節境界単位による翻訳 — 連体節について —

柏岡 秀紀，丸山 岳彦
ATR 音声言語コミュニケーション研究所

1　はじめに

講演などの独話では，「文」は長くなる傾向がある [4]．このような比較的長い文の解析では，1) 係り受け構造の曖昧さを多く含むため処理時間がかかる，2) 高い解析精度を維持することが困難である，という問題がある．一方で講演の音声翻訳において，同時通訳的な翻訳は，重要なアプリケーションであり，処理開始の遅延時間と高い精度の達成は重要なポイントである．そのため，独話の翻訳システムにおいて，「文」は翻訳の処理対象として適切な処理単位とは言い難く，文よりも短い処理単位が必要となる．我々は，同時通訳者による同時通訳データの分析 [1] や解説番組，新聞，ニュース原稿などのデータにおける言語的なまとまり (「文」，「節」など) の長さの分布等の分析を行ってきた．その結果，文よりも短く，統語的・意味的なまとまりを持つ「節」が翻訳対象として有効と考えている．そこで，「節」を見つけ出すために，節の終端位置と種類を局所的な形態素の連鎖から判断し，144 種の節ラベルを付与するツールを作成した [5]．このツールにより得られる節境界にはさまれる単位を，「節境界単位」とする．

「節」に相当する「節境界単位」は，その節ラベルに応じて大まかな訳出文の形式を推測できる．例えば，連体節を英語に訳す場合は関係詞節，条件節であれば if 節などのような形で訳出することが考えられる．このような訳出に利用できる特徴分析を行うことにより，「節境界単位」に基づく翻訳システムを構築することが考えられる．我々は，機械的に判定された節境界単位を利用し，節境界単位毎に対訳を付与したパラレルコーパスを構築している [2]．本稿では，このコーパスを分析し，節のタイプに応じた訳出の傾向と特徴について，特に連体節について報告する．

以下，まず分析対象としたデータについて述べ，「節境界単位」について概略を示し，連体節とその受け側文節を含む節との訳出の形式から 5 つのタイプに分類し，その特徴を示し，今後の分析課題について議論する．

2　対象となる独話の対訳データ

本稿で対象とするデータは，NHK の解説番組「あすを読む」の書き起こしテキストに対して作成した対訳データである．「あすを読む」は，様々な分野で注目されているトピックについて，NHK の解説委員が 10 分間で解説を行っている番組である．書き起こし作業者の主観および音声的特徴から付けられる句点により文を認定すると，1 番組には，平均 70 文弱含まれており，1 文は，約 30 単語からなる．対訳データは，解説の流れを考慮しつつ，文毎に対訳を付与したのちに，日本語の節境界単位に対応する対訳部分で分割し，節境界単位毎の対応付けを行った．節境界単位の 80% は，その対訳部分が連続して現れ，約 10% は 2 個所に別れ，約 10% は訳出として現れず，3 個所以上に訳出が別れるものは，1% 未満であった．日本語の節境界単位の認定には，次節で述べる CBAP を利用し検出した．

3　節境界単位

「節」は，統語的・意味的にまとまった「述語を中心とした」単位であり，翻訳や要約などの処理に有用な単位である．現在，対象としている講演などの独話において，構文解析が処理対象とする文は，長く，文末の判定も困難である．そこで，我々は，局所的な形態素の情報を利用し，節境界を検出する手法 (CBAP) を提案している．提案手法では，様々なデータに対して，ほぼ 97% 以上の精度で「節境界」を検出することが確認できている．この手法で節境界に対して付与している節境界ラベルを，表 1に示す．

また，以下に節境界検出の例を示す．

1. 私たち人間 / 体言止 /
2. つまり / 談話標識 /
3. 人の遺伝子を解読するという / 連体節トイウ /
4. 研究が民間企業も参加して / テ節 /
5. 激しい競争の中で今進められています． / 文末 /

節境界は，節の終端位置を示すものであり，4. のような「研究が民間企業も参加して」の節境界単位では，「研究が」という部分が「テ節」の統語的なまとまりと

表 1: 節境界ラベル（大まかな分類）

並列節	並列節
連用節	条件節, 譲歩節 時間節, 理由節, 連用節 (その他)
補足節	補足節, 引用節, 間接疑問節
連体節	連体節
その他	従属文, 体言止, 文末 主題ハ, 談話標識, 感動詞, 間投句

しては余分になる．純粋に「節」として扱うためには，このような部分を検出する必要がある．「節」を自動的に検出するには，構文解析による文の構造を把握し，節に相当する範囲を見出す手法が考えられる．構文解析においては，文を対象に行うより，節境界を手掛りにした手法で，効率が向上することが示されている [6].

