Vol.2012-NL-208 No.3
2012/9/2

統計的機械翻訳システムを利用した日本語音韻変化処理

小川 泰弘[1,a)]　外山 勝彦[1,b)]

概要：膠着語に分類される言語には，語幹に接辞が接続することにより語が構成されるという共通点があるが，音韻変化に関しては言語ごとの差異が大きい．そのため，膠着語の文生成においては言語に依存した音韻変化処理が必要である．従来は，言語ごとに規則を記述するルールベースな手法が採られてきたが，本研究では，日本語を対象として，統計的手法によりどの程度の精度が実現できるかを検証する．一方，言語学的な観点から見ると，そもそも語幹や接辞の基底形がどのような形になるかという問題がある．日本語においても複数の説が提唱されているが，どれが正しいかという決め手に欠けている．そこで，統計的な処理の立場から考えた場合，どの基底形が適切かという点についても考察する．

キーワード：音韻変化処理，日本語，統計的機械翻訳システム，派生文法，膠着語

Japanese Phonological Processing with Statistical Machine Translation Systems

Abstract: While words consist of stems and affixes in agglutinative languages, the differences among the languages with respect to phonological change are large. This result requires a specific phonological processing depending on each language for sentence generation in agglutinative languages. Although rule-based methods have been studied so far are, we adopt statistical methods and evaluate them in this paper. On the other hand, we have another linguistic problem to clarify what is a base form of a stem or an affix. Although Japanese linguists have proposed several theories about it, there is no discussion on which is right. Therefore, we also consider which a suitable base form is from the viewpoint of statistical processing.

Keywords: phonological change processing, Japanese language, statistical machine translation, derivational grammar, agglutinative language

1. はじめに

我々はこれまでに，日本語とウイグル語およびウズベク語との間の機械翻訳の研究に取り組んできた [1], [2], [3]．

日本語・ウイグル語・ウズベク語は，言語学においては膠着語に分類され，語順がほぼ同じであるなどの点で構文的類似性が高い．そのため，それらの言語間の機械翻訳においては，形態素解析の結果を逐語訳することによって，ある程度の翻訳が可能である．

しかし，音韻変化に関しては言語ごとの差異が大きい．そもそも，日本語とウイグル語の間には共通の語彙がほとんどない．ウイグル語とウズベク語には共通する語彙が多いが，ウイグル語には母音調和という音韻規則があるのに対し，ウズベク語にはそれが存在しないという差異がある．

そのため，膠着語間の機械翻訳においては，入力文解析および出力文生成のための音韻処理を言語ごとに個別に作成する必要がある．入力文に対する音韻変化処理は形態素解析の一部として処理されることが多いが，本稿では，出力文生成における音韻変化処理に着目する．

従来の音韻変化処理は，人手により規則を記述するルールベース手法が採られてきた．そこでは，音韻変化を基底形と表層形との間の変換処理と考え，例えば，2 レベル規則に基づく PC-KIMMO[4] では，人手で記述した規則を利用して双方向の変換が可能である．

それに対し本稿では，統計的機械翻訳の枠組を利用した音韻変化処理について検討する．多くの言語においてルー

[1] 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University, Japan
[a)] yasuhiro@is.nagoya-u.ac.jp
[b)] toyama@is.nagoya-u.ac.jp

ルベースの音韻変化処理は高い精度を達成できるが，統計的手法によりどの程度の精度が実現できるかを，今回は日本語を対象に検証する．統計的手法の見通しが付けば，言語ごとに規則を記述するというコストが下がるという効果も期待できる．

一方，言語学的な観点から見ると，そもそも語幹や接辞の基底形がどのような形になるかという問題がある．日本語においても複数の説が提唱されているが，どれが適切かという決め手に欠けている．そこで，言語学的な観点からではなく，統計的言語処理の立場から考えた場合，どの基底形が適切かという点について検討する．

本稿の構成は以下の通りである．まず 2 節では，本研究で使用する派生文法について概説する．3 節では基底形に関して，従来の研究の問題点と本研究での課題について述べる．4 節および 5 節は実験結果と考察について述べる．4 節では，翻訳モデルおよび学習モデルに使用するデータが与える影響について検討する．5 節では，様々な基底形を用意して，それらの比較実験を行う．6 節は本稿のまとめである．

2. 派生文法

日本語を音韻論的立場で記述する文法としては，Bloch[5] を嚆矢として様々な研究があるが，本稿では，助動詞を含めた接尾辞全体を統一的に記述している派生文法 [6], [7], [8] を基本とする．派生文法は日本語の膠着語としての性質に着目しており，ウイグル語などの他の膠着語にも応用可能であることから [1]，本稿の枠組を他の膠着語へ適用することも可能となると考えられる．以下，派生文法の概略を述べる．

2.1 連結子音と連結母音

動詞の不変化部分を語幹と呼ぶ．いわゆる一段活用動詞の場合，例えば「見る (mi-ru)」「食べる (tabe-ru)」は，不変化の部分「mi-」「tabe-」が語幹であり，その末尾は母音 i か母音 e である．このように，母音で終わる語幹を母音幹と呼び，母音幹を持つ動詞を**母音幹動詞**と呼ぶ．一方，いわゆる五段活用動詞「書く (kak-u)」の場合は，音韻論的に考えれば，「kak-」が不変化部分，すなわち語幹である．ただし，「買う (ka-u)」のようなワ行五段活用動詞の場合には「kaw-」を語幹とする．このように，末尾が子音である語幹を子音幹と呼び，子音幹を持つ動詞を**子音幹動詞**と呼ぶなお，日本語において子音幹動詞の末尾の音素は k，g，s，t，n，b，m，r，w の 9 種である．

派生文法においては，動詞の変形は動詞の語幹にいくつかの接尾辞が接続したものとして考える．そのため学校文法におけるほとんどの活用語尾や助詞・助動詞を接尾辞として扱う．それらと学校文法における活用形との対応を表 1 に示す．

