少数言語における保存活動の支援に関する研究

令和 6 年 3 月

三重大学大学院工学研究科

博士後期課程　システム工学専攻

松下真也

目次

1 章 序論

1-1 少数言語と保護活動

現在，世界で使用される言語は約 7,000 に及ぶが，話者数の不足によって消滅の危機に瀕している言語もある[1]．このような言語は「少数言語（または消滅危機言語）」と呼ばれる．言語の消滅は，言語を媒介にして継承される知識や技術，文化の消滅を意味する．そのため，言語学者らは，少数言語の保護活動を行う[2][3][4][5]．

保存活動では，減少し続ける話者の保護や言語の保存などが目的であり，対象言語を扱う人々との共同生活やインタビューなどを実施する．このとき，共同生活では食生活や宗教観などの文化を記録し，インタビューでは身近な単語を調査する．特に，少数言語に対する情報が全く無い（対象言語が文字を持っていない場合がある）場合には，図 1-1 のように手探りでインタビュー調査をせざるを得ない．そして，インタビューによって明らかにされる単語は，保護活動全体の最も基礎的な知見である．単語が十分に揃っていない場合は，文化・知識の詳細な記録・保存がより困難である．

図 1-1 身振り手振りで身近な単語を明らかにする事例

1-2 支援する活動

本研究では，保護活動の一つである「言語の保存活動」を支援する．前提として，保護活動のほとんどは手作業かつ専門的な作業であり，膨大な作業時間が必要となる．そのため，効率化や正確性の観点から，さまざまな保護活動の自動化（支援）が望まれている．しかし，本研究では，発話体系や文法・語彙の調査を優先的に行うことで，他の保護活動を円滑に進められると考える．そのため，本研究では「言語の保存活動」，その中でも発話体系の分析・文法語彙の分析の支援を対象とする．これらの活動は「少数言語の耳を頼りにした書き起こし・テキスト解析」と言い換えられ，他の保護活動と同様に手作業で行われている．これを支援し自動化することで，膨大な作業時間を減らすことができる．また，作業時間を減らすために複数人で作業する場合もあるが，作業結果の個人差が後工程に与える影響が深刻になる．そのため，本研究では，作業結果の品質を保つという観点からも，この手作業を支援する．

支援にあたっては，機械学習を用いることが主流である．機械学習では，その性能を高めるために事前知識や多量のデータを必要とする．しかし，少数言語は低資源言語とも呼ばれ，対象言語に関する(事前)知識が不足したり，対象言語で表されたデータが少なかったりする．そのため，機械学習の導入にあたり，注意が必要である．実際，少数言語は話者が少なく，文字を持たないこともあるため，多量のデータを集めたり，事前知識を得たりすることが難しい．それに対して，少数言語ではない言語(例えば英語)では，学習データとして音声データやテキストデータを収集する対象が多いことから，多量の学習データは簡単に用意できる．また，文法・語彙などに関する理解も進んでおり，機械学習適用にあたり十分な事前知識を用いることもできる．したがって，本研究では，事前知識の不足および学習データ量の不足を同時に考慮した支援をめざす．

1-3 保存活動の支援内容と目的

支援する保存活動は「少数言語の耳を頼りにした書き起こし・テキスト解析」であり，機械学習に基づく技術を用いた支援が想定される．ここで，音声を書き起こして解析するタスクの完全自動化には，音声認識技術が挙げられる．しかし，高い認識精度を得るためには，豊富な事前知識と多量の学習データが必要であるため，これらの要求を満たせない少数言語には適さない．そこで，本研究では保存活動を，「音声の書き起こし」と「書き起こされたテキストの解析」の 2 段階に分けて支援を検討する．また，段階ごとの支援から得られる知見は，音声認識技術の適用時に必要な事前知識の獲得に貢献する．そのため，段階的支援する本研究は，音声認識技術による保存活動の完全自動化を見据えた基礎的研究である．

まず，音声の書き起こし段階では，音を表す記号を用いて音声を書き起こす．この記号は，作業の進捗に合わせて記号を使い分けるが，本研究では作業者の母語など（その時点で作業者が把握しやすい音を表す記号）を想定する．この想定の下，本研究では「連続音声へ任意の記号群をラベリングする」タスクに置き換えて検討する．

このタスクを支援する技術には，音声ラベリングが挙げられる．現在は，大語彙連続音声認識エンジン Julius や Whisper を用いた手法が多く用いられている[6][7]．しかし，これらの手法には，言語固有の情報や，機械学習を用いるための多量のデータが必要となる．いずれの需要も少数言語は満たせないため，これらの手法を少数言語音声のラベリングには適さない．したがって，本研究では，少数言語音声のためのラベリング手法を検討し，書き起こしの効率化を図る．

次に，テキスト分析の段階では，単語や文法などを明らかにして記録する．手作業においては，図 1-1 の活動で得られた知見などもふまえ，記録する．このとき，書籍などが現存する場合もあるが，本研究ではそれが全く無い場合を想定する．これは，少数言語によっては，文字を持たない可能性があるためである．

これを支援する技術には，自然言語処理分野におけるテキスト分析手法が挙げられる．一般のテキスト分析は，分かち書き⇨形態素解析⇨構文解析⇨意味解析のように多段階に分けられる，一貫して言語固有の情報や機械学習を用いるための多量のデータが必要となる．そのため，いずれの段階においても，少数言語への適用は工夫が必要である．本研究では，少数言語のテキスト分析を支援するにあたって，最初のステップである分かち書きに着目する．分かち書きは「テキストはどのような塊に区切られるのか」を求めるものであり，保存活動においては，これを支援することで単語調査の効率化に貢献する．したがって，本研究では，少数言語テキストの分析における最初のステップとして，分かち書き手法を検討する．

1-4 本研究の構成

2 章では，少数言語音声の書き起こしを支援するため，Wavelet 解析および画像処理手法を利用した音声ラベリング手法を提案した．一般の音声ラベリングでは，周波数を音響特徴量 MFCC に変換し，DP マッチングを組み合わせて行う．しかし，少数言語音声において付与されるラベルが未知であるため，MFCC と DP マッチングを用いたラベリング手法では少数言語音声に適さない．そこで，本研究では，MFCC を Wavelet 解析に，DP マッチングを SIFT 特徴量に置き換えることで，少数言語のための音声ラベリング手法を提案した．

3 章では，少数言語のテキストの分かち書き手法を検討した．具体的には，事前知識なしでさまざまな言語の分かち書きを行う手法の一つである Nested Pitman-Yor Language Model（NPYLM）に着目し，不十分なデータ量での挙動を分析し，その結果に基づいて，少数言語の単語分割における NPYLM の適用上の問題点を明らかにした．

4 章では，3 章で述べた問題点の解消を目的として，「NPYLM の 2 段階適用」を提案した．分割の必要がない文字列を一文字に置き換えることで，データ量が不十分なときの NPYLM における挙動を抑制する．また，NPYLM を 2 回適用することで，その 1 回目で置き換える文字列の検出をし，2 回目で分かち書きを行う手法を提案した．

5 章では，4 章で提案した手法を改善した．4 章の手法は，3 章で述べた問題点を解消したが，別の問題を引き起こすことがあったため，これを解消する．そこで，4 章の手法における置き換えにおいて，選択的な置き換えを提案した．

6 章では全体のまとめを述べ，今後について示す．

2 章 音声の書き起こしに対する支援

2-1 はじめに

音声の書き起こしでは，音声データが入力となり，音に対応した文字を時系列で書き起こした結果が出力（テキストデータ）となる．このような入出力を想定した技術には，音声認識技術や音声ラベリング技術が挙げられる．これらの技術はいくつも存在するが，多くは適用対象の言語を固定して用いる．特に，音声データは，男女差やイントネーションといった個人差や言語差が大きいデータであることから，特定の言語に絞って用いられる．そして，高精度な出力結果を得るため，その言語固有の事前知識を必要とする．ここで，日本語話者が英語をカタカナで書き起こせるように，異なる言語であっても類似している感覚を持つことがある．そのため，他の音声認識技術（少数言語ではない別言語に用意されたもの）を用いて，部分的に書き起こせるように思える．そして，この書き起こしは完璧ではないだろうが，いくつかのそれらしい書き起こし結果を得ることは想像に難くない．しかし，既存の音声認識などでは認識精度を保つために，本来の適用が想定された言語の特徴（語彙や文法）を用いて認識結果を補正する．そのため，別言語に用意された技術を用いて少数言語音声を書き起こすことは，出鱈目な認識結果（少数言語には関係のない書き起こし）を引き起こしかねない．また，あらゆる言語を入力とすることができる技術も存在するが，処理途中で言語を特定する必要があるため，少数言語には適さない．以上より，少数言語の音声ラベリングにおいては，言語を特定せずに事前知識も必要としないラベリング手法を検討する．

検討にあたっては，少数言語の特徴である事前知識に乏しいこと，少量データであることを考慮する．まず，事前知識に乏しいことだけに焦点をあてると，教師なしラベリング手法が，少数言語におけるタスクには有力と考える．このとき，教師なしラベリング手法は，連

続音声の教師なしセグメンテーションと言い換えられ，いくつか提案されている[8][9][10]．しかし，教師なしであっても，多量の学習データを要求することから，少量データである少数言語には適さない．視点を変えて，多量の学習データを用意できる言語群からモデル（多言語モデル）を構築することで，少数言語との共通項に注目した適用も期待できる．しかし，将来的には何らかの事前知識を用いてファインチューニングは必須と考えられるため，事前知識に乏しい少数言語には適さない．次に，少量データであることも追加して議論を進めると，少量データかつ教師なしで連続音声をセグメンテーションする手法が提案されている[11]．この手法は，前述の考慮することに適するが，音声データだけでなくテキストデータも同時に入力とする必要がある．本研究では，テキストデータを出力するために，連続音声の書き起こしを位置付けている．したがって，テキストデータが全く無い状況には，この手法が適用できないため，少数言語のラベリングには適さない．

本研究では，最も基礎的な音声ラベリング手法を参考にしつつ，少数言語の条件に適するラベリング手法を提案する．前述してきた手法の多くは，音声認識またはラベリングを高精度に行うことに主眼があり，所定のラベルを想定する．しかし，少数言語においては，当てはめるラベルそのものが未知であるため，いずれも少数言語のラベリングには適さない．そこで，少数言語のラベリングでは「ある音に一致または類似箇所を連続音声から探し出す」というタスクに置き換える．これは，ラベルの把握が不十分であることをふまえ，緩やかにラベリングすることを意味する．次節以降，基礎的な音声ラベリングと比較しつつ，少数言語に適する音声ラベリング手法を提案する．

2-2 基礎的な音声ラベリング手法と少数言語のためのラベリング

本研究では，基礎的な音声ラベリング手法を参考にしつつ，少数言語の音声ラベリングを提案する．特に，本節では基礎的なラベリング手法について示し，少数言語を処理対象とした場合の課題を示す．

2-2-1 基礎的な音声ラベリング

基礎的な音声ラベリングは，連続音声に対して任意の音のマッチングを図る．このとき，音声信号そのものではなく，声道特性を表す特徴量を用いる．具体的には Mel-Frequency Cepstrum Coefficients（MFCC）と呼ばれる音響特徴量を用いる．

音響特徴量 MFCC を用いたラベリングでは，テンプレートマッチングが採用される．具体的には，MFCC のユークリッド距離を求めることで，類似度によるマッチングを実現する．テンプレートマッチングの参考として，図 2-1 を示す．まず，音の塊を特定する(音 1 ～音 3)．特定したそれぞれの音の塊と基準音声を比較し，最も類似度の高い基準音声を特定する．図では，音 2 に対して音 B が最も高い類似度を示したため，音 2 に対する音声ラベルとして B を付与する．

また，同じラベルを付与すべき音の塊同士であっても，発話時間の違いなどの時間的揺らぎが生じる．特に，MFCC では時間情報が完全には保持されないため，MFCC を用いたテンプレートマッチングでは，この時間的揺らぎによってマッチング精度が低下する恐れがある．そのため，マッチングの際には単純なテンプレートマッチングではなく，DP マッチングを用いる．図 2-2 から具体的なマッチング方法を説明する．図 2-2 では，フレームは時間情報を伴うため，同じ音でも発話時間が異なる状況を示す（ツイードの"ー"は 3 フレームであるのに対し，スイーツの"ー"は 2 フレーム）．そこで，フレーム距離を累積コストとして，最短経路問題からその状況を把握する．その結果，フレームの対応関係をふまえ，マッチングすることが可能となる．以上より，音響特徴量 MFCC と DP マッチングを用いることで，最も基礎的な音声ラベリングが実現する．

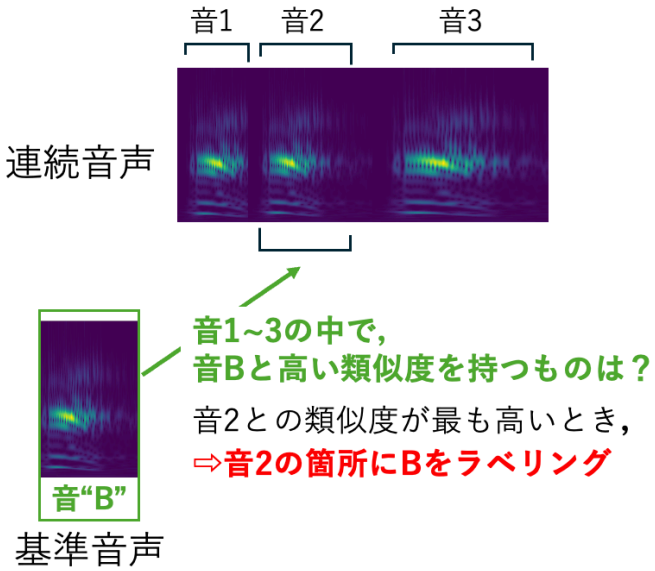


図 2-1 基礎的な音声ラベリングのイメージ

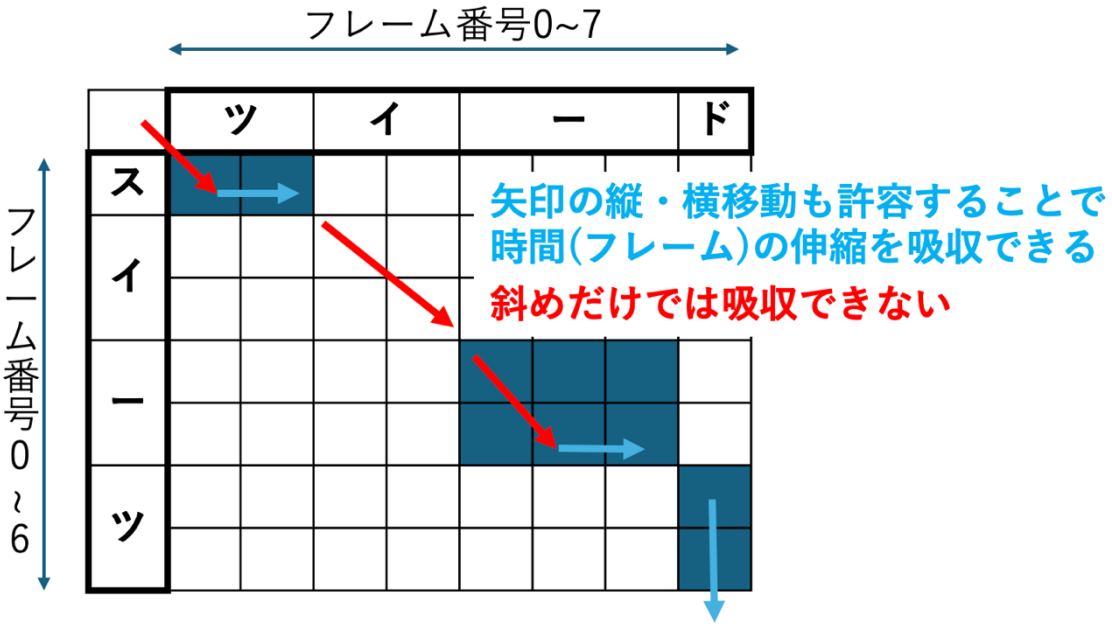


図 2-2 異なる発話時間における音声の DP マッチング
（各マスにはユークリッド距離が入り，距離行列となる）

2-2-2 少数言語を対象としたラベリング手法の検討

本節では，少数言語のラベリングにおいて，音響特徴量 MFCC と DP マッチングを組み合わせたラベリング手法を適用上の問題点を示し，その解決方針を検討する．

まず，少数言語音声における音響特徴量 MFCC の問題について述べる．MFCC は，メルケプストラム分析によって声道特性を反映する特徴量であるため，少数言語音声を書き起こす根拠に適している．しかし，DP マッチングを成功させるためには，余分な音声区間を含んだ音の塊から MFCC を求めてはならず，基準音声のいずれかに対応する音声区間だけを抽出する必要がある．余分な区間を含めて MFCC を求めると，ラベルに対して不完全な特徴量を算出し，高精度なラベリングを期待できない．このとき，解析初期の少数言語音声ではラベル境界が未知であるため，余分な音声区間を含めないように抽出することは困難である．そのため，MFCC は，少数言語音声の特徴量算出に適さない．

