J-ToBIの対話文への付与と追加タグ基準の検討

伊賀 竜彦† 河津 恵†† 斎藤 博昭† 石崎 俊††

† 慶應義塾大学大学 理工学部 数理科学科
†† 慶應義塾大学大学院 政策・メディア研究科

概 要

本研究では日本語東京方言の朗読文へ韻律構造を記述することを目的としたJ-ToBIという既存のタグセットを用いて，課題遂行型対話文にタグを付与した。J-ToBIはその文構造，単語構造を表現する意味あいが強く，対話内の感情を表すには不十分であった。当対話文に多く見られる感情を表現するために，新たに感情タグセットを定めた。さらに言い直し等の発話を示すタグ，空白時間を示すタグを設定した。4人のタグ付与者が，対話音声5分間に対してタグ付けを行い，各2人のタグ付与者間での揺れを測ることで評価を行った。既存のタグセットに関しては7割程度の一致率が見られた。

1 はじめに

対話コーパスを効果的に利用するにはその文字情報から得られる統語的解析，談話的解析のほかに，音声領域からの解析も必要とされる。対話には対話特有の韻律パターンがあるため，朗読音声とは異なる韻律情報を示した対話音声データベースを作成する必要がある。

本研究では，課題遂行型対話コーパス [1] に対して，日本語の韻律情報を効果的に付与することのできるタグセットJ-ToBI [2] を用いてタグを付与した。J-ToBIは朗読音声に付与することを対象として発案されたものであるが，自由発話にも対応することができる [2]。

付与を行う過程でどのような点で揺れが存在するのかを検証しつつ，揺れの削減を目的に，新たに必要な基準を追加した。またその対話コーパスに対し必要と判断した種々のタグを考案し，タグ付与を試行した。

2 日本語韻律構造

ここではJ-ToBIの背景となる日本語韻律構造を述べる [3]。

2.1 アクセント句

発話の音の流れをアクセントと呼ぶ。日本語のアクセントは音の高さで定義される。

日本語にはアクセント句という韻律単位が存在する。アクセント句は1モーラごとにその音の高さの高低を与えることで，生じる高低の変化の仕方により3つの型に分類される。

3つの型とは1モーラ目が低く，残りのモーラが高い0型，1モーラ目が高く，残りのモーラが低い1型，2モーラ目からNモーラ目までが高いN型である。

また，1型とN型に見られる落ち込み前のモーラをアクセント核と呼ぶ。

2.2 イントネーション句

人間の発話はその発話中幾度か呼気によって区切られる。それにともないピッチ曲線もその区切りを一つとして大きな変化を見せる。この大きな変化や呼気の存在で区切られた単位をイントネーション句と呼ぶ。

イントネーション句の区切りはアクセント句よりも大きな単位を持ち，一つのイントネーション句はその句内で複数のアクセント句を有する。

人間の発話は時間とともに下降する一つのイントネーション句に与えられる直線に，複数のアクセント句が重層したピッチ曲線によって描かれる。

3 J-ToBIタグ

ここではJ-ToBIタグ付けガイドライン [4] に示された基準を簡潔に述べる。

3.1 ワード層

発話の文字情報を単語単位に付与する。アクセント核の位置をアクセント辞書 [5] などから調べ，そのモーラの後に「’」を入れる。

また，アクセント辞書と発話のアクセント型が異なるときは単語層には辞書のアクセント型，トーン層には発話のアクセント型に沿ったタグを付与する。

3.2 トーン層

2節で示したアクセント句構造に対応する動きを記す。2モーラ目で高に移る時，またアクセント核の次のモーラで低に移る時，それぞれH-，H*+Lを付与する。H-，H*+Lが付与されるべきモーラ区間からずれて，ピッチ曲線の上昇や下降が見られるときは<，>でその事象を示す。アクセント句の開始，終了端には%L，L%を付与するが，句の1モーラ目が長母音や，アクセント核である場合には%Lの代わりに%wLを付与する。次アクセント句の1モーラ目が同様のケースで，かつアクセント句間で音声空白が存在しない時には，L%の代わりにwL%を付与する。音声からピッチ曲線が上昇を見せると想像できる場合にはH%を付与する。

