大規模アクセントラベリングコーパスの構築と それに基づくハイブリッド型アクセント結合処理

黒岩 龍 † 峯松信明 ‡ 伝康晴 * 広瀬啓吉 †
† 東京大学大学院情報理工学系研究科
‡ 東京大学大学院新領域創成科学研究科
* 千葉大学文学部
{kuroiwa,mine,hirose}@gavo.t.u-tokyo.ac.jp, den@cogsci.l.chiba-u.ac.jp

1 はじめに

日本語テキスト音声合成システムを構築する場合，一般的には，下記のような言語処理系／波形生成系が必要となる。1) 形態素解析を行ない，形態素境界を特定し，各形態素の読みやアクセント型（各形態素を単独で読み上げた際のモーラ列・音素列とアクセント型）の情報を得る，2) 無声化，連濁などの音韻処理，アクセント結合・イントネーション・継続長・パワーに関する韻律処理を行なう，3) 上記の情報を基に，波形生成を行なう。

従来筆者らは，アクセント句境界が与えられた場合に，句内のどのモーラをアクセント核として波形生成するのか（アクセント結合問題）を，各形態素（自立語／付属語）のアクセント属性を定義し，規則によってアクセント変化を記述することでシステム構築を行なって来た [1, 2]。しかし，全ての事象を規則で網羅することには限界があり（例えば [1] では副次アクセントや，付属語連鎖への対処が行なわれていない），アクセントラベルが施された大規模なコーパスを用いた機械学習・統計学習で解決を図ることも行なわれるようになった [3]。

コーパスベースの方法論を検討する場合，当然，大規模コーパスは必須のものとなるが，現時点で，高品質のアクセントラベリングが施され，研究目的で自由に利用可能なコーパスは存在していない。これらの現状を鑑み本研究では，1) アクセント結合処理モジュールを構築可能な，大規模かつ高品質なアクセントラベリングが施されたコーパスの構築と，2) それに基づく（かつ，従来の規則ベースのアクセント結合処理を踏まえた）統計的なアクセント結合処理モジュールの構築を試みる。

2 特定ラベラによる大規模アクセントラベリングコーパス

2.1 ラベリング対象とする言語事象

日本語東京方言で文を発声する際，文は幾つかのまとまりに分かれ，各々の内部では音の高さ（ピッチ）が連続的に変化する。まとまりが始まる箇所ではピッチの上昇が見られ，その後，まとまりの内部では上昇は無く，ゆるやかに下降してゆく。まとまりの内部には，語彙に依存した比較的急激なピッチの下降箇所が概ね高々1つ存在する。このようなまとまりをアクセント句と呼び，急激な下降の箇所をアクセント核と定義するのが一般的である。

しかし，実際の発声におけるピッチ変化を観察すると，明確にアクセント句の分離が困難な場合も多く，複数の句が影響を及ぼし合い，融合する現象も見られる。これについて，日本語話し言葉コーパス [4] では，後続アクセント句の句頭上昇が見られない程度まで融合が起きていれば1つのアクセント句として扱うものとするが，同程度の融合が見られても双方が核を持つアクセント句である場合は，「アクセント句には1つの核しか存在し得ない」との大前提に従い，複数のアクセント句として扱っている。

アクセント核についても，発声者や聴取者にとって知覚されているにも拘らず，実際のピッチ変化に明確に現れていない場合がある [4]。更には，発声者等の知覚も必ずしも明確ではない。

このようにアクセント句・核は明確に（物理的に）定義できるものでないが，ラベリング作業を行なう場合，何らかの定義が必要となる。本研究では下記の言葉でこれらを定義し，ラベリング対象とした。

アクセント句境界　ピッチの句頭上昇が見られる箇所をアクセント句境界とする。視覚提示される話速（約 7 モーラ/秒）に合わせて自然に読んだ場合を想定する。休止が入った場合も，通常は境界が生じる。
アクセント核　ピッチが急激に下降する箇所の直前のモーラ。句内の出現回数は制限しない。意識的に当該箇所で急激にピッチを下げ，それ以外では同じピッチで平坦に（イントネーションを除去して）読んだ場合に，違和感が生じない位置。

