	13	分掛かります	1
零	11	この間会議場の	1
ございまして	10	と申しますのは	1
なんでしょうか	8	雨だし	1
食付き	8	お電話ありがとう	1
いかが	7	由緒ある	1
ご覧いただく	6	間に合うできるだけ	1
ごゆっくり	6	円掛かります	1
のでしたら	5	、はい	1
古くから	4	左手の方すぐに	1
お二人様	4	どうですか	1
いただけます	4	、くらい	1
おあり	3	お電話かけていただく	1
かなと思う	3	ご予約なさいます	1
お出かけ	3	結構ですので	1
それともし	2	今なら	1
薬ください	2	ご心配いりません	1
時間かかります	2	ただやはり	1
のどのあたり	2	ねえ	1
ご一名様	2	ていただけませんか	1
ご覧ください	2	、でしょうか	1
ちょうどの便	2	**計**	535

表 9: 保留の箇所

箇所	出現数
申し訳ございません	25
すいません	21
申し訳ありません	8
すみません	7
とんでもございません	1
かもしれません	1
計	63

表 9 の箇所をなぜ保留にしたか説明する.「申し訳ございません」「申し訳ありません」は形容詞「申し訳ない」を無理矢理丁寧に言った表現だと考えられるが文法上正しくなく正しく判断しづらいと考えたためである.「とんでもございません」も形容詞「とんでもない」を無理矢理丁寧に言った表現だと考えられるが文法上おかしく判断しづらい．他も同様である.

5　記述文との差異の吸収方法

5.1　仮想システム

記述文と会話文の差を吸収し，記述文用の形態素解析システムを用い会話文を機械翻訳するためのシステムとして図 1 のシステムを想定する．以下に処理の流れを説明する．

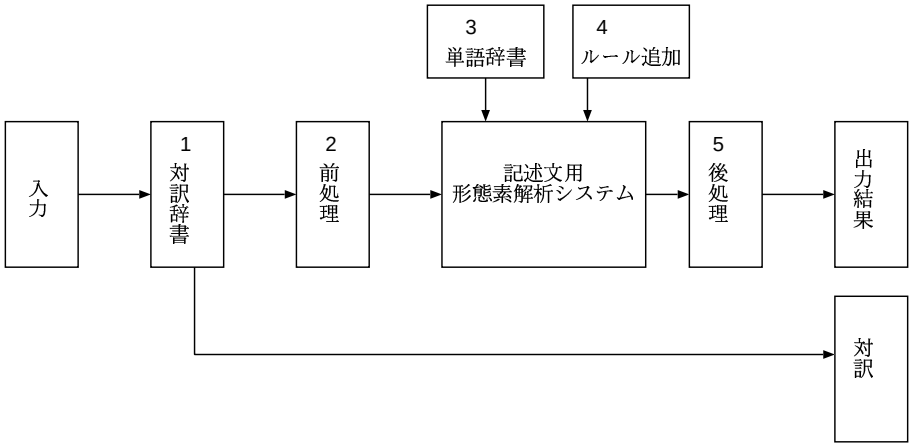

図 1: 機械翻訳のための仮想システム

- １は対訳辞書を作成し，入力文が定型的な表現ならば形態素解析すらせず直接翻訳してしまう処理である．
- ２は，会話文特有の単語だが，辞書にある単語で同様の意味をもつ単語が有れば，その単語に書き換えるいわば前処理の方法である．
- ３，４はそれぞれ，記述文用の形態素解析システムに単語を追加する方法と，会話文特有の単語の使われ方をルールに追加する方法である．
- ５は形態素解析の誤った出力結果を正しい結果に書き換えるいわば後処理の方法である．

5.2　対策を行う対象

4.2 節で誤り箇所を未知語誤り箇所と既知語誤り箇所に分けて分析した．このうち未知語誤り箇所は，今後辞書に登録するとことで対処できると考えられるので，既知語誤り箇所について考える．既知語誤り箇所は単語があるが会話文特有の使われ方をしているために誤った箇所であるので，5.1 節で示したシステムの３番のように会話文特有の使わ

