デジタル・ドキュメント 36－5
（２００２．１１．２９）

# 第 2 回国際ワークショップ“NLPXML-2002”の概要と NLP，Semantic Web の融合に向けての展開

中挾知延子[1]　野村直之[2]　浦本直彦[3]　Key-Sun Choi[4]

自然言語処理のための XML 技術（XML for NLP）ならびに XML のための自然言語処理技術（NLP for XML）をスローガンにかかげた国際ワークショップ NLPXML は第 2 回目として 2002 年 9 月 1 日，Coling2002 に併設して行われた（NLPXML-2002）．NLPXML ワークショップでは，XML と NLP を有効に組み合わせることにより，互いが抱えるボトルネックの部分を解決できる可能性を持っていることが共通の認識である．たとえば，NLP を XML に用いることでメタデータの自動生成やメンテナンスが容易になり，XML を NLP に用いることで，音声合成やコーパス処理におけるデータの抽出，整理がスムーズになる．本論文では NLPXML-2002 を振り返って，セッションごとに議論されたテーマについて発表論文の概要を述べ，それとともに Semantic Web を念頭においた次世代 Web の発展のために自然言語処理技術を駆使しようとするワークショップの動向について言及する．

# An Abstract of the 2nd International Workshop on NLP and XML with a Movement towards the Semantic Web

Chieko NAKABASAMI[1], Naoyuki NOMURA[2], Naohiko URAMOTO[3], Ken-Sun Choi[4]

The 2nd international workshop on NLP and XML (NLPXML-2002) was held in September 2002 in Taipei as a co-located workshop of Coling 2002. The main issue of NLPXML-2002 is about XML technologies for Natural Language Processing, XML for NLP, and Natural Language Processing Techniques for XML documents, NLP for XML. It is a mutual agreement among the researchers joining NLPXML that integration of NLP and XML has a promising method for solving bottlenecks the both have. For example, applying XML to NLP makes automatic generation and maintenance on meta-data easier, and applying NLP to XML makes extraction and arrangement on speech synthesis and corpus processing rapidly. This paper presents summaries of each of the presentations that took place at NLPXML-2002 and discusses research towards the achievement of the semantic web.

[1] 東洋大学，Toyo University
[2] 法政大学，Hosei University
[3] IBM，国立情報学研究所，National Institute of Informatics
[4] KAIST，NHK 技研，Korean Advanced Institute of Science and Technology，Science & Technical Research Laboratories

## 1. はじめに

NLP と XML に関するワークショップは今年で 2 回目を迎える．今回は 2002 年 9 月 1 日，台湾の台北において Coling2002（ACL の主催による第 19 回計算言語学に関する国際会議）に併設する形(NLPXML-2002) [1][2]で開催された．第 1 回目は 2001 年 11 月に東京で，NLPRS2001（環太平洋自然言語処理国際シンポジウム）に併設して行われた．第 1 回ワークショップの概要は[3]を参照されたい．第 1 回の開催から，自然言語処理のための XML 技術（XML for NLP）ならびに XML のための自然言語処理技術（NLP for XML）をスローガンにかかげた本ワークショップは，インターネット上でのデータ交換の標準フォーマットとしての地位をもはや確立したと言える XML と NLP とのシナジー効果が期待される. 例えば, 高機能なフリーの XML 関連処理系により，NLP システムの設計から実装にいたるまで、データのコンパイル，モデルの構築から，実験・評価に至る高速プロトタイピングが実現する．人間が介在する系で漸進的に曖昧性解消，高精度化する際の統一的な中間データ記述形式，コーパス記述言語として XML が果たす役割も大きい．また, XML ドキュメントの自動生成やメタデータ, オントロジの（半）自動メンテナンスなどで，産業界から多大な要請が寄せられているが，多義語の処理や制約の解決，一貫性の保持，シソーラス開発など，自然言語研究の知見が XML の実用化に貢献できそうである．