3.3 ブレークインデックス層（以下BI層）

この層は各単語の終了が2章で示したどの構造の境界と一致するかを示す。単語，アクセント句，イントネーション句，文の境界の順に1，2，3，4と付与する。自由発話に見られる，言い淀み，言い直しの箇所にはアクセント句の途中で終了の場合には1p，アクセント句が完結した後の終了には2pが付与される。

3.4 色々層

トーン層やBI層で表現できない事項で必要と思うことがあれば色々層に記す。

4 対話コーパスへの適用

ここでは当コーパスに付与する上で，追加した事項に関して述べる。

4.1 付与上の留意点

3節で示した基準によりタグ付与を行い，付与途中で結果を比較した。基準を詳細にすることで解消できる揺れに関しての問題点と付与者に指示した事項を述べる。

- 波形の開始，終了端の位置が音声と異なるときには波形に付与する。また，無声区間では波形の端の見分けが困難であったが，これは付与者の判断に任せた。
- H-，H*+Lの付与箇所が基準では曖昧であったため，示された付与例を参考に基準を定めた。H-，H*+Lは事象がおこるはずの該当モーラのピッチ曲線のピークに付与し，＜と＞は波形の該当モーラ内のピッチで事象がおこらない時のみ，事象がおこる位置に付与する。
- アクセント句1型の時，%wLとH*+Lの付与に関して，%wLが付与される波形の始めがピッチピークと重なればH*+Lは省略する。
- 「まる・がお・の」等，本来アクセント句であるものが分離されて発話されているときには個々のアクセント句と判断して，BI層に2pを付与する。
- 「まるが」のように単語の途中で発話が終了した時には1pを付与する。
- 4が付与される箇所はテキストに書きおこした時,「。」で終了すると付与者が判断した箇所や，3の発話後一定の閾値以上の空白が存在した箇所とする。
- 笑いに関してはアクセント句の構造が適用できないので，トーン層，BI層に記述は行わず，色々層に笑いである事を示すタグを付与する。
- 「あ」などの一文字発話に関しては，次に続くアクセント句との間に無音区間があるときには「あ」は一つのアクセント句と仮定する。この場合直後のアクセント句が1型の時は「あ」のアクセント型は0型，それ以外は1型とする。トーン層には開始，終了端を示すタグのみをそれぞれの型通りに付与する。無音区間が存在しない場合は「あ+直後の発話」で一つのアクセント句ととらえる。
- 平坦な長母音発話「あー」などは1つのアクセント句と仮定し，トーン層に%wLとL%を記す。

4.2 ワード層の単語

ガイドラインで示されている単語の定義は一般で認識されている単語の定義とは異なり，かつ定義自体も曖昧であった。本研究ではガイドラインの付与例を参考に，単語の定義に『用言に付加する助詞と助動詞は分けない』という点を追加した。

またアクセント辞書から判断できないアクセント句構造に関しては付与者の判断に任せた。

4.3 トーン層でのタグの省略

ガイドラインでは省略することの可能であるタグが示されていたが，省略すべきケースの基準が曖昧であったため，明確な基準を定める必要があった。

アクセント句が2型の場合にはH-とH*+Lが同じ箇所に付与される為，H-は省略する。また，アクセント句間で音声空白がないときには，前のアクセント句のL%，wL%と同じ箇所に付与される後のアクセント句の%L，%wLを省略する。

4.4 色々層への追加

筆者らは笑いには日本語のアクセント句の構造が存在しないと考え，トーン層，BI層には笑いに関しての記述を行わずに，この層で笑いが存在することを示すタグを付与する。笑いのおこる区間を <laugh，laugh> を用いて囲む。またこの層に以下の自由発話特有の情報を示す。