れ方をルールに登録する方法や，５番の誤った出力結果を正しい出力結果に直す方法が考えられる．ところで，ALT-JAWSが未知語として [*] というコードを出力すると明確な誤りであると検出できる．この場合，後の処理に大きな影響が出るので対策が優先される．よって今回は既知語誤り箇所のうち，ALT-JAWS の出力結果に [*] が出力されて明らかに誤りと検出される箇所に関して，すなわち 5.1 節で述べた 5 番の方法で対策を考える．

5.3 対策

if-then ルールによって，誤った出力結果を正しく書き換える書き換え規則を作成する．以下に書き換え規則を作成する理由を説明する．

「していただけます」という箇所を例に取ると，ALT-JAWS の出力結果は次のようになる．

し (2433, する)/ていただけ (2719, て頂く)
ます ([*]1100)

これは，文節境界が誤っておりかつ,「ます」に [*] が出力され明らかに誤りと検出できる．この箇所は本来以下のように解析されるべきである．

し (2433, する)/ていただけ (2719, て頂く)/ます (7236)

ここから，動詞の後に補助動詞「ていただけ」がきて，その後文節境界で区切られ「ます」の [*] 明らかに誤りと検出したならば正しい品詞に書き換える規則が作成できると考えられる．また，他にも「教えていただけますか」等のように「動詞＋ていただけ＋ます」という箇所は同様の誤り方と直し方が考えられた．

これらのことをふまえ，先の例をルールにすると以下のようになる．

if V/ていただけ (2719, て頂く)\n ます ([*]1100)
then V/ていただけ (2719, て頂く)/ます (7236)

これにより「ます」を正しく助詞に直しかつ，文節境界を直すことができる．

5.4 作成したルール

既知語誤り箇所 535 箇所の中で 5.2 節で述べた対象となる箇所は 204 箇所 (38.1%) であり，誤り方の種類数は 25 種類であった．表 10 に誤り方を示す．

表 10: ルールの適用できる誤り箇所

箇所	出現回数 (%)
ていただけます	79(38.7)
お〜いただけます	33(16.2)
お〜くださいませ	22(10.8)
ありがとうございます	17(8.3)
願えますか	13(6.4)
ごゆっくり	6(2.9)
ご覧いただく	6(2.9)
いただけますでしょうか	4(2.0)
おあり	3(1.5)
夜着く	2(1.0)
見れる	2(1.0)
おっしゃいました	2(1.0)
薬ください	2(1.0)
ご覧ください	2(1.0)
と申しますのは	1(0.5)
結構ですので	1(0.5)
由緒ある	1(0.5)
お電話かけていただく	1(0.5)
、くらい	1(0.5)
ご予約なさいます	1(0.5)
、でしょうか	1(0.5)
左手の方すぐに	1(0.5)
ご心配いりません	1(0.5)
ですよねえ。	1(0.5)
お電話ありがとうございました	1(0.5)
計	204

また，これらの誤り箇所に対してルールを作成し，さらに以下のグループに分類した．

- 丁寧表現
- 体言＋動詞
- 読点
- その他

以下にルールグループとその具体例を示す.

- 「丁寧表現」グループ (15 ルール)

  ルール 1 V ていただけます

  if V/ていただけ (2719, て頂く)\n ます ([*]1100)/

  then ていただけ (2719, て頂く)/ます (7236)/

  ルール 2 お〜いただけます

  if いただけ (2719, 頂く)\n ます ([*]1100)/

  then いただけ (2719, 頂く)/ます (7236)/

  ルール 3 お〜くださいませ

  if くださ (2732, 瀉だす)(2732, 下す)\n い ([*]1100)/ませ (1410)/

  then ください (2789, くださる) ませ (7239, ます)/

  ルール 4 ありがとうございます

  if ありがとう (5200)\n ござ ([*]1100)/いま (1500, 今)\n す (2436, する)/

  then ありがとう (5200)\n ござい (2183, 御座い)/ます (7236)/

  ルール 5 願えますか

  if 願え (2399, 願う)\n ます ([*]1100)/か (7700)/

  then 願え (2399, 願う)/ます (7236)/か (7700)/

  ルール 6 ごゆっくり

  if ご ([*]1100)\n ゆっくり (4190)/

  then ご (6180, 御)/ゆっくり (4190)/

  ルール 7 ご覧いただく

  if ご ([*]1100)\n 覧 (2413, 見る)\n い (2213, 居る)\n ただ (4100, 唯)\n く ([*]1100)/

  then ご覧 (1100)/いただく (2719, 頂く)