本稿では，ワークショップのセッションに沿って概要を述べていく．なお，招待講演としてセッション 2 において関洋平氏（青山学院大学）により“Multilingual Generation from XML-DB”と題して，XML データベースから XSL などを用いた多言語文書の生成を行う研究についての発表があった．関氏のワークショップへの貢献を称えるとともに，本稿では詳しい内容を掲載するには至らなかったことをご了解願いたい．また，ポスター及びプロジェクトセッションとしていくつかの大規模プロジェクトによる多言語語彙データベースや学術文書データベースにおける XML アノテーションや，言語学すなわちコーパス言語学や HPSG の視点からのアノテーションスキーマの実装，語の多義性解消を用いたオントロジのメンテナンスの提案が行われている．

## 2. セッション１：Tools and Corpora

浅いレベルから深いレベルに至るまでの目的に応じたアノテーションや，マルチモーダルなページのレイアウトをするためのアノテーションスキーマなどの，NLP におけるコーパス処理に XML は有効である．一方，NLP でよく用いられる文法を適用し，ルールによって XML 文書の変換が段階的に行われる．NLP にとってアノテーションの共有と再利用は重要であり，XML にとっての NLP ツールの標準化は欠かせないものである．

### (1) XML-Based NLP Tools for Analysing and Annotating Medical Language

**医学分野の文章を分析・アノテーションするための XML に基づいた NLP ツール**

Claire Grover, Ewan Klein, Mirella Lapata, Alex Lascarides (Univ. of Edinburgh, UK)

医学分野の自然言語文章を解析するために、XML 化したコーパスを用いた研究である．医学分野の文献アブストラクトを集めた OHSUMED コーパスである MEDLIN を XML 化し、さまざまな

レベルで自然言語処理をするために用いている. 浅いレベルである語や文の同定、品詞タグ付け、基本形への変換などの形態素解析から、深いレベルである統語解析や意味解析にいたるまでの処理に、それぞれの目的に応じた XML タグをコーパスに設定することで、処理の効率化を実現している.

**(2) A Brief Introduction to the Gem Annotation Schema for Complex Document Layout**

**多様な文章レイアウトのための GeM アノテーションスキーマ**

John Bateman(Univ. of Bremen, Germany), Renate Henshel (Univ. of Stirling, Germany), Judy Delin (Univ. of Stirling, Enterprise IDU, England)

マルチモーダルな文章レイアウトを XML で記述するためのスキーマと、そのスキーマに基づいたレイアウトに関するタグ付きコーパスの活用についての提案である。ドイツのブレーメン大学を中心とした GeM プロジェクト(Genre and Multimodality)で開発されている。GeM では、文章構造を Rhetorical, Layout, Navigation の３つの構造に分け、それらを横断するレイアウトの構成要素間のつながりや、レイアウトを決定する要因などの制約を取り入れる試みがなされている。３つの構造記述のためのアノテーションスキーマは階層構造を持ち、ベースユニットであるレイアウトの各要素の記述と、参照化されたベースユニットでグループ化されたもので構成される。スキーマでは、ベースユニットをグループ化することでさらに抽象度の高いレベルを記述している。XML 化されたレイアウトタグ付きコーパスは文章生成システムでの使用が想定されており、ユーザが使い勝手に応じてレイアウト要素を選択できたり、レイアウトの決定を支援したりすることが目的である。

**(3) Cascaded Regular Grammars over XML Documents**

**XML 文書におけるカスケード型正規文法**

Kiril Simov, Milen Kouylekov, Alexander Simov (Linguistic Modelling Laboratory, Bulgaria)

テキストコーパス処理に用いるカスケード型正規文法 CLaRK システムについての発表である。本論文でのカスケード型正規文法とは、C → R の形の規則であり、C は語句のカテゴリを示す語、R は正規表現を指す。CLaRK では XML 文書に対して正規文法が適用される。XML に対する 2 種類の適用ケースが想定され、１つはテキストノード、もう１つはエレメントノードの集合に対してである。テキストノードの場合は正規表現をベースにしたトークナイザで処理され、エレメントノードの集合の場合には XPath の表現形式でタグ付きの表現が文字の並びに置き換えられる。XML 文書を入力語の集合としてカスケード型正規文法が適用され、規則に従って変換されていく。カスケード型正規文法の適用により、XML 文書に複雑な制約付きの変換や並べ替えなどのカスタマイズが行えると述べられている。XML をルールベースで更新していく類似の研究として，ドイツ Freiburg 大学の Wolfgang May による XPATHLOG[4]がある.