表 1　動詞の語幹への接尾辞の接続例

活用形	子音幹の例	母音幹の例	接尾辞
未然形	kak-ana-i	tabe-na-i	-(a)na-i
	kak-are-ru	tabe-rare-ru	-(r)are-(r)u
	kak-ase-ru	tabe-sase-ru	-(s)ase-(r)u
	kak-ou	tabe-you	-(y)ou
連用形	kak-imas-u	tabe-mas-u	-(i)mas-(r)u
	kaϕ-ita	tabe-ta	-(i)ta
終止形	kak-u	tabe-ru	-(r)u
連体形	kak-u	tabe-ru	-(r)u
仮定形	kak-eba	tabe-reba	-(r)eba
命令形	kak-e	tabe-ro	-e/-ro
	kak-una	tabe-runa	-(r)una

ここで，終止形「kak-u」と「tabe-ru」の接尾辞はそれぞれ「-u」と「-ru」である．派生文法ではそれらをまとめて「-(r)u」と表記する．子音 r の有無は語幹の末尾に依存して決まる．例えば，語幹「kak-」に接尾辞「-(r)u」が接続した場合，子音の連続を避けるために，接尾辞の先頭の子音 r が欠落する．そのような子音を**連結子音**と呼ぶ．

一方，派生文法では表 1 に示すように，「書かない」「食べない」をそれぞれ「kak-ana-i」「tabe-na-i」と解析する．ここで，否定を表す接尾辞をまとめて「-(a)na-」と表す．先頭の母音 a は母音が連続する場合に欠落し，そのような母音を**連結母音**と呼ぶ．

以上から，派生文法においては，動詞と接尾辞の接続は以下の二つの規則にまとめられる．

接続規則 1： 連結子音を持つ接尾辞が子音幹に接続する場合，連結子音が欠落する．

接続規則 2： 連結母音を持つ接尾辞が母音幹に接続する場合，連結母音が欠落する．

以下，本稿ではこの二つの規則を**基本規則**と呼ぶ．

2.2 統語接尾辞と派生接尾辞

前節の否定の接尾辞「-(a)na-」は，「kak-ana-katta」のように，他の接尾辞が後接できる．このことは，動詞の語幹に接尾辞が接続することによって，新たな語幹が派生したと考えられる．そのような接尾辞を**派生接尾辞**と呼ぶ．

派生接尾辞に対して，新たな語幹を派生しない接尾辞を**統語接尾辞**と呼ぶ．前節の「-(r)u」はその例である．動詞の語幹に複数の接尾辞が接続する場合には，統語接尾辞が最後に接続する．派生文法のおける主な派生接尾辞および統語接尾辞を表 2 に示す．

2.3 規則変化と不規則変化

派生文法においては，基本規則によって日本語における用言の語形変化の多くを記述できるが，例外もある．本稿では，基本規則によって記述できる語形変化を**規則変化**とし，それ以外の変化を**不規則変化**と定義する．接尾辞のうち不規則変化するものは，表 2 において † で示した．

表 2　派生文法における動詞接尾辞

	役割	接尾辞	用例	
			子音幹	母音幹
	使役	-(s)ase-	kak-ase-	tabe-sase-
派	受身	-(r)are-	kak-are-	tabe-rare-
生	丁寧	-(i)mas-†	kak-imas-	tabe-mas-
接	可能	-e-†	kak-e-	–
尾	否定	-(a)na-	kak-ana-	tabe-na-
辞	希望	-(i)ta-	kak-ita-	tabe-ta-
	非完了終止	-(r)u	kak-u	tabe-ru
	完了終止	-(i)ta†	kaϕ-ita	tabe-ta
統	前望肯定	-(y)ou	kak-ou	tabe-you
	前望否定	-(u)mai	kak-umai	tabe-mai
	順接	-(i)	kak-i	tabe-
語	完了	-(i)te†	kaϕ-ite	tabe-te
	譲歩	-(i)temo†	kaϕ-itemo	tabe-temo
	却下条件	-(i)teha†	kaϕ-iteha	tabe-teha
接	開放条件	-(r)uto	kak-uto	tabe-ruto
	仮定条件	-(r)eba	kak-eba	tabe-reba
	完了条件	-(i)tara†	kaϕ-itara	tabe-tara
尾	否定	-(a)zu	kak-azu	tabe-zu
	同時	-(i)nagara	kak-inagara	tabe-nagara
	目的	-(i)ni	kak-ini	tabe-ni
辞	同時進行	-(i)tutu	kak-itutu	tabe-tutu
	前望譲歩	-(y)outo	kak-outo	tabe-youto
	命令	-e/-ro†	kak-e	tabe-ro
	否定命令	-(r)una	kak-una	tabe-runa

† は不規則変化する接尾辞である．

2.3.1　音便形

不規則変化の例としては，完了終止の統語接尾辞「-(i)ta」がある．例えば，「書く」の語幹「kak-」に接尾辞「-(i)ta」が接続する場合，基本規則に従えば「kak-ita」となるが，実際には末尾子音 k が欠落して「ka-ita」となる．これは音便と呼ばれる特別な語形変化であり，どのような変化を起こすかは末尾子音に依存する．その変化を表 3 に示す．なお，表中の ϕ は零記号で，接続によって ϕ に対応する音素が欠落したことを表している．なお，末尾子音が s の場合は，音便変化を起こさない．

「行く」と「問う」の音便形は，音便変化規則の例外である．「行く (ik-)」は語幹末子音が k であるから，「-(i)ta」が接続する場合には末尾の k が欠落して「ikita」になるはずであるが，実際には「itta」となる．同様に，「問う (tow-)」は「totta」となるはずであるが，実際には「touta」となる．

こうした音便変化は (i)t で始まる完了の統語接尾辞「-(i)te」，譲歩の統語接尾辞「-(i)temo」などでも起きるが，同時進行の「-(i)tutu」や，希望の派生接尾辞である「-(i)ta-」は音便変化を起こさない．特に，派生接尾辞「-(i)ta-」と統語接尾辞「-(i)ta」は，同形ではあるが語形変化が異なる点に注意が必要である．