本稿では，このような問題を踏まえたうえで，節境界単位を対象として特に連体節に注目した分析を行う．

表 2: CBAP による連体節ラベル

節ラベル	出現数	分析対象となる節数
連体節	5444	2796
形容詞連体節	116	
連体節トイウ	2033	714
連体節－形式名詞	1043	475
形容詞連体節－形式名詞	12	
連体節ヨウナ	180	79
連体節タメノ 連体節テノ 連体節マデノ 連体節ナドノ 連体節ホドノ	198	120

節境界単位：連体節

CBAP により付与される連体節関連のラベルと，160 番組内に含まれる連体節関連のラベルの出現数を，表 2に示す．ただし，連体節タメノ，連体節テノなど“ノ”を伴う節は，各出現数が低いため，出現数としてはその総数を示した．

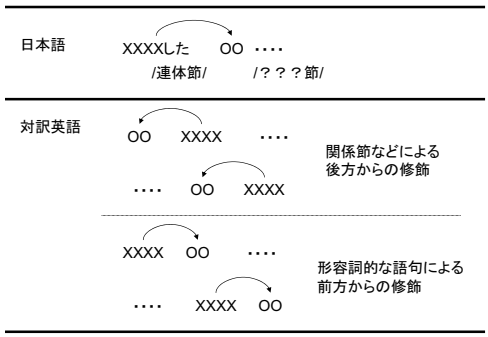
分析対象としては，対象とする連体節関連の「節境界単位」の対訳が 1 部分にまとまっており，連体節および受け側文節を含む節の対訳部分が連続して現れ，かつ，CaboCha[3] による解析結果において，連体節が次の文節に係っている節を取り上げた．表 2に分析対象となった節数も合せて示す．

ここで示す“連体節－形式名詞”は，連体節を受ける体言が形式名詞的な振舞いをするものである．形式名詞は，ChaSen により品詞が“名詞 - 非自立”となるもののうち出現形が“の，ん，こと，事，もの，もん，ところ，所，はず，つもり，わけ，訳”の単語とした．

4 対訳のタイプ

連体節の訳について考えた場合，図 1に示すように，連体節を受ける体言を後方から修飾する対訳と，前方から修飾する対訳との大きく 2 種類のタイプが考えられ，また，連体節を受ける体言を含む節が，連体節の対訳により，対訳として分割されているかどうかで 2 種類のタイプを考えることができる．

図 1: 連体節の対訳



また，これらとは別に，連体節を受ける体言が形式名詞などで対訳として表れないタイプが考えられる．そこで，この 5 つのタイプに分類し，日本語の節ラベルの種類との関連について分析を行う．その際に，日本語は，Chasen，CaboCha，対訳英語は Charniak parser を用いて解析した結果を利用した．

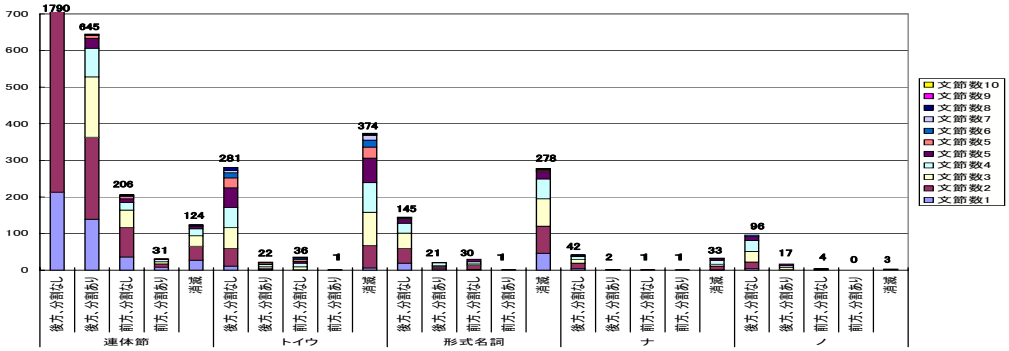
図 2に，5 つのタイプと節ラベルによる出現頻度を示す．以下に，各タイプの例を示しながら議論する．例文の 1. にあたる部分が連体節である．

後方からの修飾で受け側の節が分割されるタイプ

1. 飛行機の管制などを管理する
2. 運輸省，航空各社，飛行機の製造メーカーから対策済みの報告が出ています.