## 3. セッション 2：Document Generation

XSLT が文書生成のために有する潜在能力は大きい．多言語対応の文書フィルタや XML テンプレートからの文章自動生成における XSL での実装は，効率が良く保守性が高いことが実証されている．

### (1) Cascading XSL Filters for Content Selection in Multilingual Document Generation

**多言語文書生成における内容選択のための XSL フィルタ**

**Guilleremo BARRUTIETA**
**(Mondragon Univ., Spain),**
**Joseba ABAITUA, JosuKa DIAZ**
**(Univ. of Deusto, Spain)**

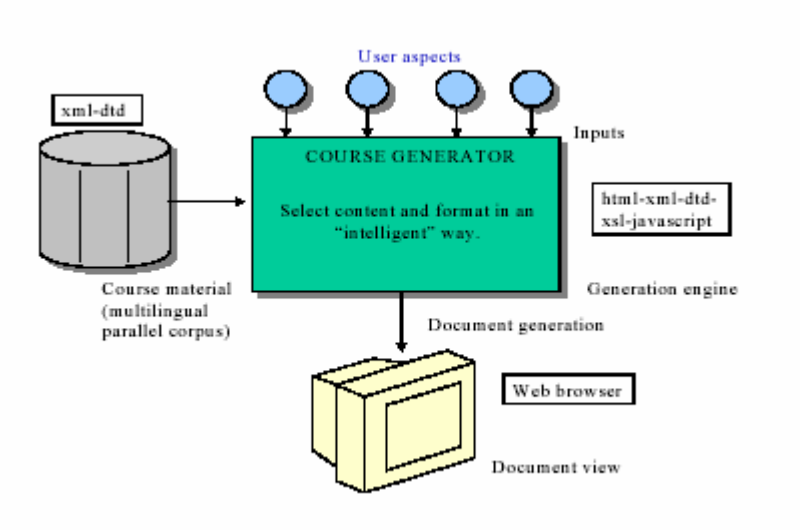


**図 1　XSL フィルタによる多言語文書生成のイメージ**

自然言語文章を RST(Rhetorical Structure Theory)に基づいて XML 化し、XSL を用いてユーザのニーズに応じてフィルタリングした文章を生成するシステム（図 1）の提案である．主として教育目的での利用が想定されている。

作成された XML ドキュメントは、実際に文書を使うユーザのプロファイルを基に CSA(Content Selection Algorithm)によってフィルタリングされる。フィルタリングの実装には XSL が用いられており, CSA による Parallel, Horizontal, Vertical という 3 種類のフィルタが XSL で実装されている．多言語とは，英語，スペイン語，バスク語の 3 種類である．

### (2) XtraGen – A Natural Language Generation System Using XML- and Java- Technologies

**XtraGen - XML と Java を用いた自然言語文章生成システム**

**Holger Stenzhorn (XtraMind Technologies GmbH, Germany)**

自然言語文章生成システム XtraGen について，文章生成のためのメカニズムとその実装について述べられている．XML を利用した文章生成には XSL が多く用いられているが、本論文では形態素をうまく操作できない、生成文のレベル分けができない、バックトラックがきかないという

XSL を用いて生成したときに生ずる問題点を克服するために独自のシステムを構築している。システムでは XML で作られたテンプレートをベースに生成が行われる。論文の後半では主として XtraGen における XML と Java の実装が説明されており、XtraGen をブースティングによる機械学習ソフト X-Booster での出力の生成に適用し、XtraGen の出力結果を評価している。

## 4. セッション 3：Discourse, Dialog and Speech

自然言語におけるマルチモーダリティを扱える XML の NLP への役割は大きい．とりわけ談話生成，音声処理のための XML によるアノテーションの利点は高く評価できる．さらに対話処理において VoiceXML を超えるものとして SALT が提案された．

