ＴＲ－ＳＬＴ－００２２

音声認識誤りが翻訳品質に与える影響の検討

Examination of the Effect on the Translation Quality by Speech Recognition Error

大田　健紘[†,‡]
Kenkoh Ohta
安田　圭志[†,‡]　菊井　玄一郎[‡]
Keiji Yasuda　Genichiro Kikui
[†]同志社大学　[‡]ATR 音声言語コミュニケーション研究所

2002 年 9 月 11 日

概要

現状の音声翻訳システムは，主に，音声認識システムと機械翻訳システムを統合することにより構築されている．本論文では，音声認識誤りが，バックエンドである機械翻訳システムへ与える影響を分析している．まず，音声認識システムＡＴＲ－ＳＰＲＥＣと，機械翻訳システムを統合したシステムによる翻訳結果の主観評価を実施し，音声認識誤りが含まれることによる機械翻訳品質の劣化の程度を調査する．次にＡＴＲ－ＳＰＲＥＣを用いて大量の認識結果を作成し翻訳を行う．そして，音声認識の正解系列を機械翻訳した結果と，音声認識結果を機械翻訳した結果をＤＰマッチングにより比較する．次に，この結果を用いてＤＰ距離を計算し，これと音声認識結果に含まれる認識誤りの特徴（品詞ごとの挿入誤り数，品詞ごとの削除誤り数など）を用いて回帰分析し，翻訳品質の劣化に関係している認識誤りについて示す．機械翻訳の評価法として主観評価（Ａ，Ｂ，Ｃ，Ｄの４ランク評価）を用いている．本研究で用いた翻訳システムはＤ３とＨＰＡＴである．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　:+81-774-95-1308

1. はじめに

現在までの音声翻訳システムの技術は，独立に研究が進められた音声認識の技術と機械翻訳の技術を単に統合することにより実現されているが，音声認識の品質と機械翻訳の品質を関連付ける研究はあまり行われていない．音声認識の研究では，単語正解精度（正解文の単語数から挿入誤り数と置換誤り数，削除誤り数を引いたものの正解文に対する割合）などの評価基準によりシステムの最適化が行われるが，このような方法が必ずしも最適な統合システムを与えるという保障はない．

しかし，現状では音声認識システムの評価基準として単語正解精度，単語正解率（正解文の単語数から挿入誤り数と置換誤り数を引いたものの正解文の単語数に対する割合），発話正解率（完全に正解文と一致した発話数の全発話数に対する割合）以外に適当な基準が存在しないため，機械翻訳からの最終出力品質を考慮に入れたシステム開発が行えていない．さらには，音声認識誤りが機械翻訳品質に与える影響についてすら，明らかにされていない．そこで，本研究では，機械翻訳の最終出力品質を考慮に入れた音声認識の評価基準を考案することを目的とし，音声認識誤りが機械翻訳品質に与える影響について分析する．

本研究では，音声認識システムＡＴＲ-ＳＰＲＥＣ（ATR-SPeech RECognition system）と機械翻訳システムＤ３（DP-match Driven transDucer）[1]，ＨＰＡＴ（Hierarchical Phrase Alignment for Pattern-based MT）[2]を用いて，音声認識結果，つまり機械翻訳システムへの入力と，機械翻訳結果との相関を調べている．これにより，どのような音声認識誤りが機械翻訳システムにとって強い悪影響を及ぼすかを明らかにすることができ，また将来的には，この知見を用いて，バックエンドの機械翻訳システムを考慮した音声認識システムの評価基準の考案に役立てることが出来ると考えられる．

本研究では，音声認識誤りを含む系列の翻訳結果を，翻訳の質の観点から評価した大量のデータが必要となる．これまでの機械翻訳の研究では，主に人手による主観評価や，翻訳自動評価[3][4][5]が用いられているが，主観評価は，人手を要するので非常にコストがかかるという問題があり，翻訳自動評価では，数百の文や発話からなるテストセットといった大きな単位での評価では主観評価と非常に近い評価結果が得られるものの，単一の文や発話を評価対象にした場合の性能は不十分である．そこで本研究では，音声認識誤りを含む系列の翻訳結果と音声認識の正解系列を翻訳した結果をＤＰマッチングにより比較することにより評価する．