次に，少数言語音声における DP マッチングの問題について述べる．DP マッチングによる特徴量の類似度計算は，図 2-2 のように発話時間には対応する．しかし，基準音声が適切に切り出されずラベルが未知な状況では，マッチングしきれない場合がある．例えば，「ツイード」と「スイーツ」の前後に続く区間が，同一の特徴を示す場合はマッチングできるが，異なる特徴である場合はできない．これは，DP マッチングでは発話時間による伸び縮みに対応するが，欠落・追加には非対応であることを意味する．特に，少数言語音声においては，欠落・追加が生じやすいため，DP マッチングは適さない．

以上より本研究では，少数言語の音声に対するラベリングの方針を変更する．新たなラベリングイメージを図 2-3 のように定めた．この手法は，音声の塊をマッチングの前に適切に切り出すことは困難なので，時間的に変化する特徴量の変化の中から，基準音声に近い部分を探すものである．さらに，マッチング自体も伸び縮み以外に欠落・追加などの変化にも許容する．

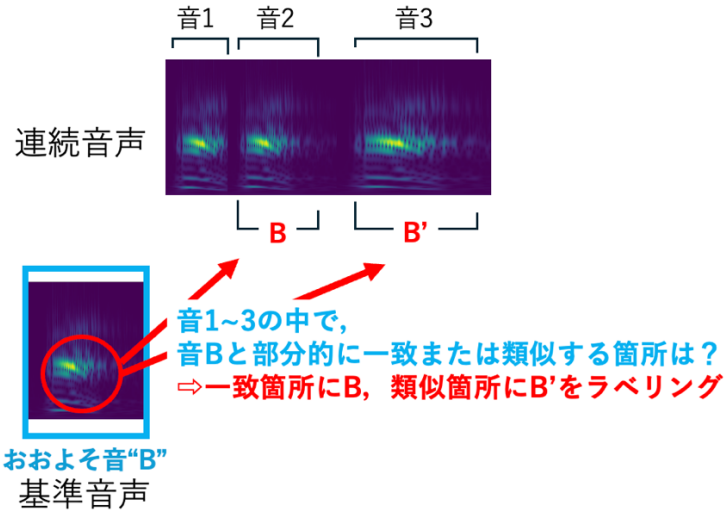


図 2-3 少数言語における音声ラベリングのイメ

2-3 Wavelet 解析と SIFT 特徴量を用いたラベリング

本節では，図 2-3 で示した少数言語のためのラベリングを実現するため，MFCC および DP マッチングを異なる手法に置き換える．具体的には，MFCC を Wavelet 解析に，DP マッチングを SIFT 特徴量によるマッチングにそれぞれ置き換える．次節以降，置き換え理由を示しつつ，Wavelet 解析と SIFT 特徴量について簡単に説明する．最後に， Wavelet 解析と SIFT 特徴量を組み合わせたラベリングにおいて，その手順を示す．

2-3-1 本研究における Wavelet 解析と SIFT 特徴量の役割

ここでは，2-2-2 節で述べた MFCC および DP マッチングの問題点の下，図 2-3 における Wavelet 解析および SIFT 特徴量の役割について述べる．

まず，MFCC から Wavelet 解析への置き換えについて述べる．前述のとおり，MFCC は少数言語音声に適さない特徴量であった．ここで，図 2-3 における基準音声に対し，曖昧なラベル境界を想定することで，少数言語音声においても特徴量を用いたマッチングが可能と考える．本研究では，曖昧なラベル境界を想定しつつ特徴量を定める手段は，Wavelet 解析が適すると考えた．Wavelet 解析は時間情報を完全に保持したままで周波数を解析できるため，基準音声で特徴を示す部分音声 A に対しては，時間情報（音声 A の発話時間）が伴うことになる．そして，この時間情報を用いることで，どの時間帯のどの特徴量がマッチングするのかを把握できる．その結果，連続音声のあらゆる時間帯に対してマッチングを図ることができる．したがって，少数言語の特徴量算出にあたって，本研究では MFCC を Wavelet 解析に置き換えた．

次に，DP マッチングから SIFT 特徴量への置き換えについて述べる．前述のとおり，DP マッチングは発話時間（伸び縮み）に強いが，本研究では発話区間の前後にも余計な区間がある場合を想定するため，DP マッチングでも解消しきれない時間的揺らぎがある．そこで，伸び縮みだけでなく欠落・追加などの変化も許容するマッチング手法について，図 2-3 より検討する．図 2-3 に示した新たなラベリングでは，連続音声から求めた時間的に変化する特徴量の中から，基準音声と部分的に一致または類似する箇所を探索する．そこで，多数の特徴量の時間変化を．縦軸に特徴量の種類・横軸に時間をとった平面上に表示した画像と見なし（Wavelet 解析によるスカログラム），画像処理技術における特徴点マッチングの技術を適用する．具体的には，画像の回転・縮尺に強い SIFT 特徴量に着目し，画像の類似性からマッチングする．特に，SIFT 特徴量におけるマッチングでは，基準音声（画像）と連続音声（画像）を特徴点で結ぶ．このとき，基準音声における特徴点の並びが，連続音声に対しても同じ並びを取るとは限らない．つまり，図 2-3 の赤枠が連続音声のある箇所に対応し，赤枠外は別の箇所に対応または非対応という結果も期待できる．その結果，基準音声と部分的な一致・類似が，SIFT 特徴量によって実現できる．

2-3-2 Wavelet 解析とは

Wavelet 解析は時間情報を完全に保持するものであり，低次元では周波数を，高次元では時間的な変化を捉える．また，フーリエ変換が正弦波の重なりであるのに対し，Wavelet 変換は任意の波形$\Psi(t)$の重なりである．特に，$\Psi(t)$は Mother Wavelet と呼ばれ，これを用いることで特徴的な波形を構成する時間を特定できる．図 2-4 より，$\Psi(t)$を$\Psi((t-b)/a)$とし，bを時間移動とすると，$\Psi(t)$は伸縮かつ移動が自由となる．したがって，Wavelet 変換とは任意の波形を，さまざまなaで定められる基底$\Psi((t-b)/a)$の重なりで表現するものである．

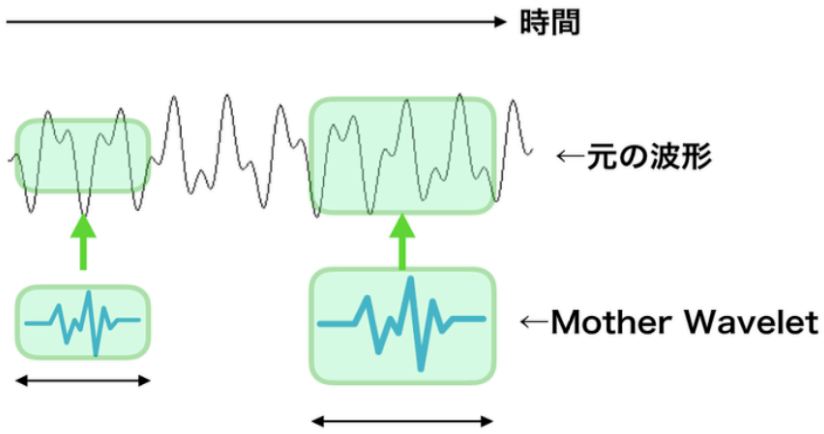


図 2-4　Mother Wavelet: $\Psi(t)$による時間の特定例

続けて，最も単純な Mother Wavelet である Haar Wavelet を参考にして，Wavelet 変換の理解を深める．Haar Wavelet を以下のような関数とする（図 2-5）．

$$\Psi(t) = \begin{cases} 1 & 0 \leq t < 1 \\ -1 & \frac{1}{2} \leq t < 1 \\ 0 & otheraise \end{cases}$$

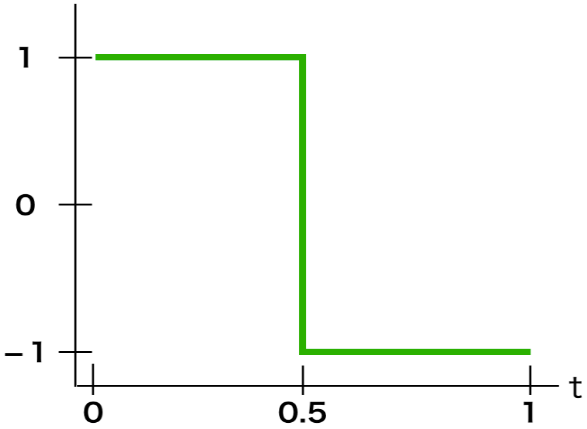


図 2-5 Haar Wavelet（$a = 1, b = 0$）

ここで，$\Psi(t)$について

$$\frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right)$$

とすると，a, bの値によってグラフの並行移動，伸縮が行われる．参考として，図 2-6 のように$(a = 2, b = 0)$，$(a = 2, b = 1)$を設定した場合を考える．最終的には以下の式（$w_{a,b} = 1$と仮定）を用いることで，図 2-7 のようにして様々な波形$x(t)$を表現することができる．

$$x(t) = w_{1,0}\frac{1}{\sqrt{1}}\Psi(\frac{t-0}{1}) + w_{2,0}\frac{1}{\sqrt{2}}\Psi(\frac{t-0}{2}) + w_{2,1}\frac{1}{\sqrt{2}}\Psi(\frac{t-1}{2})$$

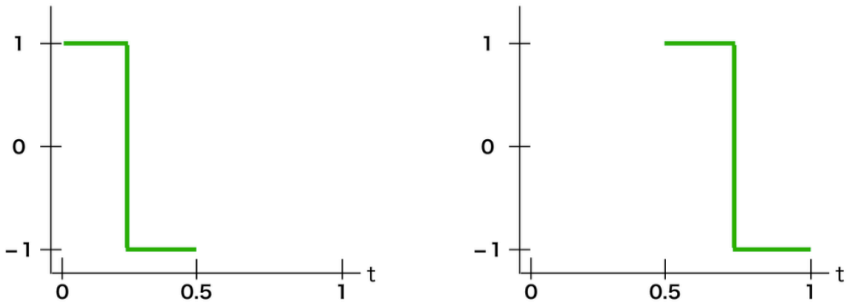


図 2-6 左）$a = 2, b = 0$　右）$a = 2, b = 1$

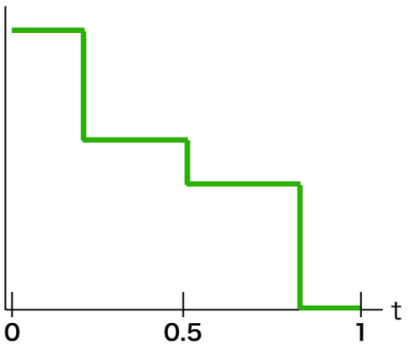


図 2-7 基底の重なりによる波形の生成

以上より，$\boldsymbol{w_{a,b}}$におけるaは波形の伸縮（周波数に関する情報）を，bは Wavelet の並行移動（時間に関する情報）を示し，aによって定められた波形が時間bにどれだけ含まれているかを$\boldsymbol{w_{a,b}}$が示す．これをふまえると，波形$x(t)$と Mother Wavelet $\Psi(t)$に対して，適当な$\boldsymbol{w_{a,b}}$を用いた解析が，以下の式で示す Wavelet 変換となる．なお，Wavelet 変換はフーリエ変換とは異なり，非定常性の波形を対象とすることができる．

$$\boldsymbol{w(a,b) = \int x(t) \frac{1}{\sqrt{a}} \overline{\Psi\left(\frac{x-b}{a}\right)} dx}$$

2-3-2 SIFT 特徴量とは

この節では，SIFT 特徴量について，以下の手順で簡単に説明する．

1. scale space の構築
2. initial keypoint (特徴点)の検出
3. unsuitable keypoint の削除
4. keypoint の位置補正
5. orientation の算出
6. keypoint descriptors の算出

ステップ 1.

scale space は解像度の異なる同一画像を scale 軸方向に重ねたものであり，SIFT においては異なる画像サイズごとに重ねる（図 2-9）．Octave 数は画像サイズに依存し，ダウンサイジング後の一辺が任意の値を下回るまで繰り返す．繰り返し回数が，Octave 数となる．ただし，[12]では Octave 数 3 および σ=1.6，$k=\sqrt{2}$（レイヤー数 5）の設定が推奨される．

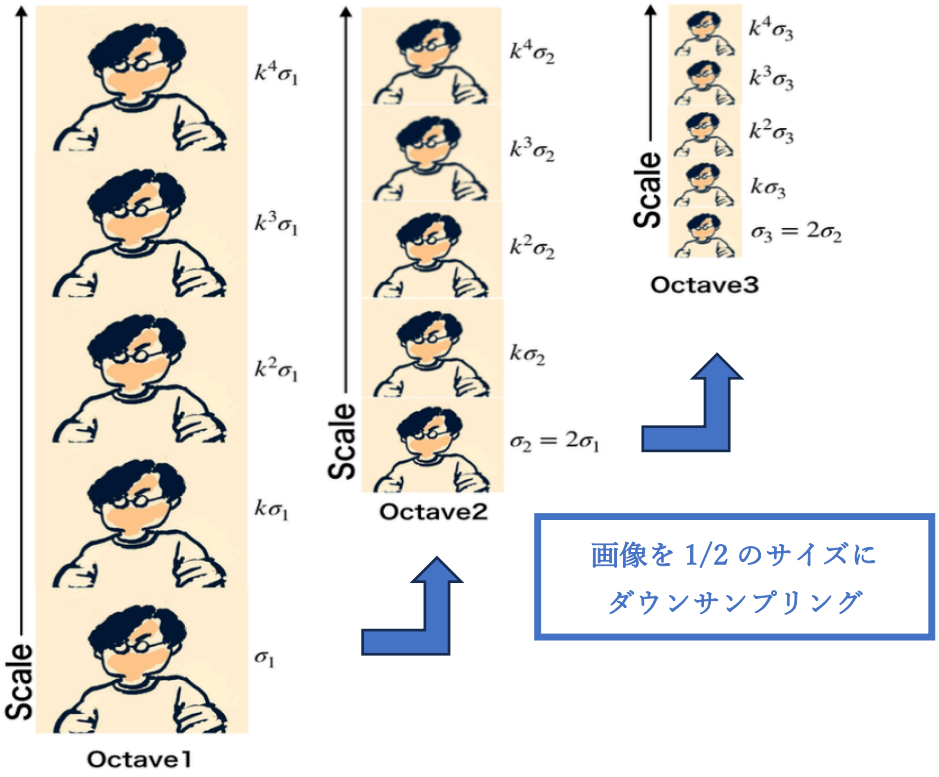


図 2-9 scale space の参考

ステップ 2~4.
initial keypoint の検出~keypoint の位置補正までにおいては，Difference of Caussian（DoG）画像を利用する．

まず，DoG 画像$D(\sigma)$は，スケールの異なるガウス関数$G(\sigma)$と入力画像Iの畳み込みにより，得られた平滑化画像Lの差分から求められる．そして，kはσの増加率であるため，段階的に大きくされた scale を用いて複数の DoG 画像を求める（参考図 2-10）．

$$D(\sigma) = \big(G(k_1\sigma) - G(k_2\sigma)\big) * I = L(k\sigma) - L(\sigma)$$

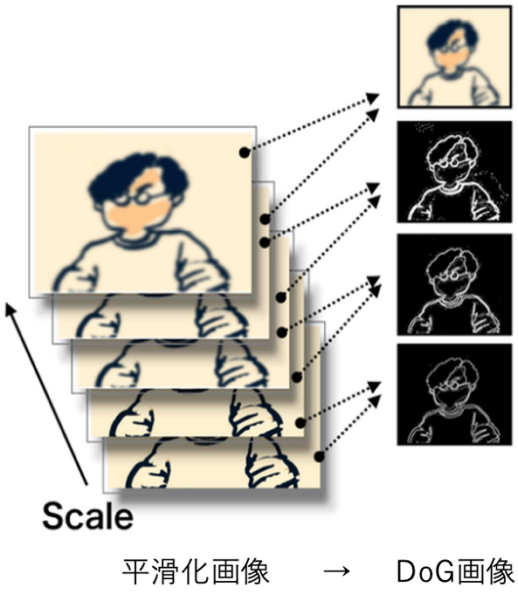


図 2-10 DoG 画像の算出イメージ

続けて，図 2-11 より，隣接する DoG 画像より，注目画素を中心に 26 近傍を比較する．そして，極値となるような近傍画素を keypoint の候補および scale として検出する（ステップ 2）．このとき，keypoint の候補に含まれるエッジ上の点や，極値であっても小さすぎる値を示す点を削除する（ステップ 3）．特に，エッジ上の点では，開口問題（局所領域における一意に対応づけが困難な問題）の影響を受けやすいために削除する．また，keypoint および scale の位置補正は，サブピクセル推定（座標をより細かく推定）によって行われ，推定後の keypoint 候補の周辺から DoG(LoG: Laplacian-of-Gaussian の近似)を出力する．出力結果が閾値以下である場合は，ノイズの可能性を考慮して削除する（ステップ 4）．なお，DoG 画像による keypoint の検出では，σは濃淡情報が多い範囲を自動的に決定するため，同一物体における撮影距離の違いが，そのまま scale の違いとなる．