表 3　動詞の音便形

語幹末子音		音便形		語例	備考
-k	-(i)ta	-ϕ	-ita	書いた	イ音便
-g	-(i)ta	-ϕ	-ida	泳いだ	イ音便＋連濁
-r	-(i)ta			切った	
-t	-(i)ta	-t	-ϕta	立った	促音便
-w	-(i)ta			買った	
-b	-(i)ta			飛んだ	
-n	-(i)ta	-n’	-ϕda	死んだ	撥音便＋連濁
-m	-(i)ta			読んだ	
-s	-(i)ta	-s	-ita	貸した	音便変化なし
ik	-(i)ta	it	-ϕta	行った	例外
tow	-(i)ta	to	-uta	問うた	

2.3.2　その他の不規則変化する接尾辞

命令形を形成する場合，子音幹には「-e」，母音幹には「-ro」という統語接尾辞がそれぞれ接続する．そこで，本稿ではこれを不規則変化する接尾辞として扱う．

可能の派生接尾辞「-e-」は子音幹にしか接続しないという点で不規則変化する接尾辞と言える．

丁寧の派生接尾辞「-(i)mas-」は，否定形を形成する場合に「-en’」という特別な接尾辞が接続する．さらに前望肯定の統語接尾辞「-(y)ou」が後接する場合，連結子音の y が欠落せず「kak-imas-you」となるので，不規則変化する接尾辞である．

2.3.3　不規則動詞

学校文法と同様に，派生文法でも「来る」と「する」を不規則動詞と考える．派生文法では，「来る」の語幹を「ko-」と考え，後接する接尾辞によって「k-」「ku-」に変化するとしている．同様に「する」の場合は，語幹が「se-」であり，後接する接尾辞によって「s-」「su-」「si-」に変化するとしている．

「来る」「する」以外にもいくつかの不規則動詞が存在する．「なさる」の語幹「nasar-」に丁寧の派生接尾辞「-(i)mas-」が接続すると，語幹末尾の r が欠落し「nasa-imas-」となる．また，命令形を形成する場合には「-e」でも「-ro」でもなく，「-i」という接尾辞が後接する．この種の動詞には他に「下さる」「仰る」「いらっしゃる」「ござる」があり，派生文法ではこれらを**変則動詞**と呼ぶ．

また，2.3.1 節で示した通り，「行く」「問う」は，音便形が特殊という点で不規則動詞である．

以上の 9 個の不規則動詞に加えて，本稿では語幹末尾が w の子音幹動詞も不規則変化すると考える．他の子音幹動詞の場合，例えば「kak-」に「-(e)ba」が接続するとそのまま「kakeba」となるが，語幹末尾が w の場合，例えば「omow-」に「-(e)ba」が接続すると「omoweba」とはならず，末尾の w が欠落して「omoeba」となる．一方，先頭が a で始まる接尾辞の場合は，「omow-」に「-(a)zu」が接

続して「omowazu」になるように，w は欠落しない．すなわち，a 以外で始まる接尾辞が接続した場合に w が欠落するという規則があるが，基本規則には含まれていない．以上のことから，語幹末尾が w の動詞を不規則動詞とする．

3. 基底形

音韻論の枠組においては，単語形成における音韻変化は，形態素の基底形が連続すると互いの影響を受けて変化し，表層形として出現すると考える．派生文法にこの枠組を適用すると，「tabe」に「ita」という基底形が後接すると i が欠落するという変化を経て表層形が作られることになる．しかし，何が基底形になるかという点に関しては，様々な説が提案されている．

例えば，Bloch[5] は過去（完了）を表す接尾辞を「-ita」ではなく「-ta」としている．また，派生文法では「来る」の語幹を「ko-」としているが，なぜ「ki-」や「ku-」ではないのかという点に関しては説明がない．

このような問題は他の言語にもある．ウイグル語にも派生文法でいう連結子音・連結母音が存在する [1] が，日本人向けのウイグル語の文法書 [9] では，連結子音・連結母音がない形を基底形と考え，例えば，語幹末が子音の動詞に，子音で始まる接尾辞が接続する場合，対応する母音が挿入されると考える．

このように，何が基底形になるかについては種々の考え方があるが，どれが良いのかという点に関しては，評価が容易でない．言語学においては，基底形を示しても，それが基底形である理由を示しているものはない．

言語処理の観点から言えば，音韻変化処理をルールベースで処理した際に，どのくらいの規則が必要になるか，もしくは規則の書き易さなどで評価できる可能性がある．しかしその場合には，基底形が変わるごとに規則を書き直す必要があり，労力が大きい．

それに対して本稿では，統計的な音韻変化処理の立場から，どの基底形が良いといえるかを評価する．すなわち各種の基底形を用意して統計的な音韻変化処理の実験を行い，その精度を比較する．

なお，表層形と表記は必ずしも一致しないが，本稿では日本式ローマ字表記を表層形と考える．

4. 音韻変化規則の学習実験

統計的機械翻訳用に公開されている各種のツールを使用して，統計的音韻変化処理について実験した．この場合，音素を統計的機械翻訳における単語，動詞句を統計的機械翻訳における文と見做すことになる．本研究では，翻訳モデルの学習に GIZA++[10]，言語モデルの学習に SRILM[11]，デコーダに Moses[12] を使用した．

それぞれの処理においては，使用する訓練データ，言語モデル，基底形など，様々なパラメータがある．これらを比較するために数多くの実験を行った．すべてのパラメータの組み合わせを実験することは容易でないため，本節以降では，本研究において最高の結果が出たものを比較元とし，比較対象となるパラメータのみを変化させた比較実験の結果を示す．そのため，これ以降のそれぞれの実験において最高の結果となったものは，いずれも同じパラメータ，同じ精度である．

まず本節では，訓練データや言語モデルを変化させることにより，どの程度の精度で音韻変化処理が達成できるかについて述べる．採用した基底形については，次節で述べる．

4.1 訓練データの比較実験

統計的機械翻訳では，翻訳モデルの学習のために，原言語文と対象言語文のペアを訓練データとして用いる．音韻変化処理の場合は，深層形と表層形のペアを訓練データとして用いることになる．本研究では，4 種類の訓練データを用意して比較実験を行った．

