統計的機械翻訳システムを利用した膠着語の音韻変化処理

小川 泰弘　外山 勝彦
名古屋大学 大学院情報科学研究科
yasuhiro@is.nagoya-u.ac.jp

はじめに

我々はこれまでに，日本語，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，こうした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は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，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．

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，言語学的な観点からではなく，統計的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．

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] 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．

表 1: 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
派生接尾辞	使役	-(s)ase-	kak-ase-	tabe-sase-
	受身	-(r)are-	kak-are-	tabe-rare-
	丁寧	-(i)mas-	kak-imas-	tabe-mas-
	可能	-e-†	kak-e-	-
	否定	-(a)na-	kak-ana-	tabe-na-
	希望	-(i)ta-	kak-ita-	tabe-ta-
統語接尾辞	非完了終止	-(r)u	kak-u	tabe-ru
	完了終止	-(i)ta†	kaϕ-ita	tabe-ta
	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-
	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
	仮定条件	-(r)eba	kak-eba	tabe-reba
	否定	-(a)zu	kak-azu	tabe-zu
	目的	-(i)ni	kak-ini	tabe-ni
	同時進行	-(i)tutu	kak-itutu	tabe-tutu
	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．

動詞の語幹と接尾辞

派生文法は，日本語の形態素を音韻単位で設定し，例えば「書かれました」は「kak+(r)are+(i)mas+(i)ta」の 4 つの形態素から構成されると考える．ここで，括弧内の音素は**連結子音・連結母音**と呼ばれ，連結子音 (母音) は，子音 (母音) に後接する場合に欠落するという連結規則 (以下，**基本規則**と呼ぶ) を提案している．この基本規則により，用言の語形変化が記述可能となり，例えば，「kak+(r)u」，「tabe+(r)u」は，それぞれ子音で終わる**子音幹動詞**「kak-」と母音で終わる**母音幹動詞**「tabe-」に，同じ接尾辞「-(r)u」が接続していると考える．なお，「-(r)are-」「-(i)mas-」のように他の接尾辞が後接する接尾辞を**派生接尾辞**，「-(i)ta」のように動詞句の末尾に来る接尾辞を**統語接尾辞**と呼ぶ．主な派生接尾辞および統語接尾辞を表 1 に示す．

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扱えるが，例外もある．本稿では，基本規則によって扱える語形変化を**規則変化**とし，それ以外の変化を**不規則変化**と定義する．

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．末尾が s の子音幹に接続する際には,「kas-ita」のようにそのままの形であるが,「kak-」に接続すると「ka-ita」に,「tob-」に接続すると「ton'-da」になるような音便変化を起こす．同様に「it」で始まる「-ite」「-iteha」なども音便変化を起こすが,「-(i)tutu」や，希望の派生接尾辞である「-(i)ta-」は音便変化しない．特に，派生接尾辞「-(i)ta-」と統語接尾辞「-(i)ta」は同形であるが語形変化に違いがある．

命令形の場合，子音幹には「-e」，母音幹には「-ro」という統語接尾辞が接続する．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である．

動詞に関しては「来る」「する」が不規則変化動詞である．派生文法では，それぞれ「ko-」「se-」を語幹とし，後続の接尾辞によって母音が変化すると考える．

その他,「行く」「問う」は，完了の「-(i)ta」が接続する音便形がそれぞれ「行った (iϕ-tta)」「問うた (to-uta)」になるという点で不規則である．

「なさる」「下さる」「仰る」「いらっしゃる」「ござる」は,「-(i)mas-」が接続する際に「nasar-imas-」ではなく「nasa-imas-」になる点，命令の接尾辞が「-i」になるという点で不規則であり，変則動詞と呼ばれる．

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．これは，a 以外で始まる接尾辞が接続した場合に w が欠落するという規則が基本規則外となるためである．

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，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」「ita」という基底形が接続すると 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が，日本人向けのウイグル語の文法書 [7] 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難しい．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．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．

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 GIZA++, SRILM, Moses といったツールを使用して，統計的音韻変化処理について実験した．