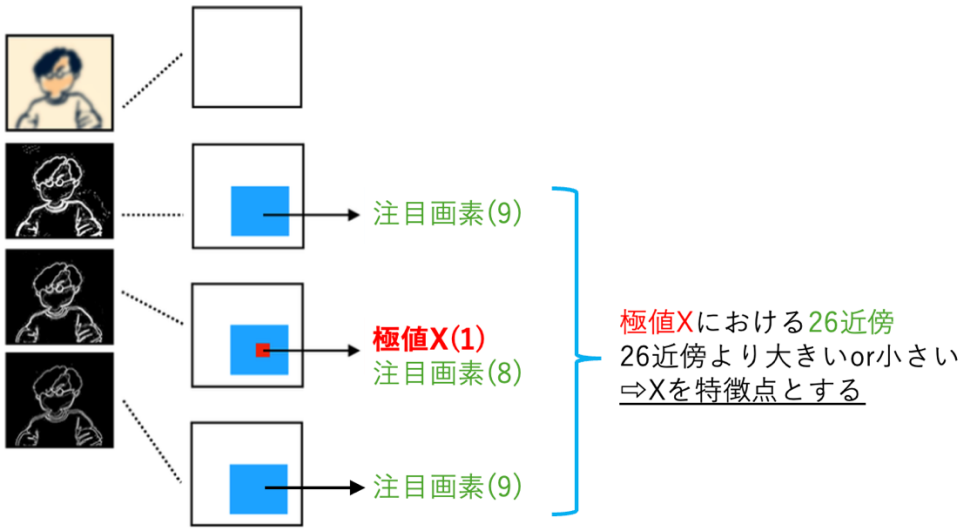


図 2-11 DoG 画像による特徴点の算出

ステップ 5.

orientation (keypoint における方向)は，算出結果を正規化に用いることで，SIFT 特徴量の回転に不変という性質に紐づく．orientation は，局所領域における平滑化画像$L(x,y)$から得られる勾配強度$m(s,y)$，勾配方向$\theta(x,y)$を以下の式より求める．

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2}$$

$$\theta(x,y) = tan^{-1}\frac{f_x(x,y)}{f_y(x,y)}$$

$$\begin{cases} f_x(x,y) = L(x+1,y) - L(x-1,y) \\ f_y(x,y) = L(x,y+1) - L(x,y-1) \end{cases}$$

図 2-12 より，求めた勾配強度と勾配方向から，重み付き勾配方向ヒストグラムを作成して，orientation を決定する．勾配方向は 36 方向に量子化し，画像の中心に近いほど勾配強度を高く重み付けする．その結果は，8 割以上の強度を示す方向を orientation とする．

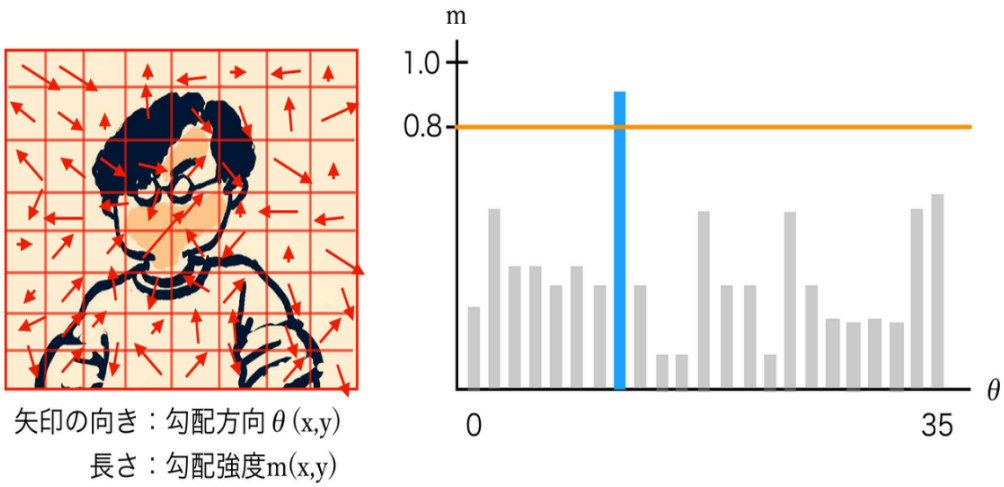


図 2-12 重み付き勾配方向ヒストグラムによる orientation の決定

ステップ 6.
orientation からなる短形領域を 16 分割した後，orientation の方向に回転させる．そして，短形領域に勾配強度と勾配方向を記述し，ブロックごとに 8 方向の勾配方向ヒストグラムを作成する．結果として，分割数と方向数の積が特徴量の次元数（16*8=128）となる．以上が SIFT 特徴量のアルゴリズムである．

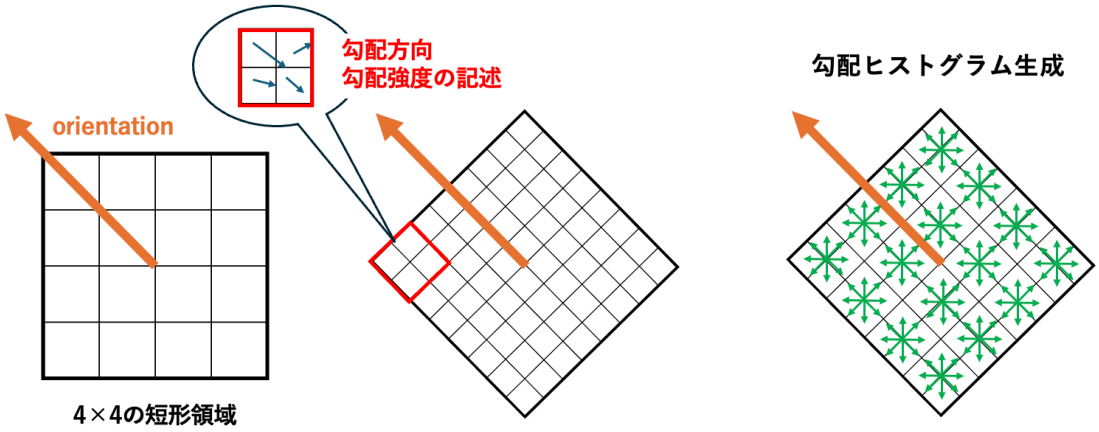


図 2-13 keypoint descriptors の算出手順

2-3-3 Wavelet 解析と SIFT 特徴量を用いたラベリング方法

ここでは，2-2-2 で述べた Wavelet 解析と SIFT 特徴量を用いたラベリングについて，その手順を述べる．まず，Wavelet 解析では，基準音声および連続音声を画像（スカログラム）に出力する（参考:図 2-14）．出力結果より，どの時間帯にどんな特徴があるかを把握できる．

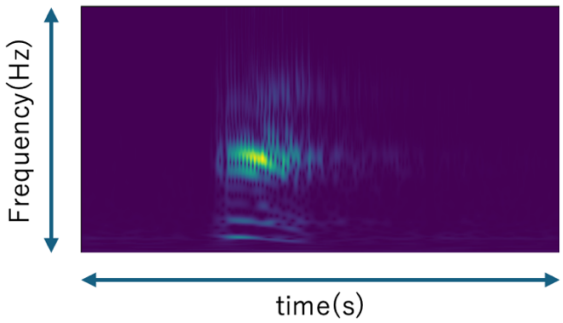


図 2-14 本研究における Wavelet 解析の出力例(700×360)

次に，Wavelet 解析から得た画像（基準音声と連続音声）に対し，SIFT 特徴量によるマッチングを図る．具体的には，画像ごとに抽出した特徴点に対し，ユークリッド距離から総当たりで比較する．そして，特徴点が対応する点の組み合わせdは以下の式より求め，最も小さいdを対応点とする．

$$d(V^{k_{I1}}, V^{k_{I2}}) = \sqrt{\sum_{i=1}^{128}(v_i^{k_{I1}}, v_i^{k_{I2}})^2}$$

（kは keypoint，v^kは keypoint の特徴量）

最後に，図 2-15 は，左右の画像は収録地点の異なる同一音声と無音区間を含み，これらの画像に対して SIFT 特徴量でマッチングさせた結果である．特に，対応点は k 近傍法（k=2）によって限定されてはいるが，対応点は部分的に存在した．これは，基準音声に対応する箇所を，連続音声に対してラベリング可能であること意味する．

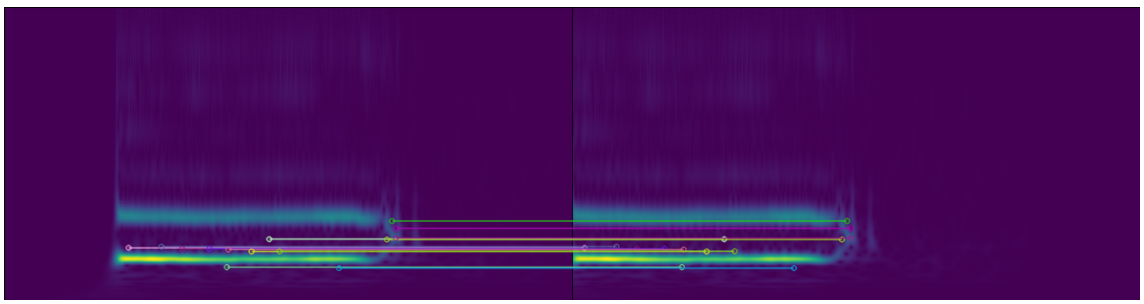


図 2-15 Wavelet 解析の結果(画像)に対する対応点の出力

2-4 Wavelet 解析と SIFT 特徴量を用いた少数言語ラベリングへの支援

2-3 節より，少数言語のラベリングを Wavelet 解析と SIFT 特徴量を組み合わせた手法で行うことを提案した．ここでは，実際の音声で，適切なマッチングが行われるかを，実験により確認する．そのために，マッチング対象における音韻情報のズレを，提案法によって検出する．このズレが検出できたとき，連続音声のどの位置からでも，基準音声に一致または類似箇所を検出できるといえる．

2-4-1 実験条件

実験では図 2-16 のようにして，人工的なズレを作成した．具体的には，日本語 50 音と英語音素（単体で発話可能な音素のみ）を 1sec で収録し，同じ音源で収録位置の異なる状況（ズレ）をペアで作成した．そして，同言同士と異なる音同士の 2 パターンから，音韻情報のズレを解消する効果を検証した．

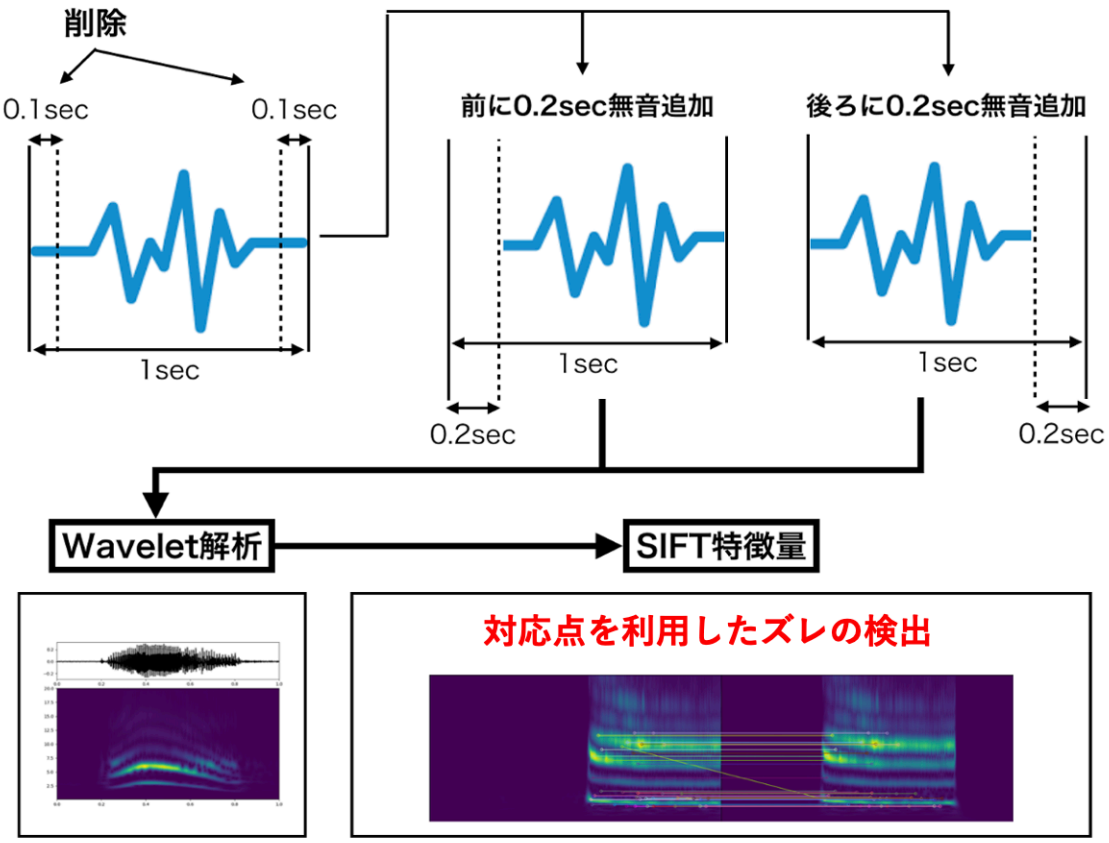


図 2-16 実験：対応点を利用したズレを検出するまでの流れ

Wavelet 解析の入力時間は 1sec であり，出力画像の横軸は 700pixel であった．このことから，画像のズレによって時間のズレを把握することができる．

ズレを検出するにあたって，Wavelet 解析はスカログラム（画像）を出力し，この出力に対して SIFT 特徴量による対応点を算出する（各特徴点は 128 次元）．ここで，SIFT 特徴量において，以下の式を満たす場合を対応点とする．

$$\frac{d_1}{d_2} < k$$

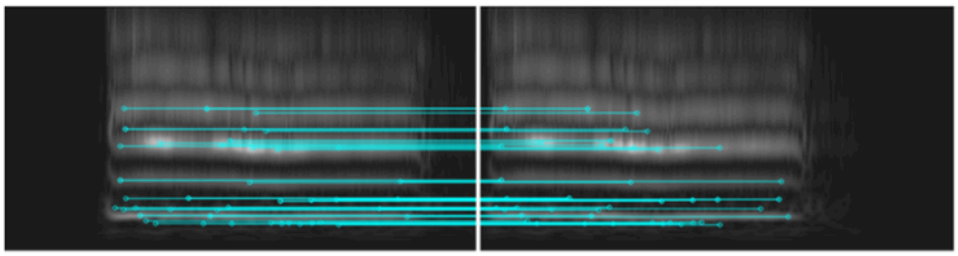
このkは任意で設定され，値が小さいほど対応点は減少しやすく，大きいほどに増加しやすくなる．当然，kの値が小さいほどに，対応点の信頼度は高くなる．本研究では$k = 2$を設定した．なお，SIFT 特徴量算出における設定パラメータは，keypoint 数:上位 500，σ:1.6，scale レイヤー数:5（σの増加率$k = \sqrt{2}$），keypoint 候補を絞り込むための閾値:0.04，エッジ上の候補を削除する際の閾値:12.1，Octave 数:3 である．また，Wavelet 解析の利用にあたっては，マッチングに必要な情報を抽出できる Mother Wavelet を選択する必要がある．本研究では，特に人間の聴覚などに関連する Morlet を Mother Wavelet に定め，Wavelet 解析を用いた．

対応点によるズレの検出は，Wavelet 解析の入力時間と出力画像の横軸で比をとることで算出する．例えば，対応点の画像的距離が 350pixel であるとき，1sec あたり 700pixel と比をとることで，対応点は 0.5sec の距離と考える．ズレを時間的に把握することで，図 2-3 のようにおおまかな音を基準音声としても，連続音声のどの時間帯に基準音声の部分的な特徴が含まれるかを判断できる．

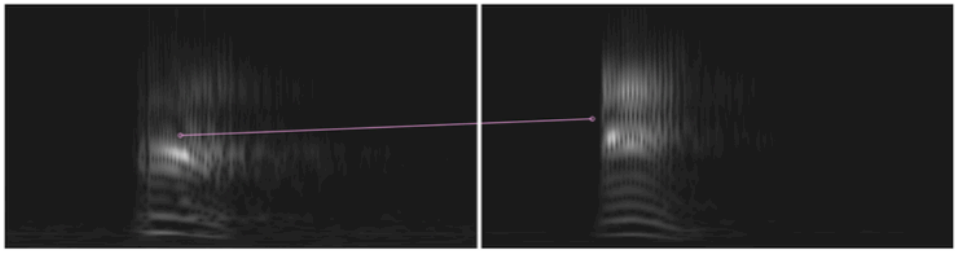
2-4-2 実験結果

対応点の算出結果を図 2-17 に示す. 同音同士の場合は複数の対応点が得られたのに対し，異なる音同士ではほとんど得られなかった．同音同士（英語音素）では対応点が，平均して 25 点得られた．これら 25 点のズレの平均を全体のズレとしたとき，146pixel のズレであった．そして，Wavelet 解析においては 1sec が 700pixel に換算されたため，146pixel は 0.2085sec のズレと言い換えられる．このとき，同音同士では図 2-16 より 0.2sec の人工的なズレが設定されるため，0.2085sec はおおよそのズレを検出できたといえる．そして，異なる音同士では対応点がほとんど得られなかったことから，図 2-3 における基準音声の部分的マッチングに対しても，不用意に対応点を示さないことがわかった．これは，基準音声に対して，おおよその音を定めることが可能になったことを意味する．

