RDBと既存のアノテーションツールによる統合的コーパス開発環境

伝 康晴
千葉大学文学部
den@cogsci.l.chiba-u.ac.jp

小磯 花絵
国立国語研究所
koiso@ninjal.ac.jp

1　はじめに

近年、コーパス・アノテーションはますます多様化し、多層的アノテーションを統合的に管理する枠組みが必要になってきている。この問題に対して、たとえばKaplan et al. (2012)は、汎用のアノテーション表現を設計することで多層的アノテーションを統合的に管理する枠組みSlateを提案している。

しかし、こういった試みの多くは書き言葉を対象としており、話し言葉、とくにマルチモーダル談話を対象とした場合、いくつかの点で不十分である。本研究では、話し言葉やマルチモーダル談話のアノテーションを念頭に置き、より汎用性の高いアノテーション表現をRDBを用いて実装し、既存のアノテーションツールとの間のインタフェースを設計することで、コーパス開発過程を統合的に管理する方法を提案する。

2　問題の所存

既存のコーパス管理環境は以下の点で不十分である。

2.1　汎用アノテーション表現

既存のコーパス管理環境の多くは、汎用アノテーション表現として、スタンドオフ形式で文書中の要素を記述するセグメントと、セグメント間の関係を記述するリンクに基づく方法を採用しており、セグメントの開始・終了位置は文書中での文字位置を利用している(Noguchi et al., 2008; Kaplan et al., 2010)。

しかし、マルチモーダル談話では、単語・文節・発話などのテキスト情報以外に、アクセント・トーンなどの韻律情報、視線・頷き・ジェスチャーなどの非言語情報も付与する必要があり、単一の文書中での文字位置では位置を記すことができず、同期された各文書（音声・映像ファイル）中での時点（ファイルの先頭からの経過時間）によって位置を記す必要がある(たとえばBrugman & Russel, 2004)。時点による位置の記述は一度決めると不変というわけではなく、位置の修正はアノテーション過程で頻繁に生じる（たとえば、アクセントの位置や単語の開始位置を微妙にずらすなど)。この際、単語と文節のように階層関係にある要素間では、一方の側での修正が他方にも影響することがある（たとえば、文節の先頭の単語の開始位置を修正すると文節の開始位置も同時に修正しなければならない)。

この問題は、セグメントの開始・終了位置を記す指標と、文書中での実際の時点への写像とを別々に表現することで解決できる(Bird & Liberman, 2001)。

2.2　既存のアノテーションツールとの接合

既存のコーパス管理環境では、汎用アノテーション表現を実装したDBと既存のアノテーションツールとの接合についてはほとんど述べていない。たとえば、Kaplan et al. (2012)は、コーパス管理環境が支援すべき項目として「他システムとの連携」や「入出力の拡張性」を挙げているが、その詳細は述べていない。

話し言葉やマルチモーダル談話のアノテーションでは、音響処理（ピッチ曲線やスペクトログラムの表示）や動画処理（スロー・コマ送り再生や複数動画の同期再生）といった機能が必要不可欠であり、それらの実装には高度に専門的なスキルが必要である。その一方で、それらの機能を備えた高機能なアノテーションツールが既に世界的に普及している。これらの既存のツールを有効に利用することが話し言葉やマルチモーダル談話のアノテーションにおいては重要な要件となる。

これら既存のツールの利用は、たんに入出力書式を変換するだけでは解決しない場合がほとんどである。たとえば、筆者らは会話中のあいづち表現の反応先（あいづち反応の契機となった他者発話中の要素）をアノテーションツールAnvilを用いてタグ付けしている。これはアノテーションDBではリンクで表現されるた

め、関係付けられる 2 つのセグメント（あいづち表現と他者発話要素）の（アノテーション DB における）index を参照する必要があるが、これは Anvil が内部的に利用している index とは異なる。よって、入出力書式の変換だけでなく、index 間の変換も必要となる。

Slate (Kaplan et al., 2012) では、アノテーション DB と密に接合したインタフェースを通じてアノテーション作業を行なっている。しかし、話し言葉やマルチモーダル談話のアノテーションでは音声や動画の再生が不可欠であり、Slate のインタフェースにそのような機能は期待できない。同様に、単一のアノテーションツールにすべてのアノテーションに必要な機能を期待することはできない。たとえば、ELAN (Brugman & Russel, 2004) は Bird & Liberman (2001) の理論に基づく極めて汎用的なアノテーション表現を提供しているが、リンクに相当するものは実装されていない。

