

# 『日本語話し言葉コーパス』

## XML 文書について

Version 1.3

菊池英明\* 塚原涉 小町守\*\*\* 山田篤† 高梨克也‡ 小磯花絵\*\*

\*早稲田大学 \*\*国立国語研究所 \*\*\*奈良先端科学技術大学院大学 †京都高度技術研究所 ‡京都大学

### 目次

1. 本文書の内容 .....	1
2. XML 文書化の目的 .....	2
3. XML .....	2
3.1. XML とは .....	2
3.2. XML の基本 .....	2
4. CSJ のための XML スキーマ設計 .....	3
4.1. 基本方針 .....	3
4.2. 各要素の内容 .....	7
4.3. 階層構造からの逸脱に対する措置 .....	19
4.4. CSJ 研究用付加情報の XML による記述例 .....	20
5. XML 文書の利用方法 .....	22
5.1. XML 文書の整形 .....	22
5.2. XML 文書に対する検索 .....	24
6. おわりに .....	34
7. 第三刷における変更点 .....	34

#### 本文書において参照するマニュアルの略称と実際のタイトル(ファイル名)

- 概観マニュアル 『日本語話し言葉コーパス』の概観(overview.pdf)
- 話者情報マニュアル 記録票データ・対話記録票データ・講演者属性データ・対話参加講演者の講演一覧の概説(data\_attribute.pdf)
- 収録マニュアル 音声収録作業の概要(recording.pdf)
- 転記マニュアル 転記テキストの仕様(transcription.pdf)
- 文節マニュアル 文節の仕様について(bunsetsu.pdf)
- 形態論マニュアル 『日本語話し言葉コーパス』の形態論情報の概要(pos.pdf)
- WDB マニュアル 短単位・長単位データマニュアル(wdb.pdf)
- 分節音マニュアル 『日本語話し言葉コーパス』の分節音ラベリング(segment.pdf)
- 韻律マニュアル 『日本語話し言葉コーパス』のイントネーションラベリング(intonation.pdf)
- 節単位マニュアル 『日本語話し言葉コーパス』における節単位認定(clause.pdf)
- 係り受けマニュアル 『日本語話し言葉コーパス』における係り受け構造付与(dependency.pdf)
- 談話境界マニュアル 『日本語話し言葉コーパス』の談話境界情報について(discourse.pdf)
- 重要文マニュアル 『日本語話し言葉コーパス』における自由要約・重要文抽出データについて(summarydata.pdf)

#### 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)には、話者情報、転記情報、形態論情報、分節音情報、韻律情報、係り受け情報など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付けて XML 形式の文書としたものが公開される。本文書では CSJ 研究用付加情報の XML 文書化の目的を述べた後、“XML”の基礎を概説し、そのうえで CSJ の研究用付加情報を表現するための XML 構造の設計について説明する。さらに、CSJ に収められる XML 文書に対する検索や整形などの利用方法を紹介する。

なお CSJ 第三刷作成（2011 年）に際して、XML 文書の仕様を一部変更したので、本文書の Versioin を 1.1 から 1.2 に変更した。第二刷までの不具合(バグ)修正を含めて、第三刷における変更については 7.参照。

## 2. XML 文書化の目的

CSJ には、講演を中心とした音声データに対しての話者情報、転記情報、形態論情報、分節音情報、韻律情報、係り受け情報など様々な研究用の情報が含まれる（概観マニュアル3.参照）。同一のデータに対して付加されたこれらの情報は相互に依存関係を持つ。異なる情報間に跨った検索や将来のメンテナンスの効率を上げるために、これらの研究用付加情報をその種類毎に独立した形で提供するだけでなく、依存関係を表現した形で提供することにした。

CSJ の研究用付加情報の相互の依存関係を表現する手段として、XML と呼ばれるマークアップ言語を用いる。XML は、構造化されたデータをテキスト形式で記述することのできる柔軟性の高い記述言語であり、多様な構造を持った情報を記述するのに向き、かつデータ構造の妥当性を検証する仕組みを持つ。また、XML は階層構造の表現を容易に実現でき、基本的には階層構造で表現できる言語学的情報の記述に向いている。例えば、転記テキストと形態論情報の関係は、転記における転記基本単位の下位階層に形態論情報における語（CSJ では短単位）を位置させ、語の属性情報として転記のテキストを持つ形で表現できる。あるいは、韻律情報と形態論情報の関係は、韻律情報のうち語境界に付加する情報を形態論情報における語の属性値として持つ形で表現できる。このように異なる種類の情報間の相互の依存関係を構造化して表現することにより、例えばある語の統語的な環境によるアクセント実現位置の分布の違いを調べるなど、種類の異なる情報に跨った検索も容易になる。XML 文書を用いて容易になる検索や分析の例を以下に列挙する。なお、提供される研究用情報の種類は音声データによって異なるため一部不可能なものもあるため注意されたい（音声データと付加される研究用情報の種類の対応については概観マニュアル3.および話者情報マニュアル2.10 参照〔〕内は分析が可能な音声データの範囲）。

- 種々の特徴分析についての話者属性（出身地や性別、年齢など）ごとの分布[全体]
- 転記タグ情報と統語構造の関係[全体]
- 分節音情報の時間情報利用による語やモーラ単位の音声へのダイレクトアクセス[コア]
- 韻律情報についてアクセント句単位での解析や F0 値の分析[コア]
- 係り受け構造や談話構造と韻律の関係[コアの一部]

など。

また、XML の関連規格である XSL や XSLT を用いれば、XML 文書から任意の形式で情報を取り出したり XML の構造を変換したりすることができる。それによって例えば、転記テキスト、短単位・長単位データベース、分節音・韻律ラベルなどをそれぞれの形式で抽出して出力したり、係り受け構造や韻律アクセント句構造などを前面に出す形に XML 文書を変換することができる。こうした検索や整形などの XML 文書利用方法について、本文書の後半で紹介する。

## 3. XML

### 3.1. XML とは

XML(Extensible Markup Language)は、SGML (Standard Generalized Markup Language)を祖とした記述言語であり、WWW で用いられる技術の標準化を行う団体 W3C(World Wide Web Consortium)によって 1998 年に勧告として発表された。データを構造化して記述できる利点により、HTML に代わる次世代の WWW コンテンツ記述言語として注目されてきたが、さらにここに来て様々な関連規格が整備され、単に WWW コンテンツ記述言語としてのみならず、汎用的なデータ記述言語としてその活躍の場を大きく広げている。最近では言語コーパスの記述言語としても使われるようになってきている。

### 3.2. XML の基本

XML で記述する基本単位は、開始タグと終了タグで内容を囲んだものであり、これを要素という。例えば”<author>菊池</author>”のように記述すれば、「菊池」という内容を持つ author 要素を記したことになる。また、このようにしてタグ名を設定することで任意の要素名を自由に導入することができる。各要素には属性とその値を表現することができる。例えば先の author 要素に値が 1 である属性 id を持たせるには、”<author id=”1”>菊池</author>”と記述する。さらに、タグで囲まれる範囲に別の要素を記すことにより、木構造を表現でき親子関係を兄弟関係を表すことができる。例えば、author 要素の子供にあたる affiliation 要素を記す場合、以下のように記述できる。

```
<author id="1">菊池英明
  <affiliation id="1">早稲田大学</affiliation>
  <affiliation id="2">国立国語研究所</affiliation>
</author>
```

#### 4. CSJ のための XML スキーマ設計

##### 4.1. 基本方針

まず始めに、XML で記述する研究用付加情報を以下に列挙する。詳細についてはそれぞれのマニュアルを参照されたい。  
 (表 1 の要素欄に示した文字は各項目先頭文字に対応)

- 話者情報 (講演話者の属性の一部) <話者情報マニュアル>
- 転記情報 (転記基本単位と時間位置、基本形表記、発音形表記、各種タグ情報、コメント情報) <転記マニュアル>
- 文節情報 (転記テキストにおける改行に相当) <文節マニュアル>
- 形態論情報 (短単位、長単位、代表形、品詞、活用の種類等) <形態論マニュアル> <WDB マニュアル>
- 分節音情報 (分節音ラベルとその時間位置、無声化の有無) <分節音マニュアル>
- 韻律情報 (トーンラベル、BI ラベルとその時間位置、単語アクセント位置等) <韻律マニュアル>
- 節単位情報 (節単位) <節単位マニュアル>
- 係り受け情報 (文節間係り受け関係) <係り受けマニュアル>
- 談話境界情報 (談話境界、談話目的) <談話境界マニュアル>
- 重要文情報 (重要文抽出データ) <重要文マニュアル>

これらの情報には、言語学的情報と音声学的情報が混在しており、例えば語の境界をはさんで音学的に融合しているために分節音ラベルが語境界をまたぐような場合など、階層構造を逸脱するものもある。しかしながら多くの用途においては言語学的な階層構造を想定することが予想されるため、CSJ の研究用付加情報を XML 形式で記述する際にはあくまで階層構造を基本とすることにした。

図 1 に、CSJ の研究用付加情報を、講演を最上層とする階層構造で表現した例を示す。この図に示したように、通常は各研究用付加情報の間の依存関係は階層構造として問題なく表現できる<sup>1</sup>。したがって、階層構造を逸脱する場合については後に説明するように特別な措置をとることとし、原則としては図のような階層構造を仮定して XML のスキーマを設計することにした。定義した XML の要素の関連図を図 2 に、それぞれの要素と属性の説明を表 1 に示す。表 1 には各属性値の例も示している。このうち“0/1”としたものについて、論理的には二値をとり得る属性であるものの、省サイズ化のため実際には“0”をとる場合は属性そのものを記述しないことにしている。



図 1. CSJ の研究用付加情報間の関係例

<sup>1</sup> なお、本文書で説明する XML 文書においては、韻律ラベルはラベルを時間的に包含する分節音単位に所属させている。韻律ラベルの種類によっては論理的な対応関係がそうした時間的な対応関係と異なる可能性があり、将来的に音声研究用派生 XML においてそれを表現することになっている。

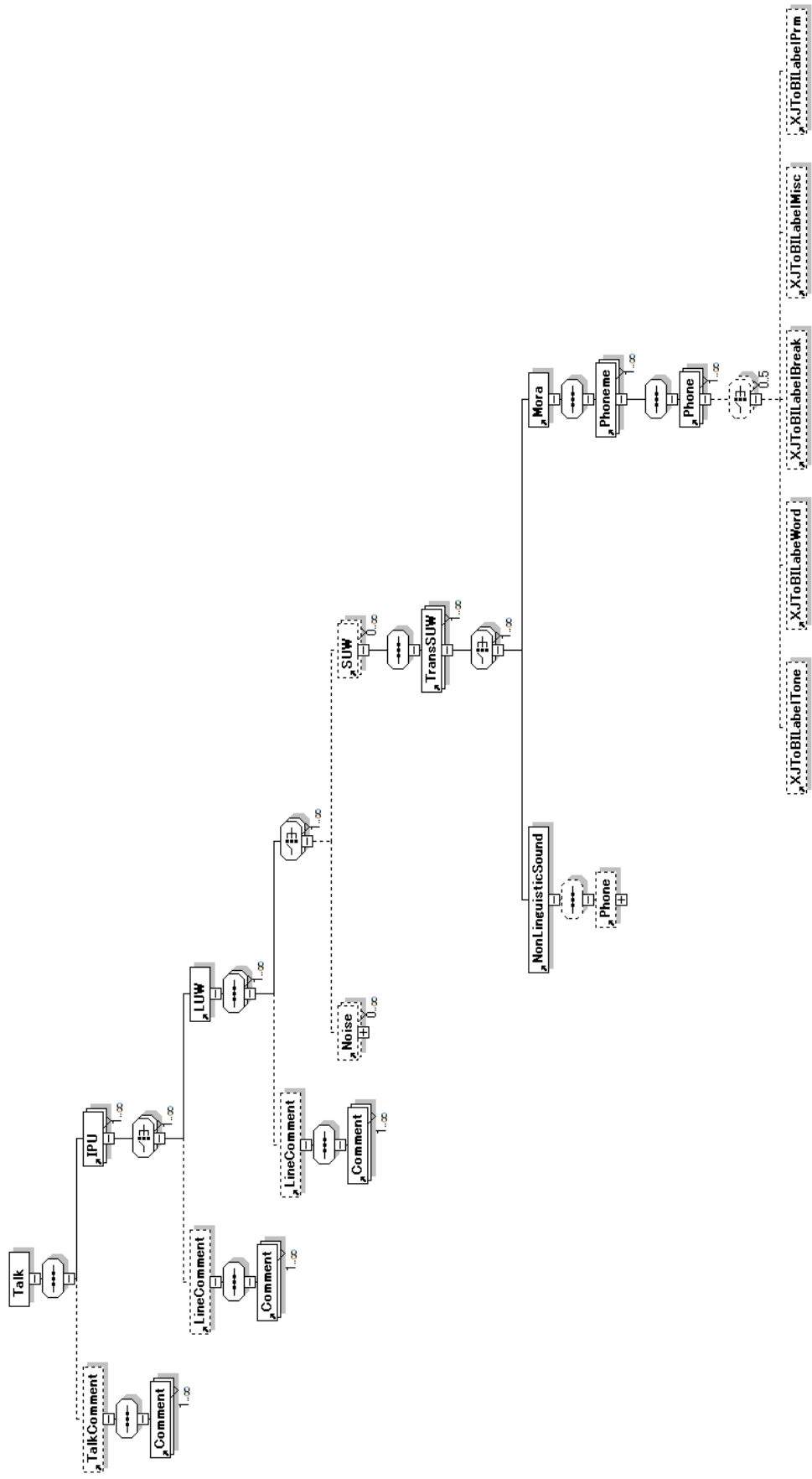


图 2. XML 元素关联图

表 1: XML 要素の内容(要素の項目の文字は p.3 で示した情報の種類に対応)

要素	属性名	説明	値(例)
<Talk>		講演(対話や朗読を含む)	
	TalkID	講演 ID	S01F0001
話	SpeakerID	講演者 ID	116
話	SpeakerSex	講演者性別	女
話	SpeakerBirthGeneration	講演者生年代(5 年刻み)	70to74
話	SpeakerBirthPlace	講演者出生地	東京都
<TalkComment>		講演単位のコメント	
<LineComment>		行単位のコメント	
<Comment>		コメント内容	
転	CommentStrings	コメント文	TYPE=O_FIL フィルター「その」 連体詞と迷う
<IPU>		転記基本単位	
転	IPUID	転記基本単位 ID	0024
転	Channel	音声チャンネル	L
転	IPUStartTime	IPU 開始時刻	00000.185
転	IPUEndTime	IPU 終了時刻	00000.702
<LUW>		長単位	
	LUWID	長単位 ID	3
形	LUWDictionaryForm	代表形(長単位)	イク
形	LUWLemma	代表表記(長単位)	行く
形	LUWPOS	品詞(長単位)	動詞
形	LUWConjugateType	活用の種類(長単位)	カ行五段
形	LUWConjugateForm	活用形(長単位)	連用形
形	LUWMiscPOSInfo1	その他情報 1(長単位)	格助詞
形	LUWMiscPOSInfo2	その他情報 2(長単位)	促音便
形	LUWMiscPOSInfo3	その他情報 3(長単位)	連語
文	IsNewLine	新転記行(文節境界)	0/1
転	LineID	転記行 ID	3
転	IsLeftOver	転記における長単位分割	0/1
<SUW>		短単位	
	SUWID	短単位 ID	3
形	ColumnID	カラム ID	2
形	SUWDictionaryForm	代表形(短単位)	イク
形	SUWLemma	代表表記(短単位)	行く
形	SUWPOS	品詞(短単位)	動詞
形	SUWConjugateType	活用の種類(短単位)	カ行五段
形	SUWConjugateType2	活用の種類 2(短単位)	カ行五段 2
形	SUWConjugateForm	活用形(短単位)	連用形
形	SUWConjugateForm2	活用形 2(短単位)	連用形 2
形	SUWMiscPOSInfo1	その他情報 1(短単位)	副助詞
形	SUWMiscPOSInfo2	その他情報 2(短単位)	語幹
形	SUWMiscPOSInfo3	その他情報 3(短単位)	言いよどみ
形	PlainOrthographicTranscription	タグ無し出現形(短単位)	行き
形	OrthographicTranscription	基本形(短単位)	行き
形	PhoneticTranscription	発音形(短単位)	イキ
節	ClauseUnitID	節単位 ID	2
節	ClauseBoundaryLabel	節境界ラベル	<テ節>
節	CU_OperationSign	節単位操作記号	:
節	CU_PreBracket	節単位前ブラケット	{
節	CU_PostBracket	節単位後ブラケット	}
節	CU_ObligateComment	節単位義務的コメント	引用節構造
係	Dep_BunsetsuUnitID	係り受け文節単位 ID	3
係	Dep_ModifieeBunsetsuUnitID	受け文節単位 ID	3
係	Dep_ObligateComment	係り受け義務的コメント	B+
係	Dep_Label	係り受けラベル	D
談	DS_Purpose	談話目的	本発表のまとめ
談	DS_SubPurpose	談話下位目的	処理結果の例示
談	DS_Comment	談話コメント	%E
談	DS_Subject1	談話作業 1	処理結果の説明
談	DS_Subject2	談話作業 2	音声資料の説明
談	DS_Subject3	談話作業 3	処理結果の提示
重	SE_Subject1_10p	重要文抽出作業 1_10p 結果	0/1
重	SE_Subject1_50p	重要文抽出作業 1_50p 結果	0/1
重	SE_Subject2_10p	重要文抽出作業 2_10p 結果	0/1
重	SE_Subject2_50p	重要文抽出作業 2_50p 結果	0/1
重	SE_Subject3_10p	重要文抽出作業 3_10p 結果	0/1
重	SE_Subject3_50p	重要文抽出作業 3_50p 結果	0/1
<TransSUW>		転記短単位	
	TransSUWID	転記短単位 ID	3
転	TagKanjiStart	漢字タグ(K)開始	0/1
転	TagKanjiEnd	漢字タグ(K)終了	0/1
転	TagKanjiMidst	漢字タグ(K)途中	0/1
転	TagKanjiOrth	漢字正書法	激
転	TagAlphabetStart	アルファベットタグ(A)開始	0/1
転	TagAlphabetEnd	アルファベットタグ(A)終了	0/1
転	TagAlphabetMidst	アルファベットタグ(A)途中	0/1
転	TagAlphabetOrth	アルファベット表記	E T
転	TagDisfluencyStart	語断片(D)開始	0/1
転	TagDisfluencyEnd	語断片(D)終了	0/1
転	TagDisfluencyMidst	語断片(D)途中	0/1
転	TagDisfluency2Start	機能言語い直し(D2)開始	0/1
転	TagDisfluency2End	機能言語い直し(D2)終了	0/1
転	TagDisfluency2Midst	機能言語い直し(D2)途中	0/1
転	TagFillerStart	フィルター(F)開始	0/1
転	TagFillerEnd	フィルター(F)終了	0/1
転	TagFillerMidst	フィルター(F)途中	0/1
転	TagIncorrectStart	発音エラー(W)開始	0/1
転	TagIncorrectEnd	発音エラー(W)終了	0/1
転	TagIncorrectMidst	発音エラー(W)途中	0/1
転	TagIncorrectNorm	発音エラー 正式発音	エビコット
転	TagBStart	知識レベル言い誤り(B)開始	0/1
転	TagBEnd	知識レベル言い誤り(B)終了	0/1
転	TagBMidst	知識レベル言い誤り(B)途中	0/1
転	TagBNorm	知識レベル言い誤り(B)正式発音	ジュータイ
転	TagMaskStart	個人名 差別語(R)開始	0/1
転	TagMaskEnd	個人名 差別語(R)終了	0/1
転	TagMaskMidst	個人名 差別語(R)途中	0/1
転	TagForeignStart	外国語・古語(O)開始	0/1
転	TagForeignEnd	外国語・古語(O)終了	0/1
転	TagForeignMidst	外国語・古語(O)途中	0/1
転	TagQuoteStart	メタ的引用(M)開始	0/1
転	TagQuoteSEnd	メタ的引用(M)終了	0/1
転	TagQuoteMidst	メタ的引用(M)途中	0/1
転	TagWrongRecitationStart	非朗読対象発話(X)開始	0/1
転	TagWrongRecitationEnd	非朗読対象発話(X)終了	0/1
転	TagWrongRecitationMidst	非朗読対象発話(X)途中	0/1
転	TagUncertainStart	語彙不明(?)開始	0/1
転	TagUncertainEnd	語彙不明(?)終了	0/1
転	TagUncertainMidst	語彙不明(?)途中	0/1
転	TagUncertainSecond	語彙不明部第二候補	実数
<Mora>		モーラ	
	MoraID	モーラ ID	3
転	MoraEntity	モーラ記号	ㇿ
転	TagPOccurrence	ポーズ<P>	0/1
転	TagPOccurrence_Start	ポーズ開始時間	00000.709
転	TagPOccurrence_End	ポーズ終了時間	00000.985
転	TagVLong	母音引き延ばし<H>	0/1