一般に，統計的機械翻訳では，パラレルコーパスからランダムに抽出した対訳文を訓練データとして用いることが多い．そこで，本研究でも同様に EDR コーパス (1.5 版)[13] から抽出したデータを 2 種類用意した．

まず，訓練データ 1 として，EDR コーパスからランダムに 1,000 文を抽出し，そこに出現した動詞句をすべて抽出したものを用意した．なお動詞句の抽出においては，サ変動詞の場合はサ変名詞の部分を除いて，「する」の語形変化の部分だけを使用した．また「書いていた」などの補助動詞が接続する場合は，分割してそれぞれ別の動詞句とした．さらに，同じ動詞句は複数回抽出した．その結果，コーパス 1,000 文から 2,580 個の動詞句が抽出できた．

訓練データを増やしたときの効果を検討するため，訓練データ 2 として，コーパス全 208,156 文に出現したすべての動詞句を使用したものも用意した．この場合，531,685 個の動詞句が抽出できた．

しかし，実はコーパス中に出現する音素の分布には偏りが大きいため，ランダム抽出では学習できない音素の組合せが発生する．子音幹動詞の場合，末尾の音素は k，g，s，t，n，b，m，r，w の 9 種であるが，語幹末尾が n となる動詞は口語では「死ぬ」1 語しかなく，今回使用した EDR コーパス中の動詞の出現比率において，0.09%を占めるだけである．また，語幹末尾が g となる動詞の出現も少なく，出現比率は 0.50%にすぎない．

そこで，コーパスから抽出したデータとは別に，動詞と接尾辞の組合せから網羅的に生成した訓練データを作成した．まず，訓練データ 3 として，動詞の語幹末尾ごとに 5 個ずつの動詞を用意した．この 5 個は EDR コーパス中の出現頻度が高いものから順に用いた．語幹末尾は子音幹動詞で 9 種，母音幹動詞で 2 種となるが，前述のように末

尾が n の動詞は「死ぬ」しかないため，これに関しては 1 個だけである．これに，不規則動詞として「来る」「する」「行く」「問う」「なさる」「下さる」「おっしゃる」「いらっしゃる」「ござる」の 9 個を追加し，合計 60 個の動詞を学習データとして採用した．また，接尾辞については，表 2 に示したすべての統語接尾辞を登録した．さらに派生接尾辞には，それに後続する接尾辞も複数追加した．その結果，接尾辞として合計 32 個を用意した．学習データは，動詞と接尾辞の組合せであり，合計 1,920 個となる．以降では，この訓練データ 3 を**組合せデータ**と呼ぶ．

さらに，接尾辞は同じままで，抽出する動詞の数を各 10 個に増やしたデータも用意し，訓練データ 4 とした．

学習された翻訳モデルの性能評価には，3 種類のテストを用意した．まず，オープンテストとして，EDR コーパスからランダムに抽出した 1,000 文中の動詞句 (平均 2,536 個) 10 セットを評価し，その精度の平均を求めた．なお，オープンテストであるから，評価に使用したデータは訓練データ 1 とは別のセットである．ただし，訓練データ 2 は EDR コーパスの全文を使用しており，オープンテストに使用したデータも含んでいるため，厳密にはオープンテストではない．表 4 における右端の「open」の欄にこの精度を記載した．

さらに，オープンテストでは音韻規則が網羅的に獲得できているかを評価できない．そこで，動詞と接尾辞の組合せから網羅的に作成した訓練データ 3 を評価テストに使用した．よって，この評価は訓練データ 3 に対してのみクローズドテストとなる．また，基本規則がどの程度獲得できたかを確認するため，規則変化する動詞と接尾辞の組合せだけに限定した場合の精度も求めた．それぞれの精度は，表 4 において「組合せ all」「組合せ regular」の欄に記載した．なお，本稿においては，末尾が w の動詞を規則変化から外したが，一方で，可能の派生接尾辞は「-e-」ではなく「-(r)e-」とし，規則変化する接尾辞とした[*1]．

また，訓練データを学習するときのオプションは予備実験により，alignment に関しては grow，reordering に関しては phrase-monotonicity-bidirectional-f-collapseff を採用した．

4.2 訓練データ比較実験の結果および考察

訓練データの比較実験の結果を表 4 に示す．size は訓練データとして使用した動詞句の数である．

オープンテストの結果を比較すると，EDR コーパスを使用した訓練データ 1 および 2 の精度が高いことが分かる．オープンテストの結果が良いということは，良く出現する動詞句は正しく処理できるということである．しかし，組合せデータに対する結果を見ると，訓練データ 1 の精度が低く，基本規則であっても学習されていないものが多い．これはコーパス中に末尾が n および g の動詞の出現が少なかったため，これらに関する規則が学習されなかったことが原因である．使用するコーパスデータを増やした訓練データ 2 では，組合せデータに対する精度も向上するが，それでも学習できない基本規則が残る．

表 4　訓練データの比較

	data type	size	精度		
			組合せ all	組合せ regular	open
1	EDR 1,000 文	2,580	79.4%	88.6%	95.1%
2	EDR 全文	531,685	92.0%	98.3%	96.9%
3	**組合せ (5)**	**1,920**	**94.4%**	**99.8%**	**93.4%**
4	組合せ (10)	3,520	94.1%	99.8%	93.8%

このことから，音韻規則を網羅的に学習するためには，動詞と接尾辞の組合せで訓練データを生成するのが良いことが分かる．これ以降も，規則を網羅的に学習できたかを重視するため，オープンテストの結果よりも，組合せデータを使用したテストの精度を重視する．組合せデータを用いたテストの精度において差が小さく，オープンテストの結果に差がある場合は，オープンテストの結果が良いものを採用する．

なお，訓練データ 3 では，使用した動詞の数は各 5 個であったが，これを各 10 個に増やた訓練データ 4 と比較すると，精度はほとんど変わらない．このことから，動詞の数は各 5 個でも充分であると言える．よって，今後の実験では訓練データ 3（組合せデータ）を用いる．