したがって、汎用のアノテーション表現を DB で管理しつつ、既存のアノテーションツールと疎かつ連続的に接合する方法が必要となる。

3 アノテーション DB

本節では、本研究で実装したアノテーション DB の構成について述べる。

3.1 汎用アノテーションの理論

汎用アノテーション表現を、Bird & Liberman (2001) の理論を基礎にして設計し、ELAN のモデル (Brugman & Russel, 2004) と Slate のモデル (Kaplan et al., 2010) を参考にして定式化した。

定義 1 時点構造 $\langle P, \leq_P \rangle$、記述構造 $\langle D, T, \sqsubseteq_T, \lambda \rangle$ 上のアノテーション文書 $\mathcal{A}$ とは、以下を構成要素とする組 $\langle I, \leq_I, J, S, L, \tau, \delta \rangle$ である。

i) 時点指標の集合 I
ii) I 上の半順序関係 $\leq_I$
iii) 記述指標の集合 J
iv) セグメントの集合 S
v) リンクの集合 L
vi) I から時点の集合 P への部分関数 τ
vii) J から記述の集合 D への関数 δ

ただし、セグメント $s \in S$ とリンク $l \in L$ は以下を満たす。

1. セグメント s は $\langle j, i_s, i_e \rangle$（$j \in J,\ i_s, i_e \in I, i_s \leq_I i_e$）の形式で、$j$ は S 中でただ一度だけ出現する。
2. リンク l は $\langle j_l, j_s, j_d \rangle$（$j_l, j_s, j_d \in J$）の形式で、$j_l$ は L 中でただ一度だけ出現する。

開始・終了点の時点指標 i_s, i_e と記述内容へのポインター j の組 $\langle j, i_s, i_e \rangle$ によってセグメントを表わす。$i_s <_I i_e$ のとき、ある幅を持った区間を占める要素（単語や発話など）に対応し、$i_s = i_e$ のとき、瞬間的に生起する要素（アクセントなど）に対応する。また、あいづち表現の反応先や照応関係のようなセグメント間のリンクは、関係付けられる 2 つのセグメントの記述指標 j_s, j_d とリンクの記述内容へのポインター j_l の組 $\langle j_l, j_s, j_d \rangle$ によって表わす。

時点指標 $i \in I$ は時間関数 τ によって文書中の時点 $p \in P$ に写像される。書き言葉では、p は文書の先頭からの文字位置であり、話し言葉やマルチモーダル談話では、p は開始点からの経過時間である。ただし、融合した語の構成語への分解（「わたしゃ」→「私 | は」）のように、実際の時点を定められない境界に位置する時点指標 i に対しては $\tau(i)$ は定義されない（図 1 参照）。

時点は一般に以下の時点構造 $\mathcal{P}$ に従う。

定義 2 時点構造 $\mathcal{P}$ とは、i) 時点の集合 P、ii) P 上の全順序関係 $\leq_P$ からなる組 $\langle P, \leq_P \rangle$ である。

すべての時点 $p \in P$ は一直線上で順序付けられる。以下では、P は非負実数、$\leq_P$ は通常の代数的大小関係とする。

セグメントやリンクの記述指標 $j \in J$ は記述関数 δ によって記述 $d \in D$ に写像される。d は属性・値対の集合 $\{a_1/v_1, a_2/v_2, \cdots\}$ によって表現される。

記述は一般に以下の記述構造 $\mathcal{D}$ に従う。

定義 3 記述構造 $\mathcal{D}$ とは、i) 記述の集合 D、ii) 層の集合 T、iii) T 上の半順序関係 $\sqsubseteq_T$、iv) D から T への関数 λ からなる 4 つ組 $\langle D, T, \sqsubseteq_T, \lambda \rangle$ である。

すべての記述 $d \in D$ は層化関数 λ によってただ一つの層 $t \in T$ に帰属させられる。たとえば、記述 {orth/私, pron/ワタシ, pos/代名詞 } は Word 層に帰属させられる。同様に、記述関数 δ と層化関数 λ の合成関数 $\lambda \circ \delta$ によって、すべてのセグメント（の記述指標）はただ一つの層に帰属させられる。層間には順序が設定でき、たとえば、Word, Bunsetsu, Utterance 層の間に Word $\sqsubset_T$ Bunsetsu $\sqsubset_T$ Utterance なる順序を設定できる。これを支配関係と呼ぶ。