2.では機械翻訳品質の評価手法である翻訳ランク評価法について述べる．さらに，ＤＰ距離が翻訳品質の劣化を表す指標として用いることが出来るかを，翻訳ランク評価法との関係を調べることにより確認する．3.では主観評価法を用いて音声認識誤りが含まれない発話を翻訳した場合と，音声認識誤りが含まれる発話を翻訳した場合を評価することにより，音声認識誤りが含まれることにより機械翻訳品質が劣化することを確認する．4.では

音声認識誤りを含む発話を大量に作成し，それらを翻訳し，この翻訳結果と正解系列をＤＰマッチングして得られるＤＰ距離を音声認識誤りの種類から回帰分析することにより，機械翻訳品質に影響を与える音声認識誤りを分析する．5.はまとめであり，本研究を総括すると共に今後の課題を述べる．

2.　機械翻訳の評価手法

2.1 翻訳ランク評価法

人手による評価手法である翻訳ランク評価法（以降，主観評価と呼ぶ）では，評価者の主観により，次の基準で４段階の翻訳品質のランク付けを行っている[6]．

Ａランク（完全訳）：訳文だけで全く問題なし．
Ｂランク（部分訳）：訳文は少し情報が欠けている．
Ｃランク（可能訳）：訳文はかなり情報が欠けている．
Ｄランク（不可訳）：訳文からは，情報が想像もできない．

人間の主観による評価であるため，異なる発話の翻訳結果間での評価基準の多少の揺れは生じる．しかし，一つの原言語発話に対して，音声認識誤りを含む系列を翻訳した結果と，音声認識誤りを含まない系列を翻訳した結果を同時に評価するため，一発話の翻訳結果内での評価の揺れはほとんど無いと考えられる．よってこの方法による評価は正確であるが，かなりの時間と労力を要するため，一度に大量の翻訳結果を評価することは出来ない．

本論文では，まず翻訳ランク評価法を用いて音声認識誤りが含まれることにより，機械翻訳品質が劣化することを確認する．

2.2　ＤＰマッチングによる評価法

ＤＰマッチングを用いて機械翻訳品質が劣化することを確認するためには，１つの原言語テキスト文に対し，翻訳の正解として１つの正解目的言語文を用意しておき，ＤＰマッチングを用いて以下に定義するＤＰ距離を計算し，この値を劣化の指標とする．

$$D_{DP} = \frac{s+i+d}{t} \qquad (3.1)$$

式（３．１）において，D_{DP}はＤＰ距離を表している．tは正解系列の翻訳結果の単語数，sは正解目的言語文とシステムによる翻訳結果をＤＰマッチングにより比較したとき

の置換語数（substitution），iは同様の方法により比較したときの挿入語数（insertion）とdは同様の方法により比較したときの脱落語数（deletion）である．ＤＰ距離は０から無限大までの値をとる．

この指標が，翻訳品質の劣化を表す指標として適切であるかを確認するために，前節で述べた人手による主観評価結果と比較する．図１に比較した結果を示す．図１は，横軸に音声認識の正解系列を翻訳した結果のランクと，音声認識誤り有りの系列を翻訳した結果のランク比較した場合のランク下降の程度を示しており，それに対応するＤＰ距離の平均が棒グラフで表され，エラーバーが標準偏差を表している．図１より，ランクの下降の程度が大きくなるほどＤＰ距離も大きくなっていることが確認できる．このことから，ＤＰ距離は翻訳品質の劣化を表す指標として適切であることが確認できた．

本論文では，機械翻訳品質に影響を与える音声認識誤りを分析する際に，大量の翻訳結果の評価が必要となるため，ＤＰ距離を用いる．

3. 音声認識誤りによる機械翻訳品質の劣化の主観評価を用いた分析

本章では，機械翻訳システムＤ３とＨＰＡＴによる翻訳結果が音声認識誤りにより，どのように影響を受けるのか確認する．具体的には，音声認識の正解系列と，音声認識誤りを含む系列をＤ３とＨＰＡＴにより翻訳する．そして，翻訳ランク評価法を用いて人手による主観評価を行う．音声認識誤りを含まない正解系列の翻訳結果を評価したものと，音声認識を含む系列の翻訳結果を評価したものを比較することにより，音声認識誤りを含む系列で評価結果が低下していることを確認する．認識に用いるテストセットは，旅行会話基本表現集（BTEC : Basic Travel Expression Corpus）[7]の３つのテストセット（set01, set02, set03）でありset01は５１０発話，set02は５０８発話，set03は５０６発話の合計１５２４発話である．

