『日本語話し言葉コーパス』節単位 XML 文書について[1]

Version 1.0

山田 篤† 高梨 克也‡

†京都高度技術研究所 ‡情報通信研究機構

1. 本文書の内容

『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese; 以下 CSJ と呼ぶ)（overview.pdf）には、話者情報、転記情報、形態論情報、分節音情報、韻律情報、節単位情報、係り受け情報、重要文情報などの様々な研究用の情報が収められ、さらにこれらの情報を相互に関係づけて XML 形式の文書としたもの（以下ベース XML と呼ぶ）が公開されている。本文書では、CSJ に収められているベース XML 文書からの派生物である節単位 XML について述べる。

2. 節単位 XML

2.1 節単位 XML とは

CSJ に収められているベース XML（xml.pdf）は 200[ms]以上のポーズで区切られた単位である転記基本単位を主な構成要素として、その下に階層的に長単位、短単位といった情報を格納している。転記基本単位は発話中のポーズに基づいて認定された単位であるので、その区切りは節単位、文節といった言語情報を担う単位とは必ずしも一致しない。この結果、ベース XML において、節単位、文節といった情報は XML の要素としては陽に現れず、転記基本単位の子孫である短単位の属性として格納されている。

一方で、研究目的によっては、転記基本単位ではなく、節単位（clause.pdf）や文節（bunsetsu.pdf）を独立した単位として取り出したいという要求がありうる。たとえば、重要文抽出は節単位、係り受け情報は文節を単位として付与されているので、これらの情報を取り出す際には節単位や文節が構成要素となっているほうが都合がよい。そこで、ベース XML 文書からの派生物として、節単位や文節を構成要素として持つ XML 文書である節単位 XML を作成した。係り受け情報や重要文情報を閲覧するための節単位 XML ビューワー（cu_viewer.pdf）はこの節単位 XML 文書を読み込んで表示を行う。

2.2 節単位 XML の構成

節単位 XML の構造を図 1 に示す。

ルート要素はベース XML と同じく Talk (講演) である。Talk はベース XML と同じ属性を持つ。

---

[1] 本文書中で参照するマニュアルは、cu_viewer.pdf 以外のものはすべて volume 1 の/DOC/に、また cu_viewer.pdf は本ディスクの/DOC/に、それぞれ収められている。

Talk の子要素として ClauseUnit（節単位）を置く。ClauseUnit は節単位 ID、談話に関する情報（discourse.pdf）、重要文選択に関する情報（summarydata.pdf）を属性として持つ。これらの情報はベース XML では当該節単位を構成する先頭の短単位の属性として格納されていた。さらに ClauseUnit に属性として、IPUID、IPUStartTime、IPUEndTime を持たせる。IPUID は当該節単位を構成する短単位が含まれる転記基本単位の ID のリスト、IPUStartTime はそのうち先頭の転記基本単位の開始時刻、IPUEndTime は末尾の転記基本単位の終了時刻である。

ClauseUnit の子要素として Bunsetsu（文節）を置く。Bunsetsu は文節単位 ID と係り受けに関する情報（dependency.pdf）を属性として持つ。これらの情報はベース XML では当該文節を構成する先頭の短単位の属性として格納されていた。

Bunsetsu の子要素として LUW（長単位）を置く。これはベース XML においては IPU（転記基本単位）の子要素として格納されていたものと同じである。

LUW の子要素として SUW（短単位）または Noise（雑音）を置く。これらはベース XML において LUW の子要素として格納されていたものと同じである。ベース XML において SUW 以下の子要素として格納されていた情報は、節単位 XML では省略する。ただし、分節音情報をもつ場合にはそこから各 SUW の開始、終了時刻を取り出し、PhoneStartTime、PhoneEndTime 属性として格納する。また、当該短単位が転記基本単位の末尾の要素である場合に IPUBoundary 属性を付与する。

なお、ここでは便宜上、文節、長単位と記述しているが、これらの単位の認定には転記基本単位の区切りが影響を与えていることに注意されたい。すなわち、本来一つの単位として認定したい文節や長単位の内部に転記基本単位の切れ目があった場合、表現上はそれぞれ二つの文節、長単位になっている。このとき、文節については分断された前半の文節の Dep_ObligateComment 属性として B+という値が、長単位の場合は後半の長単位の IsLeftOver 属性がそれぞれ設定されている。