上記のラベリング対象は，文読み上げ時の事象である。これとは別に，個々の自立語を単独で発声する場合のアクセント型もその対象とした。後述するように本ラベリングは特定のラベラによる作業となる。単独発声時のアクセント型は，アクセント辞典等に掲載されているが，アクセント感覚の個人差を含まない高品質のコーパス構築を念頭に置き，単独発声時に対するラベリングも作業項目とした。

最終的に，本研究におけるラベリング対象は，1) 文発声時のアクセント句境界とアクセント核位置，及び，2) 文中の全自立語に対する単独発声時のアクセント核位置，である。なお，付属語については，単独発声を想定すること自体が困難であり，また，不合理とも考えられるため，これらに対する単独発声時のアクセントラベリングは行なわない。

2.2　ラベラ・ラベル検査者の選定

アクセント感覚には個人差があるため，本研究では特定ラベラに全ラベリングを依頼することとした。作業量が膨大となるため，誤ラベリングも免れ得ない。そこで，付与されたラベルを検査する（誤りが含まれる可能性のある文を選定する）検査者をラベラとは別に用意した。東京生まれ・育ちであり，合唱部に所属する比較的音感の鋭い大学生 6 名に対して，日本語アクセントに対する教育を施した上で，試験により選抜し，最終的にラベラ 1 名，検査者 1 名（今後増員する可能性あり）を選定した。

2.3　使用した文セット

新聞記事読み上げ音声コーパス（JNAS）で使用されている文（毎日新聞記事から抽出した 16,178 文および ATR 音素バランス文 503 文）を用いた。既に音声コーパス用に使用されている文であり，全文に凡そ正しい読みが与えられていること，砕けた文が少なく扱いやすいことなどが選定理由である。但し，不自然な読みや誤った読みも散見されたため，事前に全文の読みを確認し，適切でないと思われる箇所に対しては，適宜，修正を施した。

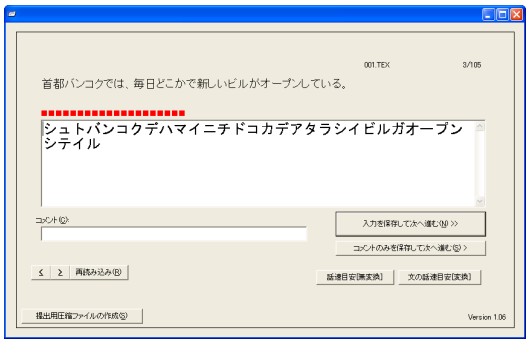

図 1: ラベリング作業用インタフェース

2.4　形態素解析

文中の自立語に対するラベリング作業であること，及び，ラベリング結果を利用する際の利便性のため，全文に対して形態素解析を行なった。形態素解析の品詞体系は UniDic に準拠した。形態素解析作業は，辞書として UniDic を用いて自動的に行なった後，用意した読みと異なった読みが付与されたものを中心に手修正した。従って，全て正確に形態素解析が行なわれている訳ではないが，読みについてはラベリング作業に用いたものと完全に同一になっている。

2.5　実際のラベリング作業

ラベリング作業は，音声を経ず，テキストに直接ラベルを付与させた。これは，音声を経ると極めて作業時間が長くなり，また，ラベラにとっての知覚的なアクセント情報を重視するためである。作業用に，図 1 に示したインタフェースを用意し，単純な操作でラベリング作業ができるようにした。

単独発声時の自立語のラベリングでは，完全な単独発声では曖昧性が残るため，名詞には助詞「が」を後続させた状態でラベリングをさせ，形容詞には名詞「こと」が後続することを想定してラベリングをさせる（「こと」自体はラベリングの対象としない）などの工夫を行なった。この作業により，

- **シュ’ト／バ’ンコクデ’ハ／マ’イニチ／ド’コカデ／アタラシ’イ／ビ’ルガ／オ’ープン／シテイル**

表 1: アクセント句を構成する形態素の数

形態素数	出現数	出現率
1	5079	17.4%
2	9829	33.6%
3	7902	27.0%
4	3972	13.6%
5	1586	5.4%
6	554	1.9%
7 以上	303	1.0%

