

# 言語の多面性を織り込んだ言語資料のデジタルネットワーク

高橋 洋成  
筑波大学 人文社会系

古代文字資料を TEI (Text Encoding Initiative) ガイドラインにしたがってエンコードし、言語学的分析のためのアノテーションを付与する場合、ツリー構造制約を持つ XML では表現しにくいパターンが生じる。近年、TEI はその解決策としてスタンドオフ・マークアップを提唱している。本稿はスタンドオフ・マークアップの過程で必要となる要素 ID を、RDF におけるリソース URI として利用することを提案する。RDF との連携を強化することにより、RDF ツールを利用したり、データを LOD (Linked Open Data) 化するなどの幅広い応用が期待できる。

## Digital Network of Linguistic Resources including Multiple Facets

Yona Takahashi  
Faculty of Humanities and Social Sciences  
University of Tsukuba

When digitalizing the ancient documents according to the TEI (Text Encoding Initiative) Guidelines, with a lot of linguistic annotations on them, there are some difficult cases for XML format because of its tree structure. TEI now recommends stand-off markup as a solution for such problems. This article suggests to utilize element-ID, required by stand-off markup, as resource-URI for RDF. It will enable us to use RDF applications for our data, or to integrate them into the LOD (Linked Open Data).

### 1. はじめに

著者はこれまで、古代中東の楔形文字文書[1]や、古代エジプトの神官文字文書の電子アーカイブ化[2]に携わり、資料に見られる文字的特徴や言語的特徴を TEI (Text Encoding Initiative) の提唱する XML タグセットを用いてマークアップする方法を模索してきた。そもそも、これらの古代文字には標準的なコードポイントが割り振られていない、もしくは、割り振られていたとしても字形の時代差や地域差、すなわち異体字の問題がつかまとう<sup>1</sup>。そのため、現状では古代文字文献における一つ一つの文字についてメタ情報を付与し、その上でテキスト解釈を施していくという文献学・言語学両面からのアプローチが必要になる。

しかしながら、文献学で扱う要素と、言語学で扱う要素とを同時にマークアップしていく場合、数々の問題が生じる。そこで本稿は、古代文字資料をマークアップする過程で生じたいくつかの問題点と、それに対する現段階での解決策を提案する。具体的には、TEI マークアップと RDF とを連携させることで、両者にとって有益な言語

データになりうることを、具体例を示しながら論じる。

### 2. 問題の所在および解決策

#### 2.1. どのような TEI 文書を生成するか

著者の関わっている古代エジプト神官文字文書の電子アーカイブシステム<sup>2</sup>は、リレーショナルデータベースを利用している[3]。また、データベースに格納されたデータを研究者の多様な関心に応じて利用可能にすべく、TEI 文書として出力することを目指している。

では、どのような TEI 文書を構築すべきだろうか。当アーカイブの画期的特徴は画像アノテーション付与方式であること、すなわち、資料画像に対して一つ一つの文字の範囲を多角形の座標データとして切り出し、その座標範囲と詳細情報とをリンクさせる方式だということである。

TEI マークアップで画像に対するアノテーションを表現するには、P5 版 (2007 年) で提唱された「ドキュメントベース」のエンコードが解決策となる。それまでの TEI は資料に内在する論理構造を抽出する「テキストベース」が主であった。もちろん、「テキスト解釈」は人文学研究における主要目的の一つであり、テキストを作成すること自体が人文学研究の成果である。しかしながら、大量の資料が存在するとき、その中にどのような論理構造を見出し、どのような特徴を機械

<sup>1</sup> Unicode 5.0 に約 1,000 字のメソポタミア型の楔形文字が収録されたが、これはシュメルのウル第 3 王朝時代の字形をもとに、より古い段階を考慮した字源的選択がなされている。そのため、後の時代に 1 文字に融合したものが、Unicode では別々の文字として収録されていることが少なくない。