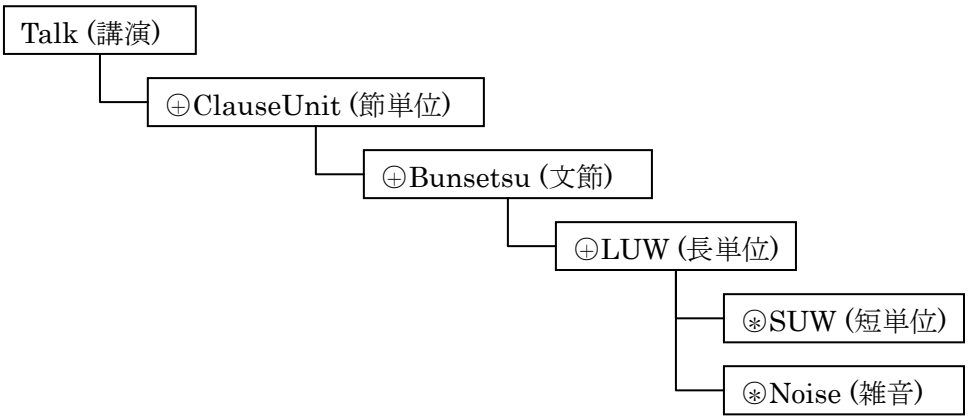

図 1: 節単位 XML の構造

節単位 XML の例を図 2 に示す。ルート要素 Talk の属性はベース XML と同じである。その子要素として ClauseUnit があり、節単位 ID が 1 の節単位は、重要文選択で Subject2 が 50 パーセントの要約率で、また Subject3 も 50 パーセントの要約率で選択したことが示されている。さらにこの節単位はベース XML 中、ID が 0006 から 0010 までの転記基本単位に対応し、その開始時刻は 00005.592、終了時刻は 00015.595 であることが記されている。

次に文節 ID が 3 の文節はこの節単位の構成要素で、文節 ID が 12 の文節に係っていくことを示している。

```
<Talk TalkID="A01F0055" SpeakerID="459" SpeakerBirthPlace="東京都"
SpeakerBirthGeneration="60to64" SpeakerSex="女">
                                    ～途中省略～
<ClauseUnit ClauseUnitID="1" SE_Subject2_50p="1" SE_Subject3_50p="1"
IPUID="0006 0007 0008 0009 0010" IPUStartTime="00005.592"
IPUEndTime="00015.595">
                                    ～途中省略～
 <Bunsetsu Dep_BunsetsuUnitID="3" Dep_ModifieeBunsetsuUnitID="12">
  <LUW LUWID="2" LUWPOS="代名詞" IsNewLine="1" LineID="002"
LUWDictionaryForm="ワタクシドモ" LUWLemma="私共">
   <SUW SUWID="1" ColumnID="001" SUWDictionaryForm="ワタクシ"
SUWLemma="私" SUWPOS="代名詞" OrthographicTranscription="私"
PhoneticTranscription="ワタクシ" PlainOrthographicTranscription="私"
APID="5" Dep_BunsetsuUnitID="3" Dep_ModifieeBunsetsuUnitID="12"
PhoneStartTime="7.207804" PhoneEndTime="7.675693" />
                                    ～以下省略～
```

図 2: 節単位 XML の例

2.3. 節単位 XML の生成

ベース XML からの節単位 XML の派生は、XML の変換処理によって実現可能であり、そのような仕組みとして XSLT（XSL Transformations）がある。実際にベース XML から節単位 XML を派生する際に用いた XSLT スタイルシート mkCUXML4base.xsl も同梱している。

3. おわりに

本稿では、CSJ のベース XML からの派生物である節単位 XML について述べた。節単位 XML では、ベース XML に格納されている研究用の情報の一部を、特定の目的に対して扱いやすい形で取り出している。ベース XML からの目的に応じた情報の取り出し方の一例として参考にしていただければ幸いである。