少数言語コーパス向け記述データの構造

大矢 一志
鶴見大学

言語資料 (コーパス) に付加されるアノテーションを共有するために提案されている複数の規格のうち，ISO24610-1, ISO24610-2(審議中), ISO24612(審議中), TEI, CES/XCES, GrAF を比較検討し，規格間にある違いは，論議が未整理である部分から生じていること，具体的には通称"standoff" と呼ばれる記述の扱いに関して更なる規格が必要であることを確認し，その案を提示する．また，本稿では，少数言語コーパスを作成する際に，現状で採りうるひとつの現実解を提示する．

Data Structure for Minority Language Corpora

OHYA Kazushi
Tsurumi University

We examine the current specifications for language annotations, especially ISO24610-1(2006), ISO24610-2, ISO24612, TEI, CES/XCES and GrAF, and confirm that there are missing specifications between them, and we need at least three new specifications for corpora with annotations. And also we show our design of DB and the data model for it, which is a resultant of the project of Online Text Database of Minority Languages in the Linguistic Dynamics Science Project(LingDy), Tokyo University of Foreign Studies.

1. はじめに

本稿は，東京外国語大学アジア・アフリカ言語文化研究所「言語ダイナミクス科学研究プロジェクト (LingDy)」で 2008 年から始められた，「少数言語の言語資料 (以下コーパス) のデータベース研究とオンライン公開プロジェクト (以下，本プロジェクト)」で得られた成果報告の一部である．人文学者が求めてきたコーパスの利用法には，現在の使われ方とは異なるものがある．近年のコーパスは，例えば，辞書で用例の順位を決めたり，自動翻訳で多くの納得を得られる候補を求める為に使われているが，これはいわば規範文法を求める使われ方に限定されたものである．一方，人文学のコーパスは，個別現象の観察・分析の記録と，そこからの単位分析[1]の資料として使われることがある．この多くは，ノートに記された個人記録といえるべきもので，かつてはカードとして，現在では電子化されたデータとして，個々の記録が分析の対象として使われている．人文学では，資料の全体から得られる結果だけではなく，その分析の元となる個別データの作成そのもの，更には，いわゆる分析対象となるデータ単位の定義自体を検討するためにも，コーパスは使われてきた [18, 17, 14].

本稿では，少数言語を対象とした言語学者がコーパスを公開する過程で必要となる検討課題のうち，世界規模のアーカイブへの参加が潮流である現時点で検討が求められる各種規格の状況を整理し，当該プロジェクトが採用すべきデータ構造を探るものである．

はじめに，コーパスと言語学の状況を確認し，課題とコーパス作成・公開への要求をまとめる．次に，影響力の強い規格の策定状況をまとめ，そこから提案間で十分に注意する箇所があることを確認する．次に，これらの検討から見えてくる技術的な課題を整理し，解決策を提示する．最後に，当プロジェクトが想定しているデザインを紹介する．

[1] どれを基本単位にするかを決めること．

2. Language Documentation

W3Cが掲げだした Semantic Web の関連技術は，コーパスへの関心を高め，結果，アノテーションの重要性を広く認識させることになった．今や関心は，コーパスの質をどのように高め，保証してゆくかという段階にまで移っている．例えば，ジャーナル[2] の創刊や，ヨーロッパにおける ISO 活動や CLARIN [3] の始動などに，それを見ることができる．コーパスの質を高めるには，アノテーションが欠かせない．そのアノテーションには，個別現象の観察・分析の記録という役割が，人文学ではある[4]．言語学でもコーパスの扱われ方は，工学の影響を受けて近年変化をしており，単純な分類は困難であるが [15]，その 1 つとして Language Documentation と呼ばれている領域がある．

近年までの言語学の動向は，70's に Chomsky 理論が統語論を席巻し，その勢いは 80's の形式言語学の発展へとつらなるが，90's になると自然言語処理研究に言語理論研究は取って代わられ，一般言語学を支えるべき理論言語学の存在意義が疑われると，一気にその勢いは無く

[2] *ACM Journal of Data and Information Quality*

[3] Common Language Resources and Technology Infrastructure

[4] 残念ながら，このような使われ方は，工学で主流ではなく，殆ど関心が払われていない．データの質を高める関心が，アノテーション付きコーパスのこのような利用面にも関心を払われることを期待したい．

なってしまった．同時に，コンピュータ・ネットワーク文化の普及により，英語を外来語とする文化の流入を迎えたかつての大言語社会でも，社会的・文化的地位の相対的低下にさらされることで，言語の多様性が評価されるようになり，結果として個別言語の記述への関心が高まっている．このような個別言語を扱う言語学のうち，広くはコーパスに基づく，狭くは現地調査によるコーパス収集から始める言語学のことを記述言語学と呼ぶとすると，現在，記述言語学で課題となるのが，コンピュータを記録に使うという，いわゆるドキュメンテーションの問題である．このような背景を元にした分野を Language Documentation[14] と呼んでいる[5]．その記録対象は広範囲に及び，コーパス (言語資料) であるから，文字のない言語も含まれるため，その一次資料には音声または音声付き動画などの，いわゆるマルチメディアが含まれてくる．主に音声を元に，記号へ転記 (transcription) され，コーパスが作成されている．