転	TagCLong	子音引き延ばし<Q>	0/1
転	TagWhisperStart	ささやき声(L)開始	0/1
転	TagWhisperEnd	ささやき声(L)終了	0/1
転	TagWhisperMidst	ささやき声(L)途中	0/1
転	TagLaughingStart	(笑)タグ開始	0/1
転	TagLaughingEnd	(笑)タグ終了	0/1
転	TagLaughingMidst	(笑)タグ途中	0/1
転	TagCryingStart	(泣)タグ開始	0/1
転	TagCryingEnd	(泣)タグ終了	0/1
転	TagCryingMidst	(泣)タグ途中	0/1
転	TagCoughingStart	(咳)タグ開始	0/1
転	TagCoughingEnd	(咳)タグ終了	0/1
転	TagCoughingMidst	(咳)タグ途中	0/1
転	TagUncertainStart	発音不明瞭(?)開始	0/1
転	TagUncertainEnd	発音不明瞭(?)終了	0/1
転	TagUncertainMidst	発音不明瞭(?)途中	0/1
転	TagUncertainSecond	発音不明瞭(?)第二候補	1
転	TagFV	母音不確定音<FV>	0/1
韻	PerceivedAcc	アクセント知覚	0/1
<Phoneme>			
	PhonemeID	音素 ID	3
	PhonemeEntity	音素記号	Kj
<Phone>			
	PhoneID	分節音 ID	1
分	PhoneEntity	分節音記号	kj
分	PhoneClass	分節音記号のクラス	Consonant
分	Devoiced	無声化の有無	0/1
分	PhoneStartTime	開始時刻	1.295305
分	PhoneEndTime	終了時刻	1.328962
分	StartTimeUncertain	開始位置不明	0/1
分	EndTimeUncertain	終了位置不明	0/1
<Noise>			
	NoiseID	ノイズ ID	3
形	ColumnID	カラム ID	2
係	Dep_BunsetsuUnitID	係り受け文節単位 ID	1
<NonLinguisticSound>			
	MoraID	モーラ ID	3
転	MoraEntity	モーラ記号	
転	TagBreath	<息>	0/1
転	TagLaugh	<笑>	0/1
転	TagCry	<泣>	0/1
転	TagCough	<咳>	0/1
転	TagNoise	<雑音>	0/1

転	TagBell	<ベル>	0/1
転	TagApplause	<拍手>	0/1
転	TagAudienceLaugh	<フロア笑>	0/1
転	TagAudienceTalk	<フロア発話>	0/1
転	TagDemoSound	<デモ>	0/1
転	TagWrongRecitation	朗読間違え	0/1
転	TagVN	応答表現うん系<VN>	0/1
転	TagWhisperStart	ささやき声(L)開始	0/1
転	TagWhisperEnd	ささやき声(L)終了	0/1
転	TagWhisperMidst	ささやき声(L)途中	0/1
転	TagLaughingStart	(笑)タグ開始	0/1
転	TagLaughingEnd	(笑)タグ終了	0/1
転	TagLaughingMidst	(笑)タグ途中	0/1
転	TagCryingStart	(泣)タグ開始	0/1
転	TagCryingEnd	(泣)タグ終了	0/1
転	TagCryingMidst	(泣)タグ途中	0/1
転	TagCoughingStart	(咳)タグ開始	0/1
転	TagCoughingEnd	(咳)タグ終了	0/1
転	TagCoughingMidst	(咳)タグ途中	0/1
転	TagCoughingMidst	(咳)タグ途中	0/1
<XJToBILabelTone>			
韻	Time	トーン記述時刻	0.233569
韻	ToneClass	トーンラベルのクラス	accent
韻	F0	ラベル位置付近の F0 値	146.0953
韻	F0Uncertain	F0 の不明確さ	0/1
韻	CategoryUncertain	カテゴリーの不明確さ	0/1
韻	PositionUncertain	位置の不明確さ	0/1
韻	Divided	複合ラベルの分割の有無	0/1
<XJToBILabelWord>			
韻	Time	ワード記述時刻	0.233569
韻	PerceivedAccPos	知覚したアクセント位置	3
韻	PerceivedAccPos2	知覚した二つ目のアクセント位置	5
<XJToBILabelBreak>			
韻	Time	BI 記述時刻	0.253141
韻	DisfluencyStart	語断片開始	0/1
韻	FillerStart	フィラー開始	0/1
<XJToBILabelPrm>			
韻	Time	Prm 記述時刻	2.513607
<XJToBILabelMisc>			
韻	Time	Misc 記述時刻	121.341629

## 4.2. 各要素の内容

図 2 に示したように、CSJ の研究用付加情報の XML 文書化に際して、講演を表す Talk 要素から韻律情報を表す XJToBILabelTone, XJToBILabelWord, XJToBILabelBreak, XJToBILabelPrm, XJToBILabelMisc 要素までを階層構造で表現する。各要素について以下に説明する。

### 4.2.1. Talk 要素

CSJ では講演(一部は対話、再朗読等)ごとに収録音声ファイルをファイル化している(講演の種類の詳細については概観マニュアル2 参照、詳細については収録マニュアル3参照)。Talk 要素にはこの講演単位についての情報や講演の話者についての情報を記録する。なお、二人の話者が存在するインタビューについては、話し手(インタビュイー)の話者情報を記す。Talk 要素は複数の IPU 要素を子要素として持つ。また、TalkComment 要素を子要素として持つ。

#### [属性]

(a) TalkID

講演の ID を記述する。講演の ID の命名規則については話者情報マニュアル2.1.1 参照。

(b) SpeakerID

講演者の ID を記述する。講演者 ID については話者情報マニュアル4.1 参照。

(c) SpeakerSex

講演者の性別を記述する。講演者性別については話者情報マニュアル4.2 参照。

(d) SpeakerBirthGeneration

講演者の生年代を記述する。講演者生年代については話者情報マニュアル4.3 参照。

(e) SpeakerBirthPlace

講演者の出生地を記述する。講演者出生地については話者情報マニュアル4.4 参照。

#### [例]

```
<Talk TalkID="S01F0001" SpeakerID="116" SpeakerSex="女"  
SpeakerBirthGeneration="70to74" SpeakerBirthPlace="東京都"/>
```

### 4.2.2. IPU 要素

転記テキストにおいて、原則として 200[ms]以上のポーズで区切られた単位を転記基本単位として認定する(転記基本単位の認定基準については転記マニュアル2 参照)。この単位を IPU(Inter-Pausal Unit)要素として表し、講演における各単位の開始時刻、終了時刻、音声ファイル中の記録チャンネル名を属性として記す。IPU 要素は一つ以上の LUW 要素を子要素として持つ。

#### [属性]

(a) IPUID

IPU 要素(転記基本単位に相当)の講演内の ID を記述する。転記基本単位の詳細については転記マニュアル2 参照。

(b) Channel

音声ファイル中の記録チャンネル名を記述する。二人の話者が存在するインタビューにおいてはこの属性により話者を区別する。記録チャンネルについては転記マニュアル1【転記基本単位情報部(単位情報部)】参照。

(c) IPUStartTime

転記基本単位の開始時刻を記述する。転記基本単位の開始時刻については転記マニュアル1【転記基本単位情報部(単位情報部)】参照。

(d) IPUEndTime

転記基本単位の終了時刻を記述する。転記基本単位の終了時刻については転記マニュアル1【転記基本単位情報部(単位情報部)】参照。

#### [例]

```
<IPU IPUID="0001" Channel="L" IPUStartTime="00000.185" IPUEndTime="00000.702" />
```

#### 4.2.3. TalkComment 要素

転記テキストにおいて、講演毎にコメントが記される(転記テキストにおける講演全体に対するコメントについては転記マニュアル1 [コメント]参照)。TalkComment 要素はこのコメント情報を Comment 要素として記述したものを子要素とする。TalkComment 要素は属性を持たない。

#### 4.2.4. LineComment 要素

転記テキストにおいて、転記基本単位に対するコメントや局所的な発話(転記基本単位の一部)に対するコメントが記されることがある(転記における転記基本単位や局所的な発話に対するコメントについては転記マニュアル1 [コメント]参照)。LineComment 要素はこれらのコメント情報を Comment 要素として記述したものを子要素とする。転記基本単位に対するコメントを記述する場合は LineComment 要素を IPU 要素の子要素とし、局所的な発話に対するコメントを記述する場合は LineComment 要素を LUW 要素の子要素とする。LineComment 要素は属性を持たない。

#### 4.2.5. Comment 要素

転記テキストにおいて、講演単位あるいは行単位にコメントが記される(転記テキストにおけるコメントについては転記マニュアル1 [コメント]参照)。Comment 要素はコメント内容を記述するための属性を持つ。

[属性]

(a) CommentStrings

転記におけるコメントを行毎に記述する。転記におけるコメントについては転記マニュアル1 [コメント]参照。

[例]

```
<TalkComment>
  <Comment CommentStrings="講演 ID:A01M0517"/>
  <Comment CommentStrings=""/>
  <Comment CommentStrings="TYPE=O_SIT 比較的長めのデモ音が多く、発話と時間的に重複すること
がある"/>
  <Comment CommentStrings=""/>
</TalkComment>
```

```
<LineComment>
  <Comment CommentStrings="TYPE=O_HYO 「延々」の言い間違いの可能性あり"/>
</LineComment>
```

#### 4.2.6. LUW 要素

形態論情報のうちの長単位を LUW(Long-Unit Word)要素とし、長単位の品詞、活用形などの情報を属性として持つ(長単位の基準については形態論マニュアル2.1 参照、長単位のデータベースについてはWDB マニュアル参照)。LUW 要素は一つ以上の SUW 要素を子要素として持つ。

[属性]

(a) LUWID

LUW 要素(長単位に相当)の IPU 単位内の通し番号(ID)を記述する。

(b) LUWPOS

長単位の品詞を記述する。長単位の品詞については形態論マニュアル3.2 参照。

(c) LUWDictionaryForm

長単位の代表形を記述する。長単位の代表形については形態論マニュアル3.1 参照。

(d) LUWLemma

長単位の代表表記を記述する。長単位の代表表記については形態論マニュアル3.1 参照。

(e) LUWConjugateForm

長単位の活用の種類を記述する。長単位の活用の種類については形態論マニュアル3.2 参照。



- (f) LUWConjugateType  
長単位の活用形情報を記述する。長単位の活用形情報については形態論マニュアル3.2 参照。
- (g) LUWMiscPOSInfo1, LUWMiscPOSInfo2, LUWMiscPOSInfo3  
長単位のその他の情報 1,2,3 を記述する。長単位のその他の情報については形態論マニュアル3.2 参照。
- (h) IsNewLine  
転記テキスト内の改行を意味する新転記行を記述する。つまりこの属性値が"1"である長単位の直前に転記テキストにおける改行が現れたことを表し、転記テキストにおける改行は文節の境界を表す。転記テキストにおける文節については文節マニュアル参照。
- (i) LineID  
転記テキストにおける行 ID を記述する。転記テキストを再現するための情報である。XML からの転記テキスト再現については5.1.1にて説明する。
- (j) IsLeftOver  
転記テキストにおける行を跨る長単位であるかどうかを記述する。転記テキストを再現するための情報である。XML からの転記テキスト再現については5.1.1にて説明する。

[例]

```
<LUW LUWID="4" LUWDictionaryForm="ハシル" LUWLemma="走る" LUWPOS="動詞"
LUWConjugateForm="連用形" LUWConjugateType="ラ行五段" LUWMiscPOSInfo2="促音便"/>
```

#### 4.2.7. SUW 要素

形態論情報のうちの短単位をSUW(Short-Unit Word)要素とし、短単位の品詞、活用形などの情報を属性として持つ(短単位の基準については形態論マニュアル2.2 参照、短単位・長単位のデータベースについてはWDB マニュアル参照)。SUW 要素は一つ以上のTransSUW 要素を子要素として持つ。

[属性]

- (a) SUWID  
SUW 要素(短単位に相当)の LUW 単位内の通し番号(ID)を記述する。
- (b) ColumnID  
短単位・長単位データベースにおいて記述する、転記テキスト内の短単位位置情報を記述する。短単位・長単位データベースにおいて転記テキストの復元に用いられる情報。短単位位置情報についてはWDB マニュアル4.1 参照。
- (c) SUWPOS  
短単位の品詞を記述する。短単位の品詞については形態論マニュアル3.2 参照。
- (d) SUWDictionaryForm  
短単位の代表形を記述する。短単位の代表形については形態論マニュアル3.1 参照。
- (e) SUWLemma  
短単位の代表表記を記述する。短単位の代表表記については形態論マニュアル3.1 参照。
- (f) SUWConjugateForm, SUWConjugateForm2  
短単位の活用形情報を記述する。短単位の活用形情報については形態論マニュアル3.2 参照。
- (g) SUWConjugateType, SUWConjugateType2  
短単位の活用の種類を記述する。短単位の活用の種類については形態論マニュアル3.2 参照。
- (h) SUWMiscPOSInfo1, LUWMiscPOSInfo2, LUWMiscPOSInfo3  
短単位のその他の情報 1,2,3 を記述する。短単位のその他の情報については形態論マニュアル3.2 参照。
- (i) PlainOrthographicTranscription  
短単位についての転記タグを除去したタグ無し出現形を記述する。短単位のタグ無し出現形についてはWDB マニュアル3.2 参照。
- (j) OrthographicTranscription  
短単位の基本形を記述する。短単位・長単位データベースにおける短単位の基本形についてはWDB マニュアル3.2 参照。
- (k) PhoneticTranscription  
短単位の発音形を記述する。短単位・長単位データベースにおける短単位の発音形についてはWDB マニュアル3.2 参照。

