2　自然言語
2 Natural Language

2-1　日本語話し言葉コーパスとその構築技術
2-1 Construction of the Corpus of Spontaneous Japanese and Annotation Techniques

内元清貴　井佐原 均　高梨克也　竹内和広　野畑 周　森本郁代
山田 篤
UCHIMOTO Kiyotaka, ISAHARA Hitoshi, TAKANASHI Katsuya, TAKEUCHI Kazuhiro, NOBATA Chikashi, MORIMOTO Ikuyo, and YAMADA Atsushi

要旨

『日本語話し言葉コーパス』を構築するにあたり、情報通信研究機構が行った情報付与について述べる。我々が付与した情報は、形態素、節単位、係り受け構造、要約、談話構造であり、これらの情報は、XML を用いて統合されている。形態素情報は、我々が提案した形態素情報付与の枠組みに基づいて人手コストを軽減することにより、転記テキストに半自動で付与した。次に、この形態素情報を用いて、続く情報付与の基礎となる単位として、節単位を認定した。続いて、これを単位として、係り受け構造、要約、談話構造に関する情報付与を行った。

This paper describes annotations for the *Corpus of Spontaneous Japanese*. The information we annotated to the corpus includes morphemes, clause units, dependency structures, summaries, and discourse structures. They are integrated in the form of XML. Morphological information was semi-automatically annotated to the transcribed text by reducing the human labor cost within the framework of morphological annotation that we proposed. Next, clause unites were detected based on the morphological information as basic units for our annotation. Then, dependency structures, summaries, discourse structures were annotated based on the clause units.

[キーワード]
話し言葉コーパス，形態素解析，節単位，係り受け構造，要約，談話構造，XML
Spontaneous speech corpus, Morphological analysis, Clause unit, Dependency structure, Summary, Discourse structure, XML

1　はじめに

本論文では、『日本語話し言葉コーパス』(*Corpus of Spontaneous Japanese：CSJ*) 及びその構築技術を紹介する。このコーパスは、科学技術振興調整費による開放的融合研究「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクト (1999 年度～2003 年度) [1] において国立国語研究所と共同で構築されたものである。CSJ は主に講演などのモノローグを対象とした自発的な話し言葉の大規模コーパスであり、このコーパスには音声データだけでなく、転記テキストも含まれる。さらに転記テキストには様々な言語情報が付与されている。図 1 に、CSJ に付与された言語情報の概要を示す。