2.1 言語学者が抱える課題

計算機を使ったコーパスの記録には長い活動の歴史があるが，かつては文字の扱いが，そして現在では，オープンレポジトリへの参加が，言語学者にとって資料の電子化の課題となっている．

コーパスには，聞き取り時に採られた 1 次資料の他，それを個人・プロジェクトベースでまとめ上げたコーパスがある．これらは，検討対象となるまでに整理された生資料になる．従来は，論文や書籍等での発表が主流であるため[6]，生資料は，発表時に紙媒体で適切な形へと書き換えられてきたが，現在のように，電子形式での編集が作業の全行程で一般的となり，さらには，オープンレポジトリへの参加が求められるようになると，電子上一貫したデータ構造で作業を進めることが期待されてきた [18]．

少数言語を扱う場合，言語学者は現地調査から対象言語を収集・記録する必要がある．この生のデータから記号へと書き起こす作業で使われる記録形式には，伝統的に “interliner annotation” と呼ばれる，例えば，

sumomo	**mo**	**momo**	**desu**
plum	also	peach	be
NOM	PRT	NOM	CPL

のような形式が採られている [18]．このようなデータを入力する統合ソフトウェアとして，Toolbox/Shoebox [7] がコミュニティスタンダードとして広く使用されている．Toolbox/Shoebox では，埋め込みタグ形式 (RUNOFF 系) のテキストデータを保存形式としている．また埋め込みタグ名に規定はなく，自由である．従って，データはプロジェクトを超えての共有が難しく，実用上は，個人レベルでのデータ管理が中心となっている．データ共有を目指したアノテーションの記述形式や，音声や動画等のマルチメディアとの関連づけを目的としたデータ形式は，これまでに数多くのものが提案されている[8]．これらの詳細は，別稿の報告書に譲る．

本プロジェクトでは，このような多様な記録形式が提案されている中で，現在 ISO で策定が進行しているアノテーションに関する規格とその関連規格類を参考に，今後の見通しと，現在採りうる最適解を探っている．

3. 関連規格の状況

現在，ISO では，コーパスのアノテーションに関する規格が精力的に作られている．例えば，ISO/TC37/SC4 では，図 1 にある規格の策定が進められている．

型番	内容	確定年
ISO24610-1	FSR(Feature Structure Representation)	2006
ISO24610-2	FSD(Feature System Declaration)	
ISO24611	MAF(Morpho-Syntactic Annotation Framework)	
ISO24612	LAF(Linguistic Annotation Framework)	
ISO24613	LMF(Lexical Markup Framework)	2008
ISO24614-1	WordSeg(Word segmentation)Basic	
ISO24614-2	WordSeg(Word segmentation)CJK	
ISO24615	SynAF(Syntactic Annotation Framework)	
ISO24616	MLIF(Multilingual Information Framework)	
ISO24617-1	SemAF(Semantic Annotation Framework)Time and events	
ISO24617-2	SemAF(Semantic Annotation Framework)Dialogue	

図 1 ISO/TC37/SC4

この他にも，確定している関連規格として，ISO10646 (Unicode)，ISO639-3，ISO15924 等がある．ISO 以外の関連規格類としては，TEI, CES/XCES などを挙げることができる．これらのうち，本プロジェクトでは，後述する本プロジェクトのデザインと関連する規格として，ISO 24610-1,2，ISO 24612, TEI P5[4], XCES[25] を検討対象とした．尚，その他の規格類については，本プロ

[5] 用語として，Language Documentation の他に，Linguistic Documentation を使う場合があるが，その違いがよく分からない．言語素性をマークアップすることを Linguistic Annotation と定義する場合もあるようだが，これも狭義の用法と思われる．

[6] 実は，今でもこれが評価対象の主流であるという問題が，言語学ではあるようで，本論で扱う様な，紙を超えたデータ編集・アーカイブを試みようとすると，それは電子形式の成果物となり，公開はデータベースまたはレポジトリへの登録となるが，それ自体は学術成果として評価されず，その前に論文・書籍など紙ベースでの発表が求められ，場合によっては，この順番を逆にすると，論文そのものが査読を通らない場合があるとのことである．データベース作成への評価が低い，または，データベースで使われる技術や利便性だけが評価の対象となり，データそのもの，またはデータ構造のデザインは評価対象にならないという，人文情報学が抱える課題は，人文学でも広く見られるのかもしれない．

[7] http://www.sil.org/

[8] e.g. TIMIT, Partitur, CHILDES, LACITO, LDC, Switchboard, TEI, ELAN, etc.

ジェクトでは扱わない[9].