3.1 主観評価を用いたＤ３・ＨＰＡＴによる機械翻訳結果の評価

本節では，人手による機械翻訳結果の評価手法である主観評価を行った結果を示す．

主観評価を行う機械翻訳結果は，音声認識の正解系列を機械翻訳した１５２４発話と，音声認識誤りを含む系列を機械翻訳した１５２４発話，そして２種類の機械翻訳システムを用いているので合計６０９６発話である．ただし，音声認識が成功すれば正解系列と音声認識結果が一致するので，それらのランクは必ず一致する．そこで，そのような発話に関しては，正解系列の翻訳結果のみについて主観評価を行った．その結果，評価文の合計は３４３１発話になる．今回用いた音声認識システムの発話正解率は約７３％である．

Ｄ３による翻訳結果の評価結果を表１に示す．次にＨＰＡＴによる翻訳結果の評価結果を表２に示す．表１，表２は，各ランクの集計を行った結果である．

3.2　音声認識誤りと機械翻訳品質の関係

本節では，表１，表２の評価結果をもとに分析を行い，音声認識誤りが含まれることにより，機械翻訳品質が劣化することを確認する．

まず，音声認識の正解系列をＤ３，ＨＰＡＴを用いて翻訳を行った結果と，音声認識誤りを含む系列を同様に翻訳を行った結果の主観評価結果を比較する．翻訳がある程度うまくいったと判断できるＡランクとＢランクの合計と，翻訳がうまくいっていないＣランクとＤランクと翻訳不可（ＮＩＬ）の合計を比較すると，すべての翻訳結果において音声認識誤りを含む系列でＡ，Ｂランクの合計が減少し，Ｃ，ＤランクとＮＩＬの合計が増加していることから，翻訳品質の劣化を確認できる．

そこで，さらに詳しく音声認識誤りによるスコアの変動を調べるために，発話単位での比較を行った．表３に音声認識の正解系列の翻訳結果から見て，音声認識誤りを含む系列の翻訳結果の主観評価によるランクが上昇，下降，変化なしの数を示す．表３より，音声認識誤りを含む系列の翻訳結果の方が，正解系列の翻訳結果のランクよりも下降するはずであるが，ランクが上昇している発話が存在する．各テストセットでランクに変動があった発話数に対する，ランクが上昇した発話数の割合をみると，Ｄ３の場合が，１６３発話中２１発話であり，ＨＰＡＴの場合が，２１６発話中２０発話であり，ともに約１割である．このことから，音声認識誤りを含む系列の翻訳結果のほうがランクが良くなる状況は非常にまれなケースであるといえる．

よって，音声認識誤りの正解系列の翻訳結果よりも音声認識誤りを含む系列の翻訳結果のほうが，品質が劣化していることが確認できる．

次に，評価結果が変化しなかった発話に注目すると，評価結果が変化しないのは，以下の３つのケースであった．

１．音声認識の正解系列と音声認識結果が一致した場合．
２．音声認識の正解系列と音声認識結果が一致せず，翻訳結果が同じになる場合．
３．音声認識の正解系列と音声認識結果が一致せず，翻訳結果も異なるが同じランクになる場合．

表４に各テストセットでの１～３の発話数を示す．ここで，特に重要になるのは１である．１の場合というのは，音声認識システムの発話正解率に一致する．

最後に，表１と表２を比較することにより，Ｄ３とＨＰＡＴどちらの翻訳品質がよいかを調べる．まず各テストセットの正解系列の評価結果を比較すると，Ａ，Ａ＋Ｂ，Ａ＋Ｂ

＋Ｃすべてのケースにおいて ＨＰＡＴのほうが，割合は大きくなっている．このことから，音声認識誤りの含まれない正解系列の翻訳については ＨＰＡＴのほうが優れていることがわかる．次に音声認識誤り有りの結果を含めて考えると，set01 のＡの割合と set03 のＡの割合において Ｄ３と ＨＰＡＴの優劣が逆転しているが，これら以外では ＨＰＡＴのほうが優勢である．しかし，正解系列の割合と音声認識誤り有りの割合を比較すると，明らかに ＨＰＡＴのほうが劣化の度合いが大きいことがわかる．このことから，ＨＰＡＴは音声認識誤りに弱いということがいえる．