データの収録と転記、形態素や韻律情報の付与

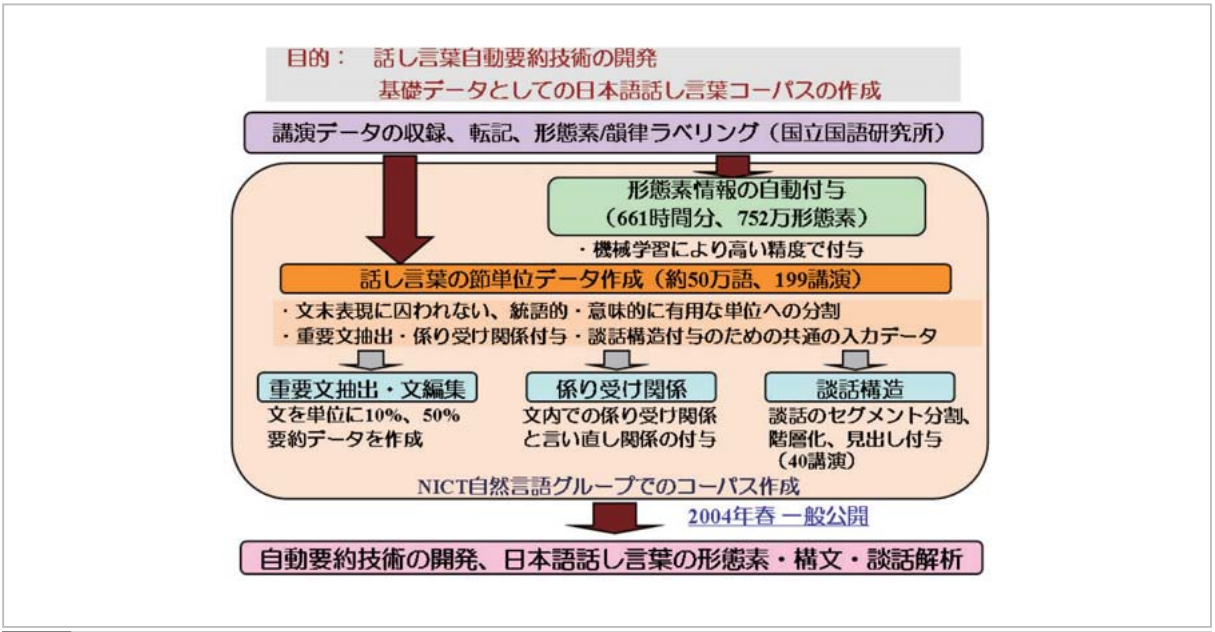


図1 『日本語話し言葉コーパス』に付与された言語情報の概要

については、国立国語研究所を中心に行われた。情報通信研究機構（旧通信総合研究所）は、転記テキストに対し、形態素、節単位、係り受け構造、要約、談話構造など様々な言語情報を付与した。形態素については、国立国語研究所が小規模の転記テキストに人手で注意深く情報を付与し、情報通信研究機構がそれを学習データとして利用して形態素解析システムを訓練し、そのシステムを用いて残りの転記テキストに対し形態素情報を付与した[2]。形態素情報付与について、より詳しくは**2**で述べる。次に、付与された形態素情報を元に、続く情報付与の基礎となる単位として、節単位を認定した[3]。この節単位の認定については、**3**で述べる。続いて、このような節を単位に、係り受け構造の付与[4]、要約データ作成[5]、談話構造付与[6]を行った。**4**で係り受け構造付与、**5**で要約データ、**6**で談話構造解析について述べる。また、**7**では、XMLを用いてこれらのデータを統合して記述・格納する仕組み[7]について概説する。

2 形態素情報付与

転記テキストには、国立国語研究所で定義された短い単位と長い単位の2種類の形態素に関する情報が付与されている。短い単位は**短単位**と呼ばれ、その定義は一般的な辞書の見出しに近い。一方、長い単位は**長単位**と呼ばれ、その定義には様々な複合語が含まれる。これら二つの単位は長さと品詞体系が異なり、長単位が短単位を包含するように定義されている。公開されたコーパス中の短単位は延べ約752万語である。一方、長単位は一つ以上の短単位から構成されるため数は2割程度少ない。これらのうち約1/8に、人手で、品詞や活用型、活用形などの形態素（形態論）情報が付与された。その約1/8における形態素情報の精度は、ランダムサンプリングによって約99.9％と推定されている。残りの約7/8については半自動で形態素情報を付与した。

本論文では、コーパスを形態素解析し整備する一連の処理を**形態素情報付与**と呼ぶことにする。この枠組みを図示すると図2のようになる。この枠組みの目的は、学習用コーパスと解析対象コーパス及び形態素解析システムが与えられたときに、少ない人的コストでコーパス全体の形態素情報の精度を向上させることにある。コーパス作成の途中では、体系や定義が変更されることが多いため、形態素解析システムはコーパスの定義変更にロバストなコーパスに基づく手法によるものを採用する。一般に形態素解析においては、未知語つまり辞書にも学習用コーパスにも現れない形態素の存在が最も問題となる。この問題に対処する