  ルール 8 いただけますでしょうか

  if いただけ (2319, 頂く)\n ます ([*]1100)/でしょ(7241, です)/う (7266)/か (7700)

  then いただけ (2319, 頂く)/ます (7236)/でしょ(7241, です)/う (7266)/か (7700)/

  ルール 9 おあり

if おあり ([*]1100)

then お (1100, 御) あり (2183, ある)

ルール１０ おっしゃいました

if おっ(2384, 織る)\nしゃ([*]1100)/いま (1500, 今)\nし (2433, する)/た (7216)/

then おっしゃい (2183, おっしゃる)/まし (7234, ます)/た (7216)/

ルール１１ ご覧ください

if ご ([*]1100)\n覧 (2413, 見る)/くださ (2732, 瀉だす)(2732, 下す)\nい ([*]1100)

then ご覧 (1100)/ください (2789, くださる)

ルール１２ と申しますのは

if と ([*]1100)\n 申し (2333, 申す)/ます (7237)\n の (1800)/は (7530)

then と (7420)\n 申し (2333, 申す)/ます (7237)\n の (1800)/は (7530)

ルール１３ お電話かけていただく

if かけ (1410, 欠け)/てい ([*]1a00)\n ただ (5100)\n く ([*]1a00)

then かけ (2413, 賭ける)(2413, 掛ける)(2213, 駆ける)(2213, 欠ける)/ていただく (2716, て頂く)

ルール１４ ご予約なさいます

if /な (3100, 無い)/さ (6280))\n いま (1500, 今))\n す (2436, する))\nでしょう ([*]1100)/か (7700)

then なさい (2383)/ます (7236)/でしょ (7241, です)/う (7266)/か (7700)

ルール１５ ご心配いりません

if ご (6180, 御)/心配 (1240)/いり ([*]1100)/ませ (1410)/ん (1800, の)

then ご (6180, 御)/心配 (1240)\nいり (2183, 要る)(1420, 煎り)(1410, 要り)(1410, 入り)/ませ (7232, ます)/ん (7196, ぬ)

- 「体言＋動詞」グループ (3 ルール)

ルール１６ 夜着く

if 夜着 (1100)/く ([*]1a00)

then 夜 (1500)\n 着く (2116)

ルール１７ 薬ください

if 薬 (1100)/くだ ([*]1a00)/さい (1260, 最)/

then 薬 (1100)/ください (2589, くださる)/

ルール１８ 由緒ある

if 由緒 (1100)/ある ([*]1a00)

then 由緒 (1100)\n ある (2187)

- 「読点」グループ (2 ルール)

ルール１９ 、くらい

if 、([P]0210)\n くらい ([*]1100)

then 、([P]0210)\n くらい (7520)

ルール２０ 、でしょうか

if 、([P]0210)\n で (5100)\n し (2433, する)\n ょう ([*]1100)/か (7700)

then 、([P]0210)/でしょ (7241, です)/う (7266)/か (7700)

- その他 (5 ルール)

ルール２１ 見れる

if 見れ (2418, 見る)\n る ([*]1100)/

then 見 (2411, 見る)/れる (7126)/

今回のルールを適用し，正しい品詞に修正すれば上のようになる．しかし，動詞「見る」は上一段活用他動詞であり，上一段，下一段，カ変の動詞は本来「られる」がつくので、本来は「見 (2411, 見る)/られる (7126)/」に直すべきだと考えられる．

ルール２２ 結構ですので

if 結構です (3226, 結構だ)\n の ([*]1100)/で (7410)(7255, だ)

then 結構です (3226, 結構だ) ので (7640)(7255, のだ)

ルール２３ 左手の方すぐに

if 左手 (1100)/の (7410)\n 方 ([*]1100)

then 左手 (1100)/の (7410)\n 方 (1800)(1100)(1100)