以上の結果より，Wavelet 解析と SIFT 特徴量を組み合わせたマッチングは，少数言語のためのラベリングが可能であることが示された．



(a)同音同士の結果



(b)異なる音同士の結果

図 2-17 対応点の算出結果

2-5 まとめ

少数言語音声の書き起こしにおいて，従来法の多くは事前知識や大量の学習データを必要とした．これは，少数言語データが不足する状況に適していないため，少数言語音声の書き起こしに対して従来の音声ラベリング手法を適用することは困難であった．従来法では，音響特徴量 MFCC に対して DP マッチングを用いることで，ラベリングを実現しているが，MFCC の算出における要件を少数言語は満たせない．そして，DP マッチングの必要性は MFCC の使用に依存するため，少数言語のためのラベリング手法を新たに提案する必要があった．

本研究では，Wavelet 解析と SIFT 特徴量を組み合わせた少数言語のためのラベリング手法を提案した．特に，音響特徴量 MFCC の代わりに Wavelet 解析を用いることで，完全な時間情報を基準となる音に紐づけることができる．この時間情報を用いることで，どの時間帯のどの特徴量がマッチングするのかを把握できた．特徴量のマッチングに際しては，Wavelet 解析を画像的に出力することで，SIFT 特徴量によるマッチングを提案した．特に，SIFT 特徴量は画像の縮尺に加えて欠落・追加にも強く，マッチングに際して対応点を出力することを利用した．その結果，少数言語の基準となる音より，一致または類似する箇所を連続音声に紐づけることができた．これは，少数言語のための緩やかなラベリングが，提案法によって可能になったことを意味する．

最後に，提案法による緩やかなラベリングは，完全な自動化ではないが，ラベリング作業の効率化には貢献すると考える．具体的には，基準となる音をおおまかにでも定めることで，連続音声を暫定的な記号でラベリングできる．そして，ラベリングされていない箇所は，新たな知見として，新規のラベルを付与するなどの工程となる．結果として，連続音声の全てを手作業で書き起こす負担の低減に対し，提案法は有効的であるといえる．

3章 書き起こされた少数言語テキストの解析

3-1 はじめに

書き起こされた少数言語テキストの解析では，単語や文法などを明らかにするが，それには段階が存在する．この段階は，日本語や英語などを対象とする一般のテキスト解析でも同様にあり，分かち書き→形態素解析→構文解析→意味解析と分別される．そして，これらの段階に対して自然言語処理技術を利用することで，作業の効率化として支援する．

少数言語のテキスト解析を支援するにあたって，本研究では分かち書きに着目した．分かち書きは，最も基礎的なテキスト解析であり，「ある文から意味のある塊に切り分ける」タスクである．特に，この塊にラベル付を行う段階が形態素解析であり，構文解析および意味解析では，塊の組み合わせ（並び）を解析する．いずれの段階においても，言語固有の知識を必要とする[27]. そして, 本研究では, 品詞などの言語固有の知識が全くない状況の下で, 少数言語におけるテキスト解析の最初のステップとして，分かち書きを支援対象とした．

少数言語に適する分かち書き手法を検討するにあたっては，言語固有の知識を必要としない手法が必要である．加えて，機械学習に基づく手法が一般であるため，本研究では教師なし分かち書き手法に着目した[28][29][30]．このうち，[28][29]では，学習パラメータを言語ごとに定めるため，言語固有の知識が全くない少数言語に適さない．対して，[30]（NPYLM: Nested Pitman-Yor Language Model）では，言語固有の知識を一切用いないで，分かち書きすることができる．これは，少数言語テキストの分かち書きに適するため，本研究では，完全教師なし分かち書き手法として NPYLM に着目した．

NPYLM では, 未知語や未知言語, 口語を対象に, 分かち書きすることができる. しかし, NPYLM は機械学習に基づくため，多量の学習用データが必要となる．少数言語の場合は, この学習データには書き起こしされたテキストが想定されるが, 量は限られる. 具体的には, NPYLM が十分な性能を示す量に対して，1/10 程度の量しか用意されない．特に，少量の書き起こしであっても，手作業で数ヶ月かかる．そのため，残り 9 割を用意することは現実的ではないため，少量データという制約の下，NPYLM を少数言語に適用する必要がある．統計的に処理をする NPYLM においては，少量データでは不十分な分割性能をもたらす．ここでいう不十分な分割性能とは，単語未満や一文字単位の分割結果を意味しており，本研究ではこのような分割を過剰分割とよぶ．特に，本研究における分かち書きは「ある文から意味のある塊に切り分ける」タスクとして扱われる．このとき，過剰分割という細かすぎる分割結果は，塊に対する意味を取りづらくなってしまう．そこで，本研究では，少数言語への適用を見据え，少量データによる NPYLM の過剰分割の改善を目的とする．本章では，少量データである場合の NPYLM の挙動を示しつつ，過剰分割の詳細と改善方針を示す.

3-2 NPYLM と NPYLM におけるデータ量と性能の関係

NPYLM とは，Hierarchical Pitman-Yor Language Model（HPYLM [31]）と Variable Pitman-Yor Language Model（VPYLM [32]）をほぼ同時に学習し，HPYLM の基底測度に VPYLM の結果が埋め込まれた言語モデルである[30]．両モデルの役割を整理すると，VPYLM は単語を，HPYLM では単語の並びを学習する．そして，いずれのモデルにおいても n-gram 言語モデルに基づき，分割結果には文法などの言語特徴も反映される．ここで，n-gram 言語モデルについて説明する．例えば，文字 n-gram における"tak"に続いて"e"(take)が出現する確率を求め，単語 n-gram でも同様に，ある単語に他の単語が続く確率を求める．そして，NPYLM では，文字および単語 n-gram モデルが Pitman-Yor 過程と呼ばれる確率過程に基づいて生成され，両モデルを階層的に学習する．その結果，与えられたテキストに対して，最適な言語モデルと分かち書きの結果を得る．NPYLM は上記のしくみによって，分割を教師なしで実施するため，事前知識が十分でない少数言語の解析に，NPYLM は有効であろう．しかし，保存活動で収集される少数言語の文章量は限定的であり，統計に基づく NPYLM は必ずしも十分な性能を発揮するとは限らない．NPYLM は半教師あり学習，意味解析などのより高度な解析に展開されており，学習データ量が不足する状況は十分に議論されていない[33]．

ここで，NPYLM の学習データ量と分割性能の対応関係について，予備実験から確認した．予備実験では，日本語テキスト（ひらがな，句点などの記号群は削除）を対象に，データ量および学習回数を変化させつつ，NPYLM の分割性能を再現率・適合率・F 値により評価した．評価にあたって，正しい分割個所は MeCab で求め，再現率は分割できた正しい箇所の割合を, 適合率は分割した箇所の正答率を示す. データ量は 100,200,300,500,1000,2000 文のデータ 6 種を，学習回数は 100,200,500,1000,2000 の 5 パターンを用意した．

予備実験の結果を図 3-1 に示す．学習回数に関しては，F 値を見る限りは性能への貢献度は低い. 対して, データ量(文数)では, 多いほどに F 値も高くなるため, データ量が NPYLM の性能へ貢献することがわかった．また，文数が 200 の場合が，本研究で扱う少数言語のデータ量に相当する．この場合において，F 値は 7 割程度を示し，単純に 3 回に 1 回の割合で誤って分割する．また，再現率と適合率では，再現率は 9 割程度を示しながらも，適合率は 6 割程度にとどまった．このように再現率が極端に高く適合率が低い場合，余計な分割個所が多いといえる（以降，この状況を過剰分割と呼ぶ).

学習回数	100			200			500			1000			2000		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
文数															
100	0.86	0.62	0.72	0.88	0.61	0.72	0.89	0.58	0.70	0.91	0.59	0.72	0.92	0.56	0.70
200	0.89	0.63	0.74	0.89	0.63	0.74	0.92	0.61	0.74	0.92	0.61	0.73	0.92	0.62	0.74
300	0.84	0.68	0.75	0.86	0.68	0.76	0.90	0.65	0.75	0.85	0.77	0.80	0.90	0.64	0.75
500	0.85	0.69	0.76	0.89	0.69	0.78	0.89	0.67	0.76	0.92	0.65	0.76	0.91	0.65	0.76
1000	0.85	0.74	0.79	0.87	0.72	0.79	0.88	0.71	0.79	0.90	0.71	0.80	0.91	0.70	0.79
2000	0.81	0.77	0.79	0.83	0.76	0.79	0.85	0.75	0.80	0.88	0.74	0.81	0.87	0.75	0.80

図 3-1 予備実験結果：分割性能に対するデータ量および学習回数の対応関係

続けて，本研究で扱う少数言語テキストと同程度の英語・日本語テキストを 10 セットずつ用意し，過剰結合が言語に依らずに引き起こされることを調査した（学習回数 500）．この調査では，日本語はローマ字で記述されたテキストを，英語はアルファベットで記述されたテキストを用いた．いずれのテキストも小文字かつ 25,000 文字のデータ量に統一され，余分な記号は削除した．この統一は，言語間の差を最小限にするための処理である．

調査の結果，再現率が高く適合率が低い状況は引き続き確認でき，少量データの場合には，言語を問わず過剰分割を引き起こしやすいことがわかった．実際の過剰分割の例を図 3-2 に示す．日本語の場合，shizuka という単語が shi/zuka という余分な分割箇所を持ち，souzou は s/ouzou という一文字単位で切れてしまう事例があった．続いて英語の場合，after や spring などで同様の事例があった．このとき，英語の ing で分割されたことは，動名詞や活用形などの事前知識があれば，いくつかの過剰分割を汲み取られる．また，英語の「-ly，-ed，re-」や日本語の「は，の，に」などはデータ量に関わらず多用されがちであるため，統計的な学習においては分割されやすいと考えられる．しかし，いずれにせよ事前知識が無い場合は，過剰な分割であること以上のことは汲み取れない．

以上より，NPYLM はデータ量が多いほどに性能を発揮し，少ない場合は過剰分割として性能が低下した．

日本語
分割対象：shizukanashinzouwosouzousurunitahenuhodoni
　　正解：shizuka na shinzou wo souzou suru ni tahe nu hodo ni
過剰分割：shi zukana shin zou wo s ouzou suru ni ta hen u h odo ni

英語
分割対象：afterthefirstrushofthedriveonespringthebossleftmarkand
　　正解：after the first rush of the drive one spring the boss left mark and
過剰分割：af ter the fir st r u sh of the drive one s p r ing the b o o s le ft mark and

図 3-2 過剰分割の実例

3-3 少数言語の分かち書きにおける過剰分割とその要因

予備実験より，少数言語のように少量データでの NPYLM は，過剰分割という状況にあった．この状況は，事前知識を持たない少数言語の分かち書きにとって，解消すべきである．本節では，解消すべき理由について述べつつ，過剰分割の要因について示す．

まず，少数言語の分かち書きにおいて，過剰分割を解消すべき理由について述べる．少数言語の分かち書きは，テキストから意味のある塊を切り出すタスクに言い換えることができる．このとき，切り出した塊は文法などの把握に役立つが，塊が細かすぎる場合は，かえってその把握が困難となる．具体的には，図 3-2 で示した s/ouzou などの一文字の分割が，過剰分割の下では大量に生じた．したがって，少数言語の分かち書きを NPYLM で支援するにあたっては，少量データに起因する過剰分割を改善する必要がある．

次に，過剰分割の要因について述べる．NPYLM は統計的に単語および単語の並びを学習するため，高頻度語と低頻度語が混在する場合は，単純に高頻度語の方が学習しやすいと考えられる．実際，NPYLM では区切りの無いテキストから単語を学習するが，記述文字の数が限られているために頻出文字列はある．そして，頻出文字列は単語である場合と，別単語の部分文字列である場合がある．例えば，頻出文字列 the があるとき，the という単語だけでなく，they など別単語の一部である可能性がある．このとき，they が低頻度語である場合は the を学習できるが，they は学習結果から棄却されやすいことが予想される．その結果，they が the/y として切り出され，過剰分割という状況になる．頻度に着目することで，過剰分割は類似文字列における頻度の競合状態が引き起こすといえる．

続けて，NPYLM の学習に着目し，過剰結合の要因について議論を進める．先の競合状態は，NPYLM において単語を学習する VPYLM で発生する．VPYLM では，Pitman-Yor 過程に基づいて，文字の並びに関する木構造を構築する．この木構造では，学習対象の文書からのサンプリングを繰り返すことで，ノードの強化や削除を行う．このとき，学習データ量が十分な場合は，高頻度語・低頻度語ともに正しくされて木構造が構築される．しかし，学習データが不十分な場合は，この木構造がうまく構築できていない可能性がある．ここで，文字の並び（単語）は確率的に学習され，長い単語であるほど確率の積が小さくなる．そのため，長い単語よりも短い単語の方が，確率的には優位にあることで正しく学習されやすい．これは，木構造が深くなりにくいことを意味する．その結果，長い単語において学習しきれなかった箇所が取り残され，先の過剰分割が引き起こされたと考えた．

以上より，少数言語の分かち書きでは過剰分割が問題であり，少量の学習データの場合は低頻度語が木構造から欠落したことに起因する．特に，欠落が生じやすい状況というのは，木構造が深くなりづらい状況と言い換えることができた．

なお，VPYLM は可変長 n-gram 言語モデルとも呼ばれ，文脈に応じた長さで単語を学習することができる．これは，実際の文書においては，長さに関わらず単語が存在することに適する．このとき，このモデルが対応する長い単語とは，高頻度なものを指しており，低頻度なものには対応しない．この低頻度かつ長い単語は，出鱈目な文字列である可能性が高いため，高頻度に伴う対応であることは当然である．また，どこからが低頻度なのかを定めることは容易ではないが，ある程度までの低頻度単語は，ポアソン補正によって学習することができる．しかし，学習データが不足する場合，低頻度の程度が悪く，補正の対象外になりやすい．特に，ポアソン分布に要するパラメータは，与えた学習データから自動的に推定する．そのため，十分なデータ量であれば汲み取れた単語 A が，少量データではこの推定に耐えきれず，所望の長い単語が学習しきれない．

3-4 過剰分割への対策方針

本節では，過剰分割への対策方針を示す．前述のとおり，学習データ量の不足する場合は低頻度語が木構造から欠落しやすく，結果として過剰分割を引き起こしていた．そして，この欠落のしやすさを，VPYLM における木構造が深くなりづらい状況と言い換えた．以下，簡単な例から，この状況について議論を進めつつ，対策方針を示す．

まず，VPYLM における木構造が深くなりづらい状況について述べる．例えば，the が高頻度語であり，that や this が低頻度語であるような英文を学習した場合を考える．この場合において，the は期待どおり観測できるが，that については，th に続く a や i を十分に観測できない．これを木構造に当てはめて考えると，the よりも that や this の方が，長い文字列であるために，もとより確率の積は小さくなる．そして，低頻度な条件が追加された場合には，欠落という結果になりやすく，最終的な木構造も深くなりづらい．特に，the という観測結果を除けば，多量に観測される th を優先的に区切ってしまう結果をもたらす．なお，短くかつ高頻度が優先的に区切られることは，通常の学習性能である．あくまで，もとから観測が困難であった長い単語が，少量データではより困難になったといえる．これは，少量データでは，期待する単語のサンプリング数も限られてしまうためである．

次に，対策方針ついて述べる．少量データにおいて，木構造を深くすることは困難である．そこで，本研究では，「置き換え」という操作を提案する．この操作では，①各ノードで扱う文字数を増やす，②部分文字列が一致する文字列において頻度が競合する状況を解消する，という二つの効果を持つ．引き続き，the と that,this の例から「置き換え」の効果を説明する．この例において，th に続く文字には e が頻度的に目立ち，a や i が相対的に目立たない．そして，単語の長さが VPYLM における木構造の深さであることから，より長いである that や this は確率的に学習しづらい．このとき，各ノードで扱う文字数を増やすことによって，それほど深い木構造でなくとも，長い単語を学習しやすくなると考えた．例えば，th を α という一文字で置き換えた場合，続く文字 at や is にかかる深さが，置き換え前よりも浅くなる．その結果，that や this が期待通り観測されるなどして，過剰分割の解消につながると考えた．また，頻度差に関しては，the と th を区別して置き換えることで，ノードの追加・削除において頻度が競合する状況を解消する．例えば，the を β，th を α と別記号へ置き変える．その結果，それぞれの木構造で学習が進むため，この状況は解消される．

VPYLM 対する置き換え操作の作用について，簡単なイメージ図を図 3-3 に示す．この図に従い，頻度が競合する状況と木構造が深くなりづらい状況は，置き換えによって共に解消される．また，this や that が th から派生が期待されるように，the から them などが新たに得られる可能性がある．このとき，the を置き換えても，the という単語が得られなくなるわけではない．後述するが，the などの所望する単語が置き換えられても，置き換え対象の 9 割程度は所望の区切りが維持されていた．これは，あらゆる長さの単語を学習できるという VPYLM の性質が，置き換え後にも維持されたことを意味する．