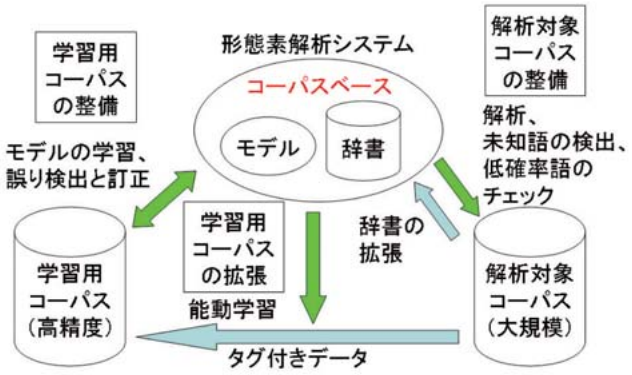


図2 形態素情報付与の枠組み

ために、これまで大きく二つの方法がとられてきた。一つは未知語を自動獲得し辞書に登録する方法であり、もう一つは未知語でも解析できるようなモデルを作成する方法である。我々は両者の利点を生かした、最大エントロピーモデルに基づく形態素解析の手法を提案した[8]。この手法で用いられるモデルは、任意の文字列について、その文字列が形態素であるときのもっともらしさを確率値として推定することができるため、未知語の問題を解決できる可能性が高い。そこで、CSJ の形態素解析にもこのモデルを採用した。さらに、本プロジェクトでは、話し言葉特有の現象に対しては、次のように対処した。

フィラーや言いよどみの存在

話し言葉に特有な現象であるフィラーや言いよどみは、任意の位置に出現する可能性があるため特定するのが難しい。CSJ ではフィラーや言いよどみには人手でタグが付与されているため、これらを削除して形態素解析し、後で挿入した。

発音形

音声認識のための言語モデルを作成するためには、形態素に関する情報の一つとして実際に発話された発音形の情報が欠かせない。しかし、辞書情報を用いて発音形の情報を補うのは無理がある。そこで、CSJ の転記テキストの発音形のフィールドと形態素解析の結果との対応をとることにより実際の発音形を付与した。

形態素情報付与の枠組みは、下記に述べるように、学習用コーパスの整備、解析対象コーパスの整備、学習用コーパスの拡張からなる。CSJ では、この枠組みで、未知語の検出と登録に 2％程度、能動学習に 1％程度人手によるチェックを行った。その結果、人手で形態素情報が付与されていない全体の約 7/8 における短単位と長単位の最終的な自動解析精度は、F 値でそれぞれ約 98.2と 96.5 程度と推測される。

学習用コーパスの整備

コーパスに基づく解析システムを用いる場合、一般に、コーパスに誤りが多いと誤りに過学習し解析精度が劣化する傾向がある。それを避けるためには、学習用コーパスの誤りを検出し修正する必要がある。CSJ では、まず人手で学習用コーパスに形態素情報を付与し、解析結果と学習用コーパスに差異がある部分に対し、それぞれモデルにより確率値を計算し、それらの値に基づいて学習用コーパスの該当部分を解析結果の対応部分で置き換えるという、コーパス誤り検出・訂正の手法を用いて誤りを検出した上で人手でチェックを行った。

解析対象コーパスの整備

解析対象コーパスに未知語、つまり、辞書にも学習用コーパスにも現れない語があると、未知語の前後も解析を誤る可能性が高くなり、未知語の数以上に誤りが増えることが多い。このような場合、解析対象のコーパスにおける未知語を検出して辞書に登録し、さらに、低確率語を人手でチェックすることによって、コーパス全体の精度を向上させられる[9]。コーパスで複数種類の形態素単位が定義されている場合でも、複数種類の形態素単位が包含関係にある場合には、最も短い単

位に関して未知語を抽出し、低確率語をチェックすれば、長い単位の精度も向上する[9]。

学習用コーパスの拡張

コーパスに基づく形態素解析システムのモデルは、一般に大量の学習用コーパスを必要とすることが多い。しかし、学習用コーパスを単純に増やしても、増やした量に比して精度の向上はわずかであることが多い。なぜなら、形態素解析のモデルでは、多くの場合、語と語の連接関係を学習しており、増やしたデータが既にモデルにとって推定が容易な連接関係であると効果がほとんどないためである。したがって、大規模な解析対象コーパスから、モデルにとって推定が難しい連接関係を多く含むような有益なデータを抽出して学習用コーパスを拡張する必要がある。それもできるだけ少ない追加で大きな精度向上が得られるようにしたい。本プロジェクトでは、能動学習により学習用コーパスを拡張することにより、人的コストを削減した[10]。