- 言い直される発話の終了時点にrを付ける
- 「えー」「あー」などの発話終了時点にfを付ける
- 「で」など接続的発話の終了時点にdを付ける

4.5 追加した層

色々層には種々のものを追加してよいが，個別に研究として使用しうる用件に関しては別の層を設けることにした。以下では追加した層を述べる。

4.5.1 重複層

当コーパスは対話音声であり，2人の対話が重複している箇所が存在する。その箇所を示す為の重複層を設け，重複区間を <ol，ol> で記した。

4.5.2 無音層

当コーパスでは，対話者同士が発話する内容を考えながら対話を進めるため，無音区間が多く現れた。この無音情報に何らかの意味があると考え，無音層を設け，<sil=*time*(msec)，sil> により無音位置を囲み，無音時間を記した。

4.5.3 感情層

韻律情報が生む感情を感情層を設け，付与する。

付与者を通じ試験的付与を行い，付与者からの聴取を行った結果から，当コーパスの対話形式に沿う感情を定めた。定めた感情を以下に示す。

平静，安堵，謙虚，疲労，落胆，不信・不満，戸惑い・困惑，驚き，恥らい，自信の欠如，自信，喜び，楽しい，遊び，呆れ，焦燥

一方，上の感情の付与をスムーズに行う目的で，以下のような発話の役割を表すタグを設定した。

質問：単純質問，確認質問
応答：賛成応答，反対応答，説明応答
説明：通常説明，受け継ぎ説明
独り言：あいづち，笑い，独り言

表 1: 付与者 A と付与者 B の比較 (トーン層)

タグ	%L	%wL	L%	wL%	HL	H-	<	>	nu	ot	all	av1	av2
%L	44	4	5	0	2	1	0	0	14	0	70	0.79	0.63
%wL	3	16	0	0	4	0	0	0	3	0	26	0.70	0.62
L%	0	0	92	1	0	0	0	0	18	2	113	0.99	0.81
wL%	0	1	4	11	2	0	1	0	6	0	25	0.58	0.44
HL	1	0	3	2	79	1	0	0	14	0	100	0.92	0.79
H-	2	0	0	0	4	44	0	0	8	0	58	0.88	0.76
<	0	0	0	0	4	3	2	0	2	0	11	0.22	0.18
>	0	0	0	0	10	2	0	0	5	0	17	0.00	0.00

表 2: 付与者 A と付与者 B の比較 (BI 層)

タグ	1	2	3	4	1p	2p	nu	ot	all	av1	av2
1	53	12	0	0	0	0	7	0	72	0.82	0.74
2	13	39	2	0	0	0	1	0	55	0.72	0.71
3	0	9	12	0	0	0	2	0	23	0.57	0.52
4	1	0	8	38	0	0	8	0	55	0.81	0.69
1p	0	1	0	0	3	0	0	0	4	0.75	0.75
2p	0	0	2	0	1	2	0	0	5	0.40	0.40

感想

宣言

感情層でのタグは，対話における感情とその発話の役割という 2 つの観点を組み合わせて記した。例えば,『自信+確認質問』のように付与する。

5 実験環境

2 人の話者による課題遂行型対話のコーパスビデオから DAT-Link [6] を用いてサンプリングレート 8KHz でステレオ音声をディジタル化した。2 話者の音声を同室内で収録したため，一方の音声には音量的に小さく相手の音声が採取されている。採取した音声波形とピッチは Entropic waves+ [7] を用いて計算，表示できるようにし，xlabel 機能を用いて，該当時間にタグを付与できるようにした。付与対象は 5 分間とし，ヘッドホンで音声を聞きながら 4 人がタグ付与を行った。

6 タグ付与結果とその評価

タグ付与者間でのタグの一致度により評価を測る。4 人間の総合的一致度の計算は困難であるため，各 2 者間で測ることにする。表 1 と表 2 は話者 A のトーン層，BI 層に付与者 A，B が付与した結果で，一例として載せる。