- (l) ClauseUnitID  
節単位 ID(0 スタートの連番)を記述する。節単位 ID については節単位マニュアル4.参照。
- (m) ClauseBoundaryLabel  
節境界ラベルを記述する。節境界ラベルについては節単位マニュアル2.参照。
- (n) CU\_OperationSign  
節単位操作記号を記述する。節単位操作記号については節単位マニュアル3.1 および 4.参照。
- (o) CU\_ObligateComment  
節単位義務的コメントを記述する。節単位義務的コメントについては節単位マニュアル3.2 および 4.参照。
- (p) CU\_PreBracket  
節単位前ブラケットを記述する。節単位前ブラケットについては節単位マニュアル4.参照。
- (q) CU\_PostBracket  
節単位後ブラケットを記述する。節単位後ブラケットについては節単位マニュアル4.参照。
- (r) Dep\_BunsetsuUnitID  
係り受け情報において、節単位内で文節に一意に付与された ID (0 スタートの連番)を記述する。
- (s) Dep\_ModifieeBunsetsuUnitID  
係り受け情報において、同一節単位内で当該文節が係っていく先の文節の ID(Dep\_BunsetsuUnitID) (ない場合がある)を記述する。
- (t) Dep\_Label  
係り受け情報における係り受け関係を示すラベルを記述する。係り受け関係を示すラベルについては係り受けマニュアル3. [1]参照。
- (u) Dep\_ObligateComment  
係り受け情報における義務的コメント(文節境界に関するラベル (B+) とその他)を記述する。係り受け情報におけるコメントについては係り受けマニュアル3. [1]参照。
- (v) DS\_Purpose  
談話境界情報における談話目的を記述する。談話目的については談話境界マニュアル2.2 参照。
- (w) DS\_SubPurpose  
談話境界情報における談話下位目的を記述する。談話下位目的については談話境界マニュアル2.2 参照。
- (x) DS\_Comment  
談話境界情報における談話コメントを記述する。談話コメントについては談話境界マニュアル2.2 参照。
- (y) DS\_Subject1, DS\_Subject2, DS\_Subject3  
談話境界情報における談話作業 1,2,3 を記述する。談話作業者については談話境界マニュアル2.2 参照。
- (z) SE\_Subject1\_10p, SE\_Subject2\_10p, SE\_Subject3\_10p  
重要文情報における重要文抽出作業 1,2,3 による 10%の重要文抽出結果を記述する。重要文抽出については重要文マニュアル3 参照。
- (aa) SE\_Subject1\_50p, SE\_Subject2\_50p, SE\_Subject3\_50p  
重要文情報における重要文抽出作業 1,2,3 による 50%の重要文抽出結果を記述する。重要文抽出については重要文マニュアル3 参照。

[例]

```
<SUW SUWID="1" SUWDictionaryForm="デ" SUWLemma="で" SUWPOS="接続詞"
OrthographicTranscription="で" PhoneticTranscription="デ<H>&gt;"
PlainOrthographicTranscription="で" ClauseBoundaryLabel="&lt;接続詞&gt;"
Dep_BunsetsuUnitID="0" ClauseUnitID="110" SE_Subject1_50p="1" SE_Subject2_50p="1"
SE_Subject3_50p="1" DS_Purpose="豊島園の駅の様子" DS_SubPurpose="夏に豊島園の駅を利用して
いる時にちょっとむかつく理由" DS_Subject1="豊島園の駅を利用する時に腹の立つこと" DS_Subject2="
豊島園の駅を利用する時に思うこと"/>
<SUW SUWID="1" SUWDictionaryForm="イク" SUWLemma="行く" SUWConjugateForm="連用形"
SUWConjugateType="力行五段" SUWConjugateForm2="連用形" SUWConjugateType2="力行五段"
SUWMiscPOSInfo2="促音便" SUWPOS="動詞" OrthographicTranscription="行っ"
PhoneticTranscription="イツ" PlainOrthographicTranscription="行っ"
Dep_BunsetsuUnitID="4" Dep_ModifieeBunsetsuUnitID="5"/>
```

#### 4.2.8. TransSUW 要素

転記テキストの基本形では、一部において短単位よりも細かい単位を対象としてタグ情報を付与しているケースがある。このようなタグ情報を正確に記録するために、転記テキストの基本形においてタグ情報が付与される範囲を一つの単位とする(転記短単位と呼ぶ)TransSUW 要素を設けた。TransSUW 要素には、そのような特別なケースだけでなく全ての場合において、転記テキストのタグ情報を記録することにした。一つのTransSUW 要素が示す転記短単位の範囲は、多くの場合、SUW 要素が示す短単位の範囲と同じであるため、TransSUW 要素は SUW 要素のただ一つの子要素となることが多い。転記テキストの基本形において短単位の一部にタグ情報が付与されているような場合に、タグ情報が付与される範囲とそれ以外に単位を分けて扱うため、SUW 要素に複数のTransSUW 要素が子要素として所属することになる。

以下に具体例を示す。例 1 では「考え」という短単位内でフィルターの記述により分割が起きている。この分割境界情報を保持するために「考え」を表す SUW 要素を「かん」「(F あー)」「が」「え」という 4 つのTransSUW 要素に分割している。また、例 2 に示すように、発音形において 1 モーラ毎に付与される可能性のあるタグ情報は Mora 要素の属性として記述するが、発音形情報であっても、基本的に短単位境界で区切って記述するものは TransSUW に記述する。これらのタグは必ずしも 1 モーラのみや 1 短単位のみには付与されずとは限らないため、複数モーラ(または短単位)に跨る可能性のあるタグには 3 つの属性名 Start, Midst, End を持たせる。例えば 1 短単位内で閉じるフィルタータグは TagFillerStart と TagFillerEnd を持つ。

```
0812 01959.782-01962.438 L:
お(K かん(F あー)が;考)えを & オカン(F アー)ガエオ
申しあげました & モーシアゲマシタ
```

例 1. SUW を分割するタグ

```
0104 00272.459-00274.670 L:
本物の & ホンモノノ
バナナだよって & (W バナ;バナナ)ダヨ(? ッテ)
```

例 2. TransSUW に記述する発音形情報

この要素は、いわば言語学的情報と音声学的情報の接点に相当する。短単位の品詞、活用の種類などの言語学的情報はこの要素の属性として記録される。そして、対応する音声に関する情報はこの要素の子要素である Mora 要素以下に記録される。音声に関する情報のうち、例えば言い誤りの有無を示すタグ(W)など、語単位で与えられるものは TransSUW 要素の属性として記録する。TransSUW 要素は一つ以上の Mora 要素あるいは NonLinguisticSound 要素を子要素として持つ。

[属性]

(a) TransSUWID

TransSUW 要素(転記短単位に相当)の SUW 単位内の通し番号(ID)を記述する。

(b) TagKanjiStart, TagKanjiEnd, TagKanjiMidst

転記短単位において漢字タグ(K)開始、終了、途中であるかどうかを記述する。漢字タグ(K)については転記マニュアル 5 参照。

- (c) TagKanjiOrth  
 転記短単位において漢字タグ(K)で示される正書法記述を記録する。漢字タグ(K)については転記マニュアル5 参照。
- (d) TagAlphabetStart, TagAlphabetEnd, TagAlphabetMidst  
 転記短単位においてアルファベットタグ(A)開始、終了、途中であるかどうかを記述する。アルファベットタグ(A)については転記マニュアル5 参照。
- (e) TagAlphabetOrth  
 転記短単位においてアルファベットタグ(A)で示される正書法記述を記録する。アルファベットタグ(A)については転記マニュアル5 参照。
- (f) TagDisfluencyStart, TagDisfluencyEnd, TagDisfluencyMidst  
 転記短単位において語断片タグ(D)開始、終了、途中であるかどうかを記述する。語断片タグ(D)については転記マニュアル5 参照。
- (g) TagDisfluency2Start, TagDisfluency2End, TagDisfluency2Midst  
 転記短単位において機能語言い直しタグ(D2)開始、終了、途中であるかどうかを記述する。機能語言い直しタグ(D2)については転記マニュアル5 参照。
- (h) TagFillerStart, TagFillerEnd, TagFillerMidst  
 転記短単位においてフィラータグ(F)開始、終了、途中であるかどうかを記述する。フィラータグ(F)については転記マニュアル5 参照。
- (i) TagIncorrectStart, TagIncorrectEnd, TagIncorrectMidst  
 転記短単位において発音エラータグ(W)開始、終了、途中であるかどうかを記述する。発音エラータグ(W)については転記マニュアル5 参照。
- (j) TagIncorrectNorm  
 転記短単位において発音エラータグ(W)で記される正確な発音を記述する。発音エラータグ(W)については転記マニュアル5 参照。
- (k) TagBStart, TagBEnd, TagBMidst  
 転記短単位において知識レベルの言い誤りタグ(B)開始、終了、途中であるかどうかを記述する。知識レベルの言い誤りタグ(B)については転記マニュアル5 参照。
- (l) TagBNorm  
 転記短単位において知識レベルの言い誤りタグ(B)で記される正確な発音を記述する。知識レベルの言い誤りタグ(B)については転記マニュアル5 参照。
- (m) TagMaskStart, TagMaskEnd, TagMaskMidst  
 転記短単位において個人名・差別語タグ(R)開始、終了、途中であるかどうかを記述する。個人名・差別語タグ(R)については転記マニュアル5 参照。
- (n) TagForeignStart, TagForeignEnd, TagForeignMidst  
 転記短単位において外国語・古語タグ(O)開始、終了、途中であるかどうかを記述する。外国語・古語タグ(O)については転記マニュアル5 参照。
- (o) TagQuoteStart, TagQuoteEnd, TagQuoteMidst  
 転記短単位においてメタ的引用タグ(M)開始、終了、途中であるかどうかを記述する。メタ的引用タグ(M)については転記マニュアル5 参照。
- (p) TagWrongRecitationStart, TagWrongRecitationEnd, TagWrongRecitationMidst  
 転記短単位において非朗読対象発話タグ(X)開始、終了、途中であるかどうかを記述する。非朗読対象発話タグ(X)については転記マニュアル5 参照。
- (q) TagUncertainStart, TagUncertainEnd, TagUncertainMidst  
 転記短単位において語彙不明タグ(?)開始、終了、途中であるかどうかを記述する。語彙不明タグ(?)については転記マニュアル5 参照。

(r) TagUncertainSecond

転記短単位において語彙不明タグ(?)で記される第二候補を記述する。語彙不明タグ(?)については転記マニュアル5 参照。

[例] (例1に対応)

```
<TransSUW TransSUWID="1" TagKanjiStart="1" TagKanjiOrth="考"/>
<TransSUW TransSUWID="2" TagKanjiMidst="1" TagFillerStart="1" TagFillerEnd="1"/>
<TransSUW TransSUWID="3" TagKanjiEnd="1"/>
```

#### 4.2.9. Mora 要素

4.2.8で述べたように、転記テキストの発音形に記される発音表記をモーラ単位に分割した単位を Mora 要素とし、モーラ記号(カナ)を属性として記す。なお、転記テキストの発音形におけるタグのうち、モーラ単位で記されるものはこの要素の属性として記録する。さらに、韻律情報において記録された知覚アクセントの位置もこの要素の属性として記録する。Mora 要素は一つ以上の Phoneme 要素を子要素として持つ。

[属性]

(a) MoraID

Mora 要素(モーラ)の TransSUW 単位内の通し番号(ID)を記述する。

(b) MoraEntity

モーラ記号を記述する。転記発音形に表記されるモーラ記号(カナ)の種類については転記マニュアル4 参照。

(c) TagPOccurrence

転記発音形において記されたポーズタグ<P>(転記基本単位内短単位中の 200[ms]以上の無音区間の存在を示す)の有無を記述する。この属性を持つ Mora 要素が表すモーラ単位の後に無音区間が現れたことを表す。タグ<P>については転記マニュアル5 参照。

(d) TagPOccurrence\_Start, TagPOccurrence\_End

転記発音形において記されたポーズタグ<P>(転記基本単位内短単位中の 200[ms]以上の無音区間の存在を示す)にあわせて記された無音区間の開始時刻、終了時刻を記述する。タグ<P>については転記マニュアル5 参照。

(e) TagVLong

転記発音形において記された母音引き延ばしタグ<H>の有無を記述する。この属性を持つ Mora 要素が表すモーラ単位の母音が引き延ばされたことを表す。タグ<H>については転記マニュアル5 参照。

(f) TagCLong

転記発音形において記された子音引き延ばしタグ<Q>の有無を記述する。この属性を持つ Mora 要素が表すモーラ単位の子音が引き延ばされたことを表す。タグ<Q>については転記マニュアル5 参照。

(g) TagWhisperStart, TagWhisperEnd, TagWhisperMidst

転記発音形においてささやき声タグ(L)開始、終了、途中であるかどうかを記述する。ささやき声タグ(L)については転記マニュアル5 参照。

(h) TagLaughingStart, TagLaughingEnd, TagLaughingMidst

転記発音形において(笑)タグ開始、終了、途中であるかどうかを記述する。(笑)タグについては転記マニュアル5 参照。

(i) TagCryingStart, TagCryingEnd, TagCryingMidst

転記発音形において(泣)タグ開始、終了、途中であるかどうかを記述する。(泣)タグについては転記マニュアル5 参照。

(j) TagCoughingStart, TagCoughingEnd, TagCoughingMidst

転記発音形において(咳)タグ開始、終了、途中であるかどうかを記述する。(咳)タグについては転記マニュアル5 参照。

(k) TagUncertainStart, TagUncertainEnd, TagUncertainMidst

転記発音形において発音不明瞭タグ(?)開始、終了、途中であるかどうかを記述する。発音不明瞭タグ(?)については転記マニュアル5 参照。

- (s) TagUncertainSecond  
 転記発音形において発音不明瞭タグ(?)で記される第二候補を記述する。発音不明瞭タグ(?)については転記マニュアル5 参照。
- (t) TagFV  
 転記発音形の母音不確定音タグ<FV>の有無を記述する。母音不確定音タグ<FV>については転記マニュアル5 参照。
- (u) PerceivedAcc  
 韻律情報において記録された知覚アクセントの位置を記述する。韻律情報における知覚アクセントの記述については韻律マニュアル4.参照。なお、同一 SUW 要素内に複数のアクセントを知覚する場合については、最初の知覚アクセント位置のみ記述する。

[例]

```
<Mora MoraID="1" MoraSUWID="1" MoraEntity="カ" PerceivedAcc="1"/>
<Mora MoraID="2" MoraSUWID="2" MoraEntity="ナ"/>
<Mora MoraID="3" MoraSUWID="3" MoraEntity="リ" TagVLong="1"/>
```

#### 4.2.10. Phoneme 要素

Mora 要素の属性 MoraEntity の値に記される、転記発音形のモーラ記号から自動的に音素単位を生成したものを Phoneme 要素とし(モーラと音素の対応関係については本マニュアル付録 1 参照)、音素記号を属性として記す。Phoneme 要素は一つ以上の Phone 要素を子要素として持つ。なお、この Phoneme 要素以下はコアと呼ばれる一部の講演データについてのみ記述される。

[属性]

- (a) PhonemeID  
 Phoneme 要素(音素に相当)の Mora 単位内の通し番号(ID)を記述する。モーラと音素の対応関係については本マニュアル付録 1 参照。
- (b) PhonemeEntity  
 音素記号を記述する。

[例]

```
<Phoneme PhonemeID="1" PhonemeEntity="r">
```

#### 4.2.11. Phone 要素

分節音ラベルの単位を Phone 要素として、ラベルやラベル区間の開始・終了時刻などを属性として記す(分節音情報については分節音マニュアル参照)。また、分節音ラベルに記された母音無声化の情報も属性として記録する。Phone 要素は韻律情報の各ラベル情報を要素とした XJToBILabelTone, XJToBILabelWord, XJToBILabelBreak, XJToBIMisc を子要素として持つ。

[属性]

- (a) PhoneID  
 Phone 要素(分節音に相当)の Phoneme 単位内の通し番号(ID)を記述する。分節音単位については分節音マニュアル参照。
- (b) PhoneEntity  
 分節音記号を記述する。分節音記号については分節音マニュアル3.参照。  
 なお、分節音関係補助ラベルで用いる"<",">"は XML において特殊な意味を持つため、代わりに"S"を用いることにする。例えば"<cl>"は"ScIS"と表現する。  
 また、無声化した母音について、"A","I","U","E","O"の分節音ラベルが用意されているが、XML 文書においては後述する Devoiced 属性で表現するため通常の母音ラベル"a","i","u","e","o"で表現する。
- (c) PhoneClass  
 分節音記号のクラスを記述する。分節音記号については分節音マニュアル3.参照。以下には、ここで用いられるクラス名と分節音記号の対応関係を示す。

```
vowel : a,i,u,e,o
```

consonant : k,g,s,z,t,c,d,a,b,F,b,p,m,y,r,w,v,kw,gw,kj,gj,sj,zj,cj,nj,hj,ky,gy,sy,zy,cy,ty,dy,ny,hy,  
Fy,by,py,my,ry  
special : N,Q,H  
others : <cl>,<pz>,<uv>,<sv>,<fr>,<fv>,<?>,<N>,<b>

(d) Devoiced

母音の無声化の有無を記述する。母音の無声化については分節音マニュアル表 1 および 5.1 参照。

(e) PhoneStartTime, PhoneEndTime

分節音の開始、終了時刻を記述する。分節音の時刻情報については分節音マニュアル4.1 参照。

なお、分節音ラベリングにおいてはラベルが示す区間の終端時刻のみが記録される。XML 文書化にあたっては、直前のラベルとともに記録されている時刻を当該ラベルの開始時刻とみなした。例外として、転記基本単位先頭のラベル内で融合する場合(ラベルの融合については分節音マニュアル5.1 参照)には開始時刻と終了時刻が同一になる。

(f) StartTimeUncertain, EndTimeUncertain

分節音の開始、終了時刻の不明確さを記述する。分節音ラベリングにおいて融合した形で付与されたラベルについて、そのラベルが示す時間区間をラベル数で等分割する形で各ラベルの開始・終了時刻を決定し、PhoneStartTime, PhoneEndTime 属性に記述する。例えば StartTimeUncertain 属性を持つ Phone 要素で表される分節音は、前の Phone 要素で表される分節音と融合したラベルにより記述されていたことがわかる。分節音ラベルの融合については分節音マニュアル5.1 参照。

[例]

```
<Phone PhoneID="1" PhoneEntity="u" PhoneClass="vowel" Devoiced="y"
PhoneStartTime="634.369018" PhoneEndTime="634.413606" StartTimeUncertain="1">
```

#### 4.2.12. Noise 要素

転記テキストの発音形に記される、言語音とは独立に生じた話者の息・咳や話者の声以外の目立つ音などの情報を記すために、Noise 要素を設ける。転記テキストの発音形に記される、言語音とは独立に生じた話者の息や咳、話者の声以外の目立つ音などの扱いについては転記マニュアル5 参照。図 2 に示すように、Noise 要素は SUW 要素と同じ階層に位置づけられている。それに伴って、属性の一部((b)に示したものは) SUW 要素と共通している。