3 節境界の認定

従来、書き言葉を対象とする場合には、情報付与の対象となる一まとまりの単位としては「文」が用いられてきた。しかしながら、自発的な話し言葉を対象とする場合、文は必ずしも自明な単位ではない。CSJ を対象とした場合、文を単位とすることには以下のような問題点がある。

- 書き言葉では書き手自身が句点によって区切りを確定するのに対して、話し言葉にはこうした情報がない。
- 独話の特徴は一人の話者が続けて話し続けることであるが、文法的に明確な文末形式が頻繁に用いられるとは限らないため、極端に長い文が生じてしまう場合がある。
- 自発的な話し言葉では、言い直し、言い換え、言いやめなどの要因により文の範囲が確定しにくい場合や語や文の断片だけで発話が構成される場合がある。

したがって、このような問題点に対処しつつ、書き言葉における文に相当するような統語的・意味的単位を何らかの方法で認定する必要がある。そこで、我々はいわゆる文に代わる単位として「節」を採用することにした。

日本語においては、述語の活用形や接続助詞などの局所的な形態素情報のみに基づいて様々な種類の節境界を自動的に検出することが可能である。我々は、国立国語研究所と共同で節境界自動検出ツール CBAP[11] を改編し、CSJ の節境界を自動的に検出するルール CBAP-csj を作成した。CBAP-csj は、ある形態素の前後 1〜3 語を読み込んで節境界の種類を判別し、その種類に応じたラベルを挿入するものである。CSJ に付与される形態素情報は「出現形 _ 品詞 _ 活用型 _ 活用形」という四組で表現されている。ルールは、登録されている節パタンに該当する形態素列を発見したらその直後にラベルを挿入するという、パタンマッチを用いた正規表現として記述されている。なお、今回の節境界認定では、こうした節境界ラベルを節直後の切れ目の大きさによって「絶対境界[ ]」「強境界/ /」「弱境界〈 〉」という三レベルに区分した。絶対境界は形式上明示的な文末表現に相当する。強境界はいわゆる文末ではないが、発話の大きな切れ目として考えられる節境界、弱境界は節境界ではあるが通常は発話の切れ目になることはないと考えられる節境界である。さらに、絶対境界と強境界のみを発話の「デフォルト境界」として採用することによって、一つ以上の節から構成される「デフォルト単位」を自動認定した。これら二種類の節境界は発話の大きな切れ目となる境界で、統語的・意味的なまとまりを備えているため、様々な分析や処理にとって有用な単位の境界であると考えられるからである。理論的には、この区別は、節境界の形態の違いから従属節を複数のクラスに分類し、それらを統語的・意味的な自立性の度合いと関連づけた南[12] の分類に基づき、これを経験的な知見によって修正したものである。こうした区分により、節の種類ごとに異なる文法的な振る舞い（主題や格要素の共有、モダリティ要素のスコープの違いなど）をあらかじめ、ある程度予測することができ、統語的・意味的に自立しないデフォルト単位が生じるのを避けることができる。

CBAP-csj は局所的な形態素列のみを参照して境界を判定するものであるため、「体言止」などの特殊な節境界は発見できず、また、言い誤り・言い差しなどのように自発的な話し言葉に特有の現象や談話構造との関係に不都合が生じる箇所には

対処できない。統語的にも意味的にもより適切な単位を認定するためには、音声情報を参照しつつ、デフォルト単位を人手修正する必要がある。そこで、我々は次の三種類の操作を定義し、修正作業を行った。人手操作基準としては、約 40 種類が定義されている。