<sup>2</sup> <https://hdb.jinsha.tsukuba.ac.jp/>

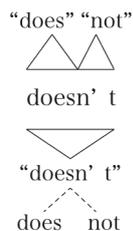


図1 文字から「語」への解釈の違い

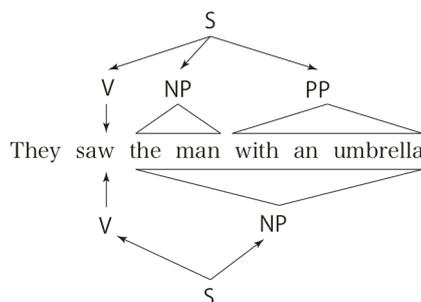


図2 構文木の解釈の違い

可読にすべきかを検討することは、現実的な時間と予算との戦いになってしまう。

そこで TEI が提唱したのが「ドキュメントベース」、すなわち、資料の「見た目」をそのままデジタル的に再現することであった。たとえば、資料に存在する編集跡や行間・余白への書き込みが「何であるか」の解釈は後回しにしつつ、編集跡や書き込みが「存在する」ことをデジタル的に共有するということである。そのための具体的な手段として、TEI ガイドラインには第 11 章「一次資料の表現」が追加された<sup>1</sup>。そこで提案されている方法は、画像化された資料に対して zone 要素で座標範囲を指定し、その範囲にある「もの」をテキストとしてエンコードするというものである。

この方法はまさに、当アーカイブで行っている画像アノテーション付与方式と同じである。したがって、現在データベースに格納されている各神官文字の座標データを、TEI の zone 要素として出力すれば良い。

## 2.2. 多様な解釈を許容する方法とは

当アーカイブは文字のみならず、言語解釈のデータ化をも目指している。画像アノテーションの示す文字の列から、言語学的な「語」の範囲を識別し、「語」に関する情報（品詞、形態分析、意味）を記述する。また、語と語は上位結節である句や節を構成し、文を形作っていく。

ここで、2つの問題がある。第1に、文字あるいは文献を構成する要素（行、頁など）と、言語を構成する要素（語、文など）は、しばしば互いの境界をまたぐということである。たとえば、言語要素としての「文」が、文献に記述された「行」の範囲に収まることはまれであり、たいていオーバーラップする。このようなオーバーラップ構造は、範囲選択&アノテーションという発想からは自然なものであるが、ツリー構造を前提とする XML では非常に表現しにくい。

第2の問題は、文字解釈や言語解釈そのものの多様性である。たとえば、図1は英語テキストの

“doesn't”という文字列から「語」の範囲を識別する事例であるが、“does”と“n't”という文字列ごとに「語」を認め、全体として2語とするマークアップする解釈と、“doesn't”を1語と見なし、その潜在的成分として“does”と“not”を想定する解釈の違いがあることを示している。また、図2は構文木を想定するとき、前置詞句（PP）を独立ノードと見なす立場と、名詞と前置詞句をまとめて一息に名詞句（NP）を想定する立場があることを示している[4]。

こうした問題に対し、近年 TEI はスタンドオフ・マークアップ（あるいはリモート・マークアップ）を提唱している。これは、情報をインラインに埋め込むのではなく、図3に示すように別々の場所に置かれた情報を互いにリンクさせることで、異なる構造を持つデータ同士を結び付ける考え方である。

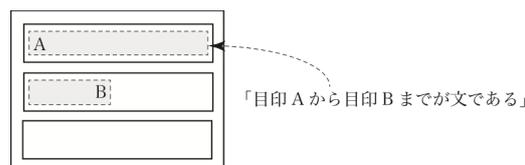


図3 スタンドオフ・マークアップの考え方

ところで、言語学資料のデジタル化に関して言えば、スタンドオフ・マークアップの考え方は様々なプロジェクトに導入されている。たとえば、ポツダム大学の Z. Amir らが開発している PAULA (Potsdamer Austauschformat Linguistischer Annotationen) も言語資料をエンコードするための XML タグセットであるが、言語分析をいくつかの層に切り分け、それぞれをファイル化するというモデルを採用している[5]。大まかには、(a) 文字列としてのデータを格納するコーパス・ファイル、(b) 文字列データを「語」の集合になるよう切り分けるトークン化ファイル、(c) 「語」の集合から構文木を構築するアノテーション・ファイルなどに分けられる。これらのファイルは XPointer を利用した内部リンクによって、互いに内部要素を参照し合っている。こ

<sup>1</sup> <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html>

のように層を分割することで、たとえば「語」の識別範囲が異なっていたり、異なる構文木を想定する場合であっても、その層を扱うファイルだけを差し替えれば済むようになる。