[属性]

(a) NoiseID

Noise 要素 (言語音以外のノイズに相当) の講演内の通し番号(ID)を記述する。

(b) その他

ColumnID、Dep\_BunsetsuUnitID については、4.2.7 SUW 要素の各項目を参照。

[例]

```
<Noise NoiseID="1">
```

#### 4.2.13. NonLinguisticSound 要素

転記テキストに記される、息、咳、拍手、ベル音などの情報を記すために、NonLinguisticSound 要素を設ける。転記テキストには言語音とは独立に生じた話者の息・咳や話者の声以外の目立つ音などと、言語音中に生じる息や笑いなどが、それぞれ異なる形式で記述される(これらの扱いについての詳細は転記マニュアル5 参照)。XML 文書においてはいずれも NonLinguisticSound 要素として表現するが、前者の場合は Noise 要素を親とし、後者の場合は TransSUW 要素を親とする。NonLinguisticSound 要素は子要素として Phone 要素を持つ。この要素は Mora 要素と同じ階層に位置づけられ、属性の一部((n)に示したものは) Mora 要素と共通している。

[属性]

(a) MoraID

NonLinguisticSound 要素の親要素 (Noise 要素か TransSUW 要素) 内の通し番号(ID)を記述する。

- (b) TagBreath  
転記テキストに記された<息>タグを記述する。<息>タグについては転記マニュアル5 参照。
- (c) TagLaugh  
転記テキストに記された<笑>タグを記述する。<笑>タグについては転記マニュアル5 参照。
- (d) TagCry  
転記テキストに記された<泣>タグを記述する。<泣>タグについては転記マニュアル5 参照。
- (e) TagCough  
転記テキストに記された<咳>タグを記述する。<咳>タグについては転記マニュアル5 参照。
- (f) TagNoise  
転記テキストに記された<雑音>タグを記述する。<雑音>タグについては転記マニュアル5 参照。
- (g) TagBell  
転記テキストに記された<ベル>タグを記述する。<ベル>タグについては転記マニュアル5 参照。
- (h) TagApplause  
転記テキストに記された<拍手>タグを記述する。<拍手>タグについては転記マニュアル5 参照。
- (i) TagAudienceLaugh  
転記テキストに記された<フロア笑>タグを記述する。<フロア笑>タグについては転記マニュアル5 参照。
- (j) TagAudienceTalk  
転記テキストに記された<フロア発話>タグを記述する。<フロア発話>タグについては転記マニュアル5 参照。
- (k) TagDemoSound  
転記テキストに記された<デモ>タグを記述する。<デモ>タグについては転記マニュアル5 参照。
- (l) TagWrongRecitation  
転記テキストに記された<朗読間違い>タグを記述する。<朗読間違い>タグについては転記マニュアル5 参照。
- (m) TagVN  
転記テキストに記された応答表現（「うん」系）を表す<VN>タグを記述する。<VN>タグについては転記マニュアル5 参照。
- (n) その他  
TagWhisperStart、TagWhisperEnd、TagWhisperMidst、TagLaughingStart、TagLaughingEnd、TagLaughingMidst、TagCryingStart、TagCryingEnd、TagCryingMidst、TagCoughingStart、TagCoughingEnd、TagCoughingMidst、TagUncertainStart については4.2.9参照。

[例]

```
<NonLinguisticSound MoraID="1" TagLaugh="1">
```

#### 4.2.14. XJToBILabelTone 要素

日本語話し言葉の韻律ラベリングスキーム X-JToBI により付与された韻律情報のうちのトーンラベル情報を、XJToBILabelTone 要素として記録する(X-JToBI については韻律マニュアル参照。韻律情報のうちのトーンラベル情報については韻律マニュアル2.3 参照)。なおトーンラベルのうち、エクステンダー記号">"は XML 文書中では記号"]"に置き換えて記述されているので、利用の際には注意されたい(エクステンダーについては韻律マニュアル 4.6 参照)。ラベルが付与された位置(時間)、ラベルのクラス、不明確さなどが属性として記録される。

[属性]

- (a) Time  
XJToBILabelTone 要素が表すトーンラベルの付与位置(時間)を記述する。ラベルの付与位置(時間)については韻律マニュアル9.参照。



(b) ToneClass

XJToBILabelTone 要素が表すトーンラベルのクラスを記述する。トーンラベルのクラスは以下の通りである。

IBT (Initial Boundary Tone): %L

FBT (Final Boundary Tone): L% (ただし Boudary Pitch Movement を構成する場合は除く), H%, HL%, LH%, HLH%

LTBPM (Low Tone of Boundary Pitch Movement): L%

PT (Phrase Tone): H-

Accent : A

Pointer : pH, pL

Extender : > (l)

Filler : FL, FH

(c) F0Uncertain

XJToBILabelTone 要素が表すトーンラベルの付与位置において F0(基本周波数)が不明確であることを示す"x"の有無を記述する。トーンラベルにおける"x"については韻律マニュアル4.参照。

(d) CategoryUncertain

XJToBILabelTone 要素が表すトーンラベルにおいてカテゴリーが不明確であることを示す"?"の有無を記述する。トーンラベルにおける"?"については韻律マニュアル4.参照。

(e) PositionUncertain

XJToBILabelTone 要素が表すトーンラベルにおいて付与位置が不明確であることを示す"! "の有無を記述する。トーンラベルにおける"! "については韻律マニュアル4.参照。

(f) Divided

XJToBILabelTone 要素が表すトーンラベルにおいて、"L%L"などのもともと融合ラベルとして表現されていたものを分割したことをこの属性で記述する。融合ラベルについて国立国語研究所報告書 No.124「日本語話し言葉コーパスの構築法」第7章「韻律情報」の7.1.8 参照。

[例]

```
<XJToBILabelTone Time="0.658911" F0="99.8448" ToneClass="fbt">L%</XJToBILabelTone>
```

#### 4.2.15. XJToBILabelWord 要素

日本語話し言葉の韻律ラベリングスキーム X-JToBI により付与された韻律情報のうちのワードラベル情報を、XJToBILabelWord 要素として記録する(X-JToBI については韻律マニュアル参照。韻律情報のうちのワードラベル情報については韻律マニュアル2.1 および3.参照)。ラベルが付与された位置(時間)、知覚アクセント位置などが属性として記録される。

[属性]

(a) Time

XJToBILabelWord 要素が表すワードラベルの付与位置(時間)を記述する。ラベルの付与位置(時間)については韻律マニュアル9.参照。

(b) PerceivedAccPos

XJToBILabelWord 要素が表すワードラベルに記された知覚アクセント位置をモーラの通し番号(Mora 要素の MoraID 属性)で記述する。なお"ma[H]'su[H]"のように、母音の引き伸ばしが行われた箇所においてアクセントによるピッチの下降が知覚されることがある。したがって"[H]"はモーラ数にカウントし、この場合は PerceivedAccPos 属性の値を2とする。

(c) PerceivedAccPos2

XJToBILabelWord 要素が表すワードラベルに記された二つ目の知覚アクセント位置をモーラの通し番号(Mora 要素の MoraID 属性)で記述する。"sa'tesa'te"のような場合、PerceivedAccPos 属性に1、PerceivedAccPos2 属性に3と記述する。

[例]

```
<XJToBILabelWord Time="0.683079" PerceivedAccPos="1">taitoru</XJToBILabelWord>
```

#### 4.2.16. XJToBILabelBreak 要素

日本語話し言葉の韻律ラベリングスキーム X-JToBI により付与された韻律情報のうちの BI(Break Index)ラベル情報を、XJToBILabelBreak 要素として記録する (X-JToBI については韻律マニュアル参照。韻律情報のうちの BI ラベル情報については韻律マニュアル2.4 および 5.参照)。ラベルが付与された位置 (時間)、知覚アクセント位置などが属性として記録される。

[属性]

(a) Time

XJToBILabelBreak 要素が表す BI ラベルの付与位置 (時間) を記述する。ラベルの付与位置 (時間) については韻律マニュアル9.参照。

(b) FillerStart

XJToBILabelBreak 要素が表す BI ラベルに記されたフィラー開始記号("<F")の有無を記述する。韻律ラベリングにおけるフィラーの扱いについては韻律マニュアル8.1 参照。

[例]

```
<XJToBILabelBreak Time="0.683079">3</XJToBILabelBreak>
```

#### 4.2.17. XJToBILabelPrm 要素

日本語話し言葉の韻律ラベリングスキーム X-JToBI により付与された韻律情報のうちのプロミネンスラベル情報を、XJToBILabelPrm 要素として記録する (X-JToBI については韻律マニュアル参照。韻律情報のうちのプロミネンスラベル情報については韻律マニュアル2.5 および 6.参照)。ラベルが付与された位置 (時間) が属性として記録される。

[属性]

(a) Time

XJToBILabelPrm 要素が表すプロミネンスラベルの付与位置 (時間) を記述する。ラベルの付与位置 (時間) については韻律マニュアル9.参照。

[例]

```
<XJToBILabelPrm Time="183.221692">PNLP</XJToBILabelPrm>
```

#### 4.2.18. XJToBILabelMisc 要素

日本語話し言葉の韻律ラベリングスキーム X-JToBI により付与された韻律情報のうちの注釈ラベル情報を、XJToBILabelMisc 要素として記録する (X-JToBI については韻律マニュアル参照。韻律情報のうちの注釈ラベル情報については韻律マニュアル2.6 および 7.参照)。ラベルが付与された位置 (時間) が属性として記録される。

[属性]

(a) Time

XJToBILabelMisc 要素が表す注釈ラベルの付与位置 (時間) を記述する。ラベルの付与位置 (時間) については韻律マニュアル9.参照。

[例]

```
<XJToBILabelMisc Time="131.361842">AYOR</XJToBILabelMisc>
```

### 4.3. 階層構造からの逸脱に対する措置

話し言葉音声においては、言語学的な階層構造を逸脱する現象がたびたび生じる。ここでは、幾つかのそうした現象に対する措置を説明する。

#### 4.3.1. 分節音ラベルの複合化

分節音ラベル作成において、母音の無声化、子音の弱化、母音連鎖や融合などによって分節音境界を定めにくい場合に、複数のラベルを一つに複合化して付与する処置を許している。これにより、ラベルの区間が明確でない分節音の単位が生じることになる。

一方、前述した階層構造から、分節音ラベルの情報を記録する Phone 要素の時間情報を参照すれば、上位階層にあたる Mora 要素や SUW 要素の、音声における開始・終了時間がわかる。したがって、大まかでも Phone 要素の時間情報を記して欲しいという要求が想定される。そこで、複合化された分節音ラベルについては、便宜上区間を均等に分割した結果の区間開始・終了時刻を各 Phone 要素の属性として記し、恣意的に与えた境界であることを別途属性で記録することにした。例えば、母音の無声化によって複合化された分節音ラベルを以下のように記述する（一部の属性を省略している）。

分節音ラベル(左からラベル区間の終端時間、ラベル)

```
244.647519 a
244.785809 s,U      #/uの無声化により複合
```

XML による分節音ラベルの複合化の記述

```
<Phone PhoneStartTime="244.564289" PhoneEndTime="244.647519" PhoneEntity="a"/>
<Phone PhoneStartTime="244.647519" PhoneEndTime="244.716664" PhoneEntity="s"
  EndPosUncertain="1"/>
<Phone PhoneStartTime="244.716664" PhoneEndTime="244.785809" PhoneEntity="u"
  StartPosUncertain="1" Devoiced="1"/>
```

#### 4.3.2. 語の発音レベルでの融合

話者の言い誤りによって、複数の語が発音レベルで融合しているように観察されるケースが頻繁に見られる。例えば、「この」と「よう(な)」の発音が融合して「コニョー(ナ)」となるようなケースである。このようなケースに対して、CSJ の転記では発音形テキストにおいて、関係する融合部分全体に(W)タグを記し「(W コニョー;コノヨウ)のように“正しいであろう発音”とともに実際の発音を表記する。一方、形態論情報は転記の基本形テキストに対して付加されるため語の融合としては扱わない。この矛盾に XML の構造上で対応するために、形態論情報は SUW 要素の属性として表わし、モーラ以下の音声学的情報は融合している範囲の先頭の SUW 要素の下位階層に記録する。残りの SUW 要素の下位階層には MoraEntity の値が” ”である Mora 要素を置く。上記の例は XML により以下のように記述される。

XML による語の融合の記述

```
<SUW SUWDictionaryForm="コノ" SUWLemma="此の" SUWPOS="連体詞" TagIncorrectNorm="コノ">
  <Mora MoraEntity="コ"/>
  <Mora MoraEntity="ニョ"/>
  <Mora MoraEntity="-"/>
</SUW>
<SUW SUWDictionaryForm="ヨウ" SUWLemma="様" SUWPOS="形状詞" TagIncorrectNorm="ヨウ">
  <Mora MoraEntity=" "/>
</SUW>
```

#### 4.4. CSJ 研究用付加情報の XML による記述例

これまでに示した仕様にしたがって、CSJ の研究用付加情報を XML で記述する例を示す。まず、同一音声に対して作成された転記、形態論情報、分節音情報、韻律情報のそれぞれを図 3 に示す。

これらの情報の XML 化にあたっては、まず始めに短単位情報を XML の形に変換する。なお、転記の情報は余すことなく短単位情報に記されているため、転記との照合は不要である。この段階で、前章で示した Phoneme 要素までが記述された XML が作成される。次の段階では分節音情報から作成した Phone 要素を Phoneme 要素の子要素として配置する。さらに韻律情報からラベル情報を Phone 要素の子要素として位置付け、知覚されたアクセントの位置の情報を Mora 要素の属性として記録するなどの処理を行う。以上の処理によって自動的に生成した XML の例を図 4 に示す。

転記テキスト

```
0091 00244.050-00245.009 L:
いつもの                & イツモノ
場所で                  & バシヨデ
0092 00245.270-00245.581 L:
(D ねろ)                & (D ネロ)
0093 00245.800-00247.076 L:
寝転がっていますと      & ネ<Q>コロガッテイマスト
```

形態論情報('いつもの'の部分に対して)

S03F0119 0091 00244.050-00245.009 L:-001-001	いつ	いつ	イツ	何時	イツ	代名詞
2001-05-14 12:01:18+09						
S03F0119 0091 00244.050-00245.009 L:-001-005	も	も	モ	も	モ	助詞
副助詞 2001-04-09 10:59:11+09						
S03f0119 0091 00244.050-00245.009 L:-001-007	の	の	ノ	の	ノ	助詞
格助詞 2001-04-09 10:59:11+09						

分節音情報('いつもの'の部分に対して)

```
244.073871 121 #
244.154540 121 i
244.187874 121 <cl>
244.240331 121 c
244.268501 121 u
244.328683 121 m
244.372218 121 o
244.418315 121 n
244.493862 121 o
```

韻律情報('いつもの'の部分に対して)

```
244.092331 115 %L
244.114585 115 A
244.494613 115 L%
```

図 3. CSJ の研究用付加情報の例  
(図中、青色で示した箇所の情報が図 4 の XML 文書内で記述されている)

```

<IPU Channel="L" IPUEndTime="00245.009" IPUID="0091" IPUStartTime="00244.050">
  <LUW IsNewLine="1" LUWDictionaryForm="イツ" LUWID="1" LUWLemma="何時" LUWPOS="代名詞" LineID="001">
    <SUW ColumnID="001" Dep BunsetsuUnitID="5" Dep ModifieeBunsetsuUnitID="6"
    OrthographicTranscription="いつ" PhoneticTranscription="イツ" PlainOrthographicTranscription="いつ"
    SUWDictionaryForm="イツ" SUWID="1" SUWLemma="何時" SUWPOS="代名詞" ClauseUnitID="44">
      <TransSUW TransSUWID="1">
        <Mora MoraEntity="イ" MoraID="1" PerceivedAcc="1">
          <Phoneme PhonemeEntity="i" PhonemeID="1">
            <Phone PhoneID="1" PhoneEntity="i" PhoneClass="vowel" PhoneStartTime="244.073871"
            PhoneEndTime="244.15454">
              <XJToBILabelTone Time="244.098746" F0="223.9560"
              ToneClass="ibt">%L</XJToBILabelTone>
              <XJToBILabelTone Time="244.114585" F0="228.1020"
              ToneClass="accent">A</XJToBILabelTone>
            </Phone>
          </Phoneme>
        </Mora>
        <Mora MoraEntity="ツ" MoraID="2">
          <Phoneme PhonemeEntity="c" PhonemeID="1">
            <Phone PhoneID="1" PhoneEntity="ScLs" PhoneClass="others"
            PhoneStartTime="244.15454" PhoneEndTime="244.187874"/>
            <Phone PhoneID="2" PhoneEntity="c" PhoneClass="consonant"
            PhoneStartTime="244.187874" PhoneEndTime="244.240331"/>
          </Phoneme>
          <Phoneme PhonemeEntity="u" PhonemeID="2">
            <Phone PhoneID="1" PhoneEntity="u" PhoneClass="vowel" PhoneStartTime="244.240331"
            PhoneEndTime="244.268501">
              <XJToBILabelWord Time="244.268501" PerceivedAccPos="1">icu</XJToBILabelWord>
              <XJToBILabelBreak Time="244.268501">1</XJToBILabelBreak>
            </Phone>
          </Phoneme>
        </Mora>
      </TransSUW>
    </SUW>
  </LUW>
  <LUW IsNewLine="0" LUWDictionaryForm="モ" LUWID="2" LUWLemma="も" LUWMiscPOSInfo1="副助詞" LUWPOS="
  助詞" LineID="001">
    <SUW ColumnID="005" OrthographicTranscription="も" PhoneticTranscription="モ"
    PlainOrthographicTranscription="も" SUWDictionaryForm="モ" SUWID="1" SUWLemma="も" SUWMiscPOSInfo1="
    副助詞" SUWPOS="助詞" ClauseUnitID="44">
      <TransSUW TransSUWID="1">
        <Mora MoraEntity="モ" MoraID="1">
          <Phoneme PhonemeEntity="m" PhonemeID="1">
            <Phone PhoneID="1" PhoneEntity="m" PhoneClass="consonant"
            PhoneStartTime="244.268501" PhoneEndTime="244.328683"/>
          </Phoneme>
          <Phoneme PhonemeEntity="o" PhonemeID="2">
            <Phone PhoneID="1" PhoneEntity="o" PhoneClass="vowel" PhoneStartTime="244.328683"
            PhoneEndTime="244.372218">
              <XJToBILabelWord Time="244.372218" PerceivedAccPos="0">mo</XJToBILabelWord>
              <XJToBILabelBreak Time="244.372218">1</XJToBILabelBreak>
            </Phone>
          </Phoneme>
        </Mora>
      </TransSUW>
    </SUW>
  </LUW>
  <LUW IsNewLine="0" LUWDictionaryForm="/" LUWID="3" LUWLemma="の" LUWMiscPOSInfo1="格助詞" LUWPOS="
  助詞" LineID="001">
    <SUW ColumnID="007" OrthographicTranscription="の" PhoneticTranscription="ノ"
    PlainOrthographicTranscription="の" SUWDictionaryForm="/" SUWID="1" SUWLemma="の" SUWMiscPOSInfo1="
    格助詞" SUWPOS="助詞" ClauseUnitID="44">
      <TransSUW TransSUWID="1">
        <Mora MoraEntity="/" MoraID="1">
          <Phoneme PhonemeEntity="n" PhonemeID="1">
            <Phone PhoneID="1" PhoneEntity="n" PhoneClass="consonant"
            PhoneStartTime="244.372218" PhoneEndTime="244.418315"/>
          </Phoneme>
          <Phoneme PhonemeEntity="o" PhonemeID="2">
            <Phone PhoneID="1" PhoneEntity="o" PhoneClass="vowel" PhoneStartTime="244.418315"
            PhoneEndTime="244.493862">
              <XJToBILabelTone Time="244.485417" F0="151.8200"
              ToneClass="fbt">L%</XJToBILabelTone>
              <XJToBILabelWord Time="244.493862" PerceivedAccPos="0">no</XJToBILabelWord>
              <XJToBILabelBreak Time="244.493862">3</XJToBILabelBreak>
            </Phone>
          </Phoneme>
        </Mora>
      </TransSUW>
    </SUW>
  </LUW>