4．機械翻訳品質に影響を与える音声認識誤りの回帰分析

本章では，前章で明らかになった音声認識誤りと機械翻訳品質の劣化の関係から，さらに詳細な分析を行う．まず，音声認識システムである ＡＴＲ－ＳＰＲＥＣによる音声認識結果の作成法を説明する．次に，ＡＴＲ－ＳＰＲＥＣを用いて音声認識誤りを含む大量の認識結果を作成し，ＤＰマッチングを用いて認識結果を Ｄ３と ＨＰＡＴにより翻訳した結果の評価を行い，その結果を述べる．最後に，音声認識結果に含まれていた音声認識誤りの種類と，ＤＰマッチングによる評価結果を用いて回帰分析し，機械翻訳品質の劣化に関係する音声認識誤りの種類を明らかにした結果を述べる．

4.1　音声認識システム ＡＴＲ－ＳＰＲＥＣによる音声認識結果の作成

音声認識誤りを含む大量の認識結果を作成するために用いた ＡＴＲ－ＳＰＲＥＣについて説明する．音声認識誤りを含む大量の認識結果を作成する際に，さまざまなパラメータを変更することにより行った．変更したパラメータは言語モデル，音響モデル，ビーム幅の３つである．言語モデルは２-gram が２つ，３-gram が１つあり，２-gram のものを単独で使った場合が２種類と２-gram と３-gram を組み合わせて使った場合が２種類の合計４種類である．音響モデルは，読み上げ音声と対話音声で共に男女の４種類がある．ビーム幅は９０，１００，１０５，１２０の４種類である．よって１発話に対して６４種類の認識結果が存在する．そして実験に用いたテストセットは４章で述べたテストセットと同様であり，５１０発話が１つ，５０８発話が１つ，５０６発話が２つ存在するため，分析に使用できるデータは合計約１３万存在する．

4.2　ＤＰマッチングによる Ｄ３・ＨＰＡＴの出力品質の評価

本節では，３章で説明したＤＰ距離を用いて機械翻訳品質の劣化を確認するとともに，大量に作成した音声認識結果をＤ３，ＨＰＡＴにより翻訳した場合にどの程度ＤＰ距離が離れたものが含まれているかを示す．Ｄ３の平均は０．４９７でＨＰＡＴの平均は０．５２６である．このことからＨＰＡＴのほうが全体としてはＤＰ距離が離れている，すなわち翻訳品質が劣化している．しかし標準偏差を見るとＤ３が０．６４９で，ＨＰＡＴは０．６０９である．このことからＤ３のほうが広範囲に広がってＤＰ距離が分布していることが確認できる．

4.3　機械翻訳品質に影響を与える音声認識誤りの回帰分析

本節では，作成した大量の音声認識結果とＤＰマッチングによるＤＰ距離を用いて，音声認識誤りの種類がＤＰ距離にどのような影響を与えるかを明らかにする．

まず，作成した全データを用いて重回帰分析を行う．重回帰分析を行う際の独立変数としては，削除された品詞（名詞，動詞，助動詞，助詞，形容詞，副詞，連体詞，接続詞，感動詞，間投詞，接頭辞，接尾辞）と挿入された品詞である．そして，従属変数としてはＤＰ距離を用いた．分析を行うためのデータの中に含まれる音声認識誤りの割合を図２に，各音声認識誤りの品詞別の内訳を図３にまとめた．

次に，重回帰分析を用いて認識誤りがＤＰ距離に与える影響を調べた結果を示す．図４に削除による影響と，挿入による影響の重回帰式の係数を標準化したものを示す．図４より分析した結果を示す．

- 挿入による影響はＤ３とＨＰＡＴでは同じ傾向が見られた．
- 削除による影響と挿入による影響を比較すると，品詞が挿入されたときの方が翻訳結果に大きな影響を与えることがわかる．
- 翻訳品質の劣化に影響を与えている認識誤りの種類は，感動詞の削除，名詞，動詞，助動詞の挿入である．

以上の分析は全データを対象としていたが，４章において音声認識誤りが含まれることにより，その翻訳結果の品質は正解系列の翻訳結果の品質よりも劣化すると述べた．そこで，正解系列の主観評価結果がＣランクやＤランクだったものに対しては，音声認識誤りが加わったとしてもこれ以上悪くなることがないため，本章の分析データとしては不適切である．なぜなら，正解系列の翻訳結果の品質が悪いので，そこに認識誤りが加わった系列の翻訳結果とのＤＰ距離を計算しても，信頼性が低いためである．このことから，回帰分析をもっと厳密に行うために対象とするデータを４章での主観評価結果がＡランクとＢランクだった発話に限定して分析を行うことにする．厳密な分析を行うことにより，大きな影響を与える品詞とそうでない品詞が明確になると期待できる．