なお，基本規則に関する規則でも精度 100%を達成できていないが，これは 2.3.2 節で述べた「-(i)mas」に「-(y)ou」が後接したときに「-(i)mas-you」となる不規則変化が原因である．これが訓練データにあるため，子音 s の後に連結子音 (y) がきても欠落しない場合が発生し，精度を下げている．訓練データを基本規則に関するものだけにすれば，多くの場合，100%の精度が達成できる．ただし，後述の言語モデルにも依存するため，言語モデルによっては，100%にならない場合がある．

4.3 言語モデルの比較実験

統計的機械翻訳の言語モデルには N グラムが用いられる．今回の実験では，言語モデルとして，最初に 12 種類のモデルを用意して比較実験を行った．その後，さらに N グラムの大きさを比較するための 3 種類のモデルを追加して比較した．

一般に，統計的機械翻訳においては，言語モデルの学習に使用するデータの量が多ければ多いほど性能が良くなると言われている．そこで，まず単純なモデルとして，EDR コーパスに出現した全文節（動詞句以外も含む）を利用するモデル 1 を用意した．

[*1] いわゆる「ら抜き言葉」に対応する．つまり本稿の枠組においては，「ら抜き言葉」は規則変化だと考える．

ただし，これには動詞句以外も含まれることから，それらを除去したモデル 2 を用意した．また，言語モデルの量を減らし，EDR1,000 文をランダムに抽出し，そこに含まれる動詞句だけからなるモデルを 2 種類用意し，それぞれモデル 5，モデル 8 とした．

なお，ここまでのモデルでは，同じ動詞句が出現した場合，重複を除去している．しかし，重複を除去しない方が学習データの量が増えることからモデル 2，モデル 5，モデル 8 に対して，重複を除去しないモデルを用意し，それぞれモデル 3，モデル 6，モデル 9 とした．

また，N グラムに関しては，モデル 2，モデル 5，モデル 8 はすべて 2-gram だけを使用している．それと比較するため，2-gram に 3-gram を加えたモデルを用意し，それぞれモデル 4，モデル 7，モデル 10 とした．

訓練データの比較実験では，EDR コーパスから作成したデータよりも動詞と接尾辞の組合せを網羅的に作成した方が性能が高かったことから，訓練データ 3 と訓練データ 4 を言語モデルの学習に使用し，それぞれモデル 11 とモデル 12 とした．

以上の 12 種類の言語モデルを比較する実験を行った．

性能評価においては，訓練データの比較実験と同じデータを使用した．今回は訓練データに組合せデータを使用しているので，組合せデータに対する精度評価がクローズドテストとなる．また，訓練データの場合と同様に基本規則に関するデータだけに対する精度も求めた．それぞれの精度は表 5 の「closed regular」および「closed all」の欄に記載した．

オープンテストは訓練データの比較実験と同様に，EDR コーパスからランダムに抽出した 1,000 文中の動詞句 10 セットを評価し，その精度の平均を求めた．この 10 セットは，言語モデルとは別にコーパスから抽出した．ただし，訓練データの比較実験と同様に，モデル 1 からモデル 4 までは EDR コーパス中の全文を使用しているため，厳密にはオープンテストではない．

4.4 言語モデル比較実験の結果および考察

言語モデルの比較実験の結果を表 5 に示す．まずモデル 1 においては，「closed regular」に示された基本規則に関する精度が他よりも低い．実際の出力結果を見てみると，連結母音が欠落する規則が学習されていない場合があった．これは，いわゆる和語動詞からなる動詞句には母音の連続がほとんどないことが原因と考えられる．動詞句において母音の連続があるのは，語幹末尾が w の動詞に接尾辞が接続した場合など例外的な状況であるのに対して，漢字で構成される単語内には母音の連続が比較的存在する．今回は動詞句の語形変化を学習するのであるから，言語モデルにも動詞句の部分だけを使うのが適切だと言える．よって，それ以外のモデルでは動詞句だけに限定して言語モデルを学習している．

表 5 言語モデルの比較

	data type	size	精度		
			closed all	closed regular	open
1	EDR 全文節	281,263	93.9%	98.6%	88.6%
2	EDR 全文	22,614	94.1%	99.8%	89.9%
3	同 重複	531,685	94.0%	99.8%	92.2%
4	同 3-gram	22,614	92.0%	99.0%	92.1%
5	EDR 1,000 文 a	1,022	90.9%	99.9%	91.2%
6	同 重複	2,580	90.7%	99.6%	91.9%
7	同 3-gram	2,580	87.7%	97.0%	91.7%
8	**EDR 1,000 文 b**	**1,063**	**94.4%**	**99.8%**	**93.4%**
9	同 重複	2,594	93.9%	99.6%	92.1%
10	同 3-gram	1,063	91.5%	97.7%	91.7%
11	組合せ (5)	1,899	94.4%	99.3%	87.3%
12	組合せ (10)	3,466	94.4%	99.5%	88.6%

表 6 N グラムの比較

N-gram	精度		
	closed all	closed regular	open
2-gram	**94.4%**	**99.8%**	**93.4%**
3-gram	91.5%	97.7%	91.7%
4-gram	90.7%	96.4%	92.7%
5-gram	90.0%	96.1%	91.4%
6-gram	90.0%	96.1%	91.4%

モデル 2，モデル 5，モデル 8 は EDR コーパスの動詞句から生成した言語モデルである．これらを比較した場合，精度が一番高かったのが，コーパス 1,000 文から作成したモデル 8 であるが，同じく 1,000 文から作成したモデル 5 との差が大きい．また，モデル 8 よりデータ量が多いモデル 2 では，モデル 8 より精度が下がっている．一般に，統計的機械翻訳では，言語モデルの学習データを増やすと性能が向上するが，今回の実験では，そのようなことは言えない．これは，表層形の記述に使用する文字種が限られていることから，ある程度の大きさの学習データがあれば，それなりのモデルが学習でき，それ以上データを増やしても効果がないことを示している．

次に，モデル 2 とモデル 3，モデル 5 とモデル 6，モデル 8 とモデル 9 の比較は，動詞句の重複を許すとどうなるかの比較である．多くの場合，オープンテストの精度は良くなるが，クローズドテストの精度は悪くなった．今回は網羅的な規則が獲得できるかという観点からクローズドテストを重視しているため，重複は取り除いた方が良いと言える．