ルール２４ ですよねえ

if で (5100)\n す (2436, する)\n よ ([*]1100)/ねえ (7700, ね)

then です (7246)/よ (7700)/ねえ (7700, ね)

ルール２５ お電話ありがとう

if あり ([*]1a00)/が (7410)\n とう (2393, 問う)

then ありがとう (5200)

5.5 机上シミュレーション

5.4 節で述べた 25 種類のルールを用い，誤った箇所に対して机上シミュレーションを行った．詳細は付録に示す．

この結果，明らかに誤りと検出される 25 種類の誤り方について机上シミュレーションで修正可能であることを示した．

6　おわりに

今回の研究は，記述文用の形態素解析システム ALT-JAWS を用いてバイリンガル旅行会話コーパスを解析し，誤った箇所を会話文特有表現のため誤った箇所とし分析を行った．分析の結果，50.0%は単語辞書には無い会話文特有の単語が原因であり，44.8%は単語辞書に単語があるが，会話文特有の使われ方をしているために失敗をしていることが分かった．

また単語が辞書にあるのに誤った箇所は，接続する単語によって決まった誤り方があり，それを if-then ルールによって正しく書き換える方法を考察した．これにより，単語に辞書があるのに誤った箇所の 38.1%を正しく直すことが出来ると期待できる．

今後の課題は書き換え規則の実装と実験，および残りの誤り箇所への対策と検討が挙げられる．

謝辞

本研究を進めるにあたり，種々の御助言をいただきました鳥取大学工学部知能情報工学科計算機工学講座池原研究室の池原悟教授，村上仁一助教授に心からお礼申し上げます．

また，終始に渡り御指導いただきました徳久雅人助手に深謝いたします．

その他，本研究に使用させて頂いた本の著者の方々，および様々な場面で御助力いただいた計算機工学講座池原研究室の皆様に深く感謝の意を表します．

参考文献

[1] 竹沢 寿幸, 白井 諭, 大山 芳之 : バイリンガル旅行会話コーパスに見られる話言葉の特徴分析, 情報処理学会研究報告, 2001-NL-141, pp.137-144, 2001.

[2] 松本 裕治, 伝 康晴 : 話し言葉の形態素解析, 情報処理学会研究報告, 2001-SLP-36, pp.9-14, 2001.

[3] T.Morimoto, N.Uratani, T.Takezawa, O.Furuse, Y.Sobashima, H.Iida, A.Nakamura, Y.Sagisaka, N.Higuchi, and Y.Yamazaki : A speech and language datebase for speech translation research, Proc. International Conference on Spoken Language Processing, pp.1791-1794, 1994.

[4] 竹沢 寿幸, 中村 篤, 隅田 英一郎 : ATR の会話音声翻訳研究用データベース, 音声研究, Vol.4, No.2, pp.16-23, 2000.

[5] 白井 諭, 横尾 昭男, 池原 悟, 奥山 信輔, 宮崎 正弘 : 多段解析法による日本語形態素解析の精度, 情報処理学会第 50 回全国大会, Vol.3, pp.37-38, 1995.

[6] 宮崎 正弘, 大山 芳史 : 日本文音声出力システムの言語処理方式, 情報処理学会論文誌, Vol.27, No.11, pp.1053-1061, 1986.

[7] 八巻 俊文, 大山 芳史, 白井 諭, 横尾 昭男 : 日英機械翻訳システム ALT-J/E の研究開発, NTT R&D, Vol.46, No.12, pp.1391-1398, 1997.

[8] 市川 孝, 見坊 豪紀, 金田 弘, 進藤 咲子, 西尾 寅弥 (編) : 現代国語辞典, 三省堂, 1988.

[9] 三省堂編修所 (編) : コンサイス外来語辞典, 三省堂, 1987.

[10] 財団法人新村出記念財団 : 広辞苑第五版, 岩波書店, 1998.

[11] 長尾 真 (編) : 自然言語処理, 岩波講座 ソフトウェア科学, 15, 岩波書店, 1996.

付録

1. 日本語文 3712 文とその形態素解析結果
2. 誤り箇所の分類
3. 机上シミュレーション