図５に主観評価結果が，Ａランクの発話とＢランクの発話のみを用いて分析した結果を示す．主観評価結果がＡランクとＢランクの発話のみを用いて分析を行った結果，より直感に合った結果が得られた．以下に図５を分析した結果を示す．

- 削除による影響はＤ３のほうが大きく，挿入による影響はＨＰＡＴのほうが大きくなっている．
- 図４では影響が小さかったＤ３場合の名詞，動詞，助詞の削除の影響が図５では大きくなっており，削除，挿入ともに影響を与える品詞の傾向はほぼ一致した．

4.4 考察

本章では，ＡＴＲ－ＳＰＲＥＣを用いて音声認識誤りを含む大量の認識結果を作成し，これらをＤ３，ＨＰＡＴを用いて翻訳した結果と，音声認識の正解系列をＤ３，ＨＰＡＴを用いて翻訳した結果のＤＰ距離が，含まれていた音声認識誤りの種類によりどのように影響を受けるかを回帰分析を用いて分析した．その結果，Ｄ３は品詞の削除に弱く，ＨＰＡＴは品詞の挿入に弱いとわかった．
さらに，ＤＰ距離に大きな影響を与えている認識誤りの種類は，名詞，動詞，助詞，感動詞の削除と名詞，動詞，助詞の挿入である．また，感動詞の削除はＤ３においては名詞の削除と同程度の影響を及ぼしていることがわかる．挿入においてはＤ３では約２倍，ＨＰＡＴでは約２倍以上，動詞よりも影響が強くなる結果が得られた．

回帰分析により機械翻訳品質への劣化を分析した結果，直感とは合わない結果であった感動詞の削除が機械翻訳品質の劣化に大きな影響を与えていることについて検討する．テストセットの中に含まれる感動詞を調べた．表５にテストセットに含まれていた感動詞を示す．表５に示したように，１発話が感動詞１単語で構成されている場合がある．そのため，１単語が認識誤りで削除されただけであったとしても，ＤＰ距離が大きくなってしまい翻訳品質に大きな影響を与えていると考えられる．

Ｄ３とＨＰＡＴでは翻訳に影響を与える音声認識誤りの種類の傾向がほぼ一致していたが，Ｄ３とＨＰＡＴで異なる点といえば，Ｄ３はＨＰＡＴより削除に弱く，ＨＰＡＴはＤ３より挿入に弱いということである．

5　まとめ

音声翻訳システムにおいて，個別に研究されてきた音声認識システムと，機械翻訳システムの間の関係を調べるためにまず，音声認識の正解系列と，音声認識誤りが含まれている系列をＤ３，ＨＰＡＴ，２つの翻訳システムで翻訳し主観評価を行った．この結果，発話単位で正解系列と音声認識誤りを含む系列のランクを比較したところ，スコアが変動し

た発話のうち約９割で音声認識誤りを含む系列でランクの低下が見られた．このことから，音声認識誤りが含まれることにより，ほとんどの発話でランクが低下することが確認された．また，音声認識誤りが含まれることによりＤ３，ＨＰＡＴによる翻訳品質は，約１割劣化がみられた．

機械翻訳結果に影響を与える音声認識誤りの種類を明らかにするために，音声認識結果に含まれる音声認識誤りの種類と，これらの翻訳結果と正解系列の翻訳結果のＤＰ距離を用いて，回帰分析を行った．その結果，削除による影響はＤ３の方が大きく，挿入による影響はＨＰＡＴの方が大きく，品詞が挿入されたときの方が翻訳結果に大きな影響を与えることが示された．また，翻訳品質の劣化に影響を与えている認識誤りの種類は，名詞，動詞，助詞，感動詞の削除と名詞，動詞，助詞の挿入であることが示された．特に名詞は削除においてはＤ３では約２倍，ＨＰＡＴでは約２．５倍以上，動詞，助詞よりも影響が強くなっている．また，Ｄ３においては，感動詞の削除も名詞の削除と同程度の影響があることがわかった．挿入においてはＤ３では約２倍，ＨＰＡＴでは約２倍以上，動詞よりも影響が強くなる結果が得られた．

謝辞

本研究を行う機会を与えてくださいました，音声言語コミュニケーション研究所　山本誠一所長，竹澤寿幸主任研究員，同志社大学工学部　柳田益造教授に感謝いたします．また，熱心に議論に応じてくださいました音声言語コミュニケーション研究所第二研究室の皆様に感謝いたします．

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

参考文献

[1] Sumita, E., “Example-based machine translation using DP-matching between word sequence”, Proc. ACL-2001 Workshop on Data-Driven Methods in Machine Translation, pp.1-8, 2001.