モデル 2 とモデル 4，モデル 5 とモデル 7，モデル 8 とモデル 10 の比較は，N グラムの比較である．2-gram だけを使用した場合と 3-gram も加えた場合とでは，いずれの場合も，前者の方が精度が高かった．特にモデル 8 に関して

表 7　連結子音・連結母音の扱いの比較

連結子音・連結母音の表現方法	精度		
	closed all	closed regular	open
大文字	**94.4%**	**99.8%**	**93.4%**
小文字	89.9%	97.0%	92.4%
削除	81.5%	87.6%	87.3%

は，さらに 6-gram まで用いる実験を追加した．その結果を表 6 に示す．表 6 からも明らかなように，N グラムとして 2-gram だけを使用した場合の精度が高い．これは，日本語の音韻変化においては，多くの場合，隣接する音素にのみ影響を受けるためだと考えられる．ウイグル語のように，離れた音素にも影響を与える母音調和がある言語においては，N の値を大きくした方が精度の向上に繋がる可能性がある．

モデル 11 とモデル 12 は組合せデータを言語モデルの学習に使用したものである．訓練データの比較実験においては，組合せデータを用いた方がコーパスから抽出したデータを使用する場合より精度が高かったが，学習モデルの場合はそういうことはなかった．

以上のことから，ある程度のサイズの言語モデルがあれば，言語モデルを大きくしても効果がなく，その差は偶然によるところが大きいといえる．

今回の実験では，結果が一番良かったのはモデル 8 であるので，これを採用した．

5. 基底形の比較実験

本節では，基底形の選択によって，音韻変化処理の性能がどのように変化するかを検討するため，基底形に関する比較実験を行った．評価に関しては，言語モデルの比較実験と同じデータを用いてクローズドテストおよびオープンテストを行った．

5.1 連結子音と連結母音の扱いの比較

まず，連結子音と連結母音をどのように扱うかを検討した．本稿では (i) のように括弧をつけて表現したが，統計的機械翻訳システムでは 1 個の音素を 1 文字で表現する必要がある．そこで，以下の三つの表現方法を考え，それらを比較した．

大文字: 連結子音・連結母音の音素に，それぞれ対応する大文字を使用．例えば「-(r)u」の基底形は「Ru」.

小文字: 連結子音・連結母音を他の音素とは区別せず，例えば「-(r)u」の基底形は「ru」のまま．

削除: 連結子音・連結母音がない形を基底形とし，接続の際に対応する音素が挿入されると考える．例えば「-(r)u」の基底形は「u」.

実験結果を表 7 に示す．これから明らかなように，連結子音・連結母音を大文字で表現する方法が一番精度が高かった．大文字で表現するのは連結子音・連結母音を通常の音素と区別することになり，小文字のまま区別しないよりも性能が良くなったと考えられる．また連結子音・連結母音を削除した場合は，語幹と接尾辞が接続する際に，必要に応じて音素を挿入する必要があるが，どの音素を挿入するかの選択が必要になり，そのために精度が下がったと考えられる．

以上を踏まえて，本研究では，連結子音・連結母音を大文字で表現する方法を採用した．

表 8　不規則動詞「来る」の基底形の比較

基底形	精度		
	closed all	closed regular	open
ko	94.3%	99.8%	93.0%
k	93.1%	99.8%	92.3%
ku	**94.4%**	**99.8%**	**93.4%**
ki	93.8%	99.8%	93.0%

表 9　不規則動詞「する」の基底形の比較

基底形	精度		
	closed all	closed regular	open
se	94.1%	99.8%	82.1%
s	93.6%	99.6%	89.0%
su	**94.4%**	**99.8%**	**93.4%**
si	94.3%	99.8%	88.6%
sa	94.2%	99.8%	89.5%

5.2 不規則動詞「来る」「する」

2.3.3 節で述べたように，派生文法においては，不規則動詞「来る」と「する」は語幹が変化する動詞とされる．派生文法では，「来る」「する」に対して，それぞれ「ko-」「se-」を基底形とするが，なぜ「ko-」「se-」であるかの説明はない．

そこで，本研究では派生文法で考える語幹「ko」「k」「ku」および「se」「s」「su」「si」を用意し，さらに「ki」と「sa」をそれぞれ加えて比較実験を行った．その結果を表 8 および表 9 にそれぞれ示す．

「来る」に対する基底形としては，「ku」と「ko」の精度が高かった．また「する」に対する基底形としては「su」の精度が高かった．これは，語幹末尾が u や o となる語幹が他にないことから，規則変化する語幹と区別可能になり，精度が上がったと考えられる．なお，「sa」に関しては，語幹末尾が a になる動詞は存在しないが，派生接尾辞の「-(a)na-」および「-(i)ta-」の語幹末尾が a であり，これらと区別ができなくなるため，精度が下がったと考えられる．

一方，「来る」の基底形に関しては，「ku」ではなく「ko」にすれば，「su」とも区別できるようになるが，基底形を

表 10　語幹末尾 w の扱いの比較

語幹末尾 w の扱い	精度		
	closed all	closed regular	open
小文字 w	**94.4%**	**99.8%**	**93.4%**
削除	92.2%	99.6%	88.0%
大文字 W	94.4%	99.8%	93.4%

表 11　変則動詞の語幹末尾 r の扱いの比較

語幹末尾 r の扱い	精度		
	closed all	closed regular	open
小文字 r	91.8%	99.8%	93.4%
削除	93.1%	99.3%	93.3%
大文字 R	**94.4%**	**99.8%**	**93.4%**

「ku」から「ko」に変えた場合，精度の差はほとんどなく，わずかに「ku」の方が精度が高かった．これは「来る」と「する」で語形変化に共通点が多いことが影響している可能性がある．

また「する」の比較においては，「se」以外ではオープンテストの結果がかなり低くなっている．これは「する」の出現がコーパス中に多いためであり，その語形変化を正しく処理できないと，オープンテストの精度の低下が大きくなる．