```

図 4. XML 文書化した CSJ の研究用付加情報の例 (一部)

## 5. XML 文書の利用方法

CSJ の研究用付加情報を XML 文書化したものの利用方法を紹介します。なお、ここで紹介する以外にも、XML の関連規格やそれに準拠したアプリケーションを用いて本 XML 文書を利用することは可能である。XML の関連規格やそれに準拠したアプリケーションについては、W3C(World-Wide-Web Consortium)の Web サイト(<http://www.w3c.org/>)にて随時情報が提供されているのでそちらを参照されたい。

### 5.1. XML 文書の整形

XML はデータの構造を任意に定義して記述することのできるデータ記述言語であり、また、その関連規格を利用すれば一度記述された構造を容易に任意の構造に変換することができる。CSJ の XML 文書には、転記情報、形態論情報、分節音情報などの様々な研究用付加情報が、相互に関係した形で記述されている。この中から必要な情報を任意の形式で出力する方法を、転記テキスト生成や分節音ラベル生成、音声研究用派生 XML 生成、また無声化母音周辺の情報抽出を例にして説明する。

ここでは、XSLT というデータ構造変換の規格を利用した XML 文書の整形・変換について紹介する。XSLT の詳細については前述の W3C のサイト関連箇所(<http://www.w3.org/TR/xslt/>)などを参照されたい。

#### 5.1.1. 転記テキストの生成

図 5 に、CSJ の XML 文書から転記テキストを抽出して出力する XSL スタイルシートを示す。

図 4 に示した XML 文書に対して図 5 のスタイルシートを適用すると、図 3 の転記テキストが再現される。XML 文書から転記テキストを再現する XML スタイルシートは /TOOL/XSL/xsl2trn.xsl に収められている。なお、同様に短単位・長単位混合形式の短単位・長単位データベースを再現する XML スタイルシートが /TOOL/XSL/xsl2sdb.xsl として収められているので参考にされたい。

```

<?xml version="1.0" encoding="EUC-JP"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="text" indent="no" encoding="EUC-JP"/>
<xsl:strip-space elements="*" />

<xsl:template match="/">
  <xsl:apply-templates select="Talk" />
</xsl:template>
(途中、省略)
<xsl:template match="IPU">
  <xsl:value-of select="@IPUID" />
  <xsl:text> </xsl:text>
  <xsl:value-of select="@IPUStartTime" />
  <xsl:text>-</xsl:text>
  <xsl:value-of select="@IPUEndTime" />
  <xsl:text> </xsl:text>
  <xsl:value-of select="@Channel" />
  <xsl:if test="not (count(LUW) = '1' and LUW/Noise)">
    <xsl:apply-templates select="LineComment" />
  </xsl:if>
  <xsl:call-template name="KihonkeiHatsunonkei">
    <xsl:with-param name="LUWs" select="LUW" />
  </xsl:call-template>
  <xsl:if test="count(LUW) = '1' and LUW/Noise">
    <xsl:apply-templates select="LineComment" />
  </xsl:if>
  <xsl:text>
</xsl:text>
</xsl:template>
(以下、省略)

```

図 5. 転記テキストを生成する XSL

#### 5.1.2. 分節音ラベルの生成

図 7 に、CSJ の XML 文書から分節音ラベル(ESPS/waves+形式)を抽出して出力する XSL スタイルシートを示す。ESPS/waves+のラベルファイルの形式は図 3 の分節音情報の例に示したように、左から時間、色番号、ラベル文字列の順に一ラベル一行で記述するようになっている。したがってスタイルシートはそれらの情報を順に出力するよう記述した。

なお、図 7 に示したスタイルシートでは分節音ラベルの融合(分節音ラベルの融合については分節音マニュアル5.1 参照)を復元しないかわりに、融合ラベルを分割したのものについては StartTimeUncertain、EndTimeUncertain 属性を利用してラベ

ルの色を変えるよう出力する。ここで示した方法で分節音ラベルを再現する XML スタイルシートは /TOOL/XSL/xsl2seg.xml に収められている。

```

<?xml version="1.0" encoding="EUC-JP"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform" xml:lang="ja">

<xsl:output method="text" indent="yes" encoding="EUC_JP"/>
<xsl:strip-space elements="*/>

<!--Extract segmental label information from XML. (Labels are not combined.)-->

<!--Print Header of "*.labs* file for X-JToBI-->
<xsl:template match="/">
signal template
type 0
comment created using xlabel Wed Feb 26 07:06:09 2003
color 121
font *-times-medium-r-*-20-*-*-*-*-*
separator ;
nfields 1
#
<xsl:apply-templates/>
</xsl:template>

<xsl:template match="IPU">
<xsl:if test="@Channel='L'">
<!--Print line of "#" and its time-->

<xsl:for-each select="descendant::Phone" >

<xsl:if test="position()='1'">
<xsl:text> </xsl:text>
<xsl:value-of select="@PhoneStartTime"/>
<xsl:text> 123 </xsl:text>
<xsl:text>#</xsl:text>
<xsl:text>&#x0a;</xsl:text>
</xsl:if>

<xsl:if test="@PhoneEndTime">
<xsl:text> </xsl:text>
<xsl:value-of select="@PhoneEndTime"/>
<xsl:choose>
<xsl:when test="@EndTimeUncertain">
<xsl:text> 117 </xsl:text>
</xsl:when>
<xsl:otherwise>
<xsl:text> 123 </xsl:text>
</xsl:otherwise>
</xsl:choose>

<xsl:choose>
<xsl:when test="@Devoiced='1'">
<xsl:value-of select="translate (@PhoneEntity, $smallcase, $uppercase)"/>
</xsl:when>
<xsl:otherwise>

<xsl:choose>
<xsl:when test="contains (@PhoneEntity, 'S')">
<xsl:text>&lt;</xsl:text>
<xsl:if test="@PhoneEntity='Sc1S'">
<xsl:text>c</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SpzS'">
<xsl:text>pz</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SuvS'">
<xsl:text>uv</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SsvS'">
<xsl:text>sv</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SfrS'">
<xsl:text>fr</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='aSfrS'">
<xsl:text>a,fr</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='iSfrS'">
<xsl:text>i,fr</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='uSfrS'">
<xsl:text>u,fr</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='eSfrS'">
<xsl:text>e,fr</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='oSfrS'">
<xsl:text>o,fr</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SfvS'">
<xsl:text>fv</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='S?S'">
<xsl:text>?</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SNS'">
<xsl:text>N</xsl:text>
</xsl:if>
<xsl:if test="@PhoneEntity='SbS'">
<xsl:text>b</xsl:text>
</xsl:if>
<xsl:text>&gt;</xsl:text>
</xsl:when>
<xsl:otherwise>
<xsl:value-of select="@PhoneEntity"/>
</xsl:otherwise>
</xsl:choose>

</xsl:otherwise>
</xsl:choose>

```

ESPPS/wave+形式のラベルファイルのヘッダを出力

IPU 要素についてのテンプレート宣言

全ての Phone 要素を対象にループ処理

PhoneEndTime 属性を持つ Phone 要素について

PhoneEndTime 属性の値を出力

EndTimeUncertain 属性を持つ場合には赤色(117)で表示するように色番号を出力し、そうでない場合は黒色(123)で出力

PhoneEntity 属性の値を出力

```

<xsl:text>&#x0a;</xsl:text>
</xsl:if>
</xsl:for-each>
</xsl:if>
</xsl:template>
<xsl:variable name="smallcase" select="'abcdefghijklmnopqrstuvwxyz'" />
<xsl:variable name="uppercase" select="'ABCDEFGHIJKLMNOPQRSTUVWXYZ'" />
</xsl:stylesheet>

```

図 7. 分節音ラベルを生成する XSL

### 5.1.3. 無声化母音周辺の情報抽出

図 8 に、CSJ の XML 文書から分節音情報を利用して無声化母音とその周辺の情報を抽出して出力する XSL スタイルシートを示す。このようにして、XML 文書内に記述された情報を任意の形式で出力することができる。

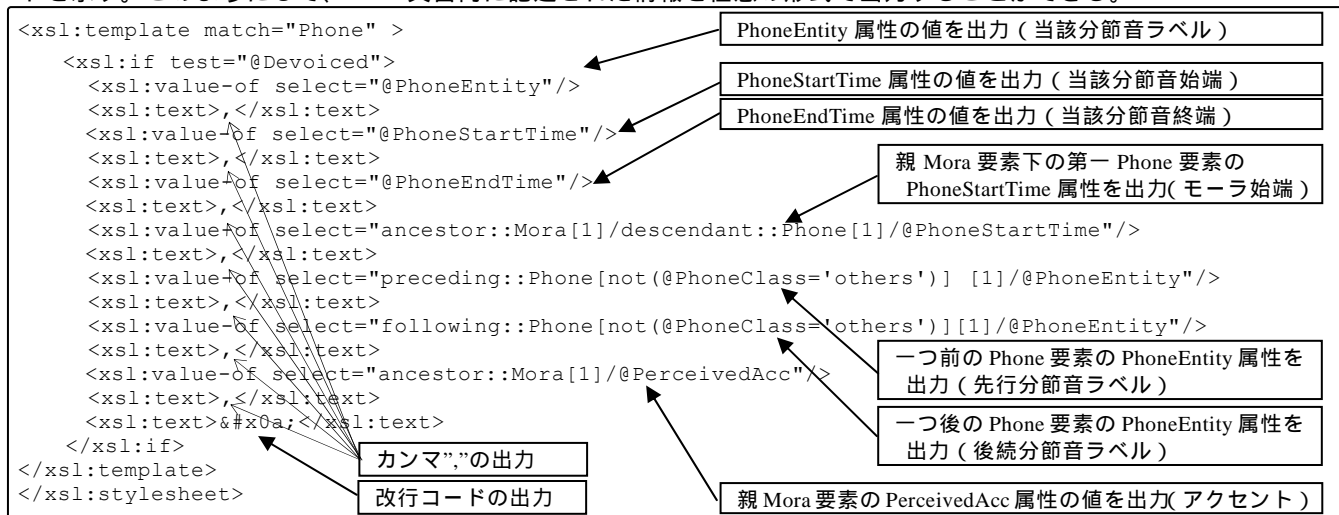


図 8. 無声化母音周辺の情報を抽出する XSL

## 5.2. XML 文書に対する検索 (第4刷サポート対象外)

今回提供する XML は形態論情報から分節音情報、韻律情報にまで至る階層構造を持っているため、これを利用して複数の階層にまたがる検索を行うことができる。たとえば、「価格が高くない」の「高い」という短単位が特定の品詞の場合にアクセント位置がどこにあったかという検索が可能である。そこで、XML に対して表示・検索を行う環境(CSJ XML Browser ver. 0.2; 以下 XML ブラウザと略す)を同梱する。XML ブラウザの基本仕様を表 2 に示す。

表 2. 入出力条件・動作条件仕様

環境	J2RE1.4.0 以上
接続先	1 XML ファイル
対象 XML	CSJ XML 形式
クエリ	XPath1.0
表示整形	XSLT 1.0
出力	CSV, HTML テキストファイル
推奨環境	P-4 2.6GHz 以上, メモリ 1GB 以上, Windows XP Home/Professional Edition

( 1.5 以上では Xalan-Java のインストールが別途必要。http://xml.apache.org/xalan-j/ を参照。)

XML の検索においては、検索条件の入力が XPath 1.0 [http://www.w3.org/TR/xpath] (最新の仕様は XPath 2.0 W3C Working Draft [http://www.w3.org/TR/xpath20/]) もしくは XQuery 1.0 [http://www.w3.org/TR/xquery/] 規格で行われることが多い。XML ブラウザは XPath 1.0 を用いて XML ファイルに対する検索を行う。なお、CSJ のウェブページを通じて今後配布予定のアップグレード版では XQuery をサポートする予定である。

以下では、Windows XP Home Edition 上での使用画面例を用いて説明する。



### 5.2.1. 起動およびファイルの読み込み

XML ブラウザは以下のフォルダ構成を前提とする。

1. 音声ファイルは .wav ファイル
2. .wav ファイルは XML ファイルと同じフォルダにある。
3. CSJ 転記ブラウザ環境 MonoForC (以下ものほしと呼ぶ)のインストールがされており、パスが設定されていること。

項目 3 に関しては、XML ブラウザからものほしを呼び出す際に必要になるが、未設定の場合でも XML の検索自体は可能である。なお、ものほしの利用方法については MonoForC マニュアルを参照のこと。

起動は、XML ブラウザのアイコンをクリックするか、コマンドプロンプトからと入力することで行う(図 9)。

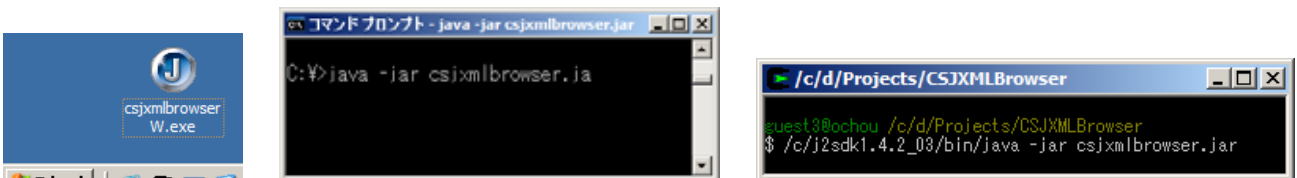


図 9. 起動方法(左:アイコンをクリックする場合、中・右:コマンドから起動する場合)

起動後の初期画面は図 10 のようになっている。起動後は 6 種類のタブ

1. ファイル選択タブ[読み込み]
2. 検索クエリ作成タブ[クエリ設定]
3. 結果 XML 表示タブ[結果 XML]
4. 表示整形用 XSL タブ[XSL でフォーマット]
5. 結果 HTML 表示タブ[結果]
6. IPU 可視化タブ[IPU]

を左から順に切り替えることで検索結果を得られるようになっている。図 10 の初期画面では「1.ファイル選択」タブが表示されている。

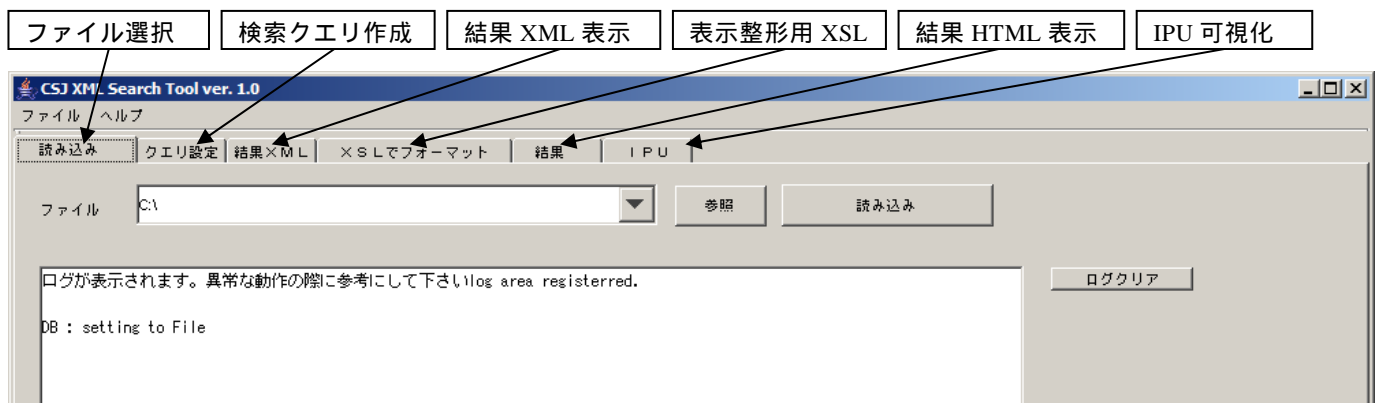


図 10. XML ブラウザ起動後初期画面

ファイルの選択は

- 方法1. 「ファイル」メニューの「XML ファイルを開く」を押す
- 方法2. ファイル選択タブ内の「参照」ボタンを押す
- 方法3. 参照ボタン左の白地テキスト部分にファイル名(パスも入力)を入力し、「読み込み」ボタンを押すことで行う(図 11)。