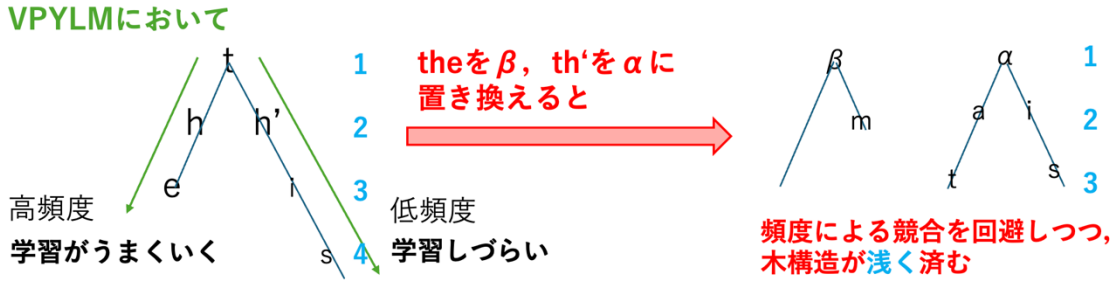


図 3-3 VPYLM に対する置き換え作用のイメージ

なお，「置き換え」は VPYLM の木が成長した後では効果が薄いと思われるので，学習前に行うこととする．置き換えにあたり注意すべきは，置き換え対象の語の選択である．本研究では，NPYLM を一度適用することで，語の候補を得ることを提案する．これまで述べてきたように，十分でない文書で NPYLM を学習した場合，得られる単語は過剰に分割されたものになりがちである．言い換えると，一度目の NPYLM の学習で得られた単語(の候補)は，高頻度語であれば所望の語そのものが，低頻度語であればその単語の断片が抽出される．したがって，the と th を区別して置き換えることが可能であり，先の効果を発揮することができると考えた．

4 章 NPYLM の 2 段階適用

4-1 はじめに

NPYLM において，学習データ量が不足する場合は，過剰分割が引き起こされていた．3 章では，この過剰分割を解消するため，「置き換え」操作を対策方針に挙げた．本章では，この操作を「NPYLM の 2 段階適用」として提案し，過剰分割の解消に有効であることを確認する．この確認において，過剰分割が解消されるとき，短い単語が減り長い単語が増加するとした．これは，過剰分割においては，短い単語が分割結果を多く占めるためである．

本章の構成は，4-2 節で提案法手順を示し，4-3 の実験より過剰分割の解消に対して提案法が有効に働くことを確認した．なお，実験は 2 つ実施した．1 つは，NPYLM の性能と学習回数の対応関係を調査するものである．特に，提案法では NPYLM を二回実施するため，素の NPYLM よりも単純に 2 倍の学習回数となり，学習時間も倍となる．そのため，提案法では学習時間に比して，性能が向上することを調査した．2 つめは，提案法が過剰分割の解消に対して提案法が有効に働くことを調査した．

4-2 提案法の手順

提案法の手順を図 4-1 に示す．提案法では，与えられたテキスト 1 回目の NPYLM で学習し，置き換え候補を得る．続けて与えられたテキスト内の置き換え候補の語を，一文字に置き換える．最後に，2 回目の NPYLM の学習を行い，過剰分割を軽減した分割結果を得る．具体的な手順を次ページに示す．

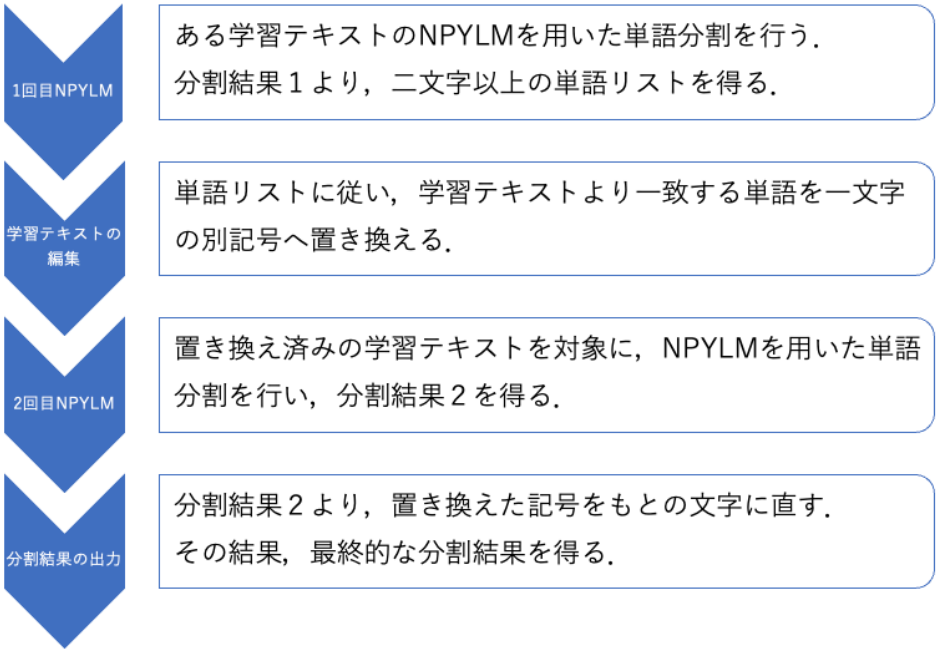


図 4-1　提案法の手順

＝＝

手順１：単語候補とその出現頻度のリスト作成

与えられたテキストで NPYLM を学習し，単語分割を実施する（1 回目の NPYLM）．この分割結果から，長さが二文字以上の単語を抽出し，単語候補（文字列）およびその出現頻度のリストを作成する．

手順２：学習用テキストの生成

頻度上位の候補から順に一語ずつ取り出し，すべての出現箇所を元文書に含まれない記号（αや①などの一文字）で置き換える．このとき，一語置き換える毎に，これまでに置き換えた文字において，与えられたテキストを占める割合を求める．この割合がP_{rep}を超えたとき，置き換え処理を終了する．なお，この置き換えは，1 回目の NPYLM で切り出された語に従って行う．すなわち，the を置き換える場合，1 回目の分割で the と切り出している部分のみが置き換え対象であり，1 回目の分割で there と切り出している部分は対象としない．

手順３：生成された学習用テキストによる NPYLM の構築

1 回目の NPYLM の結果を破棄し，置き換え後のテキストで NPYLM を学習する（2 回目の NPYLM）．

手順４：2 回目の NPYLM を用いた単語分割

学習した NPYLM を用いて置き換え後のテキストを分割し，置き換えた文字を元に戻す．これが，与えられたテキストに対する分割結果となる．

＝＝

手順に従って，学習テキストの変化を図 4-2 に示す．手順 1 では，素の NPYLM から単語分割し，二文字以上の単語リストを作成する（Targets）．ここでは，3-4 の例で示した the や th（that, this の一部）などの置き換え候補が，1 回目の NPYLM から得られると考えた．手順 2 から 3 にかけては，リストに従って別記号に置き換えられたテキストを，NPYLM で学習・単語分割を実施する．特に，手順 2 では，あらゆる候補を積極的に置き換える．このとき，the と切るべきところを誤って them に切る可能性を少しでも避けるため，長いかつ低頻度な語の一部のみを置き換え対象としたい．しかし，1 回目 NPYLM から得られた候補が，一部か否かの判断は事前知識なしでは困難である．そのため，置き換え規模を調整する P_{rep} を設けた．最後の手順 4 では，単語分割に含まれる別記号を，リストに従って元の単語に戻す．また，図上では①などで置き換えている．これは，元文書に含まれない文字であれば問題ないが，念のため全く違う記号種を用いることが好ましい（英語テキストに対し，漢字で置き換えるなど）．

なお，提案法は NPYLM を単純に適用する場合の 2 倍の学習回数を要する．学習回数の増加分の効果が得られるかどうかについては，次節で検討する．

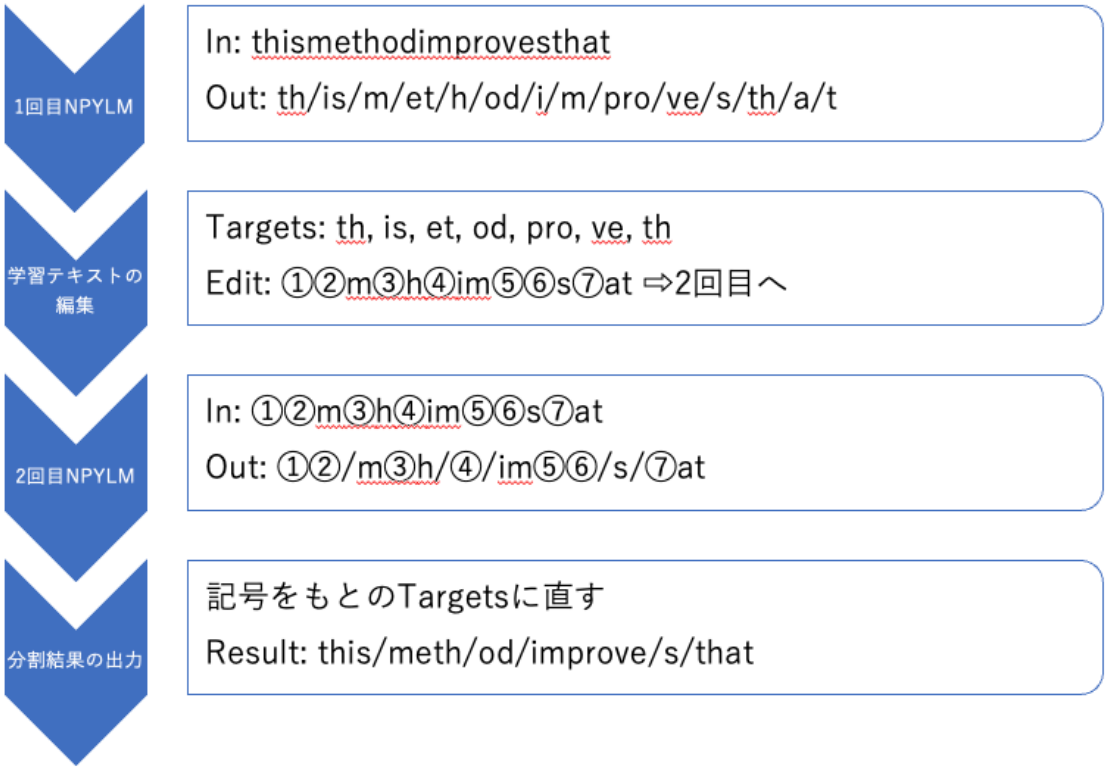


図 4-2　図 4-1 の手順に従う学習テキストの変化

4-3 実験

提案法は，過剰分割の改善を目的とした．ここでは以下 2 つの実験から，提案法の有効性を議論する．なお，NPYLM は，言語非依存な手法であるという特徴を持つ．そこで，提案法においてもこの特徴が維持されることを確かめるため，英語と日本語，少数言語から提案法の有効性を確認する．

実験 I ：NPYLM の分割性能と学習回数の関係
実験 II ：提案法の過剰分割に対する有効性の確認

本実験では，3 つの言語（日本語，英語，ある少数言語）を対象に，提案法の有効性を確認する．日本語および英語のテキストとして，英小文字のみで表記された 1 セットあたり 25,000 文字のテキストを 10 セットずつ用意した．いずれのテキストも物語文を出典としており，日本語テキストはローマ字に変換して英小文字以外を除いたものを，英語テキストは空白および不要な文字を除き英小文字に変換したものを用いた．また，本実験の少数言語テキストは英小文字のみで構成され，空白を除いた 7,601 文字のテキストを用いた．実験の対象となった少数言語は，日本語に類似した音素体系を持っており，この言語は英小文字で書き起こされた．現在，収集されたデータは専門家による解析の途中であり，一部の単語の抽出が終了した状況である．

なお，今回使用した少数言語テキストにおいて，言語学者によって分割されたものは 7,601 文字であった．これに対し，収集されたテキストデータは約 25,000 文字であったため，英語および日本語の文字数もこれに合わせて 25,000 文字とした．

評価にあたり，正しい分割箇所には，日本語テキストは原文を形態素解析器(MeCab)により分割した結果を，英語テキストは原文の空白の位置を，少数言語テキストは言語学者らにより分割された結果を用いた．分割箇所の正確さは，これらの正しい分割箇所と NPYLM による分割結果とを比較し，再現率・適合率・F 値より評価した．

4-3-1 実験 1 ：分割性能と学習回数による評価

この実験では，NPYLM の分割性能と学習回数の関係について調査した．前節で準備した英語テキスト（25,000 文字）を対象にして，ある学習回数ごとの分割性能を確認した．具体的には，同一のテキストに対して，100, 200, 500, 1000 回学習した後に，分割箇所に関する F 値を求めた．

結果の図 4-3 より，縦軸は F 値を，横軸は学習回数を示す．学習回数の増加に対して F 値の値（実線）の変化は少ないが，学習時間（点線）は学習回数と比例して増加した．つまり，提案法において 1 回目，2 回目の NPYLM の学習回数をともに n 回としたときの学習時間はほぼ変わらない．それに対して，オリジナルの NPYLM で学習回数 n 回のときから 2n 回のときにかけて，F 値の増加はほとんど無い．そのため，提案法により，学習時間に比して分割性能の向上が期待できる．

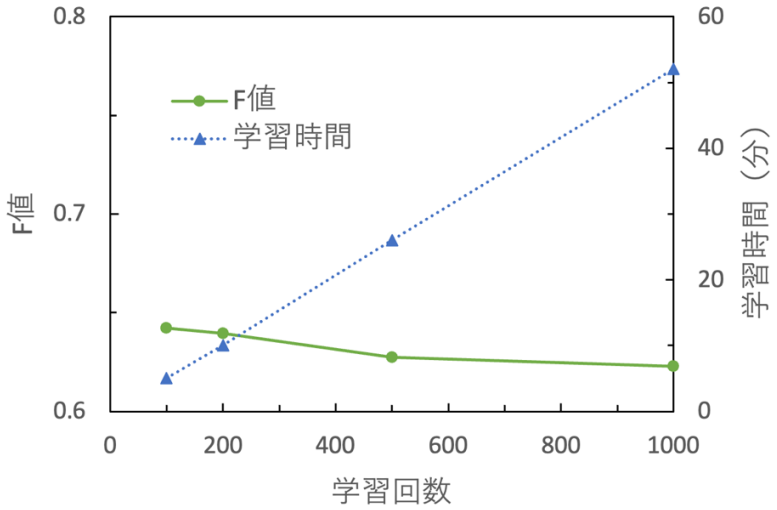


図 4-3 学習回数における学習時間および F 値の変化

4-3-2 実験 2 ：提案法の有効性および言語差の確認

この実験では，提案法の有効性を平均単語長，再現率・適合率，各単語長の分布から確認した．英語および日本語を各 10 セット，少数言語 1 セットのテキストを分割した．1 回あたりの NPYLM の学習回数は，500 回を設定した．これは，図 4-3 において，学習回数 500 回以降では，F 値の値が横ばいとなったためである．分析結果に基づいた平均単語長，再現率・適合率を表 4-1 に示す．いずれの言語も，「適用前」はオリジナルの NPYLM によるものを，「適用後」は提案法によるものを意味する．また，表の各値はP_{rep}における 10 セットの平均値である（少数言語は 1 セットのみ）．

まず，平均単語長に着目する．これは，過剰分割により短くなっていた分割後の単語長が，どの程度長くなったのかという観点であり，過剰分割の改善を間接的に評価するものである．単純に，過剰分割の場合には短い単語が多い（種類数）ため，同じ文字数である限りは短い語が減り長い語が増えることは，細かく切れすぎるという過剰分割の改善が期待できる．結果として，いずれの言語においても，提案法を適用することで平均単語長は長くなった．特に，英語と日本語の場合は，平均単語長が 2 倍以上となった．この結果は，英語・日本語それぞれの真値 6.56，5.70 文字と比べると長くなり過ぎているが，分割後の単語長という観点からは，提案法の効果が示された．

次に，適切な単語境界で分割されているかを評価するため，分割箇所の再現率・適合率に着目する．適合率は，いずれの言語においても改善した．このとき，適合率は，分割結果として得られる分割箇所のうち，正しい分割箇所の割合である．そのため，その改善は余分な分割が減少したことを意味し，過剰分割は改善されたといえる．対して，再現率は，いずれの言語においても悪化した．再現率は，分割すべき箇所のうち，実施に分割した箇所の割合であるため，その悪化は分割が不足することを意味する．しかし，単語だけでなく熟語を想定することで，必ずしもこの悪化が悪いとは言い切れない．

以上より，平均単語長が長くなったことで，過剰分割は解消された．そして，適合率の改善によって提案法の適用後には，分割すべきでない箇所を分割しづらくなったといえる．

表 4-1 提案法による分割結果の評価（$P_{rep} = 1.0$）

		再現率	適合率	平均単語長
英語	適用前	0.95	0.55	4.01
	適用後	0.66	0.71	10.96
日本語	適用前	0.95	0.55	4.01
	適用後	0.67	0.67	8.44
少数言語	適用前	0.86	0.62	3.95
	適用後	0.82	0.64	4.71

まとめ

少数言語の単語分割に対して，NPYLM の適用を検討した．少数言語のテキストは言語に関する事前知識・データ量が，英語などの既存言語と比べて不足しがちであるため，既存言語を前提とした従来の単語分割手法の適用は困難である．特に，従来法の一つである NPYLM は事前知識を用いずにテキストの単語分割を行うが，学習データが不足する場合は，テキストを過剰に分割しがちであった．そこで，学習データが不足する NPYLM による過剰分割を改善するため，「NPYLM の 2 段階適用」を提案した．

提案法では，与えられたテキストを 1 回目の NPYLM で学習し，置き換え候補を得る．続けて，与えられたテキスト内の置き換え候補の語を，1 文字の別記号に置き換える．最後に，2 回目の NPYLM の学習を行い，過剰分割を軽減した分割結果が得られる．実験では，英語，日本語，少数言語より，提案法の有効性を検証した．実験の結果，いずれの言語においても提案法は過剰分割を改善し，提案法の有効性は言語に非依存であることも確認した．このように分析対象のデータが限られている場合でも，分割性能が向上したことは，多量のデータを与えることが困難な少数言語の分析に対して大きな助けとなるだろう．

