CSJ 節単位 XML ビューワーの使用法

ver.1.0

高梨 克也† 山田 篤‡ 内元 清貴† 野畑 周† 竹内 和広†
†情報通信研究機構 ‡京都高度技術研究所

０．はじめに

・CSJ 節単位 XML ビューワーは『日本語話し言葉コーパス CSJ』に付与されている節単位，談話，重要文，係り受けの情報を表示し，各節単位の音声の再生を可能にしたものです.
・表示できるファイルはコア（対話・再朗読を除く）とテストセットの 199 講演です．節単位，重要文選択，係り受けの情報はこれらの 199 講演のすべてに付与されていますが，談話の情報が付与されているのはそのうちの 40 講演です．詳細は第 4 節及び CUList.txt を参照してください.
・これらの 199 講演について，ベース XML から CSJ 節単位 XML ビューワーでの表示用に変換した節単位 XML をこのディスクの /CU/ に含めています．また，ベース XML から節単位 XML への変換に使用したスタイルシート mkCUXML4base.xsl も/TOOL/にあります.
・以下で参照するマニュアル類について，cuxml.pdf と CUList.txt は本ディスクの/DOC/に，それ以外は volume 1 の /DOC/に収められています.

１．スタイルシート

※ブラウザでの表示目的の場合には，本節でのスタイルシートの説明は飛ばしていただいて構いません.

1.1 構造変換用スタイルシート mkCUXML4base.xsl

・用途：ベース XML（*.xml）を節単位 XML（*_CU.xml）に変換します.
・ベース XML と節単位 XML の仕様については，それぞれ xml.pdf と cuxml.pdf を参照してください.
・変換後の節単位 XML は本ディスクの/CU/に収められているため，各ユーザーがこの作業を行う必要はありませんが，mkCUXML4base.xsl を修正することによって，この節単位 XML とは若干構造の異なる派生ファイルをベース XML から派生させることもできます．ツールの改変については第 5 節を参照してください.

1.2 節単位 XML 表示用スタイルシート disp4CU.xsl

・用途：節単位 XML を HTML（*.html）形式で表示または別ファイルとして出力します.
・パラメータ（コマンドラインから使用する場合）：
disp_SE (重要文選択表示:1, 非表示:0)
disp_DS (談話情報表示:1, 非表示:0)
disp_Dep (係り受け構造表示:1, 非表示:0)
＊デフォルトはすべて表示
audio_File (再生する音声ファイルを指定)
※Windows マシンで Internet Explorer を使用する場合以外は指定しないでください.
WMP_Version (Windows Media Player のバージョンを指定)

参考：XSLT 処理系に Xalan を用いた場合のパラメータの指定方法[1]

```
java org.apache.xalan.xslt.Process -in *_CU.xml -xsl disp4CU2.xsl
-param disp_SE 1 -param disp_DS 1 -param disp_Dep 0 -out *.html
```

この場合，重要文選択と談話情報を含み，係り受け情報を含まない *.html というファイルが出力されます.

[1] XSLT の処理系としては Xalan 以外にもさまざまなものがあるが，Java 1.4 さえインストールされていれば動作可能なので，Xalan を例に挙げた.

２．CSJ 節単位 XML ビューワー viewer.html

2.1 動作環境

・ブラウザ
Internet Explorer 6 以上（Internet Explorer 5.5 で，別途 MSXML3 をインストールしてもよい）
Netscape 7.1 以上
Mozilla 1.4 以上

・音声再生（Windows マシンで，ブラウザとして Internet Explorer を使用する場合のみ）
Windows Media Player 7, 8, 9

2.2 操作

(1) viewer.html と disp4CU.xsl が同じディレクトリに置かれていることを確認してください.

(2) viewer.html をブラウザで開きます.「CSJ 節単位 XML ビューワー」というページが開かれます.

(3) 入力 XML ファイルを選択します.
－「表示するファイルの選択」の箇所の「参照」をクリックし，表示したい講演の節単位 XML を選択します[2].

(4) 表示内容の選択
－重要文，談話，係り受けのそれぞれの情報を表示するか否かについて，選択する方のチェックボタンにチェックを入れます（デフォルトではすべて表示されます）.
－各情報の表示方法については第 3 節を参照してください.