表 2: アクセント句を構成する品詞列（上位 10 種）

品詞列	出現数	出現率
[名][助]	5273	18.0%
[名]	2639	9.0%
[名][名][助]	2180	7.5%
[名][接尾][助]	1409	4.8%
[動][助動]	792	2.7%
[動]	788	2.7%
[名][名]	758	2.6%
[名][接尾]	739	2.5%
[動][助]	571	2.0%
[名][助][助]	541	1.9%
上記以外	13535	46.3%

のような文発声ラベリングと，

- **シュ’トガ**
- **アタラシ’イ**
- **スル**

のような形態素単独発声ラベリングが得られる。なお，[／] がアクセント句境界位置を，[’] がアクセント核位置を表している。ラベル検査者から誤りの可能性を指摘された文については，ラベラに確認させ，必要に応じて，訂正させた。

3 進捗状況と分析

2007 年 1 月現在，4,166 文（JNAS の先頭 40 ファイル）について，文発声・形態素単独発声の双方のラベリングが完了している。今後，用意した全ての文に対してのラベリングを目指して進めていく。作業が完了した全文に対して，アクセント句を構成する形態素数，及び，アクセント句を構成する品詞列の上位 10 種類を表 1，表 2 に示す。4 形態素以下のアクセント句が 90%以上を占めていることや，品詞列は 11 位以下の低い出現率のものの合計がおよそ半数に達することなどが読み取れる。以降の節では，構築したコーパスを用いたアクセント結合処理モジュールについて検討する。

4 CRF を用いたアクセント結合

4.1 条件付確率場

観測データ $\boldsymbol{x}$ に対する出力ラベル $\boldsymbol{y}$ を学習するに際し，CRF は $(\boldsymbol{x}, \boldsymbol{y})$ 内での連続する変数の組（y_{t-1} と y_t, y_t と x_t など）の関係についての独立した特徴（素性）f を列挙し，各素性 f の重要度を θ_f，$(\boldsymbol{x}, \boldsymbol{y})$ 内で素性 f が満たされている箇所の数を $\phi_f(\boldsymbol{x}, \boldsymbol{y})$ とおいた上で，入力 $\boldsymbol{x}$ に出力 $\boldsymbol{y}$ を割当てることの確信度として，$\sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y})$ を考え，これを，

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y} \in Y} (\exp \sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y}))}$$

として確率分布とする。正解データを与えることによる学習は，この確率をできるだけ大きくするような重要度 θ_f を探る作業となる。本研究では，CRF++[5] を用いてアクセント結合処理を実装した。

4.2 学習・推定の内容

アクセント句境界は事前に与えられているものとし，各句中のアクセント核位置について学習・推定した。これらの情報は，前節で構築したコーパスより得られる。アクセント句境界位置情報を使用してアクセント句に区切り，句に含まれる各形態素の文中アクセント型を学習・推定の対象とした。例えば「音声合成」であれば「オンセー」（無核）と「ゴーセー」（1 モーラ目に核）という文中アクセント型を，単独発声時の「音声（オ̄ンセー）」と「合成（ゴーセー̄）」から推定する枠組みである。なお，形態素境界に核が生じているもの（「流れ・を（ナガレ̄・オ）」「出来・ない（デキ̄・ナイ）」など）では，当該アクセント核は前側の形態素に属するものとした。

但し，形態素内にアクセント句境界が存在している文や複数のアクセント核を持つ形態素を含む文は文ごと除去した。最終的に，学習用 3,581 文（25,692 アクセント句），推定用 527 文（3,533 アクセント句）に分けて使用した。JNAS の 40 ファイルを，35 ファイルと 5 ファイルに分割した。

4.3 単独アクセント型を与えない学習

形態素単独発声アクセント型が与えられていない状況（[3] と同様の条件）を想定し，観測素性とし

表 3: CRF による文中アクセント型推定の結果

	すべての句		単純な句		複合名詞を含む句	
直接学習（単独型なし）	2833 /3533	82.1%	703 /822	85.9%	530 /688	77.0%
直接学習（単独型あり）	3081 /3533	87.2%	775 /822	94.3%	523 /688	76.0%
変化の学習	3137 /3533	88.8%	791 /822	96.2%	553 /688	80.4%