図 11. ファイルの選択

なお、CSJ の XML ファイルは巨大であるため読み込みには 3 ~ 10 秒程度かかる。

### 5.2.2. 検索クエリ生成タブ

ファイルを読み込み後はクエリ作成タブをクリックして開く。初期画面ではクエリ式は設定されていない(図 12)。

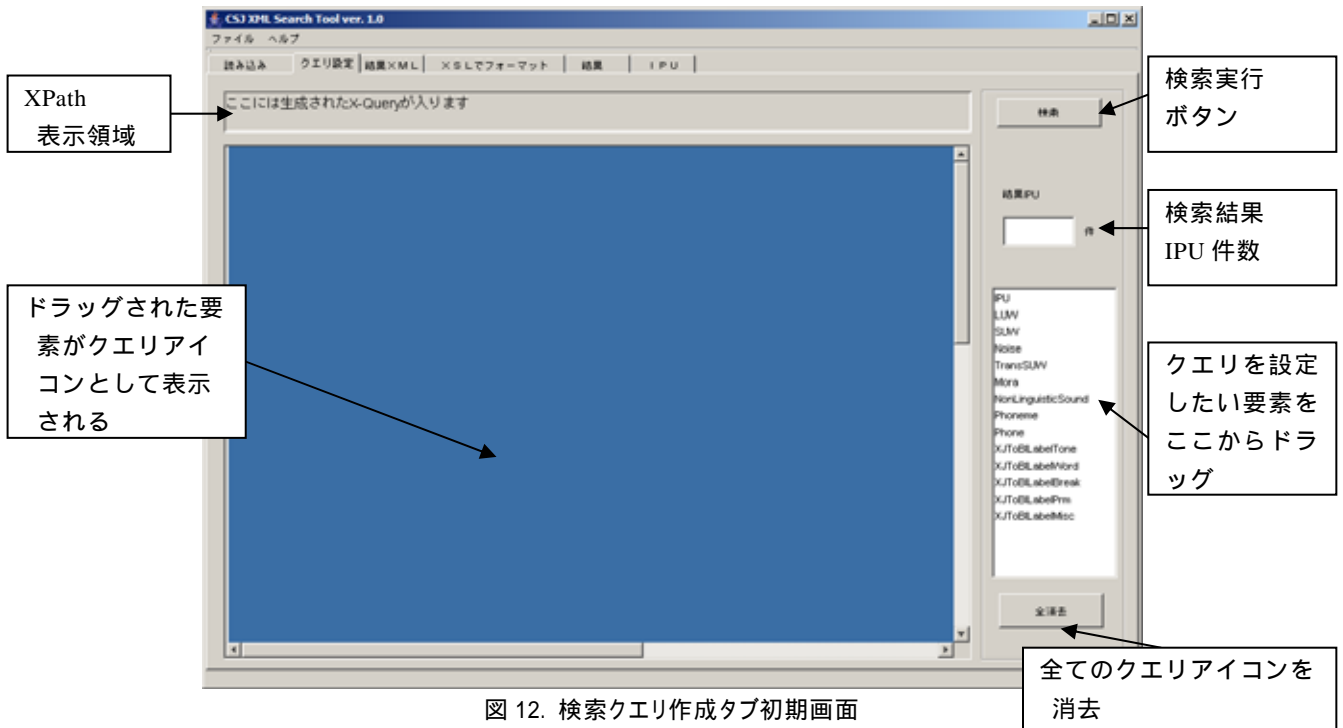


図 12. 検索クエリ作成タブ初期画面

クエリの生成はクエリアイコン(図 13)の生成とクエリアイコン間のリンク付けによって行う。生成されるクエリは XPath 1.0 形式であるが、contains()関数の簡略記号として、テキスト包含演算子(~=)を使える点が XPath と異なっている。

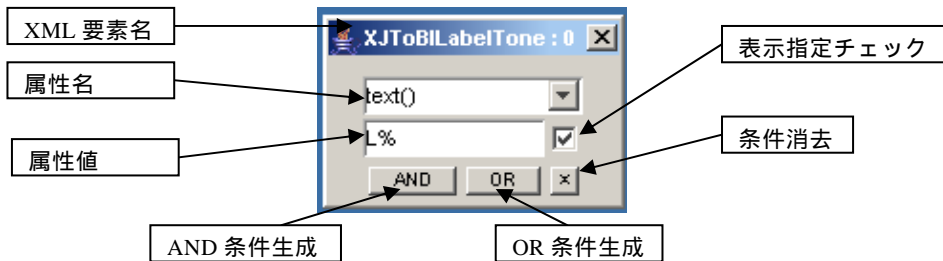


図 13. クエリアイコン基本構成図

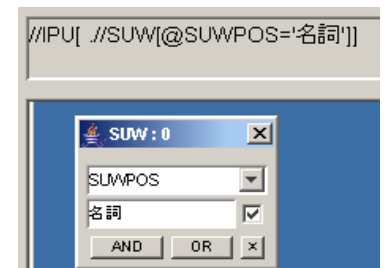


図 13-2. クエリアイコンとクエリ式

クエリアイコンの生成は画面右側の要素リストから要素をドラッグする(図 14. Step 1)か、要素リストをダブルクリックすることで可能である。次にクエリアイコンにおいて属性名を上部リストボックスから選択し、その後属性値を下部テキストボックスに入力する(図 14. Step 2)。このとき、上部 XPath 表示領域にはクエリアイコンの状態に対応する XPath 式がアイコンの状態と同期する形で表示される。例えば、SUW 要素の SUWPOS 属性を指定後、SUW 要素のテキストボックスに'名詞'を入力すると、追隨して

//IPU[//SUW[@SUWPOS='名詞']]

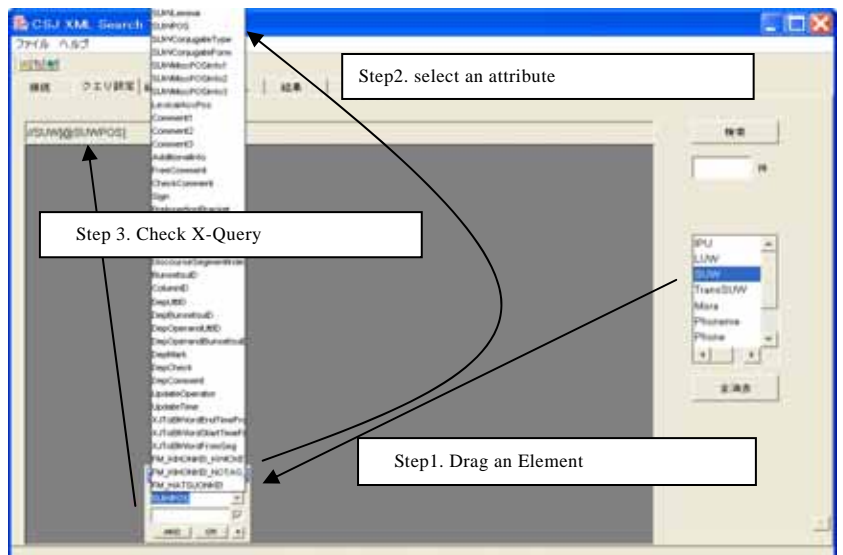


図 14. ドラッグ&ドロップによるクエリ構築

が表示される(Step 3., 図 15)。このクエリ式はリアルタイムに更新されるので、クエリアイコンとの対応を常時確認できる。

### 5.2.3. AND,OR 条件の指定

要素アイコンには AND 条件生成用チェックボックス、OR 条件生成用チェックボックス、表示指定チェックボックスが付属する。AND、OR 条件を単独で用いた場合のクエリ式を図 15 に示す。これらの条件は、それぞれある短単位中での条件指定である。これに対し、『ある IPU 内に「タイトル」という短単位があり、かつ何らかの名詞も存在する』条件指定は図 16(上)のように、また、『「タイトル」という短単位を含む IPU、または「名詞」を含む IPU』の場合は図 16(下)のようになる。

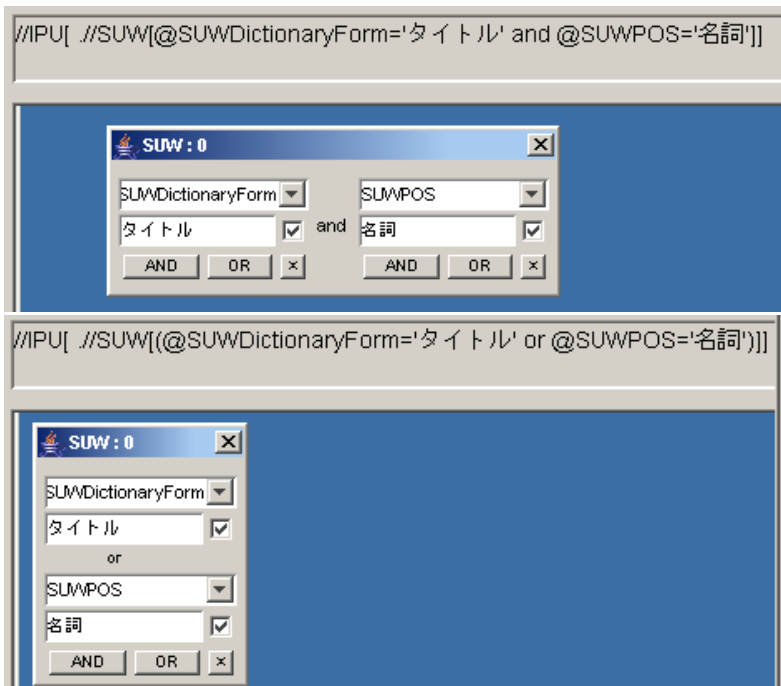


図 15. 上:AND 条件、下:OR 条件

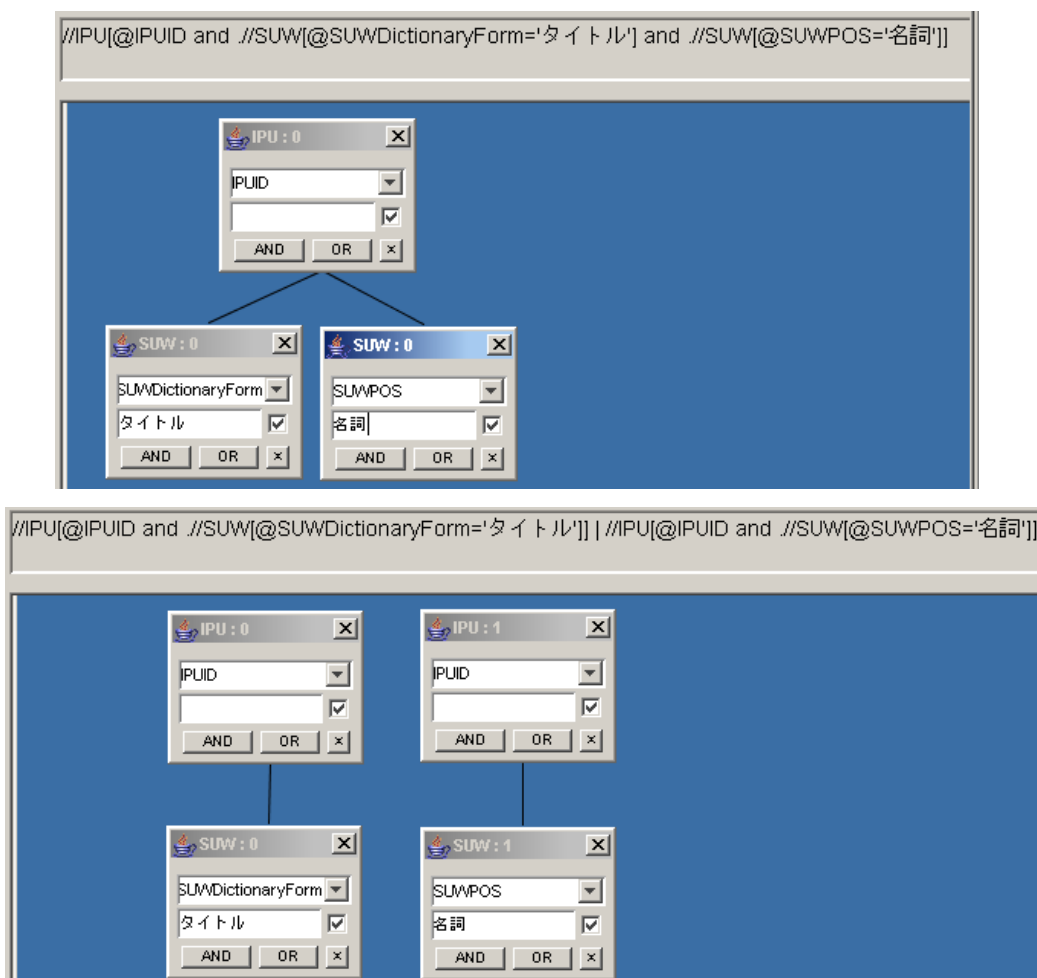


図 16. 上:1 IPU 内に「タイトル」と名詞が共起、下:「タイトル」を含む IPU または「名詞」を含む IPU

また、OR 条件は AND 条件より先に適用される (図 17(上))。これを逆転するには、図 17(下)のように配置する。

最後に表示指定チェックボックスと、要素の変更について述べる。表示指定チェックボックスをオフにされた属性制約は、検索条件としては用いられるが、その後の結果表示の際には表示されない。また、各クエリアイコンは生成時に自動的に親クエリアイコンが決められるが、これが不適切な際には親アイコンとのリンク線をクリック(アイコン自体にフォーカスが当たっている場合にはダブルクリックが必要)することで親クエリアイコンを変更することができる(図 18)。

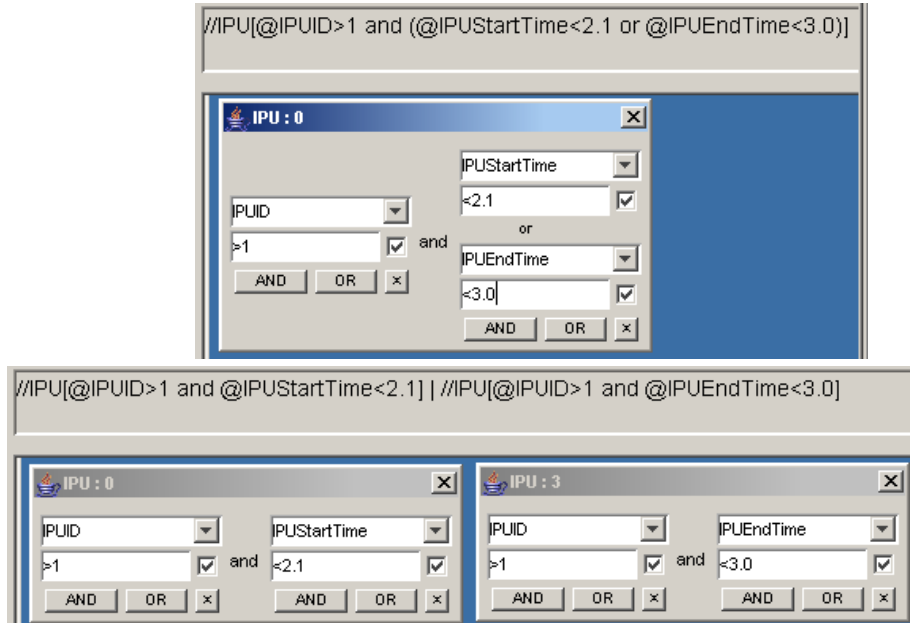


図 17. OR 条件を先に適用(上)、AND 条件を先に適用(下)

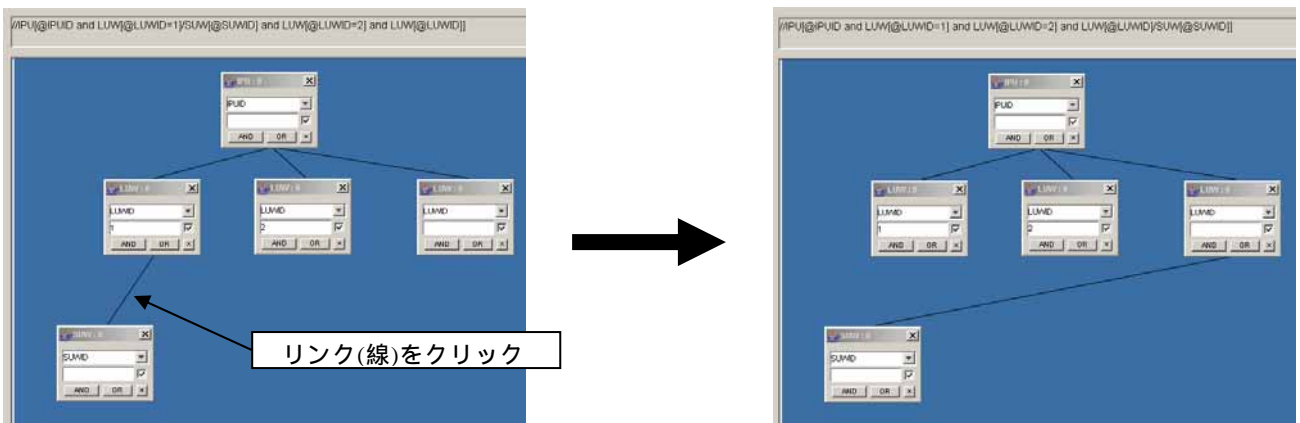


図 18. マウスクリックによる親要素の変更

これらの組み合わせでクエリ式を構成した後は、「検索」ボタンを押すことで検索が実行される。検索結果は IPU 件数として表示される。件数表示までには数秒から数十秒の時間がかかることがある。この時間は検索の複雑さと計算機の性能、そして最も大きくは検索結果件数に依存する。したがって、複雑な検索を行いたい場合には、前もって IPU < 10 などの制約条件を課しておき、結果件数を小さくすることで必要な時間を予想することをお勧めする。