(5) 音声データ再生の選択
－Windows マシンでブラウザとして Internet Explorer を使用する場合のみ, 節単位ごとの音声の再生が可能です.
－音声データを再生する場合は「音声の再生」のチェックボタンで「再生する」を選択し，「参照」から音声データ（*.wav）を指定してください.
－音声ファイルは volume 3～16 に収められている*.wav ファイルのうち，講演 ID が節単位 XML と一致するものを使用してください.

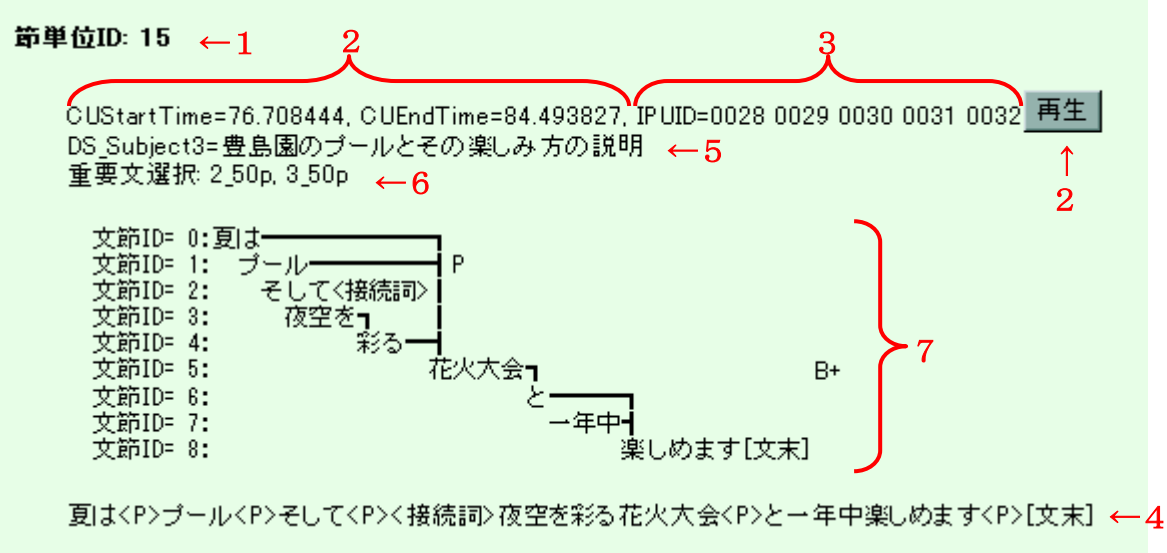
(6) 表示
－(3)～(5)の選択が済んだら，最後に「上記の設定で表示する」をクリックするとレンダリング結果が表示されます.
－(5) で音声データを「再生する」を選択した場合のみ，節単位毎に再生ボタンが表示されます.

３．表示内容の見方

・本節では 2.2 の手順に従って表示された情報の見方について説明します.
・表示されている情報がベース XML と節単位 XML でどのように格納されているかの詳細については，それぞれ xml.pdf と cuxml.pdf を参照してください．また，節単位，重要文，談話，係り受けなどの各情報の認定基準等については，それぞれ clause.pdf，summarydata.pdf，discourse.pdf，dependency.pdf を参照してください.
（cuxml.pdf 以外のマニュアルはすべて volume 1 の/DOC/にあります.）
・重要文，談話，係り受けの情報をすべて表示し，音声を再生する場合，各節単位ごとに以下のような形式で表示されます.

[2] 節単位 XML でなくベース XML を直接読み込んで表示させることも可能ですが，表示にかなりの時間がかかるため，節単位 XML の方を使用することをお勧めします.

［図１］ 節単位ごとの情報の表示



・以下，図中の１～７のそれぞれについて，第 3.1～3.7 節で順に解説します.

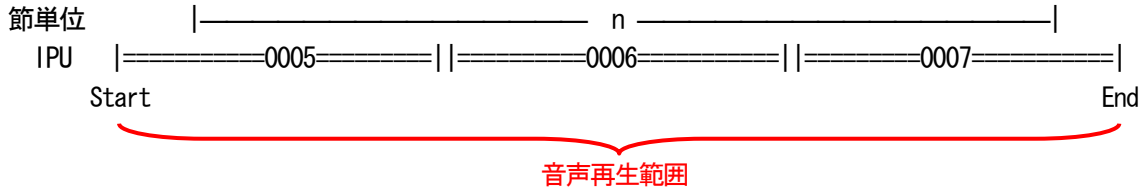
3.1 節単位 ID

・ClauseUnitID を表示しています（clause.pdf）.