- 二つ以上のデフォルト単位を「＋」でつなぐ。
- デフォルト単位を「－」で切る。
- 要素を（　）、｛　｝、《　》で囲む。それぞれ、挿入、引用、倒置を表わす。

4　係り受け構造の付与

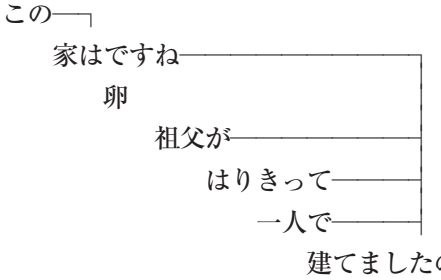
本プロジェクトでは、様々な研究開発のニーズに答えるため、CSJ においても統語構造の情報を付与することにした。コーパスは日本語が対象であるため、統語構造として文節間係り受け構造を採用した。日本語は語順が比較的自由であり、文節間の依存関係を特定するのが難しい場合が多い。しかし、文の意味を理解するためにはその依存関係を特定することが重要である。したがって、日本語の処理においては、その特定が難しいが重要である情報に特に着目し、統語構造として文節間係り受け構造を採用することが多い。我々が入手できる代表的な書き言葉のタグ付きコーパスの一つで、機械翻訳、情報抽出、要約、質問応答など様々な処理に利用されている京大コーパス[13]もこのような構造を採用している。

CSJ における文節間係り受けは原則として京大コーパスの基準に準拠するものとする。しかし、書き言葉と話し言葉では現象が異なることが多く、この基準だけではすべてを網羅することはできない。したがって、話し言葉特有の現象に対しては次のような新たな基準を設けた。

言い差し（言いやめ）

基本的に節境界認定の作業により別の節として切り出されるが、言い差し部分を越えて係り受けがある場合などは切り出されないことがある。この場合は、言い差しについては係り先なしとする。

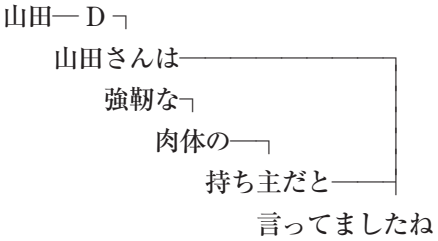
例）「卵」が言い差し



言い直し、言い換え

節単位内の言い直し、言い換えは新たに基準を設けて対応する。言い直しや言い換えにも様々な種類のものが考えられるが、CSJ における係り受け構造においては、詳細な種類の分類は行わず、言い直し、言い換えに関係する範囲を特定することに主眼を置く。言い直し関係、言い換え関係にはラベル D が付与される。

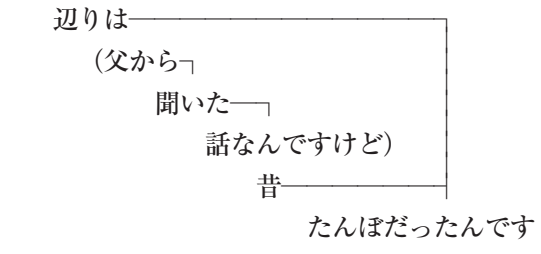
例）「山田」が「山田さん」に言い直されている



挿入構造

係り受けは挿入構造内で閉じるものとする。挿入構造は節境界認定作業により特定する。

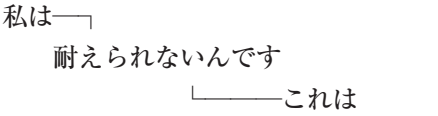
例）「父から聞いた話なんですけど」が挿入節



倒置

右から左への係り受けとする。

例）「これは」が倒置



ねじれ

発話プランの変更により、不自然な統語構造となる場合が多いので、基本的に係り先はないものとする。話題導入表現の直後など大きい切れ目に

おいては、節境界認定作業により節境界が認定され、別の単位となっている場合もある。

例)「目標は」の係り先が不自然

次の―┐
　目標は
　　マラソンで┐
　　優勝したいと┐
　　思います

実際の付与作業は独自のツールを用いて、人手で行った。一つの講演に対して、2 人の作業者が情報を付与し、1 人のチェッカーが検査をするという体制で行った。対象は節単位を認定した 199 講演分であり、対話と再朗読は対象外とした。