### 5.2.4. 結果 XML 表示タブ

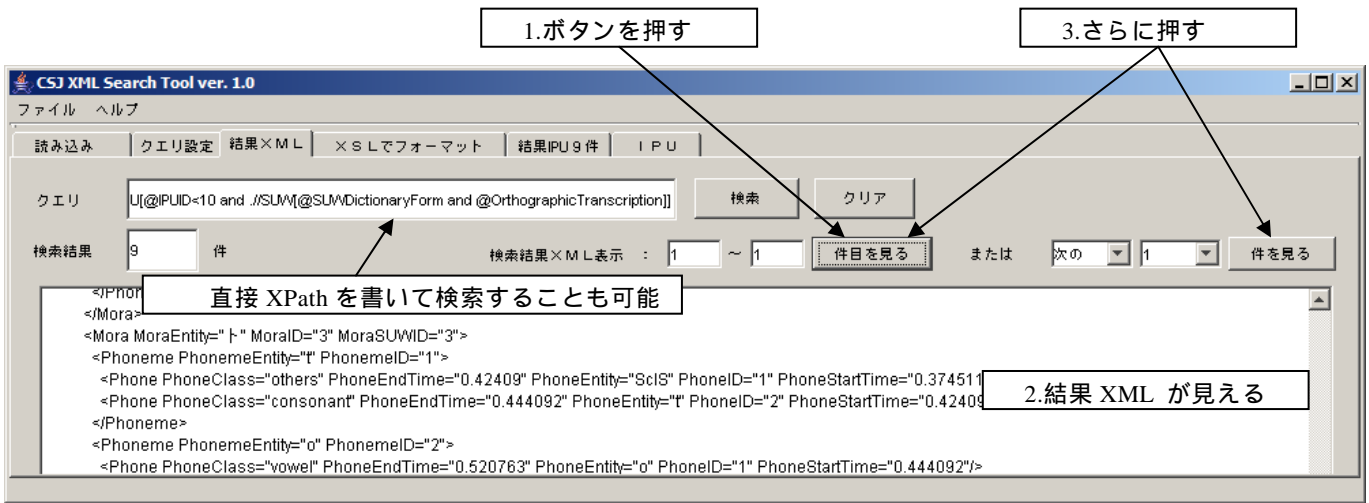


図 19. 結果 XML の表示

検索結果が表示されたら、ひとつ右のタブ「結果 XML」をクリックすることで、結果 XML 表示画面に移ることができる(XML 表示形式を必要としないユーザはこの画面を飛ばして結果 HTML 表示タブへ移ってもよい)。デフォルトでは表示画面には何も表示されていないが、「件目を見る」ボタン、または「件を見る」ボタンを押すことによって指定された範囲の XML を表示させることができる(図 19)。なお、ここで XML を表示しなくても、結果 XML をファイルへ書き出すことは可能である。

また、画面上部のクエリ入力部分に直接 XPath を入力することや、ファイルから XPath 式を読み込むことも可能である。この場合にはクエリアイコンの同時変更はされないが、簡単な XPath 式のテストや、過去に作成した XPath をファイルに保存しておき、次回に読み出して再利用するなどの利用法が考えられる(後述)。

### 5.2.5. 表示整形用 XSL タブ

「XSL でフォーマット」タブをクリックすると、XSL 表示画面が現れる(図 20)である。ここでは XSL の編集を行う。しかし、編集の必要がない場合にはこのタブを開く必要はない。しかし、本ツールは最低限の XML 関連仕様(XPath1.0, XSL1.0)も利用できるようになることも狙いの一つとしているので、自動生成された XSL スクリプトをここで自由に編集して表示情報を追加できるようにした。編集によって、たとえば「検索条件に指定したモーラの次のモーラ」などの追加情報を表示できる。デフォルトで表示される XSL は検索用クエリから自動生成されたものだが、それには修正のごく簡単なヒントがコメントとして載っているため、これを元に修正方法を習得していただきたい(図 20)。その際には XPath1.0, XSLT1.0 の関数が利用可能である。たとえば、not(), position(), last(), contains(), text(), starts-with() などを用いてより複雑な表示を行える。また、このようにして修正した XSL を保存することによって、別の機会に再利用することが可能である(後述)。編集が終了したら、適用ボタンを押すことによってこの XSL を結果 IPU に適用する。結果は「結果 IPU」タブで確認することができる。なお、この操作によって検索結果 IPU 集合が変更されることは無いので、編集 確認の作業を繰り返すことで XSL のデバッグも行える。

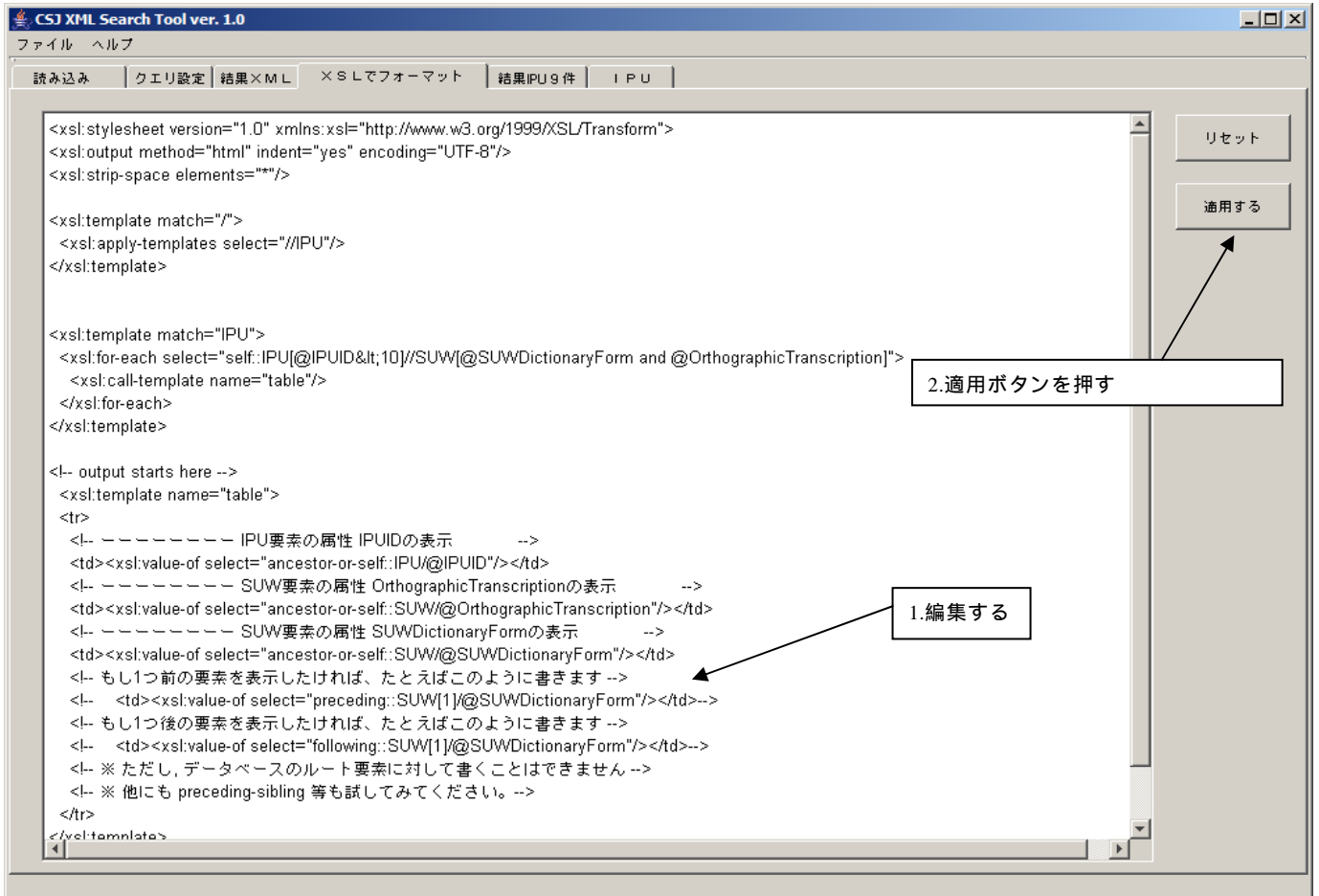


図 20. XSL 修正用表示画面

### 5.2.7. 検索結果生成・出力

先ほどの XSL を適用した結果はデフォルトでは表形式の HTML として表示される。しかし、これは CSV 形式に変更することも可能である(図 21)。CSV テキストファイルとして保存することで、他のアプリケーションで利用できる。

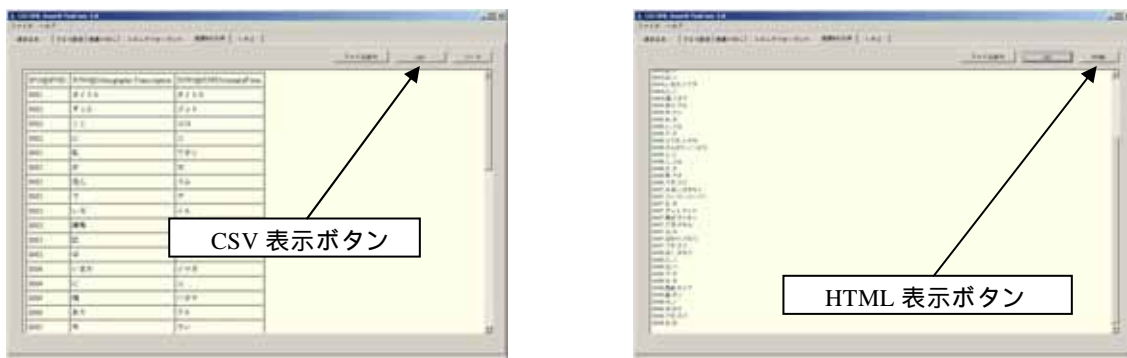


図 21. HTML 形式での結果表示(左)と CSV 形式での結果表示(右)

この変換は右上部の「CSV,HTML」ファイルを押すことで行う。

### 5.2.8. IPU 可視化タブ

一番右端にあるタブであり、検索に該当した IPU の階層構造を示す。また各ラベルをクリックすることによって該当部分のサウンドファイルを再生する。この表示画面はデフォルトでは何も表示されていないが、結果 XML 表示タブ画面と同様に「次の～件を見る」ボタンを押すことで、該当する IPU が表示される(図 22)。また、音声波形は「ものほしでみる」ボタンを押して、ものほしを起動することで見られる。ものほしについては Monoforc のドキュメントを参照されたい。

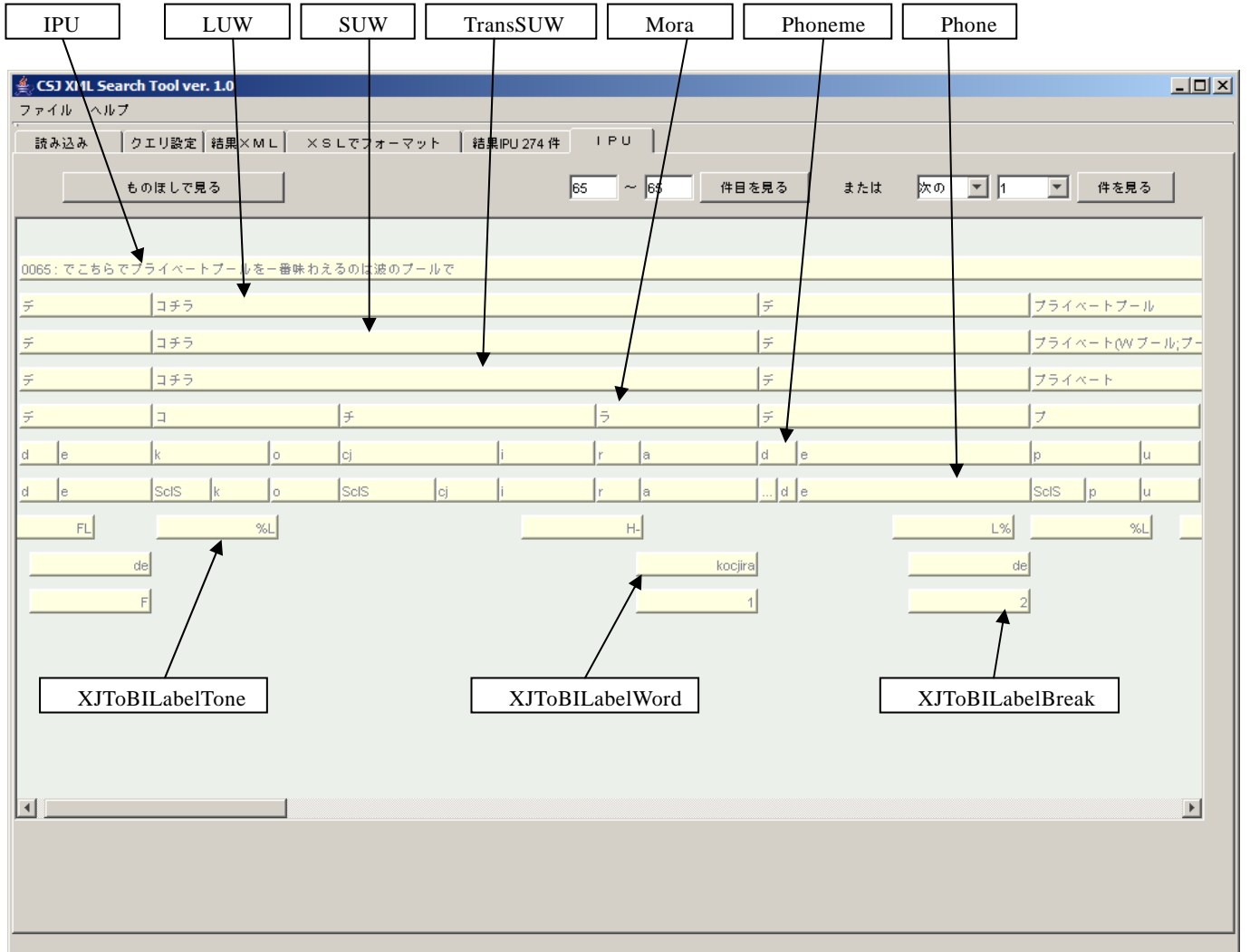


図 22. IPU 可視化画面での表示(この例では XJToBILabelPrm と XJToBILabelMisc 要素は表示されていない)

### 5.2.9. 各種ファイルの保存・読み込み

メニューの「ファイル 保存」を開くと、結果として保存できる形式が一覧できる。検索終了後であればこれらの形式で保存が可能である。保存 IPU 件数が多い場合には場合には処理時間がかかることがある。

また、「ファイル 開く」メニューでは、過去に保存した XSL ファイルや XPath 式を検索結果画面上の XPath 編集画面や XSL 修正画面などへ読み込むことができる。これによって、過去に作成した有用な XPath 式や XSL を異なる XML ファイルに対して適用することが可能である。

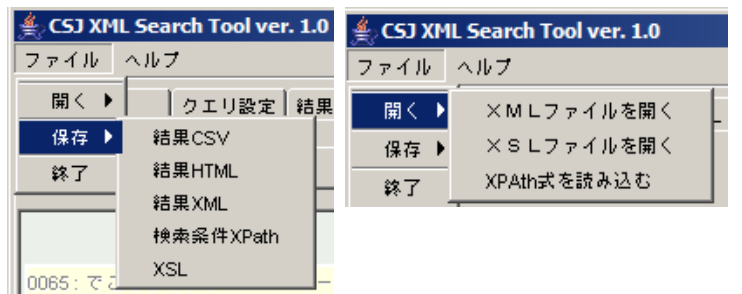


図 23. 各種ファイルの保存と再利用

特殊な利用方法としては、検索にヒットした転記を出力することが可能である。検索結果が得られたあとに、「ファイル」「開く」「XSL ファイルを開く」メニューから、同梱 XSL ファイル /TOOL/XSL/ipturn2xsl.xml を指定して、表示整形用 XSL タブで「適用」を押すことで、転記出力が得られる(図 24)。

### 5.2.10. 利用例 1. 「ジャナイ」

「ジャナイ」という表現の検索をする際には図 24 のようにクエリを構成する。前後の発音形(ここでは OrthographicTranscription 属性)も見たい場合は、XSL テキストを編集する(図 25(下))。その結果、表示情報が変更された出力を得ることができる(図 25(右))。一度検索した結果ノード集合は DOM として保持されているため、XSL を再設定して出力を変更することも可能である。

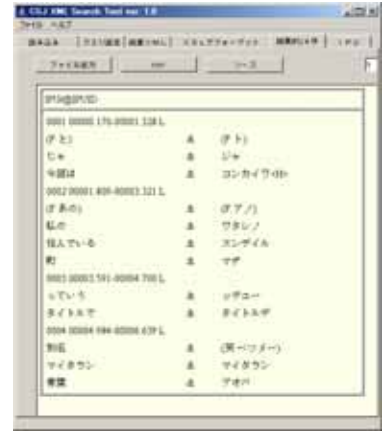


図 24. 転記の出力

The screenshot shows the CSJ XML Search Tool interface. The query editor contains the following XQuery: `/IPU[LUW[@LUWDictionaryForm]/SUW[@PhoneticTranscription='ジャ' and @SUWDictionaryForm and @SUWPOS]]`. Below the editor, there are two configuration windows: 'LUW : 0' and 'SUW : 0'. The 'SUW : 0' window shows the search criteria: `PhoneticTranscription` set to 'ジャ', `SUWDictionaryForm` and `SUWPOS` are empty, and the operators are 'and'. To the right, a table displays the search results:

助動詞	デハナイ	ダ	キ	ジャ	ナイ
助動詞	デハナイ	ダ	ニセモノ	ジャ	ナク
接続詞	ジャ	ジャ		ジャ	コレ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	デハナイ	ダ	ワケ	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン<H>	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	ン	ジャ	ナイ
助動詞	ノデハナイ	ダ	スル	ジャ	ナイ

```
<xsl:template match="IPU[//LUW//SUW/@PhoneticTranscription='ジャ']">
  <tr>
    <td><xsl:value-of select="./ancestor-or-self::SUW/@SUWPOS"/></td>
    <td><xsl:value-of select="./ancestor-or-self::LUW/@LUWDictionaryForm"/></td>
    <td><xsl:value-of select="./preceding::SUW[1]/@PhoneticTranscription"/></td>
    <td><xsl:value-of select="@PhoneticTranscription"/></td>
    <td><xsl:value-of select="./following::SUW[1]/@PhoneticTranscription"/></td>
  </tr>
</xsl:template>
```

図 25. 「～ジャ」の検索  
(左上:クエリのアイコン、下:編集後の XSL、右上:結果)

### 5.2.11. 「タカイ」のアクセント位置

次に「高い」のアクセント位置の検索の例を示す(図 26)。アクセント位置は XJToBILabelWord 要素の PerceivedAccPos 属性で指定できるので、この場合、XPath 式は以下ようになる。

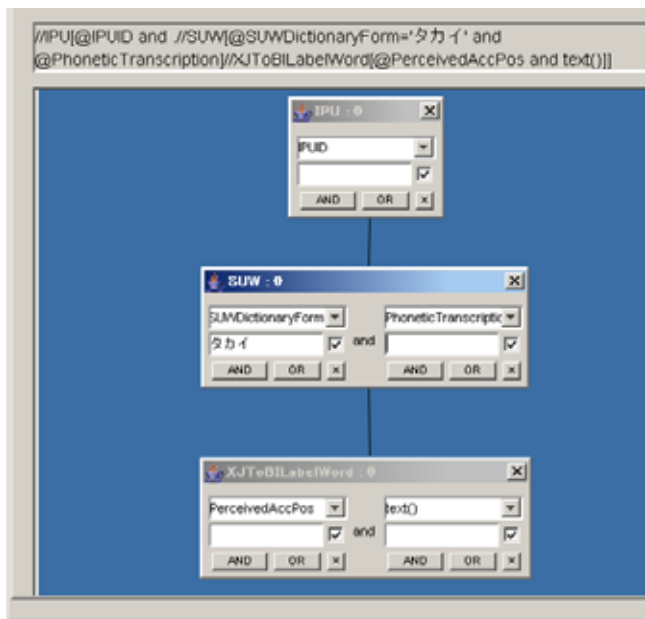
```
//IPU[@IPUID and //SUW[(@SUWDictionaryForm='タカイ'and @PhoneticTranscription)]/XJToBILabelWord [ @PerceivedAccPos and text() ] ]
```

この場合も、「タカイ」に後続する短単位を表示させるために、XSL には

```
<xsl:value-of select="./following::SUW[1]/@PhoneticTranscription"/>
```

を追加している。ここで、PerceivedAccPos 属性の値は語頭から数えたアクセント位置を示している。





0275	AC00MAR061	タカイ	タカイ	takai	2	レストラン
0277	AC00MAR061	タカイ	タカ	taka	1	サ
0277	AC00MAR061	タカイ	タカイ	takai	2	レストラン
0292	AS00MAR015	タカイ	タカ	taka	1	サ
0312	AS00MAR015	タカイ	タカイ	takai	2	ト
0041	NL00MAR081	タカイ	タカイ	takai	2	セード
0086	NL00MAR081	タカイ	タカク	takaku	1	テ
0105	PS99SEP023	タカイ	タカイ	takai	2	スージ
0385	PS99SEP023	タカイ	タカク	takaku	1	ナリ
0402	PS99SEP023	タカイ	タカク	takaku	0	オー
0404	PS99SEP023	タカイ	タカク	takaku	2	ナリ
0408	PS99SEP023	タカイ	タカク	takaku	0	オー
0409	PS99SEP023	タカイ	タカク	takaku	1	ナリ
0457	PS99SEP023	タカイ	(W アタタック;タカク)	(W atataQku)	3	ナリ

図 26. 「タカイ」のアクセント位置の検索(左: クエリのアイコン、右: 結果;数字はアクセント位置を示す)

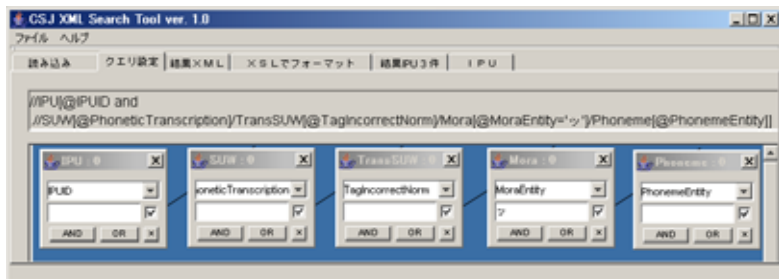
### 5.2.12. 言い誤り時の促音化「ッ」の検索

言い誤りにおける促音化の検索の場合はモーラ「ッ」を持ち、かつ、言い誤りに付与される(W)タグを表す属性 @TagIncorrectNorm があるものを検索すればよい。作成するクエリは

```
//IPU[./TransSUV[@TagIncorrectNorm]//Mora[ @MoraEntity='ッ']]
```

となる。表示指定まで含めると、アイコン配置は図 27 のようになる。この場合にも、XSL にて結果表示項目を追加することにする。追加内容は前 2 音節と後 1 音節である(図 27 上)。

```
<xsl:template match="//IPU[(@TalkID and @IPUID)]//SUV[ @PhoneticTranscription]//TransSUV[ @TagIncorrectNorm]//Mora[ @MoraEntity='ッ']//Phoneme[ @PhonemeEntity]">
<tr>
<td><xsl:value-of select="./ancestor-or-self::SUV/ @PhoneticTranscription"/></td>
<td><xsl:value-of select="./ancestor-or-self::TransSUV/ @TagIncorrectNorm"/></td>
<td><xsl:value-of select="./preceding::Phoneme[2]/ @PhonemeEntity"/></td>
<td><xsl:value-of select="./preceding::Phoneme[1]/ @PhonemeEntity"/></td>
<td><xsl:value-of select="./ancestor-or-self::Phoneme/ @PhonemeEntity"/></td>
<td><xsl:value-of select="./following::Phoneme[1]/ @PhonemeEntity"/></td>
</tr>
</xsl:template>
```



(W モッテキ,モクテキ)	モクテキ	m	o	Q	t
(W オブジェクト,オプジェクト)	オプジェクト	zy	e	Q	t
(W カイシャク,カイヤク)	カイヤク	sy	a	Q	g
(O (W ホワツ,ホワット)	ホワット	w	a	Q	d
(W セッキョク,セッキョク)	セッキョク	s	e	Q	ky
(W セッキョク,セッキョク)	セッキョク	ky	o	Q	t
(W カパレージ,カパレージ)	カパレージ	r	e	Q	qz
(W カク,カク)	カク	k	a	Q	k
(W イッシュョー,イッシュョー)	イッシュョー	e	i	Q	sy
(W (? ヅツ,ツツ)	ツツ	t	a	Q	t
(W シツ,ヒト)	ヒト	sj	i	Q	t
(F (W エーット,エーット))	エーット	e	H	Q	t
(W シュエートツ,シュエートク)	シュエートク	t	o	Q	k
(W チッケン,ジッケン)	ジッケン	cj	i	Q	k
(W コッコー,コッコー)	コッコー	k	o	Q	k
(W グループ,グループ)	グループ	r	u	Q	p
(W グルーパー,グループ)	グループ	r	u	Q	p

図 27. 言い誤りと促音化の検索例(上: 編集後の表示整形用 XSL, 左下: クエリアイコン, 右下: 表示結果例)

## 6. おわりに

CSJ の多様な研究用情報を効率よくメンテナンスし、また検索の利便性を上げるための XML の設計および利用方法について解説した。スペースの都合上、利用方法については簡単に触れる程度にとどめた。今後、CSJ の利用に関するチュートリアル等を企画し、その中で XML 文書の利用方法を紹介することを予定している。

## 7. 第三刷における変更点

### 7.1 種々の情報修正に伴う XML 文書修正

第三刷において、転記、短単位、分節音ラベル、韻律ラベルなどの種々の情報の修正が行われた。これらの修正はいずれも XML 文書に反映させた。

### 7.2 バグの修正

#### 7.2.1 対話データのインタビュアー(Lチャンネル)側音声ラベル情報修正

対話データには 2 名の対話者の発話の情報が存在するが、音声ラベルの情報はインタビュイー(Rチャンネル)についてのみ公開する予定であった。しかしながら第二刷まで、インタビュアー(Lチャンネル)の音声ラベルが十分に検証されていないまま XML 文書内でのみ公開されてしまっていた。今後、十分に精度を検証したうえで公開を目指すとして、現時点では XML 文書内に記されたインタビュアーの音声ラベルの情報(Phone 要素以下)を削除することにした。

#### 7.2.2 対話データの話者情報修正

対話データには2名の対話者の発話の情報が記述されているが、7.2.1に述べたとおり、Phone要素以下で記述される音声ラベルの情報についてはインタビュイーの情報のみが記述される。したがって、Talk要素のSpeakerID, SpeakerBirthPlace, SpeakerBirthGeneration, SpeakerSexの各属性にはインタビュイーの情報を記載すべきであるが、第二刷までのXML文書にはインタビュアーの情報が記載されていた。これをインタビュイーの情報を記載するよう修正した。

#### 7.2.3 Phone 要素の情報修正

Phone 要素には分節音ラベルの情報が記述されるが、分節音ラベルとして提供される"\*\*.seg"のファイル内の情報と整合がとれていないケースが多数あった。分節音ラベルを修正したケースとあわせて、XML 文書内の Phone 要素の情報を、分節音ラベルファイル(\*.seg)と完全に整合がとれるよう検証して修正した。

#### 7.2.4 Phone 要素の PhoneID 属性

発声中に観察された声帯のきしみやボーカルフライなど、独立の音素と認定されないケースにおいて、Phone要素にPhoneID属性が付けられていなかった。全てのPhone要素についてPhoneID属性を持つように修正した。

## 7.3 仕様の変更

### 7.3.1 Divided 属性の付加(第二刷での対応)

第二刷において、韻律ラベルのトーン層の複合ラベルを表現するために XJToBILabelTone 要素に Divided 属性を追加した。第二刷における仕様の変更はこの点のみである。

### 7.3.2 Noise 要素の属性

Noise要素の一部で、統語情報が付与されているケースがある。しかしNoise要素は言語を構成する要素ではないため、これらの属性を持つべきではない。調査の結果、これらの属性は直後に現れるSUW要素が持つべき属性であることが分かった。従って、以下の属性についてはNoise要素から削除し、直後に現れるSUW要素に付与する変更を行った。

CU_PreBracket
ClauseUnitID
DS_Comment
DS_Purpose
DS_SubPurpose
DS_Subject1
DS_Subject2
DS_Subject3
Dep_Label
Dep_ModifiedBunsetsuUnitID
Dep_ObligateComment
SE_Subject1_10p
SE_Subject1_50p
SE_Subject2_10p
SE_Subject2_50p
SE_Subject3_10p
SE_Subject3_50p

### 7.3.3 SUW 要素の ClauseUnitID 属性

SUW要素のClauseUnitID属性は節単位IDをあらわしており、節単位マニュアル4. に記されているとおり「節単位の範囲を表す「節単位ID」は当該節単位の最初のSUW 要素の属性として記述されて」いる。これについて、全てのSUW要素がClauseUnitID属性を持つように仕様を変更した。

### 7.3.4 <FV>の表現

転記情報における<FV>が、IPU 要素内で独立のLUW 要素とされているケースと、LUW 要素内で独立のNoise 要素(SUW 要素相当)とされているケースにおいては、“ (?) ” と同じようにMora 要素で表現することにした。

## 7.4 未修正の問題

### 7.4.1 促音周辺の表現

分節音ラベルにおける促音の表現については、後続する子音によってラベリング方法が異なる(分節音マニュアル4.2 B) 参照)。この基準にしたがうと、「イッテ」のような発音に対して以下のような分節音ラベルが付与されることがある。

|i|<pz>,<cl>|Q,t|e| (分節音ラベル(\*.seg))

この場合、実態としては、“i”の終端から“t”の始端までの区間を、“Q”と“<pz>”と“<cl>”が分け合うと考えるのが妥当である。したがってXML 文書内では時間情報を持つPhone 要素として表現するため、出現順序に基づいて以下のようなラベル列と見なすことになる。

|i|Q|<pz>|<cl>|t|e| (Phone 要素(\*.xml 内))

|i|Q| t|e| (XJToBILabelWord 要素(\*.xml 内))

このとき、“Q”、“<pz>”、“<cl>”の順に記述するか、“<pz>”、“Q”、“<cl>”の順に記述するかが問題になるが、第二刷では前者が41回、後者が285回とゆれていた。修正のコストを考えて以下のような“<pz>”、“Q”、“<cl>”の順に統一して記述することにした。

|i|<pz>|Q|<cl>|t|e| (Phone 要素(\*.xml 内))

|i| Q| t|e| (XJToBILabelWord 要素(\*.xml 内))

なお、この方法で表現する際、<pz>を「ッ」のMora 要素の最後に所属させるか次のMora 要素の先頭に所属させるべきか検討したところ、次のモーラの先頭に所属させているケースは7ケースしかなかった。したがってこれらの修正も含めて、「ッ」の最後に所属させることにした。

なお、「(D ケッ)」のようなケースでは、Q,<pz>の順とする。

## 謝辞

第三刷の仕様変更およびバグ修正は、理化学研究所西川賢哉氏、野口広彰氏との共同作業による。ここに記して感謝の意を表す。

## バグ情報報告に際してのお願い

各要素にはIDを持たせています。バグ情報の報告時にはそれぞれの要素名とIDをお知らせください。

例) Talk/@TalkID="S01F011", IPU/@IPUID="0123", SUW/@SUWID="3", この品詞は接続助詞ではないか?

付録1:現代日本語のモーラとPhoneme要素PhonemeEntity属性値の対応  
 (上段:モーラ、下段:PhonemeEntity属性値(“+”は複数の要素があることを示す))

直音系列モーラ					拗音系列モーラ			周辺のモーラ								
ア	イ	ウ	エ	オ	ヤ	ユ	ヨ	イエ								
a	i	u	e	o	y+a	y+u	y+o	y+e								
カ	キ	ク	ケ	コ	キャ	キュ	キョ	クワ								
k+a	kj+i	k+u	k+e	k+o	ky+a	ky+u	ky+o	kw+a								
ガ	ギ	グ	ゲ	ゴ	ギャ	ギユ	ギョ	グワ								
g+a	gj+i	g+u	g+e	g+o	gy+a	gy+u	gy+o	gw+a								
サ	シ	ス	セ	ソ	シャ	シュ	ショ	シエ	スイ							
s+a	sj+i	s+u	s+e	s+o	sy+a	sy+u	sy+o	sy+e	s+i							
ザ	ジ	ズ	ゼ	ゾ	ジャ	ジュ	ジョ	ジエ	ズイ							
z+a	zj+i	z+u	z+e	z+o	zy+a	zy+u	zy+o	zy+e								
タ	チ	ツ	テ	ト	チャ	チュ	チョ	ティ	トゥ	チェ	ツァ	ツイ	ツェ	ツォ	テュ	
t+a	cj+i	c+u	t+e	t+o	cy+a	cy+u	cy+o	t+i	t+u	cy+e	c+a	c+i	c+e	c+o	ty+u	
ダ			デ	ド				ディ	ドゥ	デュ						
d+a			d+e	d+o				d+i	d+u	dy+u						
ナ	ニ	ヌ	ネ	ノ	ニャ	ニユ	ニョ	ニエ								
n+a	nj+i	n+u	n+e	n+o	ny+a	ny+u	ny+o	ny+e								
ハ	ヒ	フ	ヘ	ホ	ヒャ	ヒユ	ヒョ	ヒエ	ファ	フィ	フェ	フォ	フュ			
h+a	hj+i	F+u	h+e	h+o	hy+a	hy+u	hy+o	hy+e	F+a	F+i	F+e	F+o	Fy+u			
バ	ビ	ブ	ベ	ボ	ビャ	ビユ	ビョ	ヴァ	ヴィ	ヴ	ヴェ	ヴォ				
b+a	b+i	b+u	b+e	b+o	by+a	by+u	by+o	v+a	v+i	v+u	v+e	v+o				
パ	ピ	プ	ペ	ポ	ピャ	ピユ	ピョ									
p+a	p+i	p+u	p+e	p+o	py+a	py+u	py+o									
マ	ミ	ム	メ	モ	ミャ	ミュ	ミョ	ミエ								
m+a	m+i	m+u	m+e	m+o	my+a	my+u	my+o	my+e								
ラ	リ	ル	レ	ロ	リャ	リュ	リョ									
r+a	r+i	r+u	r+e	r+o	ry+a	ry+u	ry+o									
ワ				ヲ				ウィ	ウエ	ウォ						
w+a				w+o				w+i	w+e	w+o						

撥音	促音	長母音(第二要素)
N	Q	H