3.2 時間情報と音声再生

・ベース XML 及び節単位 XML に PhoneStart/EndTime の値がある講演（CUList.txt で core=1 の 177 講演）についてはその値を，これがない講演（core=0 の 22 講演）には IPUStart/EndTime を表示しています．各値の定義については xml.pdf を参照してください.
・どちらの種類の値についても，数値は 60 進法ではなく秒を単位とした 10 進法で表示されており，また桁数合わせのための 0 は削除しています[3].
・「再生」ボタンをクリックすると，Windows Media Player が自動的に起動され，各節単位の音声が再生できます.
・ただし，転記基本単位 IPU の範囲が節単位の範囲とクロスする場合が多いため，IPUStart/EndTime を用いている場合には，音声再生の範囲が節単位の範囲とは正確に一致しない箇所があります[4]．こうした場合，音声再生の範囲は当該節単位を含むよう広めに設定してあります．なお，IPUStart/EndTime を用いている場合には，開始・終了時間の直後に * を付けています.

［図 2］ IPUStart/EndTime を用いた場合の節単位 n の音声再生の範囲



[3] ベース XML 及び節単位 XML では，IPUStart/EndTime では IPUStartTime="00001.075"のように桁数を合わせるために 0 が挿入されていますが，PhoneStart/EndTime では PhoneStartTime="3.151094"のように，桁数合わせがありません．従って，IPUStart/EndTime については，表示が XML での格納とは異なることになります.
[4] また，こうした場合，人名等の個人情報が含まれている場合には，これらの個人情報部分を含む転記基本単位全体の音声がマスキングされています.

・3.3 転記基本単位 ID

・当該の節単位に含まれる短単位 SUW を含む転記基本単位の ID を表示します.
・当該の節単位が複数の転記基本単位にわたるものである場合には，これらの転記基本単位の ID が列挙されます.
・転記基本単位境界が節単位境界と一致しない場合には，図２と同様に，IPUID は広めに取られます.

3.4 節単位情報（本文）

・節単位（clause.pdf）の言語情報を表示したものです.
・節単位は以下の 3.5～3.7 の情報付与のための基礎となる単位です.
・ベース XML 及び節単位 XML から，当該の節単位に含まれる短単位 SUW の OrthographicTranscription（transcription.pdf）を獲得し，並べて表示したものです[5].
・当該の短単位 SUW が節単位認定についての情報を持っている場合には，これらの情報を以下の順に並べて表示します.

CU_PreBracket, OrthographicTranscription, ClauseBoundaryLabel, CU_PostBracket, CU_OperationSign

・節単位の途中に転記基本単位の境界がある場合には，その位置に <P> が挿入されています[6].
・節単位認定についての情報のうち，CU_ObligateComment は，以下のように，節単位の末尾にまとめて表示されます.

［図３］ 節単位情報（本文）の表示例

それから<P><接続詞 L>(F えー)(D かつ)<P>各月齢児群における<P>((F えー)ちょっと今日は言いませんでしたけれども/並列節ケレドモ/)+{原型を<P>(F えー)有意に聞いた[文末]:変形を有意に聞いた}という<トイウ節>人数比を見ますと<P>/条件節 ト/;挿入節;引用節構造

黒：OrthographicTranscription，緑：ClauseBoundaryLabel，赤：CU_Pre/PostBracket，
青：CU_OperationSign，紫：IPUBoundary，橙：CU_ObligateComment

3.5 談話境界情報

・談話境界情報（discourse.pdf）は談話セグメントの冒頭の節単位に付与され，当該セグメントの内容と範囲を表すものです.
・談話境界認定に関する情報を DS_Purpose，DS_SubPurpose，DS_Comment，DS_Subject1，DS_Subject2，DS_Subject3 の順に，それぞれ独立の行として表示します.
・談話境界認定に関する情報が付与されているのは第４節の /ds/ （CUList.txt の set=ds）の 40 講演です．これらの 40 講演以外のものについては，2.2 の (3) で談話情報の「表示」「非表示」のどちらを選択しても表示結果は同じになります.
・また，以下の箇所では DS_Comment が表示されていませんが，このように，値が空の場合には属性名の表示も省略されます.