表は横軸の付与者 B の結果を正解として比較した，縦軸の付与者 A の一致数と率を示している。誤差を 100msec まで許可し，誤差内に正解のタグが存在した時，一致個数をカウントした。存在しなかった時には一番近い位置のタグを誤答とし，揺れ方を測った。all は全ての個数であり，av1 は何らかのタグが付与された中での一致率，av2 は何も付与されていない事も考えた上での一致率である。各付与者間同士の平均で既存の層で 7 割程度の一致度が得られた。

この比較方法は，あくまで一方の話者がどれくらい相手のタグと比較して同じタグを付与できたかという事を表す。一方からの一致度合を測るのみでは付与者間でのタグ一致度が高い，低いとは言えず，双方の一致度を用いて評価する必要がある。

以下では各層ごとの揺れの発生具合と，揺れた原因を述べる。

6.1 トーン層と BI 層

6.1.1 アクセント句の切り分け

全体を通じて L%に関する一致度の高さが得られた。ワード層で各単語の終了時点が示されているので時間的な揺れは考えられない。揺れがおこる原因は少なくともアクセント句以上であるという区切りの成否にある。BI 層の 1 のタグの一致度を見ると，話者 A，B 共に 1 の一致度合も高く，アクセント句以上の切り分けは問題ない。

6.1.2 ピッチ曲線の読みとり

H-，H*+L の付与の一致度が高く，これは発話が韻律通りであればワード層の情報から決定され，さらにアクセント句の切り分けに成功していることに理由がある。av1 に比較し，av2 の低下が見受けられた。このタグにおいて av1，av2 の差が顕著であるということは，同モーラ内での付与を試みているにも関わらず，ピッチ曲線を読みとれていないことを示す。

さらにピッチ曲線の読みとりの必要がある，<，> のタグの一致度の顕著な低さが各付与者間で見受けられた。

6.1.3 イントネーション句の分離

アクセント句の切り分けに成功している事を踏まえて 2，3 の一致度合を見ると，話者 A で若干一致度合が低かった。理由は話者 A の発話が相手に説明するという，途中に思考をはさんだ途切れがちな発話であるからで，2，3 の区別が困難であったと思われる。

一方，話者 B は話者自身の話し方が話者 A に比較し一息で話す話し方であり，イントネーション句の区切りを示す 3 タグが少なく，データとして不足がちであった。

付与者同士の比較結果を観察するに，付与者 A，B 間での比較が高い一致度を示したが，別の付与者の一致度を測る上で，対象を付与者 A とした時と付与者 B とした

時の数字に開きが表れた。また，少数ではあるが 1p，2p との混同も見受けられる。逆に 3 から眺める 2 のタグへの混同度合も高くある。さらに 3 の一致度は 2 の一致度に比較し低い。

以上を踏まえるとイントネーション句の分離は困難であったと考えられる。

6.1.4 1 型と長母音

%L に比較し，%wL に関しての一致度がやや低い。%L が付与される状況下で，1 型，長母音であれば%wL が付与できるという%wL の基準から，付与されるべき状況が増えるにつれこの一致率は近づくと考えられる。%L を他のタグと比較すると，wL%との混同がよく見られる。「音声空白」感覚の異なり，波形の端の見分けの困難さが揺れを生むと想像できる。これはイントネーション句の分離に関しても同様のことがいえる。音声空白に関して閾値を設ければ解決はできるが，あくまで当コーパスに対しての閾値である形にしておくのがよい。

6.2 色々層への追加

笑いに関する揺れは重複層や無音層同様に一方の話者の笑いがもう一方の話者の笑いと重なったり，波形的に見分けるのが困難であったりと，笑いの開始終了位置の判定が一致しないことが原因である。しかし無音層では時間的に厳密な正確さを要求する必要はない。その区間で何らかの意味を持つ発話が存在しないことを示せば良いので，このために無音層やトーン層と見比べながらの付与をすべきである。