### (1) XML/XSL in the Dictionary: The Case of Discourse Markers

**辞書における XML と XSL の役割－談話マーカの事例から**

**Daniela Berger, David Reitter, Manfred Stede (Univ. of Potsdam, Germany)**

自然言語文章を談話マーカでタグ付けして XML 化することにより、人間に対してはビューを提示して可読性の向上させる一方で、文章自動生成や文章理解システムへの適用も提案した研究である。ここでの談話マーカとは、主に接続詞ならびに接続の働きをする語を指す。本論文では、談話マーカで XML 化した辞書 DiMLex と XSL を用いて、HTML 文書へ変換することで人間への可読性の向上を図っている。また、文章自動生成・文章理解システムについてはシステムの表現形式である Lisp プログラムへの変換を XSL で行っている。論文では、文章自動生成・文章理解システムにおいては談話マーカについての文法と意味両面の記述が別々に用意されるべきであり、その考えに基づいた DiMLex におけるエントリ項目の最適な構成が今後の課題であると述べられている。**XiSTS のシラブル辞書の内容例**

### (2) XiSTS – XML in Speech Technology Systems

**XiSTS – 音声処理システムにおける XML**

**Michael Walsh, Stephen Wilson, Julie Carson-Berndsen (Univ. College Dublin, Ireland)**

音声認識システムで用いるシラブル辞書に XML を用いた研究である。論文では音声認識システムを構成する３つのサブシステムが述べられている。3 つとは、Time Map モデルをベースにした音韻認識システム LIPS、XML 化された特徴ベースの辞書 REFLEX（図 2)、人間の音声技術者によって別の音声の特徴記述セットを変換して新しい辞書を生成するシステム T-REX である。LIPS では音素配列オートマトンが Network Generator によって生成された後 XML 化され、そのオートマトンにしたがって音素のパージングが行われ、候補の音素列が決定される。それらの音素列は REFLEX への入力となり、XML 化された特徴ベースの音韻辞書が生成される。音素の特徴の記述は T-REX でユーザによってカスタマイズされ新たに辞書が生成される。音声認識システムにおける音声認識、音声合成、辞書生成のフェーズに XML を適用することの有用性が述べられている。

```
<syllable>
  So:n
  <onset type="first">
    <segment phonation="voiceless"
             manner="fricative" place="palato"
             duration="null">S</segment>
  </onset>
  <nucleus type="first">
    <segment phonation="voices" manner="vowellike" place="back"
             height="mid" roundness="round" length="tense"
             duration="null">o:</segment>
  </nucleus>
  <coda type="first">
    <segment phonation="voiced" manner="nasal" place="apical"
             duration="null">n</segment>
  </coda>
</syllable>
```

**図 2　XiSTS のシラブル辞書の内容例**

### (3) SALT: An XML Application for Web-based Multimodal Dialog Management

**SALT - Web 上でのマルチモーダル対話処理のための XML アプリケーション**

**Kuansan Wang (Microsoft Corporation, USA)**

Web 上での分散環境におけるマルチモーダルな対話処理システムのための XML ベースの言語 SALT についての発表である。SALT(Speech Application Language Tags)はプログラミング言語に依存せず、HTML や XML 文書に対話処理のためのインタフェースとして埋め込むことができる。ここでいうマルチモーダルとは、GUI 環境においてユーザがさまざまな手段で入力を行うという意味であり、対話処理においては、音声での入力、テキストでの入力、マウスでのイベント通知など、会話の内容をシステムに伝えるために行うユーザのアクションの多様性を指す。SALT の特徴として、Web ページ単位の制御フローを保っていること、データと表示を切り離すことでモジュール性を高めていること、対話の解釈にセマンティックオブジェクトを用いたオブジェクト指向モデルで行っていることである。論文の後半では SALT の実装についての説明と拡張性について述べられている。SALT の拡張として、対話の種類による制約を組み入れて応答を形成させたり、テレフォニーデバイスへ対応させたり、応答の内容を意味的に豊かなものにさせたりできることがあげられている。

## 5. セッション 4：Semantic Web

NLP におけるアノテーションや中間生成結果にメタデータを使うならば，汎用性や流通性の面から考えると Semantic Web での標準スキーマへの準拠が適している．生成の対象をメタデータにした場合でもそれらの生成やメンテナンスに NLP は有効である．過去に NLP のパラダイムにおいて研究されてきた多義性解消（WSD）や大規模オントロジの洗練やメンテナンスの手法は，Semantic Web での Web オントロジにおいても十分役立つと考えられる．また，エージェントのプロファイルの記述には IR での知見も取り込むことで洗練が期待できる．

### (1) Annotating the Semantic Web Using Natural Language

**自然言語を用いたセマンティックウェブへのアノテーション**

Boris Katz, Jimmy Lin (MIT Artificial Intelligence Laboratory, USA)