[2] Imamura, K., “Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT”, Proc. TMI 2002, pp.74-85, 2002.

[3] Papineni, K. et al., “Bleu: a Method for Automatic Evaluation of Machine Translation”, Proc. ACL, pp.311-318, 2002.

[4] NIST, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, http://www.nist.gov/speech/tests/mt/mt2001/resource/, 2002.

[5] 安田圭志，菅谷史昭，竹澤寿幸，山本誠一，柳田益造, 対訳コーパスを用いた翻訳品質自動評価法, 情報処理学会論文誌，Vol.43，No.7，pp.2108-2216, 2002.

[6] Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S., "Solutions to problems Inherent in spoken-language translation: The ATR-MATRIX, MT-Summit VII, pp.229-235, 1999.

[7] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S., "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world", Proc. LREC2002, pp.147-152, 2002.

表１　Ｄ３による翻訳の主観評価結果

	A	A+B	A+B+C
set01 正解系列	0.661%	0.804%	0.859%
set01 音声認識誤り有り	0.614%	0.737%	0.778%
set02 正解系列	0.657%	0.789%	0.819%
set02 音声認識誤り有り	0.612%	0.740%	0.766%
set03 正解系列	0.678%	0.781%	0.832%
set03 音声認識誤り有り	0.634%	0.731%	0.787%

表２　ＨＰＡＴによる翻訳の主観評価結果

	A	A+B	A+B+C
set01 正解系列	0.676%	0.882%	0.925%
set01 音声認識誤り有り	0.594%	0.767%	0.816%
set02 正解系列	0.709%	0.862%	0.894%
set02 音声認識誤り有り	0.656%	0.799%	0.835%
set03 正解系列	0.688%	0.822%	0.877%
set03 音声認識誤り有り	0.625%	0.745%	0.802%

表３　正解系列と音声認識誤りを含む系列の主観評価の変化

D3	set01	set02	set03	HPAT	set01	set02	set03
ランク上昇	11	4	6	ランク上昇	10	4	6
ランク変化無	437	463	460	ランク変化無	420	452	436
ランク下降	62	41	40	ランク下降	80	52	64

表4　ランクに変化が無かった発話の詳細

D3	set01	set02	set03
音声認識結果も 翻訳結果も同じ	363	399	377
音声認識結果が異なるが 翻訳結果が同じになる	34	31	38
音声認識結果が異なり 翻訳結果が異なる	40	33	45
HPAT	set01	set02	set03
音声認識結果も 翻訳結果も同じ	361	399	377
音声認識結果が異なるが 翻訳結果が同じになる	20	13	21
音声認識結果が異なり 翻訳結果が異なる	37	40	38

表5　テストセットに含まれる感動詞

一発話が感動詞だけの場合の語　：　おやすみなさい　ありがとう　すみません こんにちは　はい　おめでとう　いいえ　ごめんなさい　おはよう　どういたしまして ありがとうございました　失礼しました　かしこまりました
一発話内に感動詞以外の品詞が含まれる場合の語　：　もしもし　ああ　うん　さあ えー　どうぞ　あのう　そう　こんばんは　申し訳ありません　おや　あれ　そうね 申し訳ございません　えっ　もしもし　いえ　ごちそうさま　お陰様　ほれ　あらら おい　そりゃあ　恐れ入ります