また，音声空白に関して付与する 3 種類のタグに関してはタグ付与時に示した基準の表現が曖昧であったため，付与者にその意味がしっかりと伝わらなかった。

6.3 無音層

この層ではまず付与対象となる音声波形の開始終了時点が雑音と混同され，無音の判定位置があいまいになった。さらに，空白に関する閾値を設けていなかったために，付与者の『聞いた感じ』という主観も含まれた結果になった。この空白閾値の問題は他の層のタグ結果にも影響を及ぼしている。

6.4 重複層

重複層を設けたが，発話が重複している箇所が少なく，重複タグのデータ数としても少なかった。また，存在している箇所でも重複時間が短く，無音層や笑い同様，波形の見分けが出来ずに音声的にはっきり重複している箇所でも揺れがおこった。付与には雑音のさらに少ない音声が望ましい。

また，重複層の付与は別の波形として出力される 2 話者の音声波形を見比べ，音声を聞き比べながらの付与であったため，重複の見分けが困難であったといえる。

6.5 感情層

感情タグに関しては揺れが多かった。以下にはその揺れの種類と理由を述べる。

両話者の発話で言葉の持つ表現に依存して考えた感情と韻律からとらえる感情との間で揺れが生じた。

さらに話者 A に関しては途切れ途切れの発話が多く，また，全文が『平静』ではあるが，文末で発声音量の低下が見られる箇所が多かった。それが『平静』,『困惑』,『自信の欠如』間の揺れを生んだ。また，同様に断定的な発話は『平静』と『自信』の揺れを生んだ。

話者 B に関しては質問する立場で，かつはっきりした物言いであるからか，揺れが A よりも少なく，感情タグの評価としても満足できるものであった。しかし文末の動きで与える感情の捉え方が異なるという点は話者 A と共通している。

また,『不信』,『困惑』間の判断しにくい感情の揺れが見られた。タグの設定に問題があったと言える。

感情と他のタグを比較しながらでは，同一文内での空白 (言い淀み) の前後の発話形態 (速度，音量) の変化によって，あとの要因が感情に強くあらわれている。この場合には，空白を『困惑』『自信の欠如』と取る文全体の感情と，空白が示す思考後の発話形態を重く見た『自信』『平静』間との揺れが表れている。

当コーパスに付与した感情タグの結果から考えるに，その言葉の意味が持つ感情と，韻律から感じる感情との違いからおこる揺れが見られた。感情の感知は発話末の存在で左右され，さらに言い淀みから生まれる音声空白は揺れを大きくする事が解った。

7 おわりに

対話音声に J-ToBI を付与する試みを行った。既存の層に関しては既存の基準に関する解釈の揺れや，基準への言及の不十分な事がタグの揺れにつながったので，基準を追加することで揺れを抑えた。結果として低い数値は BI 層の 3 タグが顕著で、自由発話におけるイントネーション句の区切りが困難であるという課題を残した。

新たに設定した無音層，重複層では，波形開始，終了端においてのずれが多く見られたが，時間的な制限に厳しい要求をしない研究に利用するのであれば，問題ないタグ付与が行える。

感情層に関しては，さらに多くのコーパスに付与を試み，付与データを作成することにより，コーパスにあわせた，また本研究の付与でも問題となった言葉にあわせた感情出現パターンを構築する必要がある。

参考文献

[1] 平成 9 年度 対話処理技術専門委員会活動報告 (日本電子振興協会 自然言語処理システムに関する調査報告書 所収)

[2] ニック・キャンベル: Tones and Break Indices (ToBI) システムと日本語への適用，日本音響学会誌 53 巻 3 号，pp.223-229，1997

[3] 古井貞熙: “音声情報処理”，森北出版，1998

[4] J. J. Venditti: Japanese ToBI Labelling Guidelines, Ohio-State University, Columbus, U.S.A., 1995

[5] 日本語発音アクセント辞典，NHK 出版，1998

[6] http://www.tc.com/

[7] http://www.entropic.com/