5 章 選択的な置き換えによる NPYLM の 2 段階適用

5-1 はじめに

従来法「NPYLM の 2 段階適用」では，過剰分割の改善を平均単語長らにより確認することができた．しかし，長くなった単語の正しさまでは評価しきれておらず，出鱈目に長い単語である可能性は否定できなかった．特に，再現率および適合率が大きく減少しつつ，平均単語長が大きく改善する事例があった．特に，再現率および適合率の大幅な減少は，ほとんど分割されず，数少ない分割もほとんどが誤りであったことを意味する．そのため，分割結果は，ほとんどが出鱈目に長い単語である可能性が高い．本研究では，これを過剰結合と呼び，従来法に起因する現象であることを確認した．

従来法は過剰分割に対して有効であったが，特定のテキストに対して過剰結合が発生することは好ましくない．そこで，テキストに関わらず過剰分割を解消するため，置き換えの規模$\boldsymbol{P_{rep}}$を調整することで，過剰結合の発生を抑制することを検討した．これは，従来法の「置き換え」という単純な操作において，「置き換え過ぎた」ことで過剰結合をもたらしたと考えたためである．そして，置き換え過ぎた場合では，置き換えるべき語とそうでない語の混在が想定された．そこで，本研究では，置き換えるべき語のみを選択的に置き換える方法を必要とした．このとき，$\boldsymbol{P_{rep}}$の調整は，選択的な置き換えに相当すると考え，調整後の分割結果より，過剰結合の解消を示唆する結果が得られた．この結果を受け，置き換え語の長さに着目し，新たに提案法を示した．

また，従来法では過剰分割に対する性能を平均単語長より確認したが，分かち書きにおいては，出鱈目に長い塊であってはならない．先の過剰結合は，ほとんど分割されないという意より出鱈目という表現をしたが，どのような長さの塊に対しても出鱈目であってはならない．そこで，新たな分割指標を設け，従来法と提案法を評価した．特に，提案法では過剰分割を解消しつつ，過剰結合が抑制されることを確認した．

この章の構成を示す．5-2 節では$\boldsymbol{P_{rep}}$の調整による置き換え候補の単純な選択を検討し評価する．続く 5-3 節では，少数言語の分かち書きに対する支援にあった評価をするため，4 章とは異なる評価指標を再定義する．あわせて，従来法も再評価する．5-4 節では再評価の結果から残りの問題点を議論し，5-5 節で問題点の改善手法を提案する．最後に，5-6 節で提案法が過剰分割と過剰結合に対して有効に働くことを確認する．

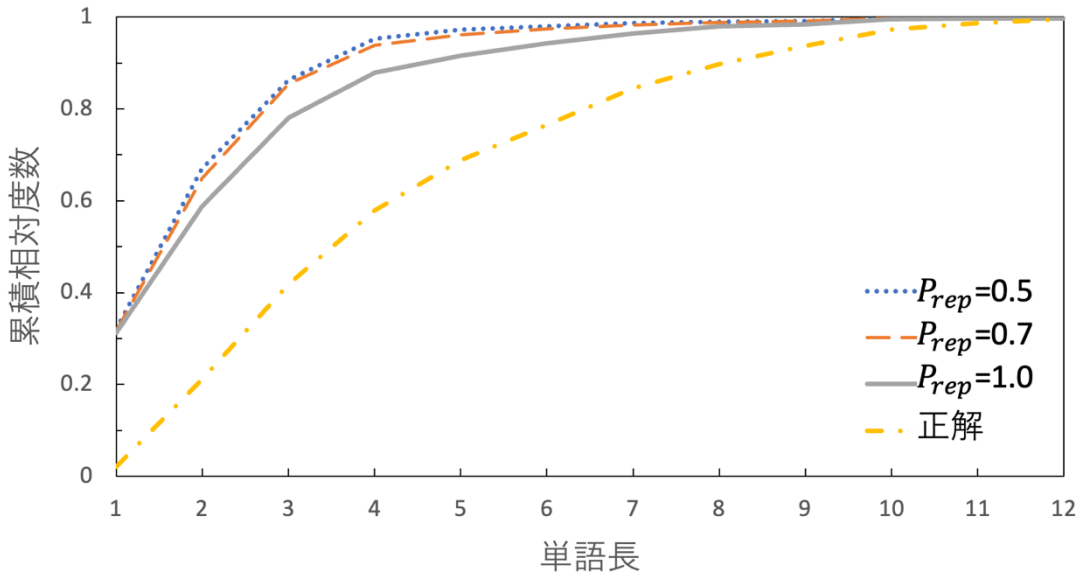
5-2 $\boldsymbol{P_{rep}}$の調整による限定的な置き換え

実験 2 より，提案法の適用後は分割すべきでない箇所を分割しづらくなったが，再現率の低下が示すような過剰結合も引き起こしてしまった．特に，過剰結合は置き換え操作が引き起こした状況であるため，置き換え操作を選択的に行う必要がある．しかし，少数言語への適用を見据えては，事前知識を用いないで置き換え対象を選択することは困難である．そこで，本節では，従来法の手順 2 で示した$\boldsymbol{P_{rep}}$を調整することで，選択的な置き換えに相当する操作を実施する．そして，その結果を受け，選択的な置き換えを新たに提案する．

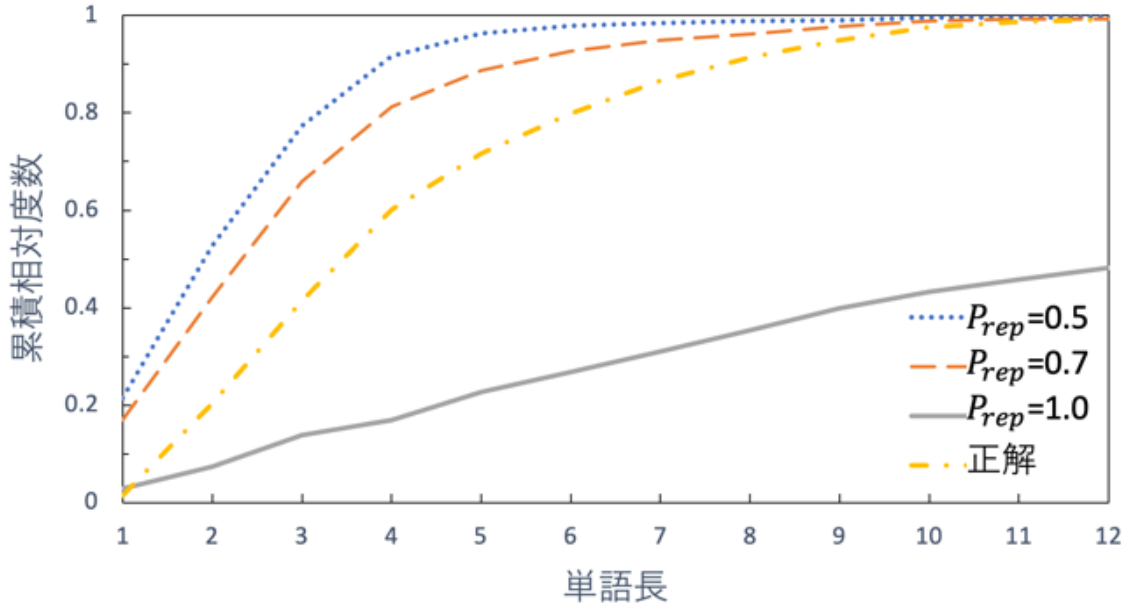
5-2-1 過剰結合であった事例と$\boldsymbol{P_{rep}}$の調整

実験 2 では過剰結合をもたらすデータセットが，英語では 10 セット中 2 セット，日本語では 10 セット中 3 セットを観測した．特に，実験 2 では$\boldsymbol{P_{rep}} = 1.0$を設定したため，1 回目 NPYLM より二文字以上の単語が，全て置き換え対象となっていた．これは，3-4 節で述べた「短い単語と単語の部分文字列（the と this の部分文字列 th）を分けて学習する」に対して，$\boldsymbol{P_{rep}} = 1.0$によって事前知識なしでも，置き換えに取りこぼしが無いようにするためである．しかし，5-1 で述べたとおり，過剰結合をもたらすデータセットにおいては，「置き換え過ぎた」と考える．そこで，$\boldsymbol{P_{rep}}$を調整することで，過剰結合が発生しないことを確認する．

まず，$\boldsymbol{P_{rep}}$の調整にあたって，過剰結合をもたらすデータセットの判別は，図 5-1, 5-2 の(a), (b)を参考にした．これらの図は，横軸を抽出した単語の長さ，縦軸を単語の出現頻度の累積相対度数として，分割結果より得られた単語を集計したものである．それぞれ，P_{rep}が 0.5, 0.7, 1.0 の場合の結果と，正しい単語分割の場合を含む．これらの図において，分割結果が正しい分割に近づくほど，結果の曲線（P_{rep}）が正しい場合の曲線（一点鎖線）に近づく．このとき，分割が過剰であるならば，短い単語が多い．そのため，単語長と頻度の比較においては，正しい場合の曲線の上方に結果の曲線が位置する．逆に，分割が不足であるならば長い単語が多く，正しい場合の下方に結果の曲線が位置することになる．いずれの言語でも，(a) P_{rep}の拡大に伴って分割が改善する場合より，P_{rep}の拡大に伴って正しい場合の曲線に近づいた．このことから，過剰分割は解消されたといえる．これに対して，(b)特定のP_{rep}で分布が悪化する場合より，$P_{rep} = 1.0$のとき正しい場合の下方に結果が位置した．これは，$P_{rep} = 1.0$で過剰結合に陥ったことを意味する．したがって，特定のP_{rep}で分布が悪化した場合を，過剰結合をもたらすデータセットと判別した．

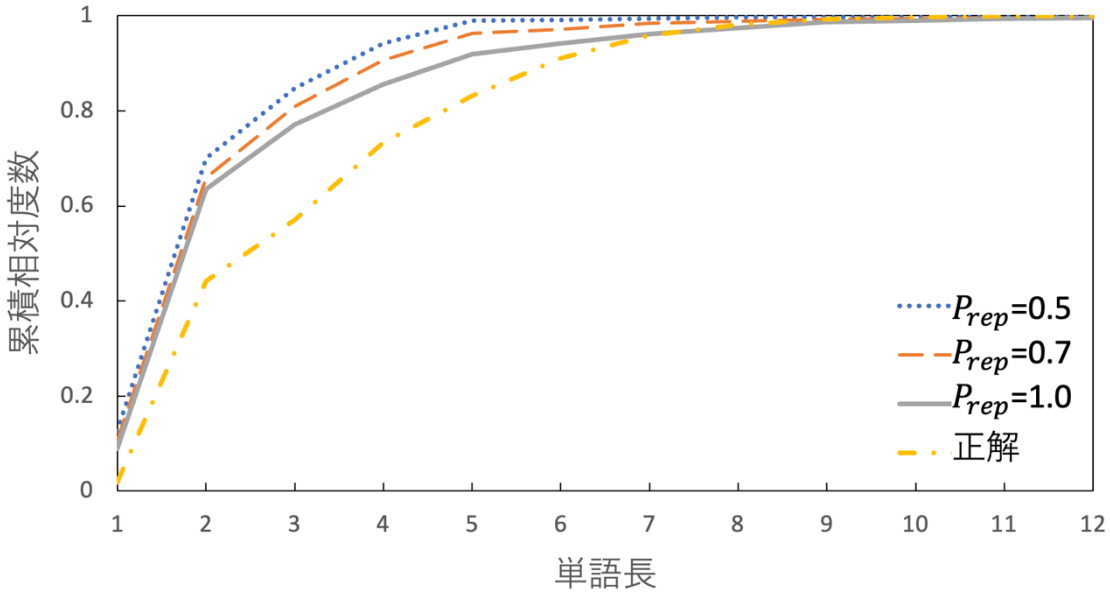


a）P_{rep}の拡大に伴って分布が改善する場合

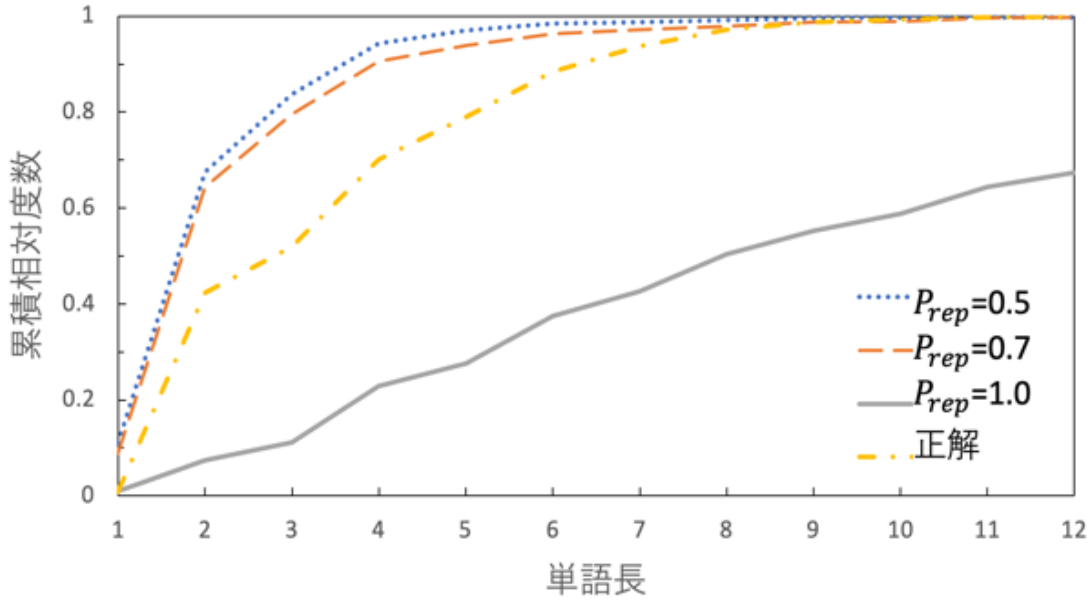


b）P_{rep}の拡大に伴って分布が悪化する場合

図 5-1 各P_{rep}における単語長の変化（英語）



a）P_{rep}の拡大に伴って分布が改善する場合



b）P_{rep}の拡大に伴って分布が悪化する場合

図 5-2 各P_{rep}における単語長の変化（日本語）

次に，事前知識なしで過剰結合を検出する方法について述べる．図 5-1，5-2 より，P_{rep}を高く設定することで，過剰分割を解消する効果は高まるが，過剰結合をもたらす事例もあることが示された．そして，過剰結合である場合はP_{rep}を減らし，分割し直すことを試みる．しかし，過剰結合であるかどうかは，正しい分割という事前知識が必要であった．そこで，分割結果（$P_{rep} = 1.0$）における短い単語に着目し，過剰結合かどうかを事前知識なしで判断する基準を検討した．

短い単語（単語長 5）が分割結果（$P_{rep} = 1.0$）を占める割合を調査したところ，図 5-3 で示す結果が得られた．この図では，縦軸を単語長 5 における累積相対度数として，英語および日本語の各 10 セットについて$P_{rep} = 1.0$における結果である．この結果より，分布が悪化する場合と改善する場合の間には，短い単語（5 文字以下）がテキストを占める割合に差があることがわかる．そこで，単語長 5 における累積相対度数が 0.5 を下回ったセットについて過剰結合と判定した．

なお，ここで短い単語を 5 文字以下と定めた理由は，他のP_{rep}では単語長 5 以降で出現頻度の累積相対度数がほぼ横ばいになるたったためである．

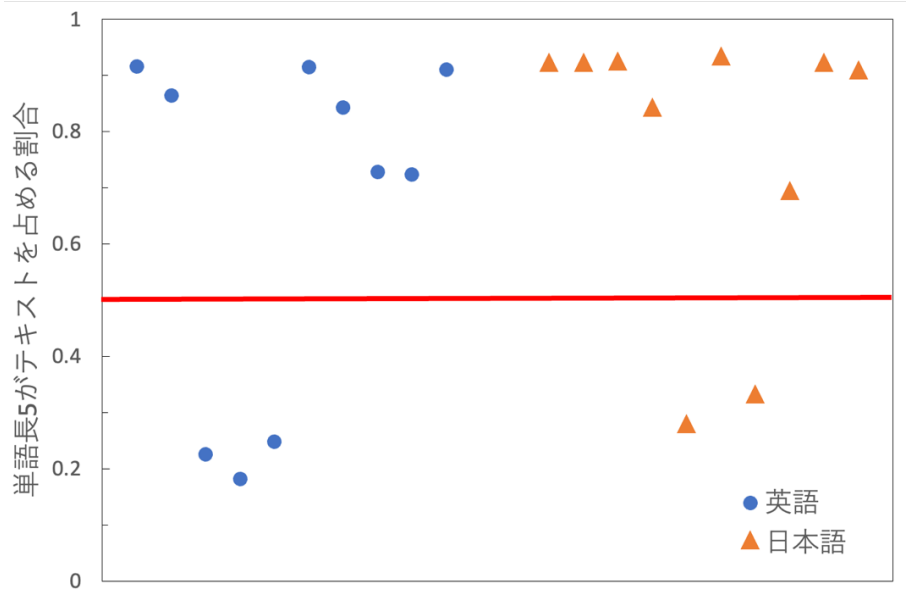


図 5-3 短い単語が分割結果を占める割合の調査結果

5-2-2 過剰結合に対するP_{rep}の調整結果

図 5-3 を参考にして，過剰結合をもたらすデータセットを検出することができた．これに対し，特定の$P_{rep}(=1.0)$を避けるために$P_{rep}=0.7$として再度分割を行った．このようにP_{rep}調整を含めた提案法を「P_{rep}調整版」として，表 4-1 に適用した結果を表 5-1 に示す．この結果より，P_{rep}を調整することで，再現率の低下を緩和しつつ平均単語長は改善することが示された．