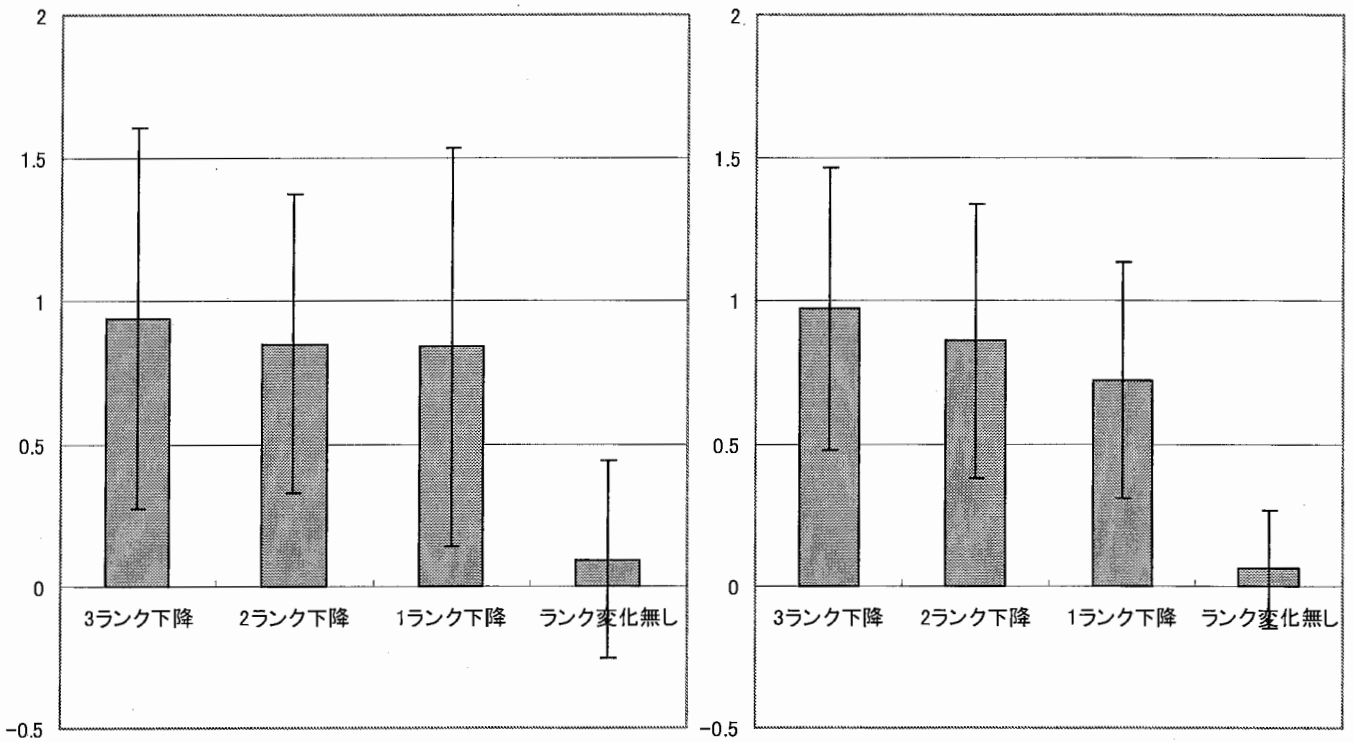

図１　主観評価結果とＤＰ距離の関係（左：Ｄ３　右：ＨＰＡＴ）

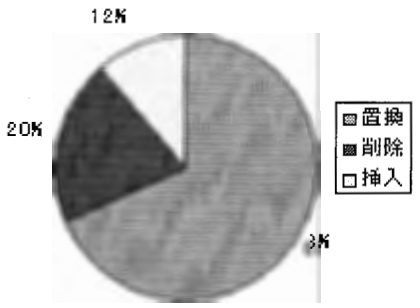

図２　含まれる認識誤りの割合

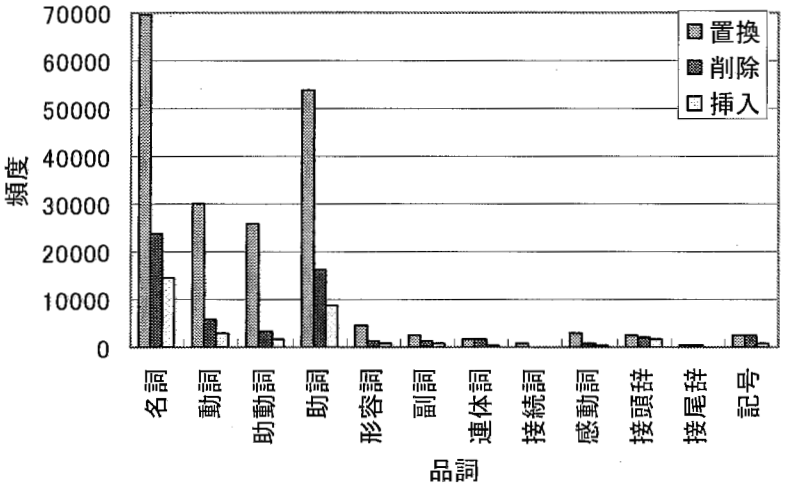