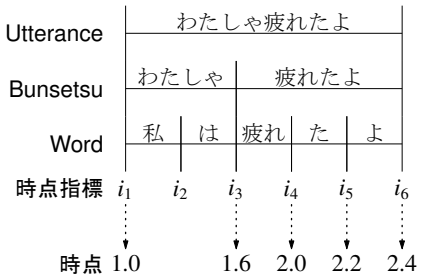

図 1　多層的アノテーションの例

多層的アノテーションの例を図 1 に示す。図中で「私」「わたしや」などのラベルが付いている区間がセグメントであり、各セグメントの開始・終了点は区間の両端に付いているバーの先にある時点指標で示されている。支配関係にある 2 つの層（たとえば Word と Bunsetsu）に帰属するセグメント間（たとえば「私」と「わたしや」）では、（半順序関係 $\leq_I$ に基づく）時点指標区間の包含関係によって、要素間の階層関係が表現される。時点指標（の一部）は時間関数（図中の矢印）によって文書中の時点に写像される。時点軸上の包含関係は必ずしも要素間の階層関係を表現しない。たとえば、頷きを表現する Nod 層があり、時点区間 [1.7,2.3] に頷きのセグメントがあったとしても、Nod と Bunsetsu の間には支配関係が設定されていないため、この頷きと「疲れたよ」の間には階層関係はない。

3.2　RDB による実装

3.1 項の汎用アノテーション表現を RDB によって実装した。実装には SQLite を用いた。図 2 にクラス図によるモデル表現を示す。このモデルは、ELAN と Slate のモデルの折衷案とみなすこともできる。

セグメントとリンクはそれぞれ Segment と Link で表現し、属性・値対によるこれらの記述を Attributes で表現する。時点指標は TimePoint で表現し、時点への写像は属性 time_value で表現する（値を持たない場合もある）。Segment と Link は層を表現する Tier に帰属し、Tier 間の支配関係は Dominance で表現する。

4　既存のアノテーションツールとの接合

本節では、3 節のアノテーション DB を既存のアノテーションツールと接合する方法について述べる。既存のアノテーションツールとしては、筆者らが日常的に利用している ELAN・Anvil・Praat を取り上げる。

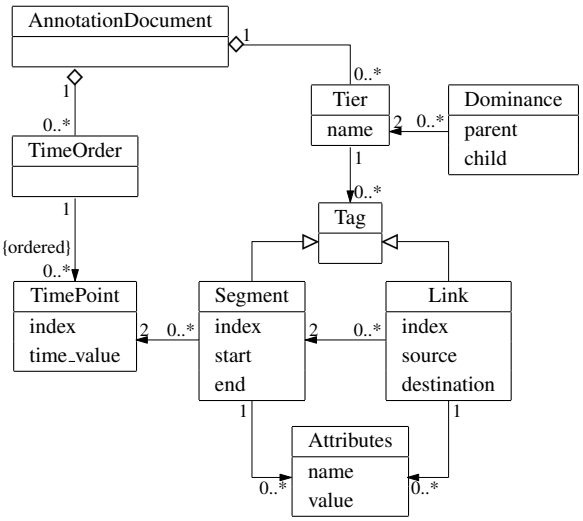

図 2　クラス図によるモデル表現

4.1　ELAN

ELAN[*1]は高機能なビデオ・アノテーションツールであり、ジェスチャー研究などで世界的に広く利用されている。アノテーション表現として、本研究とほぼ同様のものを用いており（ただし、リンクは実装されていない）、入出力書式は図 2 と類似のモデルを XML で表現したものとなっている。したがって、アノテーション DB と ELAN との接合は直接的である。

ただし、大きな相違点として、図 1 の「わたしや」→「私 | は」と「疲れたよ」→「疲れ | た | よ」の分節化は別の方式で実現される。後者のようにセグメントの時点がすべて定まる場合は “Time subdivision” と呼ばれる方式で、前者のように中間の時点が定まらない場合は “Symbolic subdivision” と呼ばれる方式で実現される。両者の混在は許されないため、この例では、Word 層を二重化し、「疲れ | た | よ」のような “Time subdivision” に基づく分節化を行なったのちに（この段階では「わたしや」は分節化しない）、「私 | は」のような “Symbolic subdivision” に基づく分節化を行なう（図 3 参照）。