3.1 Feature Structure

Feature Structure(素性構造, 以下 FS) とは, 元は, 言語学の統語分析で使われる (弁別) 素性の集合を指すものである. FS の記述規則に関する規定には, TEI P3(1994)[10], TEI P4(2002)[11], TEI P5(2007)[12]がある. P4 と P5 では, FS の使い方に若干の修正が施されている[13]. P5 の内容の一部は, ISO24610-1(2006) Language Resource Management – Feature Structures – Part1: Feature Structure Representation に, 規定の一部として取り込まれている[14][15].

TEI と ISO24610-1 では一貫して, FS を"general purpose data structure" としている. ISO24610-1 では, 更にその意味モデルを「素性-素性値」ペアの集合と規定している. また, このモデルの表記法として, 表, グラフ, XML の 3 つを規定し, XML による表記法には, TEIP5 にある 3 つの scheme が "standard" として導入されている (図 2). 規格では b) の手法を第一候補として推奨している. ちなみに, TEI P5 では, 更にこれに XInclude を併用する手法を推奨している (図 3). FS で使用可能なデータ型等の一覧を規定する仕組みが, TEI と ISO24610-2 Language resource management – Feature structure – Part2: Feature system declaration で用意されている. TEI では, TEI Header 部での宣言が想定さているが, ISO24610-2 の草稿では, 記述インスタンスの場所を特定するような定義は避けられているという違いがある.

```
a)<f name="x" value="y"/>
b)<f name="x">
    <value type="string">y</value>
  </f>
c)<f>
    <name>x</name>
    <value type="string">y</value>
  </f>
```

図 2

```
b')<f name="x">
    <xi:include ref="y" xpointer="z"/>
  </f>
```

図 3

FS は,「素性-素性値」ペアの集合を示す汎用の scheme として使うことが可能であることから, この応用範囲は統語・言語素性に留まらず, scheme が未定の対象の記述での使用が可能である[16].

3.2 CES/XCES, LAF/GrAF

言語に関するアノテーション記述一般に関する規格類として, 上記 FS の他に, CES(1998), その XML 対応版である XCES(2003)[25] がある. CES/XCES は, アノテーションやその対象となるデータ (primary data) の scheme の他, これらの 2 つの記述の関連付けのリンクも定義している. この意味では, TEI と同様に, 包括的な言語素性を記述する scheme を策定したものといえる[17]. 現在では, CES/XCES を開発してきた N.Ide 氏が, この発展版として LAF[18][8] を提案し, ISO24612[9] として審議中である. LAF とは, 言語学で必要なアノテーション (Linguistic annotaton) を表示する"general framework" とある. 現行の記述上からは, 前述の FS 間とで, 想定する記述対象の差が明確ではない. さらに N.Ide 氏は, LAF をモデルの定義とし, その XML 記述 (scheme) として GrAF(2007)[11] を提案している. GrAF とは, グラフ構造のデータをモデルとする XML の scheme を定義するもので, 例えば, 要素 node と edge を使い, それを表現している. LAF/GrAF には既に応用例が報告されており, 例えば [16][12] 等がある.

3.3 規格間の整理

FS を規定する ISO24610-1 と TEIP5, 並びに CES/XCES, LAF/GrAF の提案間では, 以下の点で違いを見ることができる.

(1) 記述対象の違い
(2) 用語の違い
(3) モデルの違い
(4) 記述構造 (scheme) の違い
(5) 記述方法の違い (instantiations 間の関係)

記述対象として, FS は, 名称からは統語素性を対象とすることが伺われるものの, 規格上は汎用としている. CES/XCES では対象を, 素朴にコーパスとしている. 一方, LAF/GrAF は, 言語素性を対象としている. 例えば, その応用として, GATE や UIMA 等の自然言語処理向けコーパス作成環境間で, データの共有を促す pivot data としての利用が提案されている [12].

微妙ではあるが重要な違いとして, 用語の違いがある. ISO24610-1 は, TEI P5 から XML Scheme の箇所を転用しているものの, ISO では"standoff" を使用し, TEI ではハイフン付きの"stand-off" を使用している. ちなみに, CES/XCES を扱う論文では"stand(-)off" を通称の一例として挙げることが多く, LAF/GrAF を扱う論文は"stand-off" を使うことが多い. 後述するように, この

[9] これらは少数言語を検討対象から外していることから, 当プロジェクトでは扱わない.
[10] Ch.16 Feature Structures
[11] Ch.16. P3 の内容とほぼ同じ.
[12] Ch.18
[13] 主な違いは, FS と関連するインスタンス間のリンクの扱い方にある. TEI P4 では, HyTime を起源とするリンクを提案してきたが, P5 ではこれを捨て, リンク構造に過度に依存しない scheme を採用したことが影響したものである. これは, ある意味, TEI がマークアップ言語の記述の可能性を提案する姿勢から, 現場での利用寄りに立ち位置を修正したともいえる.
[14] 具体的には, P5 Ch.18.1 から 18.10 までの内容が, ISO24610-1 の Ch.5 から Ch.5.10 までの内容となり, P5 Ch.11 の内容が ISO24610-2(草案) の Ch.8 から Ch.8.5 までの内容になっている. ISO24610-2 は審議中である.
[15] この発行時期の差は, TEI P5 の確定版の発行が伸びたことによるもの.
[16] TEI ではそのような利用を薦めている.
[17] CES での論議は TEI にフィードバックされるとの記述がある [24].
[18] Language Annotation Framework