5　各種要約データの作成

従来、計算機による自動要約の手法は、重要文あるいは重要部分の抽出を基本としている。すなわち、「要約」とは重要部分の抽出の集合とみなすことが多い。このような背景から、本プロジェクトにおいても、CSJ 内の談話を「要約」した重要文選択データを作成したが、今後の自然言語処理技術の発展に寄与すべく、重要文選択以外の方法による「要約」データも作成した。具体的には、以下の 3 タイプの要約データを作成した。なお、これらのデータは節単位を認定した 199 講演分に対して提供されている情報である。

重要文選択データ

重要文選択データは、それぞれの講演について、要約率 50 %と 10 %の重要文選択を行うことにより作成した。ここで、要約率とは、例えば 10 %の要約率を指定された作業者は、与えられた転記テキストの文字数で全体の 10 %分だけになるよう転記テキスト中の単位を選択することを示す。重要文選択の際、選択に使用する単位は、前節で述べた節境界情報を利用した「節」である。まず要約率 50 %の重要文選択を行い、この中から更に元講演に対する要約率が 10 %になるように重要文選択を行った。

自由要約データ

重要文選択データとは別に転記テキストから直接書き言葉の文章の形式に講演を要約した自由要約データも作成した。自由要約データもそれぞれの談話につき、要約率 50 %と 10 %の 2 種類のデータを作成した。要約率 50 %のデータ作成では編集操作を限定し、重要な部分の抽出と各部分内での表現の変更のみで要約作成を行った。要約率 10 %の場合は、基本的には要約率 50 %の場合と同様の操作を中心にデータ作成を行うが、それでは必要な内容を十分に含めることができない場合には、自由な表現の書換えや部分の入替えを許した。

文編集データ

文編集は、自由要約データとは異なり、重要文選択データに対して、作業者が、特定の言語操作のみを用いて重要文を書き換え、簡潔な要約を作成したデータである。すなわち、計算機が自由要約を自動生成する課題を考える上で、既存の重要文選択手法との溝を埋める中間的な課題と言える。

作業者が重要文選択されたデータについて行うことができる言語操作は、単語や文節を削除することが基本である。転記テキストにない新しい語や表現を挿入する操作は、それを行わないと文が非文法的であるときのみに限定した。また、このような言語操作により重要文選択データを書き換える目的は、重要文選択データの要約としての読みやすさを維持しながら、冗長性を排除し、より簡潔な要約を作成することに主眼を置いた。

6　談話構造解析

本節では、CSJ に付与した談話構造に関するタグとその付与方法を紹介する。本プロジェクトで行った談話構造タグ付与は、Grosz と Sidner の談話構造理論(以下 GS)を背景としている[14]。GS では、話し手の意図や目的が談話の表層的な言語構造に反映されると見なす。GS における話し手の意図ないし目的とは、以下にかかわるものである。

- なぜ(他の行動ではなく)談話という言語行動によって事をなそうとしているのか
- なぜ(他の内容ではなく)この談話の内容を伝達しているのか

さらに、談話全体が一つの目的を持つだけではなく、談話を構成する複数の談話セグメントも、談話全体が果たす目的の部分目的となり得る談話

セグメント目的(以下、談話目的)を持つと考える。

GS を談話モデルとして用いた先行研究は幾つか存在する。我々はそのような研究の中で実際のデータに談話構造を付与した Nakatani らのマニュアル[15](以下、IAD)に注目した。そして、IADを CSJ の談話に適用する際の問題点を整理し、IAD 拡張することによって CSJ の談話に談話構造タグを付与することにした。IAD では、談話構造タグ付与のための作業を(1)セグメント境界の特定、(2)セグメント間の階層関係の特定、(3)談話セグメント目的の記述、の三つに分けている。しかし、IAD による談話タグ付けを予備的に行った結果、IAD では上の各作業をどの順番で行うかについては明示していないことから、例えば、セグメント境界とセグメントの階層関係を同時に特定しようとして作業が混乱し、作業者間でタグ付け結果に相違が生じやすいことが分かった。