人間がセマンティックウェブに自然言語を通じて質問し、満足のいく回答を得るためのアノテーションのしくみを RDF に実現したシステム START を通じて、セマンティックウェブによって人間が多大な利益を享受するための方法を模索している。START では 3 つの手法が提案されている。1 つ目は RDF のプロパティとして、自然言語で記述された、想定される質問のパタンとそれに対する応答が組み込まれていることである。2 つ目はメタデータとしてのアノテーションスキーマであり、ユーザの発する多様な自然言語の質問に応答できるための記述が用意されている。3 つ目はプランスキーマであり、投げかけた質問に対してどのような手順で応答を導出するかについて、人間が通常用いている方法を手続き的に記述したものである。START は 1993 年以来 Web で公開されており、現在は地理データ、映画情報、人物録などにおける質問応答システムとして活用されている（図 3）。

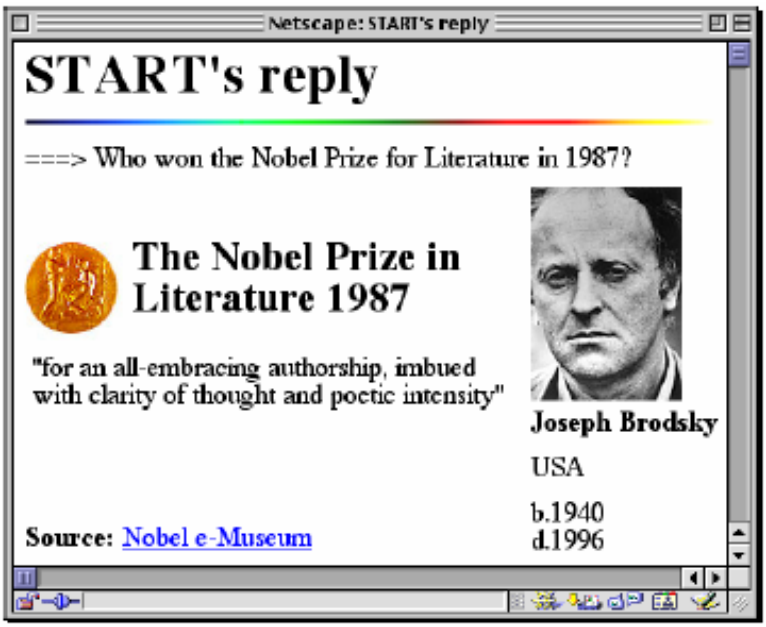


**図 3 START システムの応答画面例**

## 6. おわりに

ワークショップは最後の Semantic Web セッションにおける全体討論“The Roles of Natural Language and XML in the Semantic Web”で幕を閉じた．次世代 Web の発展のために NLP を駆使しようという参加した研究者たちによる強い動機が感じられた．必ずしも研究のためばかりでなく，産業界の発展にも寄与すべく進められているこの NLPXML 研究コミュニティの活動は評価に値する．

なお，第 3 回 NLPXML ワークショップは 1 年も経たずして，2003 年の 4 月にハンガリーのブダペストで EACL2003（11th Conference of the European Chapter of the Association for Computational Linguistics）に併設されて開催される予定である[5]．“Language Technology and the Semantic Web”と題された第 3 回のワークショップが盛大に開催されることを期待したい．

## 参考文献および参照 URL

本稿で述べた論文はすべて[1]の文献に掲載されている．

[1] Graham Wilcock, Nancy Ide and Laurent Romary eds. Proc. of the 2nd Workshop on NLP and XML (NLPXML-2002).

[2] 2nd Workshop on NLP and XML. http://www.ling.helsinki.fi/~gwilcock/NLPXML/

[3] 野村直之，中挾知延子，Key-Sun, Choi：第１回国際ワークショップ"NLP and XML"の概要とマルチモーダル・デジタル・ドキュメントの ISO 標準について．情報処理学会ディジタルドキュメント研究会，Vol.2002,No.28,pp.55-62. 2002

[4] LoPiX. http://www.informatik.uni-freiburg.de/~may/lopix/

[5] EACL 2003 Workshop. http://www.cs.vassar.edu/~ide/events/NLPXML3.html