それぞれの処理においては，使用する基底形，モデル，パラメータに様々な種類がある．これらを比較するために数多くの実験を行ったが，紙面の制限から興味深い点についてのみ結果と考察を示す．

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．

訓練データの比較実験

統計的機械翻訳では，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．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が、出現する音素の分布の偏りが大きいため，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，EDR コーパス (1.5 版) 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，訓練データの作成においては，コーパスから抽出したもの (データ 1，2) と動詞と接尾辞の組合せから網羅的に生成したもの (データ 3，4) を比較した．

EDR コーパスからは，ランダムに 1,000 文を抽出し，そこに出現した動詞句を使用したもの (データ 1) と，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 (データ 2) を用意した．

組合せでデータを作成する際には，動詞の語幹末尾ごとに EDR コーパス中の出現頻度が高いものから 5 個ずつ採用した．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 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 1 に示したすべての接尾辞と「-(r)uto」,「-(i)tara」,「-(i)nagara」,「-(y)outo」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，合計 32 個登録した．学習データは，動詞と接尾辞の組合せで，合計 1,920 個となる．このデータを以降**組合せデータ**と呼ぶ (データ 3)．同様に，抽出する動詞の数を各 10 個に増やしたデータ 4 も用意した．

性能評価はオープンテストとし，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 (open)．しかし，これだけでは音韻規則が網羅的に獲得できているか評価できないため，組合せデータに対する精度も求めた (組合せ all)．

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 (組合せ regular)．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]．

なお，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．

[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．

表 2: 訓練データの比較

	data type	size	組合せ all	組合せ regular	open
1	EDR 千文	2,580	89.8%	78.6%	94.3%
2	EDR 全文	531,685	90.6%	98.2%	95.4%
3	**組合せ**	**1,920**	**92.6%**	**99.6%**	**90.9%**
4	組合せ (10)	3,520	92.5%	99.5%	91.5%

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 2 に示す．size は訓練データに含まれる動詞句の数である．

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データ 1，2 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，データ 1 の精度が低く，基本規則でも学習されていないものがある．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のが原因である．使用するコーパスデータを増やしたデータ 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．

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，組合せデータに対する精度を重視する．

なお，訓練データに組合せを用いたデータ 3 と 4 の比較では，元となる動詞の数を各 5 個から各 10 個に増やしても効果はなかった．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にはデータ 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「-(i)t」で始まる接尾辞に規則変化するものと不規則変化するものの 2 種類があるためである．組合せ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．

言語モデルの比較実験

言語モデルの学習には，EDR コーパスから作成したデータ (モデル 1〜6) と，訓練データの実験で使用したデータ 3(モデル 7) を用意し，合計 7 個のモデルを比較した．なお，モデル 3 以外では重複を除き，モデル 4 以外では言語モデルに 2-gram を使用した．

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．次に，動詞句以外を除去したモデル 2 を用意し，これについては，重複を除去しないモデル 3 と言語モデルに 3-gram まで含めたモデル 4 も用意した．コーパスの量を減らし，EDR コーパスからランダムに抽出した日本語文 1,000 文において出現した動詞句を用いたものを 2 種類用意し，モデル 5，6 とした．

組合せデータは訓練データの実験に用いたものと同じである．よってモデル 7 では，訓練データと同じものを言語モデルの学習に用いることになる．

表 3: 言語モデル用データの比較

	data type	size	closed all	closed regular	open
1	EDR 全文節	281,263	91.8%	98.5%	88.9%
2	**EDR 全文**	**22,614**	**92.6%**	**99.6%**	**90.9%**
3	同 重複	531,685	92.2%	98.9%	93.9%
4	同 3-gram	22,614	89.2%	92.8%	92.1%
5	EDR 千文 a	1,022	89.4%	99.2%	92.2%
6	EDR 千文 b	1,063	92.9%	99.3%	94.2%
7	組合せ	1,899	92.7%	98.7%	86.5%

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，クローズドテストになる (closed all)．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 (closed regular) も求めた．

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 3 に示す．

まずモデル 1 においては基本規則だけの精度が低く，連結母音が削除され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 (例外は，語幹末尾が w の動詞に接尾辞が接続した場合など) のに対して，漢字で構成される単語内に，母音の連続があることが原因と考えられる．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いえる．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．