ここまでの分割状況について，実際の分割結果（図 5-4, 5-5）から振り返る．まず，図 5-4 は，P_{rep}の拡大に伴って分布が改善する場合の結果であり，いずれの言語も適用後には正しい分割箇所に近づいている．具体的には，英語の適用前でも the や of, drive などの正しい分割はあったが，af/ter や le/ft などは過剰分割の状況にあった．対して，適用後には適用前の正しい分割に加え，先のような過剰分割における余分な分割が無くなった．全ての状況が解消されたわけではないが，一部でも解消されていることから，過剰分割は改善の傾向にあるといえる．日本語においても，適用前の正しい分割はそのままに，shin/zou などから余分な分割が無くなった．また，適用前の s/ouzou が適用後には wosouzou となり，助詞＋動詞の状況にある．これは，厳密な単語ではないが，単語の意を取り違えるほどの状況でもないと考えられる．そして，少数言語においても，正解より細かく分割されていた箇所が，的用語には解消された．

次に，図 5-5 は，特定のP_{rep}で分布が悪化する場合の結果であり，いずれの言語においても，適用前の状況は図 5-4 と相違ないが，適用後には分割箇所が極端な減少を示し，却って過剰結合であるといえる．しかし，この過剰結合については，P_{rep}を調整することで回避できることは，表 5-2 から示された．結果として，適用前よりも正しい分割に近づく結果が得られた．具体的には，英語における適用後の 2 単語目は長すぎる語であったが，調整後はよりより正しい分割結果となった．また，適用後にある every は，調整後でも正しく分割された．これは，P_{rep}の調整が単純に過剰結合を回避するだけでなく，過剰分割の改善効果も維持していることを意味する．日本語の場合も同様にして，調整は効果的であった．

以上より，提案法はP_{rep}の調整を一部で必要とするものの，過剰分割を改善したといえる．そして，これらの改善はいずれの言語でも確認された．特に，英語と日本語は文法的に全くことなる言語であることから，提案法は言語に非依存で過剰分割へ有効的であるといえる．ただし，提案法によって再現率は調整前後で改善したが，適合率は未だ低い．これは，解消しきれなかった過剰分割の存在を示唆する．

表 5-1 提案法による分割結果の評価（P_{rep}調整版）

	再現率	適合率	平均単語長
英語	0.89	0.64	5.66
日本語	0.80	0.64	6.33
少数言語	P_{rep}を調整したケースなし		

英語
正解　after the first rush of the drive one spring the boss left mark and
適用前　af ter the fir st r u sh of the drive one s p r ing the b o s s le ft mark and
適用後　after the first r u sh of the drive one s p r ing the bos s left mark and

日本語
正解　shizuka na shinzou wo souzou suru ni tahe nu hodo ni
適用前　shi zukana shin zou wo s ouzou suru ni ta hen u h odo ni
適用後　shizukana shinzou wosouzou suru ni ta hen u hodo ni

少数言語
正解　ho naq taang habe moq mitonyaung hiro joq ho majo baqan
適用前　ho na q taa ng ha be moq mito nyaung hi ro joq ho m a jo b aqan
適用後　ho na q taang habe moq mitonyaung hiro joq ho m a jo b aqan

図 5-4 分布が改善するときの分割結果

英語
正解　every one knows that mineral coal is dug out from the crust of the
適用前　e ver y one know s that miner al coal is d ug out f rom the crust of the
適用後　every oneknowsthatmineralcoalisdugoutfromthecrustofthe
調整後　every one know s that mine r al coal is d ug out from the crust of the

日本語
正解　kore ha touji toshite mezurashii koto de mo nai
適用前　ko re ha touji to shi te me zura shii koto de mo nai
適用後　kore hatoujitoshite mezurashiikotode monai
調整後　ko re ha touji toshite mezurashiikotode mo nai

図 5-5 分布が悪化するときの分割結果

5-2-3 提案法の多段階適用と$\boldsymbol{P_{rep}}$調整による課題

本節では，$\boldsymbol{P_{rep}}$調整における課題について述べつつ，提案法の 3 回目以降の多段階適用の必要性について述べる．

まず，2 段階適用は有効的に働いたが，一部ではP_{rep}調整が前提となった．ここで，置き換え操作におけるP_{rep}調整の役割を考察し，課題に言及する．前述のとおり，P_{rep}を 1.0 から 0.7 に調整することで，過剰結合は解消されることが示された．この調整にあたって，P_{rep}は頻度上位から優先的に置き換えられていたことから，P_{rep}が 0.7 のときには頻度下位が置き換え対象外となった．もとより，NPYLM を占める長い単語は，低頻度であることが多い．そのため，ここで置き換え対象外となった単語は，長い単語である可能性が高い．言い換えると，過剰分割の可能性が低い単語を，$P_{rep} = 1.0$では置き換えたことになる．特に，提案法は過剰分割の解消が目的であるため，過剰分割でないものは置き換えるべきでは無いと考えられる．

実際に，分割不足を示すP_{rep}は 1.0 で固定ではなく，0.95 や 0.9 などでも分割不足が観測された．このことから，過剰分割でないものを置き換えなくなるまで，P_{rep}の調整が必要であり，この除外がP_{rep}調整の役割といえる．しかし，どの程度の除外を要するのかは，図 5-3 のようにして都度確認が必要となるため，いくらか手戻りが発生することは避けられない．したがって，P_{rep}調整ではなく，より明確に「置き換えるべきものを置き換える」という手順が次の課題となる．

次に，提案法では，過剰分割の解消を目的に，1 回目の NPYLM で置き換え候補を抽出し，2 回目の NPYLM で置き換え後のテキストを分割した．この拡張として，2 回目の結果をもとにさらに置き換えを行い，3 回目の NPYLM を実行することも可能である．しかし，単語の長さという観点からは，3 回目以降の NPYLM は必要ないと考えた．

P_{rep}調整を含む 2 段階適用の結果である表 5-1 より，平均単語長は英語が 5.66 文字，日本語が 6.33 文字であった．このとき，正しい分割結果における平均単語長は英語が 6.56 文字，日本語が 5.70 文字であった．いずれの言語においても，提案法の結果と正しい結果との誤差は 1 文字以内であり，長さという観点からは 3 回目以降の NPYLM は必要ないと考える．

5-3 新たな分割指標と従来法の再評価

本節では，少数言語の分かち書きタスクに合った評価指標について述べつつ，従来法による分割結果をこの指標に基づいて再評価する．

まず，少数言語のための評価指標について述べる．少数言語の分かち書きタスクにおいて，有益な分割結果とは「意味を持った塊の抽出」である．従来法においては，この塊を単語単位に固定していたため，熟語やフレーズといった塊までは評価しきれていなかった．ここで，この塊について簡単に議論する．例えば，単語同士が結合することで，熟語およびフレーズ相当が得られた場面を考える．意味を持つか否かという観点であれば，このような場面も保存活動に貢献すると考えられる．実際，得られた塊が単語や熟語，フレーズのいずれであろうとも，後工程に控えた構文解析や意味解析にも十分に貢献し得る．例えば，how という単語と howareyou (how are you)というフレーズのいずれか，もしくは両方が得られても意味を持つという点では同じである．特に，単語単位であれば解体素解析に，フレーズ単位であれば構文解析に貢献できる．したがって，「意味を持った塊」には単語および熟語・フレーズといった複合語も含め，新たな評価指標とする．

次に，従来法を新たな評価指標で再評価する．従来法の適用後には長い単語が量産され，平均単語長らが改善することで，過剰分割は改善されたといえた．このとき，量産された長い単語に関して，熟語・フレーズ相当として意味を持つかどうかを確認した．その結果，熟語・フレーズ相当（以降，複合語）も得られていたが，意味を持たず中途半端に長いものも発生してしまった．ここでいう中途半端な成長とは，単語から熟語に成長し損ねた状況を指す．例えば，/helikestoreadabook/は理想として/he/likes/to/read/a/book/に分割されるはずが，/he/li/ke/s/rea/dab/oo/k/と過剰分割されたとする．そして，この過剰分割を解消するために従来法を適用したところ，/helikes/readaboo/k/という結果が得られた場合を考える．このとき，/helikes/のように結合する場合は複合語として許容できる．しかし，/readaboo/k/のように中途半端に結合する場合は，本来は単語に成長するはずの/k/が取り残されるなどして，read にとっては過剰結合となってしまう．そして，従来法の適用前でも，こうした状況は過剰分割に隠れて存在していたが，全体的に過剰分割が解消されることで，/readaboo/などがより目立つようになってしまった．特に，/readaboo/k/の状況は，少数言語の分かち書きタスクにおいて，意味が汲み取りづらい点から好ましくない．

続けて, /readaboo/の状況が, 従来法によってどれだけ引き起こされているかを調査した. その結果を表 5-2 に示す. この表は, 25,000 文字からなる英語テキストの結果（P_{rep}調整の必要が無かったデータ）であり, 6 文字以上の単語について分類している. なお, このテキストにおける平均単語長の真値は約 6 であったことから, 上記の長い語を 6 文字以上と定めた. この表より, 従来法の適用前でも, 中途半端な語は存在していたが, 過剰分割という状況に圧されて, 目立っては存在しなかった. 対して, 従来法の適用後には, 正しい単語が 10 倍以上に増え, 過剰分割は「意味のある塊」を持って改善していた. しかし, この改善は置き換えた語の成長が背景にあることから, 中途半端な語も数倍に増えてしまった. このとき, 中途半端な語が増えたことにより, 従来法における再現率の低下が引き起こされたと考えられる. 以上より, 従来法は過剰分割を解消した結果として, 有益な塊を増加させたが, 同時に中途半端な語も量産してしまった. 特に, 中途半端な成長（過剰結合）は他の単語が成長する機会を阻害するため, 有益な結果ではない. ここでいう阻害とは, /readaboo/k/における/book/が得られなくなったことを意味する.

したがって, 従来法は適用前後で得られる単語が長くなるという効果を発揮していたが, 少数言語の分かち書き結果として常に有益な単語ばかりではなかった.

表 5-2：6 文字以上の抽出単語に関する評価（英語）

	抽出単語種類数	正しい単語	正しい複合語	中途半端な語
NPYLM	23	9	5	9
従来法	262	140	61	61

5-4 過剰結合における置き換えるべきでない語の調査

5-2 節および 5-3 節をふまえ，再評価で確認した中途半端に長い単語の発生を抑制するためには，置き換え対象を慎重に選ぶ必要があった．そこで，本節では置き換えるべき語が何かを，予備実験から確認した．

予備実験では，1 回目 NPYLM が示す単語のうち，正解単語と不正解単語の 2 パターンをそれぞれ置き換え対象とし，結果を再現率・適合率・平均単語長により評価した．なお，従来法は過剰分割を解消するにあたって，短い単語を成長させる効果を持つ．そのため，ここでの正解単語は「これ以上の成長を必要としない語」，不正解単語は「未だ成長を必要とする語」と位置付ける．

予備実験で用いたテキストは，25,000 文字からなる英語と日本語を 10 セットずつである．従来法($P_{rep} = 1.0$)も合わせた結果を表 5-3 に示す．この表より，いずれの言語でも不正解単語を置き換えた場合には，適合率および平均単語長で最も大きな改善を得た．特に，適合率は分割箇所の正答率を意味するため，指摘された過剰結合は不正解単語を置き換えた場合に最も少ない．対して，正解単語を置き換えた場合，平均単語長においては中程度の結果が得られたが，適合率はいずれの言語においても悪化する結果となった．このように適合率が悪化しながら平均単語長が長くなることは，出鱈目に長い単語の発生が危惧される．

表 5-3 置き換え対象を変化させた場合の比較結果

	置き換え対象	再現率	適合率	平均単語長
英語	$P_{rep} = 1.0$	0.95	0.55	4.01
	正解単語	0.94	0.53	4.33
	不正解単語	0.93	0.57	5.01
日本語	$P_{rep} = 1.0$	0.95	0.55	4.01
	正解単語	0.83	0.57	4.95
	不正解単語	0.86	0.59	5.39

過剰結合を回避するためには，正解単語の置き換えはリスクであると考えたが，平均単語長が中程度の結果であったことをふまえると，P_{rep}の調整を要するほどに長い単語が増えたとは考えづらい．そのため，正解単語においては，置き換え効果がさほど大きくないと解釈し，置き換えの対象外とするほどに強い制約は必要がないと考えた．したがって，置き換えるべき語としては，不正解単語を積極的に置き換えるべきことはわかった．ただし，正解や不正解の同定は事前知識なしでは実施できないため，教師なしでも得られる情報を用いて，不正解単語の積極的な置き換えを検討する．

続けて，1 回目 NPYLM による置き換え候補に対して，単語の長さの観点から選出することを検討した．特に，この単語の長さは 1 回目 NPYLM の結果に基づくため，教師なしで得られる．

従来法で指摘してきた過剰分割は，単語未満の分割結果が過剰に存在すると言い換える．このとき，前節の予備実験で設定した不正解単語のうち，長さという観点からは短いものが多を占めると考えられる．実際，前節で用いた正解・不正解単語について，英語のある 1 セットより単語長ごとの分布を確認した（図 5-6）．その結果，単語長が短ければ，不正解単語の種類数は多いことがわかった．つまり，短い単語という観点からは，不正解単語に近づくことはできた．また，長いものを置き換えることには，過剰結合というリスクがある．単純に，置き換えでは一文字でそれ以上の文字数を扱うため，長い単語がさらに長くなる危険性がある．したがって，置き換えるべき語としては，所定の長さ以下の単語を対象とすれば良いと考える．

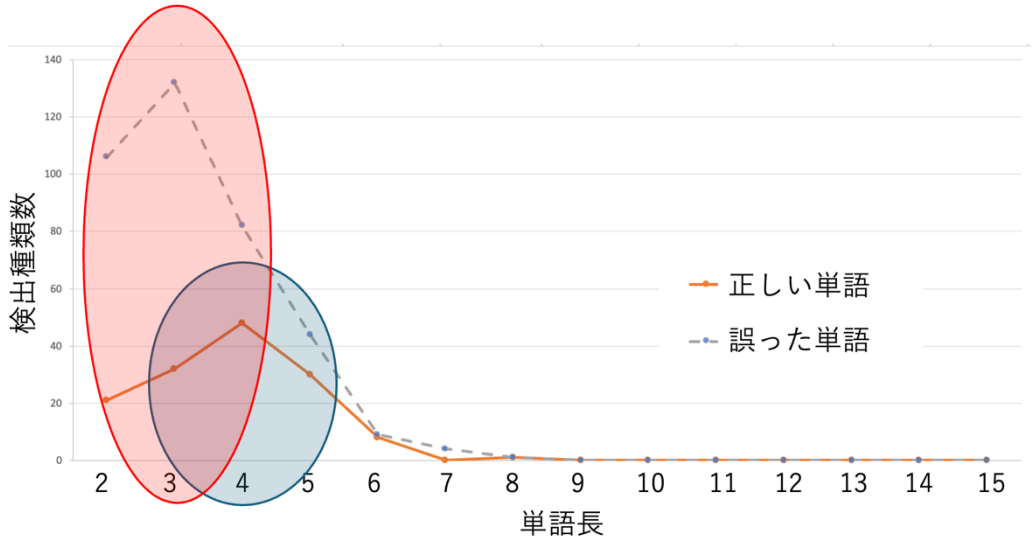


図 5-6 単語長ごとの異なる単語種類数の分布（1 回目の結果）

最後に，所定の長さ以下の単語を置き換えたとき，図 5-6 で示すように正解単語も多く含まれる．このとき，表 5-3 で述べたように，正解単語は置き換えても，出鱈目に長い単語がそれほど発生しないと予想した．ここで，実際に置き換えられた正解単語が，置き換えの前後で余分な結合がどの程度生じたのかを確認した．表 5-4 は，$P_{rep} = 1.0$において，適用前後（1 回目 NPYLM の結果と二回目 NPYLM の結果）で正解単語の一致率を調査した結果である．調査対象は，P_{rep}の調整が必要なかったデータセットである．表 5-4 より，1 回目で得られていた正解単語のうち 2 回目に重複するものは減少するが，8~9 割程度は余分な結合に変化していないことがわかる．対して，残りの 1 割程度は，より長い単語への変化が想定される（置き換え操作では，それ以上細かく分割されないため）．そして，平均単語長の伸びは，不正解単語を置き換えた場合よりも緩やかであったことから，正解単語を置き換えても，深刻な過剰結合は引き起こさないといえる．

以上より，選択的な置き換えは，所定の長さ以下の単語を置き換えることを提案する．

表 5-4 $P_{rep} = 1.0$において 1 回目から 2 回目に引き継がれた正解単語

セット	1	2	3	4	5	6	7
1 回目の正解	140	198	118	244	214	209	135
2 回目で重複した正解	129	183	111	219	171	166	127
一致率	0.92	0.92	0.94	0.90	0.80	0.79	0.94