スタンドオフ・マークアップを徹底している PAULA ではあるが、神官文字文書や楔形文字文書のように文字情報を重視するマークアップを目指すにはタグセットが不足しており、当システムへの採用は見送った。しかしながら、言語分析の層を分割する PAULA の考え方は、大いに参考になるものである。

### 2.3. 露出する ID と URI 設計について

TEI 文書でスタンドオフ・マークアップを実現する場合、リンクの終端となりうる要素に `xml:id` 属性で ID を付与する必要がある。その結果、要素 ID は TEI 文書の URI の一部として `example.xml#ID` のように露出することになる<sup>1</sup>。それゆえ、ID の付与は TEI 文書の内部設計のみならず、URI 設計にも関わる問題である。

このように URI 設計を念頭に置くのであれば、ID 付きの URI を RDF 用のリソース URI として転用可能にすれば良い。また、RDF 自体は有向グラフモデルであり、XML のツリー構造制約から自由である。資料に対するアノテーション記述は RDF ベースで行い、それを TEI 文書に組み込めるようにすれば、スタンドオフ・マークアップの負担が軽減されるように思われる。

さらに、作成した RDF 文書は Linked Open Data (LOD) にも利用することができる。近年、ポツダム大学の Ch. Chiarcos らが中心となって Linguistic Linked Open Data (LLOD) 構想が立ち上がっている [6]。言語資料のデジタル化自体は世界中で行われているが、それらを積極的に共有・活用していくには、このような LLOD 構想は大事な試みである。当アーカイブのデータも、LLOD と連携していける形を目指したい。

こうした発想に基づき、次節では「ドキュメント指向」な TEI 文書におけるスタンドオフ・マークアップと、RDF によるアノテーション記述とを比較し、どのような連携の仕方があるかを検討する。

## 3. TEI と RDF のインタラクション

本節では、神官文字文書 BM10221 から具体的なマークアップ例と、それに対する RDF 表現を挙げていく。なお、RDF の記述には Turtle 構文を用いる。また、あらかじめ次のような URI 接頭辞を宣言しているものとする。

```
@prefix rdf: <http://www.w3.org/1999/02-22-rdf-syntax-ns#> .
```

<sup>1</sup> <http://www.w3.org/TR/xptr-framework/>

```
@prefix dct: <http://purl.org/dc/terms/> .
```

```
@prefix BM10221-3: <https://hdb.jinsha.tsukuba.ac.jp/gallery/bm10221/3#> .
```

```
@prefix tei: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-> .
```

この中で `tei:` 接頭辞はややトリッキーである。これは TEI ガイドラインの HTML 版が拡張子なしでも取得可能であることを利用したもので、たとえば `tei:TEI` は `<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-TEI>` に展開される。これを TEI 要素を示す URI として利用する。また、文中で `#ID` のように書いた場合、その ID を持つ要素を指す。

### 3.1. 異なる構文解釈を許容する方法

TEI のタグセットを用いた「語」と「句」のマークアップの典型例は、次のようなインライン・マークアップである。

```
<s>
  <phr xml:id="phr-1">
    <w xml:id="w-1" lemma="pA">...</w>
    <w xml:id="w-2" lemma="mr">...</w>
  </phr>
</s>
```

しかし、この方法では異なる句構造の解釈が生じたときに対応しにくい。そこでスタンドオフ・マークアップの出番となる。

```
<s>
  <phr xml:id="phr-1">
    <span from="#w-1" to="#w-2"></span>
  </phr>
  <w xml:id="w-1" lemma="pA">...</w>
  <w xml:id="w-2" lemma="mr">...</w>
</s>
```

`phr` 要素を置く場所は、スキーマに反しない限りはどこでも良い。重要なのは、`phr` 要素が `w` 要素の ID を参照することで、XML のツリー構造制約に関わらず、仮想的に語群を取り込んでいるということである。この方法であれば、句の範囲として別の可能性を挙げることもできる。次の例では、語 `#w-1` から `#w-2` までを 1 つの句 `#phr-1a` とする解釈と、`#w-1` から `#w-12` までを 1 つの句 `#phr-1b` とする解釈があり、それぞれの蓋然性が 50 パーセントずつであることを表している。

```
<s>
  <phr xml:id="phr-1a" exclude="#phr-1b">
    <span from="#w-1" to="#w-2"></span>
  </phr>
  <phr xml:id="phr-1b" exclude="#phr-1a">
```