図３　各認識誤りの品詞別の内訳

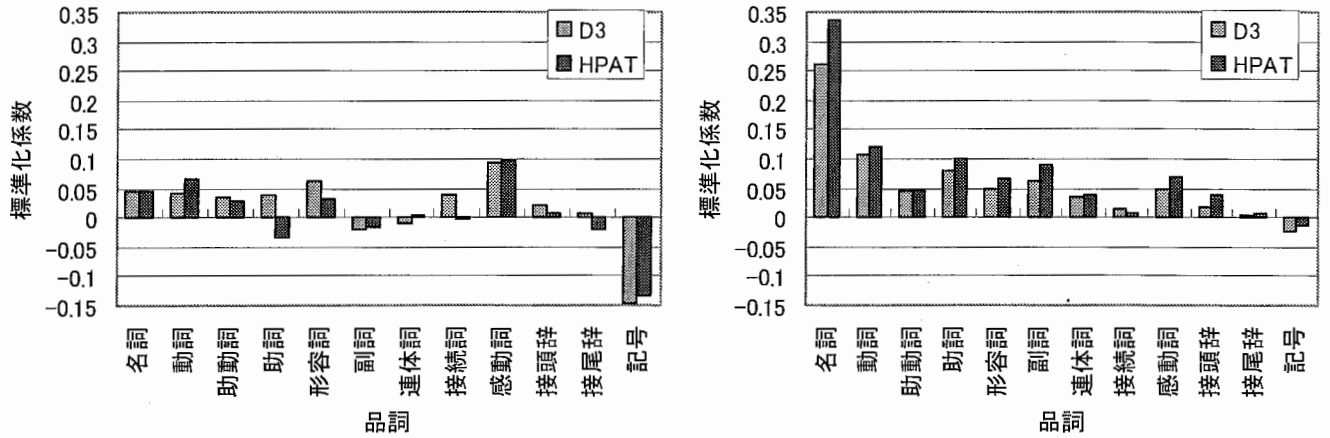

図４　全データでの削除による影響（左）と挿入による影響（右）

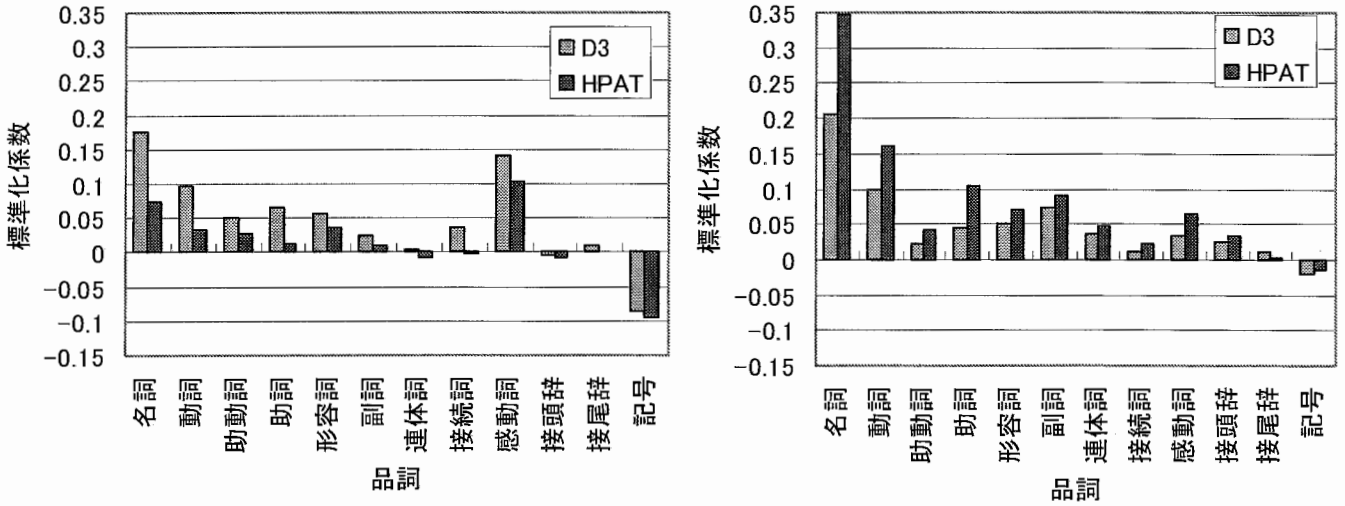

図５　正解系列の主観評価結果がＡランクまたはＢランクの発話での
削除による影響（左）と挿入による影響（右）