このような層の二重化（とその逆変換）が ELAN とのインタフェースでは必要となる。

4.2　Anvil

Anvil[*2]もビデオ・アノテーションツールである。ELAN の登場以降は利用機会が少ないが、ELAN に

[*1] http://www.lat-mpi.eu/tools/elan/
[*2] http://www.anvil-software.de/

```
<TIER TIER_ID="Word1" PARENT_REF="Bunsetsu"
      LINGUISTIC_TYPE_REF="word1">
 <ANNOTATION>
  <ALIGNABLE_ANNOTATION ANNOTATION_ID="a4"
    TIME_SLOT_REF1="ts1" TIME_SLOT_REF2="ts2">
   <ANNOTATION_VALUE>わたしゃ</ANNOTATION_VALUE>
  </ALIGNABLE_ANNOTATION>
 </ANNOTATION>
 ...
</TIER>
<TIER TIER_ID="Word2" PARENT_REF="Word1"
      LINGUISTIC_TYPE_REF="word2">
 <ANNOTATION>
  <REF_ANNOTATION ANNOTATION_ID="a8"
    ANNOTATION_REF="a4">
   <ANNOTATION_VALUE>私</ANNOTATION_VALUE>
  </REF_ANNOTATION>
 </ANNOTATION>
 <ANNOTATION>
  <REF_ANNOTATION ANNOTATION_ID="a9"
    ANNOTATION_REF="a4" PREVIOUS_ANNOTATION="a8">
   <ANNOTATION_VALUE>は</ANNOTATION_VALUE>
  </REF_ANNOTATION>
 </ANNOTATION>
 ...
</TIER>
<LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="word1"
    CONSTRAINTS="Time_Subdivision"
    TIME_ALIGNABLE="true"/>
<LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="word2"
    CONSTRAINTS="Symbolic_Subdivision"
    TIME_ALIGNABLE="false"/>
```

図 3　ELAN における層の二重化の例

はないリンク・アノテーションの機能があり（Slate の共参照タグ付けと同様にマウスクリックで参照先を選択可能）、筆者らはあいづち表現の参照先をはじめ、話者移行や照応関係のタグ付けで Anvil を利用している。

Anvil もスタンドオフ形式のアノテーションを XML で表現しているが、ELAN と比べると本研究のアノテーション表現との相違点が多い。とくに、リンクを表現するためのセグメントの参照に内部的に用いている index を利用しており、アノテーション DB で用いている index との相互変換が必要である。また、時点指標は用いず、直接、時点によって開始・終了点を表わしているため、時点指標と時点との写像を管理する必要がある（時点が変更された場合に対応するため）。

4.3　Praat

Praat[*3]は音声分析・アノテーションツールであり、話し言葉の音声学的アノテーションで標準的なツールとなっている。筆者らは、分節音・単語境界や韻律情報のラベリングで利用している。

Praat の入出力書式は独自のものであるが、スタンドオフ形式と等価である。しかし、やはり時点指標は用いず、時点によって位置を表わしているため、Anvil と同様な対応が必要である。とくに Praat での作業では、時点を操作したり、セグメントを追加・削除することが多いため、時点指標を表わすための Tier を特別に設け、時点への写像を管理する。

[*3] http://www.fon.hum.uva.nl/praat/

5　おわりに

本研究では、さまざまなアノテーション方式を包含する汎用アノテーション表現を RDB を用いて実装し、既存のアノテーションツールとの間のインタフェースを設計することで、コーパス開発過程を統合的に管理する方法を提案した。筆者らは、本手法を用いて、『千葉大学 3 人会話コーパス』と『日本語話し言葉コーパス』（RDB 版）の開発を進めており、分節音情報・形態論情報・韻律情報・発話単位・あいづち表現・視線・頷き・話者移行・照応関係など、極めて多彩なアノテーションを行なっている。将来的には、アノテーション・タスクや作業者などのコーパス作成プロセスの管理まで視野に入れている Slate などを適宜拡張し、アノテーション DB として用いることも検討したい。

参考文献

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, **33**, 23–60.

Brugman, H., & Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the 4th International Language Resources and Evaluation Conference (LREC 2004)*, 2065–2068. Lisbon, Portugal.

Kaplan, D., Iida, R., & Tokunaga, T. (2010). Annotation process management revisited. In *Proceedings of the 7th International Language Resources and Evaluation Conference (LREC 2010)*, 3654–3661. Valletta, Malta.

Kaplan, D., Iida, R., Nishina, K., & Tokunaga, T. (2012). Slate: A tool for creating and maintaining annotated corpora. *Journal for Lanuage Technology and Computational Linguistics*, **26**, No. 2, 91–103.

Noguchi, M., Miyoshi, K., Tokunaga, T., Iida, R., Komachi, M., & Inui, K. (2008). Multiple purpose annotation using SLAT: Segment and link-based annotation tool. In *Proceedings of the 2nd Linguistic Annotation Workshop*, 61–64. Marrakech, Morocco.