上記を踏まえて，本研究では「ku」および「su」を基底形として採用した．

5.3　語幹末尾が w の動詞

2.3.3 節で述べたように，本研究では，a 以外で始まる接尾辞が接続した場合に w が欠落することから，語幹末尾が w の動詞を不規則動詞とする．

これに対する基底形としては，以下の三つの表現方法を考え，それらを比較した．

小文字: 語幹末尾の子音 w をそのまま小文字で表記する．例えば「omow-」の基底形を「omow」とする．

削除: 語幹末尾の子音 w を削除する．例えば「omow-」の基底形を「omo」とする．

小文字: 語幹末尾の子音 w を大文字 R で表記する．例えば「omow-」の基底形を「omoW」とする．

これらの比較結果を表 10 に示す．表 10 を見ると，子音末尾の w を削除した場合の精度が低かった．これは w を削除することによって他の動詞と区別できなくなるものが存在することが原因と考えられる．例えば，「言う (iw-)」の動詞の末尾の w を削除すると「i」となり，母音幹動詞の「居る (i-)」と区別がつかなくなる．

一方，末尾の w が小文字でも大文字でも精度に差はなかった．これは，動詞の末尾以外に w が出現することが少なく，w を大文字にしても効果はなかったと考えられる．

以上を踏まえて，語幹末尾が w の動詞に対しては，語幹末尾をそのまま w としたものを基底形として採用した．

5.4　変則動詞

2.3.3 節で述べたように，「なさる」「下さる」などの変則動詞も，語幹末尾の r が欠落するなどの点で不規則動詞である．これも先述の語幹末尾が w の動詞と同様に，以下の三つの基底形を用意して比較した．

小文字: 語幹末尾の子音 r をそのまま小文字で表記する．例えば「nasar-」の基底形を「nasar」とする．

削除: 語幹末尾の子音 r を削除する．例えば「nasar-」の基底形を「nasa」とする．

小文字: 語幹末尾の子音 r を大文字 R で表記する．例えば「nasar-」の基底形を「nasaR」とする．

これらの比較果を表 11 に示す．語幹末尾が w の場合と異なり，変則動詞の場合は 末尾子音 r を削除した方が，削除しない場合よりも精度が高かった．これは変則動詞の場合，末尾 r を削除すると，「nasa」「kudasa」「ossya」「irassya」「goza」となり，すべて語幹末尾が a となる．先述した通り，語幹末尾が a となるのは派生接尾辞の「-(a)na」と「-(i)ta-」だけであり，動詞には存在しない．よって，変則動詞の末尾の r を削除したことにより，他の語幹末尾が r の動詞と区別できるようになり，精度が上がったと考えられる．

一方，末尾の r を大文字の R にすると，さらに精度が上がった．これは，R にすることにより他の語幹末尾が r の動詞と区別できるようになったことに加えて，r が欠落しない通常の接尾辞が接続する際には，r を挿入する規則より， R を r に変化させる規則の方が学習しやすいという点が考えられる．挿入するよりも変化させる方が精度が高いのは，5.1 節で述べた連結子音・連結母音の表現の場合と同じである．

以上を踏まえて，変則動詞に対しては，語幹末尾を R にしたものを基底形として採用した．

5.5　「-(i)ta」などの音便変化する接尾辞

2.3.1 節で述べたように，完了終止の「-(i)ta」や「完了」の「-(i)te」などの統語接尾辞は，表 3 に示すような音便変化を起こす．これに対処するための基底形として，変化した語幹の末尾と接尾辞を連結させた「ita」「ida」「tta」「n'da」と，それぞれの先頭を大文字にして他の音素と区別できるようにした「Ita」「Ida」「Tta」「Nda」を用意した．さらに，先頭の音素を削除した「ta」「da」と，それらの先頭を大文字化した「Ta」「Da」を用意した．

以上の 12 種類の基底形を比較した結果を表 12 に示す．表 12 を見ると，最後の 3 種類の基底形「Ta」「da」「Da」の精度が高い．ここで，「Ta」と「Da」の場合の結果は全く同じである．これは統計的機械翻訳システムが，「T は t の大文字である」といった情報を持っていないことが原

表 12　-(i)ta などの基底形の比較

基底形	精度		
	closed all	closed regular	open
ita	88.9%	98.5%	88.3%
Ita	87.4%	98.4%	84.2%
ida	89.0%	98.4%	87.3%
Ida	90.2%	99.7%	87.0%
tta	91.4%	99.8%	93.7%
Tta	91.6%	99.8%	93.5%
n'da	86.8%	99.8%	85.3%
Nda	93.0%	99.8%	93.2%
ta	93.3%	99.8%	94.0%
Ta	94.0%	99.8%	94.3%
da	**94.4%**	**99.8%**	**93.4%**
Da	94.0%	99.8%	94.3%

因である．システムから見れば，T であっても D であっても，他の箇所では使用されていない文字という意味しかなく，同じ扱いになる．よって以降では「Da」を無視して「Ta」についてのみ議論する．なお，「Ita」と「Tta」および，「Ida」と「Nda」で精度に違いがあるのは，I が連結母音として他の箇所でも使用されているからと考えられる．

音便変化が他の接尾辞では起きない特別な変化なので，そのための専用の文字 T を導入した「Ta」の精度が高いのは，これまでの他の基底形の比較実験と同じである．

しかし，クローズドテストの結果を見れば，「da」の方がさらに精度が高い．基底形が「Ta」の場合の結果を比較すると，「Ta」の場合には失敗していた「isog- + -(i)ta → isoida」および「torikum- + -(i)ta → torikun'da」となる変化が，基底形を「da」にすることで正しく処理できていた．

ここで，表 3 を改めて見ると，統語接尾辞「-(i)ta」が「-ida」もしくは「-ϕda」となり da が出現するのは，語幹末尾が g および b，n，m の場合だけである．4.1 節で述べたように，語幹末尾が n および g の動詞がコーパス中に出現する比率は少ないが，実は b および m も少ない．こうした動詞の出現比率は，g，b，n，m の 4 種類を合計しても 4.2% に過ぎない．そのため，da の出現が言語モデルの段階で上手く学習できていなかったと考えられる．オープンテストにおいては，基底形に「Ta」を選んだ方が「da」より良くなるのも，この考察を裏付ける．つまり，上記の 4 種類の動詞の出現比率が少ないため，それ以外の動詞の語形変化を正しく生成できる「Ta」の方がオープンテストの結果が良くなったのであろう．