て，前後 2 形態素を含めた 5 形態素について，[基本形/基本形読み/書字形/品詞/活用型]，[品詞]，[品詞 (大分類のみ)]，[活用型 (大分類のみ)]，[活用形 (大分類のみ)]，[モーラ数] と当該形態素の文中アクセント型との関係を与えた。また，遷移素性として，連接する形態素の文中アクセント型同士の関係を与えた。CRF++で学習・推定した結果が表 3 上段である。約 82% のアクセント句について正しい推定が得られている。{ 名詞, 動詞, 形容詞, 形状詞 }+{ 助詞, 助動詞 } の 2 語で構成された単純なアクセント句を対象とした場合，正解率が高く，逆に名詞の連続（複合名詞）を含む句では率が低くなる。なお，複数の核を持つアクセント句については，最初の核（主アクセント）が一致していれば正解と見なした。全ての核が一致した場合のみを正解とすると，最大で 3%程度正解率が低下する。

4.4 単独アクセント型を与える学習

前節での素性に加え，観測素性に [単独発声アクセント型] を加えて同様の学習を行なった。推定結果は，表 3 の 2 段目に示している。

以上の型推定は，文中アクセント型を直接学習・推定の対象としていたため，単独発声型と文中型がいずれも “1” である場合と “2” である場合は別々の事象として扱われる。そこで，単独型から文中型への「型変化」の様子を学習・推定の対象とすることで，類似する現象を共通のものと捉えられるようにした。具体的には次のような学習・推定を行なった。

単独型が有核の場合，文中型が有核であれば，単独型からの変化量を表すラベル（“[0]”, “[+1]”, “[+2]”, ...; “[−1]”, “[−2]”, ...）を学習対象とし，文中アクセント型が無核であれば，無核を示すラベル（“non”）を学習対象とした。一方，単独型が無核あるいは，値を持たないものに対しては，文中型を直接の学習対象とした。推定結果より，機械的に文中型に相当する数値に復元する。表 3 の下段に結果を示す。

文中型を直接推定する場合，単独型未使用時と使用時では多くの違いが見られ，特に単純な句においては 86%から 94%になるなど顕著な違いが見られる。しかし，複合名詞を含む句では正解率に向上が見られない。単純な句では単独型がそのまま文中型となることが多いのに対し，複合名詞においては型変形が頻出することによると考えている。

単独型から文中型への型変化を推定の対象とした場合では，文中型を直接推定する場合と比較し，全体的に正解率の向上が見られる。単純な句に対して 4%ほどの不正解が見られるが，これらの中にも，アクセントの揺れの範囲に納まると考えられるものが多数存在し，推定性能としては十分に高いものと考えている。それに対し，複合名詞を含む句では依然 20%程度の不正解が見られる。複合名詞の型変形に対して，[1] では規則として実装している。規則構築の際の議論を，より明確に CRF 学習に反映することで，正解率の向上が期待できる。本稿では頁数の関係で詳述しないが，規則を熟考して導出される観測素性を用いることで，91.7%, 96.4%, 85.6%まで向上させることに成功している。

5 まとめ

コーパスベース及び規則ベースのハイブリッド型アクセント結合処理モジュールの構築を念頭に置き，高品質なアクセントラベリングが施されたコーパスを構築すると共に，CRF を用いたアクセント結合処理を実装した。複合名詞の問題など未解決の問題があるものの，高い推定精度を得ることができた。

参考文献

[1] 匂坂芳典, 佐藤大和: “日本語単語連鎖のアクセント規則”, 電子通信学会論文誌, vol.J66-D, no.7, pp.847–856, 1983.

[2] N. Minematsu, R. Kita, and K. Hirose, “Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion,” Trans. IEICE, vol.E86-D, no.3, pp.550–557, 2003.

[3] 長野徹, 森信介, 西村雅史: “N-gram モデルを用いた音声合成のための読みおよびアクセントの同時推定”, 情報処理学会論文誌, vol.47, no.6, pp.1793–1801, 2006.

[4] 国立国語研究所報告 124「日本語話し言葉コーパスの構築法」, 独立行政法人国立国語研究所, 2006.

[5] 工藤 拓: “CRF++: Yet Another CRF toolkit” http://chasen.org/˜taku/software/CRF++/