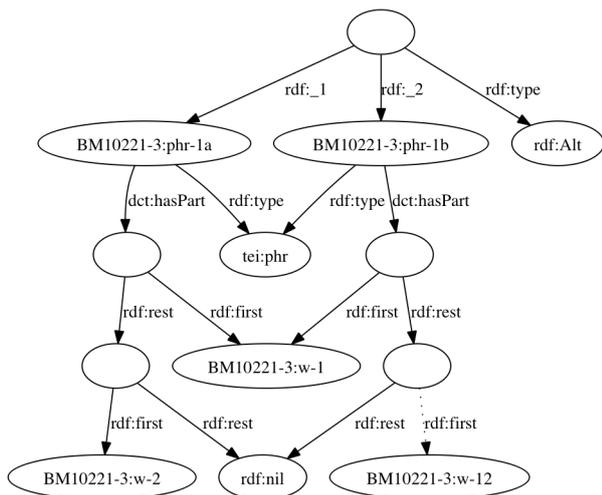


図 4 異なる句構造の可能性の列挙

```
<span from="#w-1" to="#w-12"></span>
</phr>
<alt target="phr-1a phr-1b" mode="excl"
  weights="0.5 0.5"/>

<w xml:id="w-1" lemma="pA">...</w>
<w xml:id="w-2" lemma="mr">...</w>
</s>
```

このようにデータ構造を柔軟に表現できるスタンドオフ・マークアップであるが、その反面、データ編集の見通しは非常に悪く、編集支援ツールの充実が強く望まれる。そこで、データを RDF で記述し、RDF ツールを利用してグラフ描画すれば、データ構造を視覚的に確認できるようになる。上記例のような選択枝を RDF で表すには `rdf:Alt` を使うことができる。また、「含んでいる」ことを表すのに、ここでは Dublin Core の `dct:hasPart` を用いる。

# 図 4 を参照

```
[
  rdf:_1 BM10221-3:phr-1a ;
  rdf:_2 BM10221-3:phr-1b
] a rdf:Alt ; .

BM10221-3:phr-1a
a tei:phr ;
dct:hasPart (
  BM10221-3:w-1 BM10221-3:w-2) .

BM10221-3:phr-1b
a tei:phr ;
dct:hasPart (
  BM10221-3:w-1 BM10221-3:w-12) .
```

この RDF 記述からは図 4 のようなグラフが描かれ、データ構造を視覚的に把握するという用途には十分と言えよう。

### 3.2. 「語」に関する情報の共有化

TEI で言語情報を扱う方法、とりわけ「語」の情報を扱うためのタグセットやデータ型は非常に限られている。このことは人類の言語の多様性を考えればやむをえないことであり、必要ならば British National Corpus (BNC) のように<sup>1</sup>、個別言語研究の側で TEI タグセットを拡張するという方針も妥当である。

だがそれでも、ある言語現象をコーパス横断的に検索したいという場合、言語記述のための標準的なデジタル語彙のある方が望ましい。近年、ワシントン大学の S. Farrar らが、危機言語記述のための言語学用語オントロジー GOLD (General Ontology for Linguistic Description) を開発しており<sup>2</sup>、これを TEI マークアップに組み込むことを検討する。

結論から言えば、GOLD 語彙の組み込みは非常に容易である。TEI で語を表す `w` 要素や形態素を表す `m` 要素は、分析情報を格納するための `ana` 属性を持つ。そして、`ana` 属性の取りうる値は URI のリストである。つまり、GOLD の語彙 URI を `ana` 属性に並べれば事足りるのである。

```
<s>
  <w xml:id="w-1" lemma="pA" ana="
http://purl.org/linguistics/gold/Definit
eArticle
">
  <span from="#p-3-1-1"
    to="#p-3-1-2">pA</span>
  <m ana="
http://purl.org/linguistics/gold/Masculi
neGender
http://purl.org/linguistics/gold/Singula
rNumber
">pA</m>
</w>
  <w xml:id="w-2" lemma="mr" ana="
http://purl.org/linguistics/gold/CommonN
oun
">
  <span from="#p-3-1-1"
    to="#p-3-1-2">mr</span>
  <m ana="
http://purl.org/linguistics/gold/Masculi
neGender
http://purl.org/linguistics/gold/Singula
rNumber
">mr</m>
```