DS_Purpose=練馬区の町の様子
DS_SubPurpose=練馬区の町の様子
DS_Subject1=練馬区がいかにいいところかの説明
DS_Subject2=自分が住んでいる練馬区ののんびりとした町の様子
DS_Subject3=練馬区の特色の紹介

---

[5] ベース XML と同様，個人情報は×××などで伏せられています.
[6] ただし，転記基本単位境界が短単位 SUW 中にある場合には，この短単位が C タグでマークされるため（transcription.pdf），<P> は挿入されていません.

3.6 重要文選択情報

・重要文選択情報（summarydata.pdf）は 3 人の被験者が各節単位の内容の重要度を判断し，50%もしくは 10%の節単位を選択した結果を表します.
・重要文選択に関する情報を以下のように 1 行にまとめて表示します.

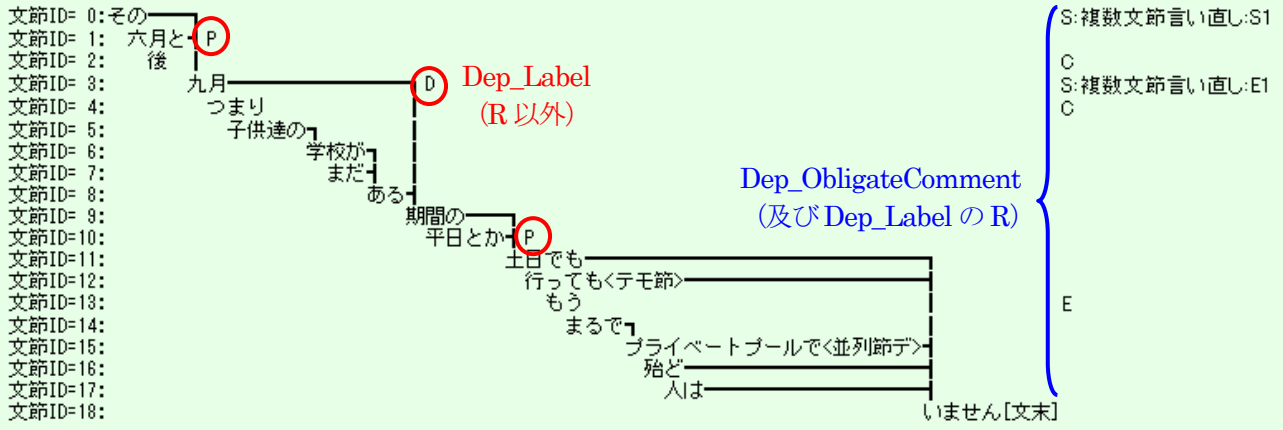
重要文選択: 1_10p, 1_50p, 2_10p, 2_50p, 3_10p, 3_50p

・6 種類の値がすべて空の場合，この行全体の表示が省略されます.

3.7 係り受け構造情報

・係り受け情報（dependency.pdf）は節単位の範囲内で文節間の係り受け関係などを記述したものです.
・文節は転記ファイル中の転記行に相当します（transcription.pdf，bunsetsu.pdf）.
・文節 ID は各節単位内で 0 からの通し番号によって与えられています.
・文節ごとに独立の行として表示されます.
・各文節の表示形式は節単位情報の表示（3.4）とほぼ同様の規則に従うものですが，転記基本単位境界を表す<P>は表示されません.
・係り受け関係は文節間のアークによって表現されます.
・Dep_Label は原則として各アークの肩の位置に表示されますが，表示の都合上，倒置を表す R ラベルだけは Dep_ObligateComment と同じ位置に表示されます.
・Dep_ObligateComment と Dep_Label の R は対応する文節の右方に表示されます.