小さな用語の違いには，規格の位置づけの違いが反映されている．

モデルの違いとして，ISO24610-1 では，「(素性) 名前-値」対の集合をモデルとして想定している．TEI と CES/XCES では，モデルの規定はない．LAF(/GrAF) では，(有向) グラフ構造のモデルを想定している．

記述構造 (scheme) の違いとして，TEI と ISO24610-1 の FS では，図 2 にある scheme を採り，LAF/GrAF では，図 4 のような記述構造を採っている[19]．この記述で

```
<anchor id="x"/>here<anchor id="y"/>
<sink id="s1" start="x" end="y"/>
<edge from="a1" to="s1"/>
<node id="a1">[FS]</node>
```

図 4

は，各行の記述がそれぞれ Primary Data, 単位化 (Segmentation/Location Data)，リンク (Link Data)，アノテーション[20]を表現している．ちなみに，TEI ではグラフ構造を記述する scheme として，要素<node><arc>を用意している (図 5) [21]．

```
<node id="n1"></node>
<node id="n2"></node>
<arc from="n1" to="n2"/>
```

図 5

記述方法の違いとして，各規格の scheme から作られる実記述 (instantiations) [22]は，同一 scheme 下でも多様となる可能性の他にも，それらの相互関係の取り方に思想の違いがある．具体的には，“stand(-)off annotation” style と呼ばれる記述方法の捉え方の違いによるものである．

4. Stand(-)off Annotation Style

“Stand(-)off annotation” style とは，アノテーションが付加されるテキスト (一次データ/primary data) 中にアノテーションを直接埋め込ま (markup) せずに分けて記述する，記述スタイルの総称である[23]．近年この用語は，アノテーションの書き方として様々な論文で見ることができる．これらを参照する際に注意すべきことは，primary data とアノテーションを分けて記述するスタイルには，多様な実現型 (instantiations) があることである．

TEI が採る standoff の書き方は，P4 と P5 とでは先述の通り，異なっているが，最新の P5 では，standoff は XInclude による実現を，例えば図 3 のように推奨している．ここでは，アノテーション中に primary data を埋め込む記述が採られている．埋め込み処理以前の記述の状態を観察すれば，primary data とアノテーションの 2 つの instantiations は，リンクパス 1 で関連付けられている構造になっている．

CES/XCES では，standoff の記述として，例えば図 6 の様な記述が例として挙げられている[24]．primary data

```
0                   1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
 M y   d o g   h a s   g a s .

<struct id="s1" from="0" to="2">
  <feat name="pos" value="adj"/></struct>
<struct id="s2" from="3" to="6">
  <feat name="pos" value="nom"/></struct>
<struct id="s3" from="s2" to="s1" type="gen">
```

図 6

には，整数値によるロケーションアドレスが仮想上あるとし，それを要素 struct の属性 from と to で参照，“span(データ単位)“を形成している．また，要素 struct は，データ単位を形成する struct 同士を関連付ける記述としても使うことができる．要素 struct は単位化，リンク，アノテーションの 3 情報を同時に記述し，TEI の記述スタイルと同様，リンクパス 1 で 2 つの instantiations が関連付けられている．

LAF/GrAF では図 4 のように，primary data, 単位化，リンク，アノテーションを示す 4 つの instantiations から構成される，primary data からアノテーションまではリンクパス 3 で関連付けられている構造を採っている．

各 XML instantiations から構成される構造物が示すモデルは，standoff スタイルによる記述が採用するリンクの素性で分析することができる．

(1) リンクの起点 (reference-referent 関係)
  a) primary data 中に参照先を記述
  b) primary data 以外に参照先を記述

(2) 参照手法 (IDREF の種類)
  c) ID 参照
  d) ロケーション参照

---

[19] GrAF 自体の scheme は規格が策定中のため不詳．

[20] この用語・分類は本稿のもの．

[21] $SCHEME_{GrAF}[arc/edge] \subset SCHEME_{TEI}$.

[22] 本稿では，XML による記述一般を，“instantiations” と表現する．これは，XML Infoset に留まらず，どの部分記述をも含む実記述データを示している．後述するように，DTD 等の schema 言語を前提とする instance とは異なり，scheme を想定しない記述も含みたいことから，instance とは異なる用語を使用した．他の論文でも指示に苦労しているようで，fragments などもこれに使われている．