<sup>1</sup> <http://www.natcorp.ox.ac.uk/docs/URG/>

<sup>2</sup> <http://www.linguistics-ontology.org/>

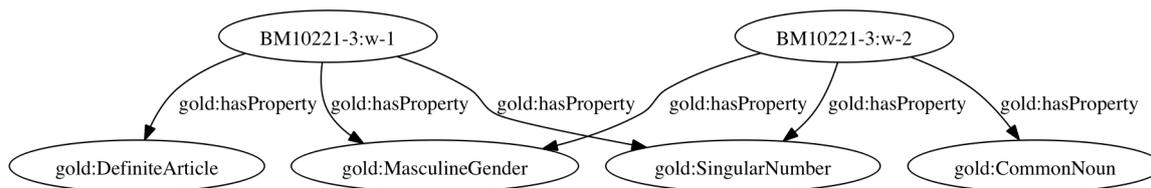


図 5 文法に関する共通語彙

```
</w>
</s>
```

RDF としての記述も容易であり、図 5 のようなグラフが描かれる。なお、gold: 接頭辞の展開形については前述の TEI マークアップにおける URI を確認されたい。

# 図 5 を参照

```
BM10221-3:w-1
gold:hasProperty gold:DefiniteArticle ;
gold:hasProperty gold:MasculineGender ;
gold:hasProperty gold:SingularNumber .
```

```
BM10221-3:w-2
gold:hasProperty gold:CommonNoun ;
gold:hasProperty gold:MasculineGender ;
gold:hasProperty gold:SingularNumber .
```

語 #w-1 の品詞は定冠詞 (gold:DefiniteArticle)、文法的性は男性 (gold:MasculineGender)、文法的数は単数 (gold:SingularNumber) となっている。語 #w-2 についても同様に、品詞は名詞 (gold:CommonNoun)、文法的性・数は男性・単数であることが記述されている。

ここで、ana 属性に GOLD 語彙を羅列したり、RDF 述語が全て gold:hasProperty であるため、語彙の意味が不明瞭になっていないかという疑問が生じるかもしれない。しかし、たとえば gold:MasculineGender という語彙は gold:NumberProperty クラスの下位クラスとして定義されているため、それが「文法性」を表すものであることはオントロジーによって保証されている。同様に、gold:CommonNoun が gold:PartOfSpeechProperty すなわち品詞の下位クラスであることも保証されている。

TEI タグセットにおける ana 属性と、URI によって表される GOLD 語彙は、非常に相性が良いと言えよう。

### 3.2. 「ドキュメントベース」マークアップと RDF

2.1 節で述べたように、神官文字文書アーカイブでは資料画像に対する文字ごとの座標が格納されている。それらのデータを TEI 文書として構築すると次のような形になる。

```
<surface>
<graphic url="BM10221-3.jpg"/>
```

```
<line xml:id="line-1">
<zone points="
3687.075,1156.45
3671.15,1167.575 ...">
<c xml:id="c-1-1"
corresp="characters.xml#c-221"
rend="color:red"
type="Phonetic">pA</c>
</zone>
<zone points="
3649.05,1199.525
3633.4,1214.875 ...">
<c xml:id="c-1-2"
corresp="characters.xml#c-192C"
rend="color:red"
type="Phonetic">A</c>
</zone>
...
<zone ...>
<c xml:id="c-1-51"
corresp="characters.xml#c-188B"
rend="color:black"
type="Determinative"></c>
</zone>
</line>
</surface>
```

この例では、文字 #c-1-1, #c-1-2, ....., #c-1-51 までが第 1 行 #line-1 である。各文字は色付けされており (rend 属性)、字典リソース (characters.xml) の対応する解説にリンクするとともに、親である zone 要素によって資料画像 (BM10221-3.jpg) の座標範囲に関連づけられている。次の RDF 記述では、標準語彙のないプロパティ、すなわち文字色を表す :color、字典参照を表す :sign を独自プロパティとして扱っている。

```
BM10221-3:line-1
a tei:line ;
dct:hasPart (BM10221-3:c-1-1
BM10221-3:c-1-2
... BM10221-3:c-1-51) .
```

```
BM10221-3:c-1-1 # 文字 1
a tei:c ;
:color "red" .
```

```

tei:zone
  "3687.075,1156.45
  3671.15,1167.575 ..." ;
:sign <characters.xml#c-221> ;

BM10221-3:c-1-2 # 文字 2
a tei:c ;
:color "red" .
tei:zone
  "3649.05,1199.525 3633.4,1214.875 ..." ;
:sign <characters.xml#c-188B> ;