[図 4] 係り受け情報の表示例



3.8 表示フォントについての注意

・ブラウザで MSP ゴシックのようなプロポーショナル（等幅）でないフォントが選択されていると，表示がずれる場合があります.
・Windows 環境ではフォントの大きさによって文字サイズが揃わないことがあるため，係り受け表示のフォントは 9 ポイントで固定してあります．もしこれを変更したい場合には，viewer.html の先頭にある以下の部分のフォントサイズを書き換えてください.

```
<style type="text/css">
<!--tt { font-size: 9pt }-->
</style>
```

４．volume 18 のディレクトリ構成

/DOC/
cuxml.pdf（節単位 XML 仕様）
cu_viewer.pdf（節単位 XML ビューワーマニュアル：この文書）
revision_data.pdf（文編集データ仕様）
CUList.txt（節単位 XML ファイルリスト）

/TOOL/
mkCUXML4base.xsl（ベース XML→ 節単位版 XML の変換用 XSL）
disp4CU.xsl（節単位版→表示用 HTML の変換用 XSL）
viewer.html（節単位 XML ビューワー）

/SR/（文編集データ XML：revision_data.pdf 参照）
講演 ID/
10PER/
50PER/

/CU/（ベース XML から派生した節単位 XML：CUList.txt 参照）
ds/ （コア&談話情報あり：40 講演）
A01F0145, A01M0025, A01M0056, A01M0070, A01M0137, A01M0157, A02F0116, A03F0072, A03M0005, A03M0059, A03M0138, A05F0043, A05M0031, A06F0028, A06F0073,
S00F0209, S00F0210, S00M0065, S00M0071, S00M0117, S00M0213, S01F0157, S01F0166, S01F0183, S01M0051, S01M0227, S02F0100, S02F0189, S02M0011, S02M0161, S03F0119, S03F0214, S03M0089, S03M0098, S03M0194, S04F0013, S05F1600, S05M0412, S05M0613, S06F1034
core/ （コア&談話情報なし：137 講演）
A01F0055, A01F0067, A01F0122, A01F0132, A01F0143, A01M0007, A01M0015, A01M0020, A01M0021, A01M0030, A01M0048, A01M0065, A01M0074, A01M0083, A01M0096, A01M0097, A01M0099, A01M0103, A01M0110, A01M0115, A01M0131, A01M0133, A01M0140, A01M0142, A01M0147, A02F0038, A02F0082, A02M0076, A02M0098, A02M0107, A03F0108, A03F0109, A03F0153, A03M0004, A03M0010, A03M0018, A03M0045, A03M0061, A04M0026, A04M0047, A05F0039, A05F0154, A05F0502, A05M0002, A05M0040, A05M0068, A06F0049, A06F0075, A06F0120, A06F0128, A06M0092, A07F0844, A11M0369, A11M0469, A11M0846,
S00F0014, S00F0031, S00F0041, S00F0066, S00F0082, S00F0083, S00F0088, S00F0131, S00F0134, S00F0173, S00F0177, S00F0197, S00M0025, S00M0053, S00M0075, S00M0112, S00M0115, S00M0153, S00M0199, S00M0218, S00M0221, S01F0006, S01F0038, S01F0050, S01F0074, S01F0151, S01F1522, S01M0005, S01M0091, S01M0101, S01M0182, S01M0205, S01M0225, S01M0706, S02F0012, S02F0094, S02F0113, S02F0121, S02F0129, S02F0180, S02F0852, S02M0043, S02M0068, S02M0076, S02M0092, S02M0103, S02M0191, S02M0198, S02M0245, S02M1698, S03F0062, S03F0072, S03F0108, S03F0133, S03F0184, S03F0224, S03F0232, S03F0314, S03F0383, S03F1477, S03F1577, S03M0003, S03M0046, S03M0106, S03M0141, S03M0201, S03M0317, S03M0996, S03M1133, S04F1495, S05F0463, S05F1041, S05F1517, S05M1236, S05M1505, S05M1666, S06F0167, S06F1566, S06M0373, S06M0894, S06M0895, S07M0833
TS_A/ （非コア&テストセット&人手形態素あり：11 講演）
A01F0001, A01F0034, A01F0063, A01M0141, A03M0106, A03M0112, A05M0011,
S00F0148, S00F0152, S00M0008, S00M0070
TS_B/ （非コア&テストセット&人手形態素なし：11 講演）
A02M0012, A03M0016, A03M0156, A04M0051, A04M0121, A04M0123, A06F0135, A06M0064,
S00F0019, S00M0079, S01F0105

５．著作権

・mkCUXML4base.xsl，disp4CU.xsl，viewer.html 及び関連マニュアルの著作権は Studio ARC と独立行政法人情報通信研究機構にあるものとします．詳細は下記の BSD スタイルのライセンス規定に従ってください．
・その他のデータファイルの著作権については『日本語話し言葉コーパス』に準じます．