一番性能が良かったのがコーパス 1,000 文から作成したモデル 6 であるが，同じく 1,000 文から作成したモデル 5 と差が大きい．また，モデル 6 よりデータ量が多いモデル 2 ではモデル 6 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無いことを示している．

また，モデル 2 と 3 の比較に見られるように動詞句の重複を許すと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が悪くなった．今回は網羅的な規則が獲得できるかを重視して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．

一般に言語モデルについては N-gram を用いるが，N のサイズをいくつにするのが適切かという問題がある．モデル 2 と 4 を比較すると N の値が 2 のときに，すなわち 2-gram だけを使用した方が結果が良かった．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．

表 4: 異なる基底形の比較

基底形	closed all	closed regular	open
派生文法	84.6%	99.0%	72.7%
提案基底形	**92.6%**	**99.6%**	**90.9%**

訓練データでは組合せデータを用いた方が性能が良かったが，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，その性能は偶然によるところが大きいといえる．一番結果が良かったのはモデル 6 であるが，偶然だと考えられるので，モデル 6 とほぼ同等で，基本規則に関してはモデル 6 以上の性能を示したモデル 2 を今回の実験では採用した．

基底形の比較実験

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．紙面の制限のため省略するが，この他にも各種の比較を行った．例えば，平仮名表記とローマ字表記の比較では，ローマ字表記の方が精度が高かった．

また，連結母音・連結子音に関しては，以下の三つを比較し，結果が良かった 1. を採用している．

1. 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」．
2. 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．
3. 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」．

想定した基底形

今回の実験では，それぞれ以下の基底形を比較した．なお**太字**で示したのは派生文法における基底形である．

「来る」: ko，ku，k，ki

「する」: se，su，s，si，sa

「なさる」などの変則動詞: nasar，nasa

「思う」などの末尾が w: omow，omo

完了の「-(i)ta」など: Ita，ita，Tta，tta，Ta，ta，Ida，Da，da，Nda，n'da

命令の「-e/-ro」: e，**ro**, i, E

紙面の制限により，すべての基底形の比較結果を記述することはできないため，今回の実験で一番良い精度を示した基底形 (上記で 下線 を引いたもの) を**提案基底形**とし，これと派生文法における基底形とを比較した．なお，派生文法では命令の「e」「ro」のいずれが基底形かは述べていないが，予備実験で精度の良かった「ro」を採用した．

基底形比較実験の結果および考察

基底形の比較実験の結果を表 4 に示す．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．派生文法の基底形のうち，精度低下の一番の原因は「する」の基底形「se」であった．これは末尾が e となるため，末尾が e の母音幹動詞と区別がつかない．基底形を「su」にすれば，他の動詞と区別できるようになり，精度が向上した．その場合「来る」の基底形も「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「ku」から「ko」に変えた場合，クローズドテストの出力結果に差異はあるが，精度は同じだった．よってオープンテストの結果が良い「ku」の方を採用した．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のかもしれない．同様の理由で,「なさる」などの変則動詞も末尾の r を削除することにより精度が上がった．

完了の「-(i)ta 」の基底形は「Ta」が精度が高く．これは Bloch[5] の主張に近いといえる．これも，T という他で使用されていない文字を使用することにより，精度が向上したと考えられる．

以上から，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り，例外的な変化をする場合には，その変化に固有の文字を使用するのが良いと言える．

本稿では 4 節の実験でも，提案基底形を用いている．

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，訓練データと言語モデルの学習データとも，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, ムフタル, 杉野, 外山, 稲垣: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, ムフタル, 杉野, 稲垣 : 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, 福田, 外山: 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, K.: Two-level model for morphological analysis, IJCAI-83, pp.683–685, (1983).

[5] Bernard, B.: Studies in Colloquial Japanese, Part I, Inflection, In *Journals of the American Oriental Society*, Vol. 66 pp.97–109 (1946).

[6] 清瀬: 日本語文法新論–派生文法序説–, 桜楓社 (1989).

[7] 竹内: 現代ウイグル語四週間, 大学書林 (1991).