[23] この”stand(-)off” という用語には少し注意が必要である．アノテーションの対象とアノテーションを分けて記述するという手法は，“separate markup” とか “remote markup” など，様々な表現がこれまでにされてきたが，このアイディア自体は，HyTime(1992) にある Location Address Element に見ることができる．TEI P3(1994) にあるリンク類は，この HyTime を参考に検討されたものである．CES は，この TEI P3 を参考にしている．このような書き方を”standoff” と表現したのは 1997 年の [21] であるとされている．N.Ide 氏は 2000 年の [6] で”satndoff” を使うものの，2001 年の [7] 以降は “stand-off” を併用している．TEI P4(2002) では”stand(-)off” という用語は使われていない．TEI P5(2007) では “stand-off” が使用されているものの，それを導入した ISO 24610-1 では “standoff” に書き換えられている．ちなみに，N.Ide 氏がまとめる ISO 24612(LAF) の草案では “stand-off” が使用されている．ちなみに，GATE[5] 等に影響を与えた TIPSTER の ver.2.1 は 1996 年の発行である．わたしたちが使うとすれば “standoff” を使うべきなのかもしれない．

[24] 文献 [10] にある例を修正した．

(3) リンク構造 (リンクパスの数)

e) direct link

f) indirect link

a) は，primary data 中に，URL を値として持つ IDREF 属性を持つアノテーションをマークアップする，現在まで主流の手法である．standoff スタイルの視点からいえば，mono XML document instance を想定する記述法といえる．b) は，単位化要素，リンク，アノテーション等の，primary data 以外の記述で，IDREF が宣言されているものである．例えば，TEI P5 で採られている，アノテーション中に XInclude を使い，primary data を埋め込む手法などがこれに相当する．standoff スタイルの視点からいえば，multiple XML instantiations を想定する記述法といえる．

c) は，いわゆる通常の IDREF(S) で，その値には URL を取るものである．d) は，ID 不要の IDREF(S) によるもので，例えば，XPointer により所在が指定された，ID を持たない部分記述を参照する手法が，これに相当する．

e) は，IDREF の参照先に所望のデータがある，いわゆるリンクパスの数が 1 となるリンク関係で構成されている，リンク元とリンク先の記述の構造である．f) は，このリンクパスの数が 1 を超える記述の構造で，例えば，IDREF が参照する ID を持つ要素にも IDREF があり，その参照先である ID に，所望のデータが記述されている構造である．

この 6 つのリンク素性を使い，standoff スタイルを採る先の規格類を弁別してみると，TEI は，b,c,d,e の素性を，CES/XCES は，b,c,d,e,f の素性を，LAF/GrAF は，b,c,d,f の素性を持っている．これにより，3 つの規格類に共通するリンク素性 b が，tandoff スタイルを特徴付ける弁別要素になっていることが分かる[25]．また，リンク素性 e と f は，3 つの規格類で持ち方が異なっていることが分かる．このリンク構造に関する素性の差に対応することが，各規格で作られたデータの可換性を保証する，重要なポイントになっている．

5. Location Address Element

先のリンク素性で，リンクパスが 1 を超えるいわゆるリンク構造には，必ず Location Address Element が使われている．Location Address Element とは，従来のリンクでは IDREF 属性が担っていた所在情報を，要素が担うものである．この概念自体は，standoff スタイルが論議される以前の HyTime に見ることができる[26]．例えば，図 6 中の，属性 id に値 s1 を持つ要素 struct は，それ自体は primary data をリンクパス 1 で参照しているが，属性 id に値 s3 を持つ要素 struct からすると，属性値 s1 を持つ要素 struct は，Location Address Element になっている．すなわち，所有関係"gen" を成す記述実体"My" と"dog" を知るには，属性値 s1 を持つ要素 struct が持つ所在情報を解決する必要がある．

このような Location Address Element を扱うには，リンク先にある要素が，さらにリンク情報を担うものなのか，またはそこが最終的なアノテーション記述があるものなのかを，事前に知る必要がある．例えば，standoff スタイルのデータとして図 7 を想定してみる．ここでは，先の図 6 とは異なり，便宜上，TEI のスタイルに従い，primary data 中に要素 anchor[27]を挿入している．

```
<standoff>
 <anchor id="a1"/>Fly<anchor id="a2"/>me<anchor id="a3"/>
 to<anchor id="a4"/>the<anchor id="a5"/>Moon<anchor id="a6"/>
 <location id="l1" from="a1" to="a2"/>
 <location id="l2" from="a2" to="a3"/>
 <annotation id="an1" target="l1">
   <f name="pos">nom</f>  </annotation>
 <annotation id="an2" target="l2">
   <f name="pos">pronom</f> </annotation>
</standoff>
```

図 7

この記述は，primary data とアノテーションの間にはリンクパス 2 の構造になっている．このデータに対して，XQuery を使い，指定した primary data のアノテーションを得るには，図 8 の様な問い合わせをすることになる．

```
let $target := for $i in //text() where $i = "me" return $i
let $postAll := $target/following-sibling::node()
let $postNode := $postAll[position()=1]
let $preAll := $target/preceding-sibling::node()
let $preNode := $preAll[last()]
let $loc := for $i in //location
    where ($i/@from=$preNode/@id and $i/@to=$postNode/@id)
    return $i
let $ans := for $i in //annotation where $i/@target=$loc/@id
    return $i
return $ans
```