```

ところで、上記の文字情報は文献に記載された言わば可能態としての文字であるが、文字は「読み」ないし「意味」が選択されることで実現態となり (c 要素の type 属性)、言語的な「語」を読者に想起させるものとなる。そのことを示すのが次に挙げる RDF 記述である。

```

BM10221-3:c-1-1
  :sign_function :Phonetic ;
  :sign_phone "pA" .
BM10221-3:c-1-2
  :sign_function :Phonetic ;
  :sign_phone "A" .
...
BM10221-3:c-1-6
  :sign_function :Determinative ;
  :sign_note "pyramid".
BM10221-3:c-1-7
  :sign_function :Determinative ;
  :sign_note "house".

```

標準化されていない独自プロパティとして、文字の実現態としての機能を表す `:sign_function` と、選択された「読み」を表す `:sign_phone`、そして「意味」を示す `:sign_note` を使っている。

文字を記述するための語彙も将来的に標準化していかなければならないが、その第一歩として考えるべきことは、文字の可能態と実現態とを区別する必要性である。文字の可能態は字典作成などに必要な側面であり、他方、文字の実現態は言語機能に関わる部分である。それゆえ、たとえば文字の「読み」の候補が複数あるような場合には、実現態の部分だけを入れ替えられるような仕組みにするのが望ましいであろう。

#### 4. おわりに

本稿は、古代文字資料を TEI に準拠してエンコードし、言語学的なアノテーションを付与していく際に生じうるいくつかの問題を取り上げた。そして、XML では扱いづらい部分に RDF を利用することで、作業の効率性を高めるとともに、作成されるデータの見通しが良くなる可能性について論じた。

人文学研究者の理想のツールの 1 つは、「オーバーラップを気にせず、資料に対して自由にアノテーションを付与することのできるもの」であろう。とりわけ、文献学や言語学の資料では、様々な階層の要素が重なり合っていることが普通である。近年における TEI の「スタンドオフ・マークアップ」や「ドキュメントベース」の考え方は、資料に対する自由なアノテーション付与の方法が求められていることを反映しているように思われる。そうであるならば、「もの」に対するメタ情報を付与するための汎用的な枠組みである RDF、および RDF ツールを活用し、相互に連携していくことが今後の人文学研究にとって大いに有益であろうと思われる。

#### 謝辞

本研究は科学研究費「基盤研究(C):高細度画像と XML データを用いた古代エジプト語文書の言語記述アーカイブズの構築」代表:永井正勝(課題番号:24520452)、および「基盤研究(C):アノテーション付与型画像データベースシステムのための汎用プラットフォーム構築」代表:和氣愛仁(課題番号:25330395)の助成によるものである。関係各位に謹んで感謝の意を表す。

#### 参考文献

- 1) 高橋洋成:アマルナ文書の電子化—文字研究・言語研究を目指して—, 情報処理学会研究報告, 人文科学とコンピュータ研究会報告 Vol.2013-CH-99, No.6, pp.1-7 (2013) .
- 2) 永井正勝・和氣愛仁:古代エジプト神官文字写本を対象とした言語情報表示システムの試作, 人文科学とコンピュータシンポジウム論文集, Vol.2012, pp.225-230 (2012).
- 3) 和氣愛仁:RDB と CMS を用いたアノテーション付与型画像データベースシステムの構築—データ構造とインターフェイスの標準化を目指して—, 情報処理学会研究報告, 人文科学とコンピュータ研究会報告, Vol.2013-CH-99, No.7, pp.1-8 (2013).
- 4) Bański P.: “Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless.” in Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies, Vol. 5 (2010). 入手先 (<http://www.balisage.net/Proceedings/vol5/html/Banski01/BalisageVol5-Banski01.html>) (参照 2013-10-20).
- 5) Amir Z., Florian Z. and Arne N.: PAULA XML: Interchange Format for Linguistic Annotations. 入手先 (<http://www.sfb632.uni-potsdam.de/en/paula.html>) (参照 2013-10-20).
- 6) Chiaros C., Nordhoff S. and Hellmann S.: Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata, Springer (2012).