この問題に対処するため、談話セグメントの同定作業を以下の二つの作業に分割した。

作業 1)　一つの談話を階層性のない小説の章のような談話セグメントに分割する。この作業は音声を聞きながら行い、一つの談話は複数人の作業者で分析する。作業者間が安定して認定した談話セグメントをセクションと呼ぶ。

作業 2)　節認定作業で得られた節をまとめあげて、内容上一貫性のある節の連鎖のパターンを発見する。この作業で認定した談話セグメントをエピソードと呼ぶ。

これにより、談話全体を大きく分割する談話セグメント階層と、節のレベルを意味的にまとめる談話セグメント階層の認定を、作業のレベルで分離した。この結果として、セクション境界をまたいでエピソードが認定された例は非常に少なくなった。また、二つの作業の結果の整合性を保つために、作業 2 において談話目的を記述しその談話目的に基づいてセクションの談話目的を設定した。この結果、公開データでは、一つの談話は複数のセクションからなり、各セクションには一つ以上のエピソードが存在する。

7　XML を用いたアノテーションの統合

CSJ にアノテーションとして付与された様々な情報は、XML を用いて統合し記述・格納した。このためにまず、転記テキストに対して付与された文節、係り受け、節境界、重要文、談話構造に関する情報の XML 化と相互の関連の記述を行い、次に、国立国語研究所により別途付与された音韻的なアノテーション情報との結合を行った。言語的な情報に対しては、講演を節単位、文節といった階層構造で表現し、各節単位に談話や重要文の情報を、また文節に係り受けの情報を持たせた。これを一定長以上の無音区間によって分割された転記基本単位を構成要素とする構造と結合する際には、転記基本単位が節単位や文節と交差する可能性があるため、情報損失なしにこれらを結合する手段として、節単位や文節といった単位を階層の中では明示的には表現せず、これらが持っていた情報をすべて当該単位の構成要素のうち、先頭の短単位に持たせるという方法をとった。この基本転記単位を構成要素とする構造から節単位、文節といった階層構造を復元する際には、転記基本単位の境界を越えてすべての短単位について、節単位ないし文節に関する属性を持っているものから後続する短単位のうち、節単位ないし文節に関する属性を持たないものを集めることになる。XML を用いることにより、論理的に異なる層に属しているデータや、互いに依存関係を持つデータを効率的に表現することができた。

作成された XML データの利用形態としては、単一講演内のデータを対象とする場合と、複数講演を横断的に調べる場合がある。単一講演を対象とする場合は、もとの XML インスタンスから目的に応じた情報や構造を抽出し、別の形式に変換することで利用しやすくなる。XSLT などの XML 関連技術を用いることにより、これは容易に実現できる。複数講演を対象とする場合は、データベースシステムが必要となるが、現状ではネイティブ XML データベースを用いて XML データのまま格納する方法、データを関係データベースのデータ構造に置き換えて格納する方法、バックエンドとして関係データベースを用いるがフロントエンドでは XML データを用いる方法がある。

8 まとめ

本論文では、情報通信研究機構（旧通信総合研究所）が『日本語話し言葉コーパス（CSJ）』に対して行った情報付与の概要を述べた。プロジェクト終了後、CSJ に付与された情報を学習に用いて、文境界認定や形態素解析、係り受け解析など自然言語処理技術の精度向上を図る[16][17]とともに、CSJ の作成に伴って開発されたツール群の充実を図るなど、引き続き話し言葉を対象とした情報付与技術や解析技術の研究開発を行っている。

参考文献

1 古井，前川，井佐原，“科学技術振興調整費開放的融合研究推進制度―大規模コーパスに基づく『話し言葉工学』の構築―”，日本音響学会誌，Vol.56，No.1，pp.752-755，2000.