図 8

ここでは，記述されているリンクとは逆の方向に，リンクパスを 1 つずつ辿っている．このような問い合わせには，ID/IDREF 情報を解決した先にある要素が，リンク要素であることが前提になければ，全てのリンクパスを解決することはできない．参考として，アノテーションから primary data へと正の方向でリンクパスを辿る様子を図 9 に示す．

このようなリンクアドレスの解決に対応するために，XQuery を拡張する試み [1] もあるが，この場合でも，リンクパス数と，途中の Location Address Element の scheme は既知であることが前提となっている．

リンク先にある要素がリンク情報を担うものであるかは，scheme の定義を見ても分からない．同様に，リンクパス上の instantiations の構造は，scheme 定義からは分からない．また，リンクパス数を知る手段がない．リン

[25] c と d は参照手法の弁別に寄与していない．

[26] HyTime に従えば，Location Address Element を「所在番地要素」と訳出できるが，英語表記の方が意味をよく伝えるので，これを使う．

[27] TEI でいう milestone 要素で，ID 属性を持つ空要素．

```
let $target := for $i in //text() where $i = "nom" return $i
let $ann := $target/../../@target
let $loc := for $i in //location where $i/@id=$ann return $i
let $start := for $i in //anchor where $i/@id=$loc/@from
    return $i
let $end := for $i in //anchor where $i/@id=$loc/@to return $i
let $ans := for $i in $start/following-sibling::text()
            for $j in $end/preceding-sibling::text()
    where $i eq $j and string-length(normalize-space($i)) > 0
    return $j
return $ans
```

図 9

クパス 2 の図 7 向けに用意された XQuery 式は，例えば，リンクパス 3 の図 4 向けには使えない．このような状況下では，standoff スタイルで記述された instantiations は十分に扱うことができない．すなわち，現行規格群の元で Language Documentation を実践するには，不足している規定がある．

6. 必要なもの

standoff スタイルで記述された instantiations を扱うには，少なくとも，新たに 3 つの規格群が必要である．

(1) XML データパイプライン処理インタフェース
(2) multiple instantiations のモデル
(3) そのモデルを構成するプラン

(1) のパイプライン処理インタフェースは，XML データに限定しないコーパス向けのものは，既に GATE や UIMA がデファクトスタンダードとして使用されている．もしこれが，XML 関連規格として成る場合，XSLT などで扱われている Dynamic Context を汎用な対象として扱うことができるようになる．これにより，現在は，ロケーション解決や埋め込み処理など，伝統的に想定されてきた暗黙の処理候補以外の処理も，instantiations 上から示すことができるようになる．例えば，Location Address Element を扱う規定も，ここで明示的に示すことが可能になる[28]．

(2) の複数の instantiations から構成されるモデルは，パイプライン処理インタフェースも素性として含むことができるもので，参照モデルとしての利用を想定している．例えば，リンク情報などを伴う参照モデルがあれば，リンクパスの渡りも，処理上，再帰的に定義することが容易になる．各種処理が素性として定義されることで，単体の XML document instance を対象とするだけでなく，複数の instantiations から構成される抽象データモデルを，処理過程の時系列をも含めてモデル化することができる．

(3) のモデルプランは，パイプライン処理インタフェースを素性として含む instantiations を指定し，参照モデルとなる抽象データモデルを構成するプランを示すものである．これにより，既存の XML データを変更することなく所望するデータを再構成できる他，対象とする資源を限定することができる．

これらにより，standoff スタイルで書かれた instantiations の構造が多様であっても，モデルプランを通して，現行規格下の処理系でも対応できる XML インスタンスを得ることも可能になる．

これら multiple instantiations から成るモデルとそのモデルプランは，HyTime において Grove と Grove Plan として示された考え方を踏襲し，発展させたものである[29]．

6.1 理想のデータの持ち方

上記の 3 つの規格群が整うことで，例えば以下のことも期待している．

- アノテーション種毎に XML document instance を作成し，これを関連させる．このような複数のレイヤを自由に記述上，扱うことができる．
- location address element を変更するだけで，基本単位の再構成ができるようになる
- 最小単位は location address element 上で自由に構成でき，リンクパス数も自在であれば，記述を重ねる (新たなデータを書き起こす) だけで，構造を変化させることができる[30]．

このような環境が実現すると，いわゆる生産的アノテーション行為を支援するシステムが実現することになる．生産的アノテーション行為とは，マークアップすることにより構造が見えるという、書くこと本来の活動と同じ生産活動のことで，人文学でいう，書きながらの分析にあたる行為である [19]．コンピュータを記録に使う Documentation において，資料の観察・分析をアノテーションとして記録することが可能になれば，人文学研究にとって大きな助けとなる．

7. 対 応 策

コーパスのアノテーションに関連する規格類には，前述のように不備があり，先述した理想の規格群も，現状では用意されていない．それを見越して，いま採りうる作成データのデザインとして，例えば，リンクパス 2 の standoff スタイルによるアノテーション記述が考えられる．

standoff スタイルの scheme は規格毎に異なってはいるが，stanodff スタイルの重要性は変わらない．多レイヤに対応し，木構造を超える構造 (e.g. overlap) にも対応し，比較的フラットな構造を作ることができるなど，少数言語のコーパスには利点が多い[31]．