今回はオープンテストの結果よりもクローズドテストの結果を重視するため，基底形として「da」を採用した．

5.6　命令の統語接尾辞「-e/ro」

最後に命令形の統語接尾辞について比較した．2.3.2 節および 2.3.3 節で示したように，命令形の場合，子音幹には「-e」，母音幹には「-ro」，変則動詞には「-i」が接続する．そこで，基底形としてはこの 3 種類と，さらに他のどこでも使用されていない文字として「E」を用意した．比較実験の結果を表 13 に示す．

表 13　命令の統語接尾辞の基底形の比較

基底形	精度		
	closed all	closed regular	open
e	93.7%	99.5%	93.4%
ro	93.6%	99.8%	91.6%
i	93.4%	99.5%	93.3%
E	**94.4%**	**99.8%**	**93.4%**

表 14　提案基底形と派生文法との比較

基底形	精度		
	closed all	closed regular	open
派生文法	84.4%	98.2%	74.2%
提案基底形	**94.4%**	**99.8%**	**93.4%**

これまでの実験と同様，他で使用されていない文字 E を用意すると，精度が高くなることが分かる．よって，命令の統語接尾辞の基底形として「E」を採用した．

5.7　派生文法との比較

本研究で比較した基底形をまとめると以下のようになる．

「来る」：　ko, k, **ku**,ki
「する」：　se, s, **su**, si, sa
「思う」などの末尾が w の動詞：　**omow**, omo, omoW
「なさる」などの変則動詞：　nasar, nasa, **naraR**
「-(i)ta」などの音便変化する統語接尾辞：　ita,　Ita, ida, Ida, tta, Tta, n'da, Nda, ta, Ta, **da**, Da
命令の統語接尾辞「-e/-ro」：　e, ro, i, **E**

太字で示したものが，本研究において最も高い精度を示した基底形であり，以降では**提案基底形**と呼ぶ．

ところで，これまでの実験では，提案基底形から一つの基底形だけを変化させ，他の基底形は固定して比較したが，複数の基底形を同時に変化させて比較することも考えられる．そこで，本研究が参考にした派生文法における基底形との比較を試みた．上記において 下線 を引いたものが派生文法において基底形とされているものである．ただし，派生文法では命令の統語接尾辞「-e」と「-ro」のいずれが基底形かは述べていないため，表 13 で基本規則に関する精度が最も高かった「ro」を採用した．

提案基底形と派生文法における基底形との比較実験の結果を表 14 に示す．この結果を見ると，オープンテストの精度に大きな差がある．これは，5.2 節で述べたように，出現回数が多い「する」の音韻変化処理の誤りが多いためである．

また，訓練データの比較実験（表 4）や言語モデルの比

較実験（表 5）と比較した場合，派生文法における基底形の方が精度が低い．すなわち，訓練データの量や言語モデルの選定と比べて，基底形の違いが精度に大きな影響を与えることが分かる．

よって，統計的音韻変化処理においては基底形の選択が重要である．また，その際の基底形の選び方としては，例外的な変化を規則に対処するためには，その変化に固有の文字を使用するのが良いと言える．

6. おわりに

本稿では，統計的機械翻訳の枠組を利用した音韻変化処理について検討した．その結果，翻訳モデルと言語モデルの学習のために使用する訓練データに関しては，ある程度の量があれば，それ以上増やしても効果がないことが分かった．これは逆に言えば，データを増やしても精度向上が期待できないということであり，統計的手法の限界を示したと言える．

また，様々な基底形につい統計的な観点から比較し，基底形の違いが統計的音韻変化処理に大きな影響を与えることを確認した．また基底形の選択に関する指針も得た．

現在，ウイグル語とウズベク語に関しても同様の実験を進めている．また，今後，例外的な規則は人手で記述し，それ以外は統計的な手法で獲得するなどのハイブリッドな手法についても検討する．

謝辞 本研究は，日本学術振興会科学研究費補助金若手研究 (B)（課題番号 22700143）の補助を受けている．

参考文献

[1] 小川泰弘, ムフタル・マフスット, 杉野花津江，外山勝彦, 稲垣康善：派生文法に基づく日本語動詞句のウイグル語への翻訳, 自然言語処理, Vol. 7, No. 3, pp.57–78 (2000).
[2] 小川泰弘, ムフタル・マフスット, 杉野花津江, 稲垣康善: 日本語–ウイグル語間機械翻訳におけるウイグル語音韻変化処理の形式化, 言語処理学会第 8 回年次大会講演論文集, pp.29–32 (2002).
[3] 小川泰弘, 福田ムフタル, 外山勝彦: 日本語対訳辞書拡張のためのウイグル語からウズベク語への翻字手法, 言語処理学会第 14 回年次大会講演論文集, pp.472–475 (2008).
[4] Koskenniemi, Kimmo: Two-level model for morphological analysis, IJCAI-83, pp.683–685 (1983).
[5] Bernard Bloch: Studies in Colloquial Japanese, Part I, Inflection, In *Journals of the American Oriental Society*, Vol. 66 pp.97–109 (1946).
[6] 清瀬義三郎則府: 日本語文法新論–派生文法序説–, 桜楓社 (1989).
[7] 清瀬義三郎則府: 日本語学とアルタイ語学, 明治書院 (1991).
[8] Kiyose, Gisabruo N.: Japanese grammar –A new approach–, Kyoto University Press (1995).
[9] 竹内和夫: 現代ウイグル語四週間, 大学書林 (1991).
[10] Franz J. Och and Hermann Ney: A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, Vol. 29 No. 1 pp.19–51 (2003).
[11] Andreas Stolcke: SRILM – An Extensible Language Modeling Toolkit, In *Proceedings of the International Conference on Statistical Language Processing*, Vol.2 pp. 901–904 (2002).
[12] Philipp Koehn and Hieu Hoang: Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.868–876 (2007).
[13] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書, 日本電子化辞書研究所 (1996).