2 K. Uchimoto, K. Takaoka, C. Nobata, A. Yamada, S. Sekine, and H. Isahara, "Morphological Analysis of the Corpus of Spontaneous Japanese", IEEE Transactions on Speech and Audio Processing, Vol.12, No.4, pp.382-390, 2004.

3 高梨，丸山，内元，井佐原，“『日本語話し言葉コーパス』における節境界認定”，平成 15 年度国立国語研究所公開研究発表会予稿集，pp.33-34，2003.

4 内元，丸山，高梨，井佐原，“『日本語話し言葉コーパス』における係り受け構造付与”，平成 15 年度国立国語研究所公開研究発表会予稿集，pp.35-36，2003.

5 野畑，内元，高梨，井佐原，“『日本語話し言葉コーパス』における要約データの作成”，第３回話し言葉の科学と工学ワークショップ講演予稿集，pp.99-104，2004.

6 竹内，森本，高梨，小磯，井佐原，“『日本語話し言葉コーパス』における談話構造タグの仕様”，平成 15 年度国立国語研究所公開研究発表会予稿集，pp.37-38，2003.

7 山田，高梨，内元，竹内，野畑，森本，井佐原，“『日本語話し言葉コーパス』におけるアノテーション統合”，第３回話し言葉の科学と工学ワークショップ講演予稿集，pp.33-38，2004.

8 内元，関根，井佐原，“最大エントロピーモデルに基づく形態素解析―未知語の問題の解決策―”，自然言語処理，Vol.8，No.1，pp.127-141，2001.

9 K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara, "Morphological Analysis of a Large Spontaneous Speech Corpus in Japanese", ACL, pp.479-488, 2003.

10 K. Uchimoto, and H. Isahara, "Morphological Annotation of a Large Spontaneous Speech Corpus in Japanese". IJCAI, pp.1731-1737, 2007.

11 丸山，柏岡，熊野，田中，“節境界自動検出ルールの作成と評価”，言語処理学会第９回年次大会発表論文集，pp.517-520，2003.

12 南，“現代日本語の構造”，大修館書店，1974.

13 黒橋，長尾，“京都大学テキストコーパス・プロジェクト”，言語処理学会第３回年次大会発表論文集，pp.115-118，1997.

14 B. J. Grosz and C. L. Sidner, "Attention, intention, and the structure of discourse", Computational Linguistics, Vol.12, No.3, pp.175-204, 1986.

15 C. H. Nakatani et al, "Instructions for annotating discourse", Technical Report, 21-95, Center for Research in Computing Technology, Harvard University Press, 1995.

16 下岡，内元，河原，井佐原均，“日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化”，自然言語処理，Vol.12，No.3，pp.3-17，2005.

17 R. Hamabe, K. Uchimoto, T. Kawahara, and H. Isahara, "Detection of Quotations and Inserted Clauses and Its Application to Dependency Structure Analysis in Spontaneous Japanese", COLING-ACL, pp.324-330, 2006.

内元清貴（うちもときよたか）
知識創成コミュニケーション研究センター自然言語グループ主任研究員（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ主任研究員）　博士（情報学）
自然言語処理

井佐原均（いさはらひとし）
知識創成コミュニケーション研究センター自然言語グループリーダー（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループリーダー）　博士（工学）
自然言語処理

高梨克也（たかなしかつや）
京都大学学術情報メディアセンター電子化・ディジタルアーカイブ研究分野特任助教（元情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員）
コミュニケーション科学

竹内和広（たけうちかずひろ）
大阪電気通信大学情報工学科講師（元情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員）　博士（工学）
自然言語処理

野畑　周（のばたちかし）
マンチェスター大学コンピュータサイエンス学科リサーチアソシエイト（元情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員）　博士（工学）
自然言語処理

森本郁代（もりもといくよ）
関西学院大学法学部外国語研究室准教授（元情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員）　博士（言語文化学）
会話分析

山田　篤（やまだあつし）
京都高度技術研究所主席研究員（元情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員）　博士（工学）
自然言語処理