リンクパス 2 のモデルを採ることで，リンクパスが 1 を超える場合に location address element の扱いが規定されていないという問題が生じる．それでも，リンクパス 2 を採る理由として，アノテーション種毎に XML document instance が用意される時，ロケーションデータは共

[28] リンク素性に関する考察には，拙稿 [20] がある．

[29] これらの規格群は，マークアップ言語一般の基本定義と強く関連するものになることも期待している．

[30] もちろん，その構造は単純ではなく，必要以上に複雑になることもある．

[31] 従来の木構造ベースの scheme を基本に内容を分析するという作業の重要性は変わらないと考えるが，それに固執することもないと考えている [19]．

有される可能性が高いことが挙げられる．何故ならば，1) アノテーション自体が共有される場合が容易に想定されること[32]，2) 基本単位を定義するロケーションデータの ID は共有されやすいこと，3) 単位境界変更 (demarcation change) にこのデザインの方が対応しやすい，というメリットがあるからである[33]．

8. 本プロジェクト

2008 年 12 月から始められたもので，2009 年度は全体のデザインの策定するために，調査，元データを整理し，2010 年度に DB を作成，2011 年に公開・オープンレポジトリへの登録を計画している．対象とする少数言語は，ユカギール語，アリュートル語，イテリメン語，ユピック語，ホジュン語，テディム・チン語，シベ語を予定している．現在までに，データ単位と語彙集を決め，Toolbox からのデータ交換の準備を整え，公開するメタデータ項目を検討し，その scheme を確定した．本年度中には，オープンレポジトリに公開する英語版用コンテンツが作成される予定である．

8.1 システムデザイン

システムデザインを決めるにあたり考慮したことは，言語学者が普段整理・記録するデータがそのまま乗る DB を作り，DB 中のデータのメンテナンスも，極力手数を減らすことである．

データ入力には，記述言語学者がその利用法を積極的に学習している入力支援ソフト Toolbox を使用する．プロジェクトの立ち上げ時には，XML エディタを使った入力までの支援を求められたが，Language Documentation の現状[34]や言語学者の学習負担を考慮し，Toolbox による入力を選択した[35]．Toolbox が吐き出す，RUNOFF 系のマークアップ埋め込み型データを，XML データへと変換する．XML データの scheme は，FS とリンクパス 2 の standoff スタイルによる，オリジナルのスキームを作成する予定である[36]．アノテーション部では FS に倣う構造を採る予定である．ロケーションの指定は，LAF/GrAF や TEI の Graph にある scheme に倣う構造を採る予定である．primary data 部では，いわゆる milestone タグを使う予定である[37]．メタデータには，TEI Header を採用する予定である．言語学者が用意するメタデータには，資料別のものと，共有する話者情報の 2 種類があるが，公開するメタデータは，各資料別のものを作成する[38]．オープンレポジトリへの参加は，TEI Header scheme のデータから，OLAC ならびに SOAS が規定する scheme のメタデータへ変換し，それを登録する．ちなみに，データ本体の scheme には規定がないため，オリジナルスキームによるデータで登録する予定である．

データ公開用の DB には，Berkeley DB XML(DBD XML) を使う予定である．使用する DB には，1)index を作らずとも検索できる環境があること，2) サービスとして動かす必要がないこと，3)XQuery に対応していることを求めた．処理速度については，データ量が少ないので，考慮していない．検索インタフェースについては，現時点で未定である．音声データについては，ロケーションデータを元に部分音声を切り出すツール sclip(sound clip)[39]を使い，公開用の音声データを作成中である．

9. さいごに

本稿では，アノテーションに関する現行規格類には，standoff に関して規定が十分ではないところがあり，少なくとも 3 つの新たな規格群が必要であることを提案した．その中では，敢えて，standoff スタイル向けの scheme を新たな規格として求めることはしていない．確かに，アノテーション入力支援用の多くのソフトウェア (e.g. ELAN, GATE, etc.) では，既に独自の standoff スタイルの scheme を使用し，それが皮肉にもコーパス共有の妨げとなっている．本プロジェクトの目標が，少数言語コーパスを国際レポジトリへ登録することであるからすると，そのような scheme 規格は多くの事前検討を不要とし，コーパス作成・共有に直接貢献するのは，明らかである．本稿で 3 つの規格群を提案した背景には，本プロジェクトに加えて，マークアップ言語の記述に関する一連の研究が目指す，マークアップ言語一般の理論を求める研究がある．本稿では，現実解を提示することもさることながら[40]，本プロジェクトを基礎研究向けの記述実験と位置づけた場合の検討結果を提示した．本稿の成果は，東京外国語大学アジア・アフリカ言語文化研究所 LingDy プロジェクト[41]から受けた，贅沢で寛容な支援を元に生まれたものである．関係各位に心から御礼を申し上げたい．