5-5 提案：選択的な置き換えの手順

この提案法は，従来法における手順 2 を変更したものであり，短い単語の基準を暫定的に 6 と設定した．この設定にあたっては，日本語と英語の平均単語長の真値が，いずれも 6 程度であったことによる．他の言語でも同様の値になるかは，別途に検証を要する．しかし，全ての言語を調査することは現実的ではないため，ここでは定性的に議論する．平均単語長はそのままの意味で値を返すが，真値を求めるためには豊富な事前知識が必要である．例えば，漢字や英語の接頭辞，活用形などを事前に知り得るならば，平均単語長をより正確に求めることができる．しかし，少数言語で正確な値を出すことは困難であるため，ここは言語学者らによる意見を参考にされたい．例えば，収集した音声を言語学者らが書き起こす際には，あらゆる知見を用いて慎重に行う．具体的には，発話の時間的間隔や特定の発話の頻度などであり，これらを総合することで，ある既存言語に発話体系が似ているなどと推測することができる．そして，この推測の段階で，主だった単語長は把握することが可能であると考えられる．このようにして，テキストから直接推測することは現状困難であるが，あらゆる手作業と並行することで提案法に要するいくらかの数値は用意されるだろう．

手順 1 ：単語候補とその出現頻度のリスト作成

与えられたテキストで NPYLM を学習し，単語分割を実施する（1 回目の NPYLM）．この分割結果から，長さが二文字以上の単語を抽出し，単語候補（文字列）およびその出現頻度のリストを作成する．

手順 2 ：学習用テキストの生成

頻度上位の候補から順に一語ずつ取り出し，すべての出現箇所を元文書に含まれない記号（α や①などの一文字）で置き換える．このとき，単語長 6 文字以下の語を置き換える．なお，この置き換えは，1 回目の NPYLM で切り出された語に従って行う．すなわち，the を置き換える場合，1 回目の分割で the と切り出している部分のみが置き換え対象であり，1 回目の分割で there と切り出している部分は対象としない．

手順 3 ：生成された学習用テキストによる NPYLM の構築

1 回目の NPYLM の結果を破棄し，置き換え後のテキストで NPYLM を学習する（2 回目の NPYLM）．

手順 4 ：2 回目の NPYLM を用いた単語分割

学習した NPYLM を用いて置き換え後のテキストを分割し，置き換えた文字を元に戻す．これが，与えられたテキストに対する分割結果となる．

5-6 実験：【提案法】選択的な置き換えによる NPYLM の 2 段階適用

従来法では，特定のデータにおいて過剰結合を引き起こした．過剰結合を回避するため，P_{rep}の調整という対策を講じた．しかし，調整には手戻りが発生してしまうため，より単純な対策として，所定の長さ以下の単語を置き換える方法を提案した．そこで，本節では，この提案法では，選択的に置き換えることで，過剰分割および過剰結合の解消に対して有効に働くことを確認する．

5-6-1 実験条件

前提として，従来法はすでに過剰分割の解消に効果を発揮するが，過剰結合という悪化をもたらす事例があった．提案法では，過剰分割を解消しつつ過剰結合を抑制する手法として，選択的な置き換えを実施する．

実験条件について述べる．25,000 文字からなる英語テキストと日本語テキスト（ローマ字）を図 5-7 のような状況で 1 セットずつ用意し，提案法を適用した．評価においては，素の NPYLM，従来法，提案法の 3 つのモデルを学習し，分割結果から得られる単語の正しさの比率を比較する．なお，正しさの単位には単語と複合語を据えるため，再現率らの代わりに比率を用いる．この比率においては，単純に正しい単語（意味を持った塊）がどれだけえられたのかを示すため，支援する保存活動（単語などを探る）にとっても分かりやすい単位といえる．図 5-8 より，この比率の算出方法について述べる．この図より，従来法では単語単位に絞って正しさを評価していたが，新たな評価指標として前述のとおり，「意味を持つ塊の抽出」という観点からは単語単位は条件として狭すぎる．そこで，緩やかに評価するために複合語も許容して，「意味を持つ塊がどれだけ得られたか」を評価する．

なお，図 5-8 に示された ourprevious と method は，ぞれぞれ複合語と単語単位を示すが，今回は proved も正解とする．Ground truth と比較すると improved から im が欠損してしまっているが，現存する単語を正しいとするならば，proved も正解といえる．加えて，将来的な構文解析などでは proved を正解と扱うことは躊躇われるが，現時点では単語同士の関係性を解析するような支援ではないため，今後の課題として今回は議論の対象外とした．

English
Original text: we improved our previous method
Ground truth: we/improved/our/previous/method
  # “/” is desired segmentation
Unsegmented text: weimprovedourpreviousmethod

Japanese
Original text: 我々は従来法を改良した
Segmented by MeCab: 我々/は/従来/法/を/改良/し/た
Ground truth: wareware/ha/jurai/hou/wo/kairyo/si/ta
  # Romaji text with segmentation
Unsegmented text: warewarehajuraihouwokairyositа

図 5-7 実験用テキストの詳細

Original text : we improved our previous method
Ground truth : we/improved/our/previous/method
Segmentation result : wei/m/proved/ourprevious/method (5 words)
Correct segmentation : proved, ourprevious (our + previous), method (3 words)
Ratio of correct words = 3/5 = 0.6

図 5-8 評価における比率の算出方法

5-6-2 実験結果

素の NPYLM・従来法・提案法を対象に，「意味を持つ塊がどれだけ得られたか」の確認結果を表 5-6 に示す．この表が示すように，提案法が最も高い正答率をもたらした．これは，従来法と比較して「選択的な置き換え」が有効的に働いたことを意味する．特に，過剰分割や過剰結合のいずれの場合においても，正答率は期待できない．そのため，最も高い正答率をもたらす結果が，過剰分割および過剰結合に対して最も有効に働くといえる．

なお，提案法が最も良い結果であっても，正答率は 6 割を下回るため，さらなる精度向上が必要である．もっとも，保存作業は試行錯誤で行われるため，試行錯誤に際して予備知識をいくらか得ることが期待される．そのような場合，教師なし手法を半教師あり手法へ拡大させることができ，されなる精度向上に期待できる．

表 5-6 意味を持つ塊の抽出率の比較

		正解単語種類数	単語種類数	正答率
英語	素の NPYLM	149	518	0.29
	従来法	337	738	0.46
	提案法	374	743	0.52
日本語	素の NPYLM	96	483	0.20
	従来法	195	478	0.41
	提案法	435	792	0.55

5-7 まとめ

従来法「NPYLM の 2 段階適用」は過剰分割を解消したが，過剰結合を引き起こした．過剰結合の解消には，P_{rep}の調整によって対策した．しかし，調整には手戻りが発生するため，より手軽な「選択的な置き換えによる NPYLM の 2 段階適用」を提案した．

提案法の評価にあたって，少数言語の分かち書きタスクにとって有益な分割について議論し，「意味を持つ塊の抽出率」という評価指標を設定した．評価の結果，提案法は従来法とは異なり，過剰分割と過剰結合の両方に対して，有効に働くことを確認した．

6 章 総論

6-1 本研究で得られた成果

本研究では，少数言語に対して言語学者らが行う保護活動のうち，「言語の保存活動」に着目して支援を検討した．保存活動においては，収集した音声を書き起こし，書き起こし結果（テキスト）から文法や単語などの知識を分析し，記録するものであった．そして，言語学者らは，この活動の自動化を望んでいる．自動化を検討するにあたって，この一連の操作におけるデータ変換を辿ると，音声認識技術やいくつかの自然言語処理技術が相当した．しかし，これらの技術は，事前知識および大量のデータを必要としており，少数言語はいずれも不足しがちであった．

そこで，本研究では，教師なしかつ少量のデータをふまえた支援を検討した．具体的には，音声の書き起こしは音声マッチングとして，書き起こされたテキストの分析を教師なし分かち書きとして検討した．音声マッチングにおいては，音響特徴量 MFCC に与えるデータに課題があり，Wavelet 解析と SIFT 特徴量によって，少数言語のための音声ラベリングを新たに提案した．実験の結果，提案法は少数言語音声を緩やかにマッチングすることができた．これにより，把握している限りの音を，連続音声に対して自動的にラベリングでき，対象言語の音素などが同定されきっていない段階からでも適用できる．そして，ラベリングされていない箇所に関しては，新規ラベルを当てはめるなどの工程が想定される．総じて，言語学者らの知見を未だ必要とする手法ではあるが，連続音声を全て手作業で書き起こす必要は無くなった．この点から，提案法は，少数言語音声の書き起こしの支援といえる．教師なし単語分割においては，主なテキスト分析である形態素解析・構文解析・意味解析の全てに先立つものとして，「テキストから意味を持つ塊を抽出する」タスクを定めた．そして，

NPYLM の 2 段階適用を基本的なアイデアとして，このタスクの自動化を提案した．この提案は，言語学者らによるテキスト分析の効率化を図る．加えて，「意味を持つ塊」という知見は，保存活動の様々な段階に貢献する．例えば，音素などを詳細に同定しようとする場合，活用形や文法といった意味を定める音へも言及される．このとき，意味が変化するタイミングから音を同定するが，これは意味を持った塊同士を比較して得る．したがって，この提案によって得られた結果は，テキスト分析の後工程だけでなく，保存活動の全体の効率化が期待できる．

6-2 今後の課題

本研究では，音声の書き起こし，書き起こしテキストの分析という 2 段階に分け，支援を検討した．それぞれの段階で細かな課題は残されているが，全体的には半教師あり手法の適用を今後の課題としたい．その理由には，言語学者らによる作業と支援は並行して行うことが挙げられる．特に，データ量の問題は時間的な制約が多いため別枠となるが，事前知識としては作業途中でもいくらか得られるはずである．そこで，半教師あり手法によって，あらゆる段階でさらなる精度向上を今後の課題としたい．

本研究では，手作業を支援するにあたって，手作業の工程に従い，音声を書き起こす段階と書き起こしたテキスト解析の 2 段階を支援した．このとき，書き起こしをせずに，直接に音声を切り出すタスクとしても支援となりうる．例えば，連続音声から切り出された塊に対して，後工程で表音記号などを当てはめて記録する．ただし，本研究と同様に，事前知識および学習データの不足は共通する．また，音声を直接切り出すタスクでは，テキスト変換における音響的特徴量の欠損が解消されうる．そのため，切り出す単位と音響特徴量の対応関係についても検討し，このタスクによる支援も議論を進めていく．

謝辞

本論文は，著者が三重大学大学院工学研究科博士後期課程時に行った研究をまとめたものである．本論文を進めるにあたり，懇切丁寧なご指導を賜った三重大学の高瀬治彦教授，静岡理工科大学の友次克子教授，高野敏明准教授に感謝いたします．

最後に，本論文をまとめるにあたり，助言，討論，その他お世話になったすべての方々に感謝いたします．

参考文献

[1] Krauss, Michael , ‘‘The world’s Language in crisis,’’ Language Volume 68, Number 1, March 1992, pp. 4-10, DOI: 10.1353/lan.1992.0075（1992）

[2] Christopher Moseley. ‘‘UNESCO Atlas of the World's Languages in Danger,’’ UNESCO（2010）

[3] 大角翠【書評・紹介】ニコラス・エヴァンズ著，大西 正幸，長田俊樹，森 若葉 訳.‘‘危機言語:言語の消滅でわれわれは何を失うのか,’’ 京都大学学術出版会(2013). 言語研究, 144: 119-127 (2013)

[4] デイヴィッド・クリスタル著，斎藤 兆史/三谷 裕実 訳.‘‘消滅する言語 人類の知的遺産をいかに守るか,’’ 中央公論新書（2004）

[5] 坂本 邦彦.‘‘危機言語研究の現在——ユネスコアトラスに関する C. モーズリーの論考——,’’ 尚美学園大学総合政策研究紀要第 28 号，pp. 87-99,（2016）

[6] 菊池英明，前川喜久雄，五十嵐陽介，米山聖子，藤本雅子.‘‘日本語話し言葉コーパスの音声ラベリング,’’ 音声研究, Vol.7, No.3, pp.16-26（2012）

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” Tech. Rep., OpenAI（2022）

[8] H. Kamper, A. Jansen, and S. Goldwater. ‘‘A segmental framework for fully-unsupervised largevocabulary speech recognition,’’ Comput. Speech Lang., 46(C)（2017）

[9] H. Kamper, K. Livescu, and S. Goldwater. ‘‘An embedded segmental k-means model for unsupervised segmentation and clustering of speech,’’ Proc. of ASRU（2017）

[10] F. Kreuk, J. Keshet, and Y. Adi. ‘‘Self-supervised contrastive learning for unsupervised

phoneme segmentation,'' Proc. of Interspeech（2020）
[11] Alexei Baevski, Wei-Ning Hsu, Alexis CONNEAU, Michael Auli. ''Unsupervised Speech Recognition,'' Advances in Neural Information Processing Systems 34, Proc. of NeurIPS (2021)
[12] DAVID G. LOWE. ''Distinctive Image Features from Scale-Invariant Keypoints,'' International Journal of Computer Vision 60(2), 91–110（2004）
[13] 荒木 雅弘. ''フリーソフトで作る音声認識システム パターン認識・機械学習の初歩から対話システムまで(第 2 版) ,'' 森北出版株式会社（2017）
[14] 安藤 彰男 編著. ''基礎音響学,'' コロナ社（2019）
[15] 金谷 健一. これなら分かる応用数学教室 最小二乗法からウェーブレットまで'', 共立出版（2003）
[16] 松坂 喜幸. ''ディジタル画像処理'', 公益財団法人 画像情報教育振興協会（2015）
[17] Rafael C. Gonzalez and Richard E. Woods. ''Digital image Processing fourth edition,'' Pearson Education Limited (2017)
[18] 宮澤幸希, 三浦英朗, 菊池英明, 馬塚れい子. ''連続音声から 音韻カテゴリ獲得モデルに関する考察,'' 第 25 回人工知能学会全国大会講演論文集 1D2-1（2011）
[19] 川原 繁人. ''ビジュアル音声学,'' 三省堂（2018）
[20] 日本音響学会 編. ''音響学入門ペディア,'' コロナ社（2017）
[21] 加藤 重広・安藤 智子. ''基礎から学ぶ音声学講義,'' 研究社（2016）
[22] 原島 博. ''信号処理教科書 不規則信号とフィルター,'' コロナ社（2018）
[23] 中川 聖一 編著. ''音声言語処理と自然言語処理,'' コロナ社（2013）
[24] 一般社団法人 日本音響学会 編. ''実験音声科学 音声事象の成立仮定を探る,'' コロナ社（2018）
[25] Stevens, K.N.(1998). ''Acoustic phonetics,'' Cambridge:MIT Press（1998）
[26] 永田 靖・棟近 雅彦, ''多変量解析法入門,'' サイエンス社（2001）
[27] Pak lrina, Teh Phoey Lee. ''Text Segmentation Techniques: A Critical Review,'' Innovative Computing, Optimization and Its Applications: Modelling and Simulations, Volume 741, pp.167-181（2018）
[28] Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, Vasudeva Varma, ''Attention-Based Neural Text Segmentation,'' Proc. of ECIR2018, pp.180-193（2018）
[29] Yan Shao, Christian Hardmeier, Joakim Nivre, 2018 ''Universal Word Segmentation: Implementation and Interpretation,'' Transactions of the Association for Computational Linguistics 6, pp.421–435（2018）
[30] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. ''Bayesian unsupervised word segmentation with nested pitman-yor language modeling,'' Proc. of ACL-IJCNLP, pp.100-108（2009）

[31] Yee Whye The. ‘‘A Hierarchical Bayesian Language Model based on Pitman-Yor Processes,’’ Proc. of COLING/ACL, pp.985-992（2006）
[32] 持橋 大地, 隅田 英一. ‘‘Pitman-Yor 過程に基づく可変長 n-gram 言語モデル,’’ 情報処理学会論文誌 48 (12), pp.4023-4032（2007）
[33] Ryo Fuji, Ryo Domoto, and Daichi Mochihashi. ‘‘Nonparametric Bayesian Semi-supervised Word Segmentation,’’ Proc. of TACL, pp.179-189 (2017)
[34] C.M. ビショップ 著, 森田 浩/栗田 多喜夫/樋口 知之/松本 裕治/村田 昇 監訳. ‘‘パターン認識と機械学習 上 ベイズ推論による統計的予測,’’ 丸善出版（2012）
[35] Jenny R. Saffran, Elissa L. Newport, Richard N. Aslin. ‘‘Word Segmentation: The Role of Distributional Cues,’’ Journal of Memory and Language, Volume 35, Issue 4（1996）
[36] R. G. Casey and E. Lecolinet, ‘‘A survey of methods and strategies in character segmentation,’’ in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 7, pp. 690-706（1996）
[37] Sennrich, Rico, Barry Haddow, and Alexandra Birch. ‘‘Neural machine translation of rare words with subword units,’’ arXiv preprint arXiv:1508.07909（2015）
[38] Kamper, Herman, Aren Jansen, and Sharon Goldwater. ‘‘Unsupervised word segmentation and lexicon discovery using acoustic word embeddings,’’ IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.4, pp. 669-679（2016）
[39] Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. ‘‘Contextual dependencies in unsupervised word segmentation,’’ Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics（2006）
[40] Yarowsky, David. ‘‘Unsupervised word sense disambiguation rivaling supervised methods,’’ 33rd annual meeting of the association for computational linguistics,（1995）