[32] 例えば，POS など用語集として用意できるもの．

[33] すると，結局は location address element にある問題をそのまま負うことになる．アプリケーションを多数揃える苦労よりも，primary data やアノテーションデータ自体の保存性を採ることになる．

[34] 別稿の報告書でまとめる予定．

[35] 言語学者の中には Toolbox からの卒業を主張する方もいる．わたくしも同意見だが，今回のデザインで Toolbox を採用した背景には，入力中のコンコーダンス自動作成機能など，人文研究者の細かい要求に応えている Toolbox から離れることは，現時点では，人文学研究の効率を落とすのではないかと判断したことがある

[36] 今回のプロジェクトでは，TEI scheme は採用しない，できるだけフラットな構造となるオリジナルスキームを採用する．

[37] これについては，現在，検討を進めている．

[38] 話者情報のファイルは独立させない．1 資料 1XML データファイルになる．

[39] 本プロジェクトで作成．公開予定．

[40] この詳細は別稿の報告書で記す．

[41] LingDy プロジェクトは文部科学省特別教育研究経費を受けたもので，その正式名称は「急速に失われつつある言語多様性に関する国際研究連携体制の構築」である．本プロジェクトは，LingDy プロジェクトの活動のひとつである．

参 考 文 献

[1] W.Alink, R.Bhoedjang, A.deVries, and P.Boncz, 2006, "Efficient XQuery Support for Stand-Off Annotation", *Proc. of XIME-P 2006*

[2] P.Bański and A.Przepiórkowski, 2009, Stand-off TEI Annotation: the Case of the National Corpus of Polish, *Proc. of Linguistic Annotation Workshop*

[3] S.Bird and M.Liberman, 1999, A Formal Framework for Linguistic Annotation, *Technical Report MS-CIS-99-01*, University of Pennsylvania

[4] L.Burnard and S.Bauman eds., 2007, *The TEI Guidelines P5*, TEI

[5] H.Cunningham, 2000, Software Architecture for Language Engineering, Ph.D. thesis, University of Sheffield

[6] N.Ide, P.Bonhomme, and L.Romary, XCES: An XML-based Encoding Standard for Linguistic Corpora, *Proc. of LERC 2000*

[7] N.Ide and C.Macleod, 2001, The American National Coupos: A standardized resource for American English, *Proc. of Corups Linguistics*

[8] N.Ide, L.Romary, and E.de la Clergerie, 2004, International Standard for a Linguistic Annotation Framework, *Journal of Natural Language Engineering*

[9] N.Ide, 2006, Linguistic Annotation Framework, ISO/TC37/SC4 N311

[10] N.Ide and K.Suderman, 2006, Merging Layered Annotations, *Proc. of Merging and Layering Linguistic Information*

[11] N.Ide, and K.Suderman, 2007 GrAF: A GrAF-based Format for Linguistic Annotations, *Proc. of the Linguistic Annotation Workshop*

[12] N.Ide and K.Suderman, 2009, Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA, *Proc. of Linguistic Annotation Workshop*

[13] ISO, 2006, *ISO/DIS 24610-1 Language Resource Management – Feature Structures – Part1: Feature Structure Representation*

[14] J.Gippert, N.P.Himmelmann, and U.Mosel eds., 2006 *Essentials of Language Documentation*, Mouton de Gruyter

[15] 風間喜代三ほか,1993,『言語学』 東京大学出版会

[16] M.Kountz, U.Heid, and K.Eckart, 2008, A LAF/GrAF based Encoding Scheme for underspecified Representatioins of syntactic Annotations, *Proc. of LREC 2008*

[17] 松村一登, 2006,「文字化された言語資料の少ない言語とテクストのマークアップ」『TEI Day in Kyoto 2006 報告書』, 京都大学

[18] K.Ohya, 1999, "Introduction to new text-based data management – For those who are tired of re-writing a list of corpora on the occasion of presentation–",*Journal of Chiba University Eurasian Society* No.2, Chiba University

[19] 大矢一志, 2006,「マークアップの課題を syntax から見た分類と解決のステップ」,『TEI Day in Kyoto 2006 報告書』, 京都大学

[20] 大矢一志, 2008,「リンク構造で関連性を表現する為のリンク要素間リンクの制御」『情報処理学会研究会報告』 Vol.2008, No.100, 情報処理学会

[21] H.S.Thompson and D.McKelvie, 1997, Hyperlink semantics for standoff markup of read-only documents, *Proc. of SGML Europe '97*

[22] K.Wörner et.al., 2006, Modelling Linguistic Data Structures, *Proc. of Extreme Markup Languages*

[23] A.Witt et.al., 2009, Multilingual language resources and interoperability, *Language Resources and Evaluation* Vol.43, No.1, Springer Netherlands

[24] CES, 1996, Corpus Encoding Standard Document CES1 Ver.1.2

[25] XCES, 2003, XCES 1.04, http://www.xces.org/