2a. the Ministry of Transport
1. that controls air traffic,

図 2: 連体節ラベルの分類タイプ別出現頻度



2b. airline companies and aircraft manufacturers have reported that they have completed the required measures./

最も一般的に出現するパターンである．どの節ラベルにおいても，大きな割合を占めており，対訳としては，関係詞節 (句)(SBAR) の形式に訳されることが最も多く，次いで，動詞句 (VP)，前置詞句 (PP) の形式に訳されることが多い．また，連体節トイウにおいては，動詞句 (VP)，前置詞句 (PP) の形式より，節相当 (S) の形式で訳されることが多い．

後方からの修飾で受け側の節が分割されないタイプ

1. 消費者と事業者との間で結ばれる
2. 契約 を幅広く対象とすると

2. the law covers a wide range of contracts
1. to be made between consumers and businesses.

一般的な連体節では，かなりの割合で出現するが，他の節ラベルでは殆ど表れない．訳出される場合にも，関係詞句 (SBAR) として訳されることが約 30% を占め，次いで 20% 程度が，名詞句 (NP) で始まるパターンとなっている．この 20% の大半は，関係詞が省略されたために関係詞句 (SBAR) としてまとまらなかったものと推定される．

前方からの修飾で受け側の節が分割されるタイプ

1. 随分違った
2. アプローチ を取ってます.

2a. are taking
1. quite different
2b. approaches./

相対的にこのように訳されることは余りなく，例に示すような形式が殆どである．形容詞的な振舞いを行う名詞句 (NP) の構成成分として訳出されることが多い．

前方からの修飾で受け側の節が分割されないタイプ

1. その海軍の主力である
2. 原子力潜水艦 が沈没をして

1. the main power of that navy,
2. the nuclear power submarine, sank and

図 2を見れば明かだが，この形式で訳出が行われることは殆どない．訳出される形式は，受け側の前方に表れることから，この節を受ける体言とともに，名詞句 (NP) を構成する形容詞句あるいは，この例の様な形式で訳される．

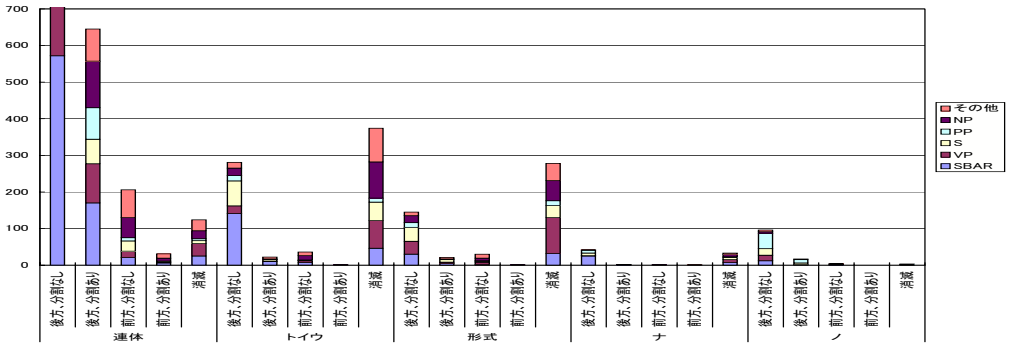
受け側の節が表れないタイプ

0. 強いアメリカに信頼を寄せて
1. 多くの人が株式投資に向かった
2. 結果 です.

(the increasing stock prices is)
1. that so many people have jumped into the stock market,
0. putting trust in the strong United States./

全体的に比較的多く表れている．その殆どは，この節を受ける体言が形式名詞の場合である．図 2では，連体節トイウの場合に，かなりの頻度で受け側の節が表れていないものが見受けられる．これは，節ラベル付与の判定において，“トイウ” が連接する節では，後続する体言（受け側の体言と推定される）の種別を考慮していないためである．実際に連体節トイウのラベルを持ち受け側の節が表れないタイプを調べると，2，3 の例外を除いて全て形式名詞の条件を満すものであった．また，連体節

図 3: 連体節ラベルの分類タイプ別にみた訳出形式



ラベルの場合に表れているものは，先に示した形式名詞の出現形とはことなる体言であるが，“問題，道，場合，形，状況，状態” など形式名詞に含めても問題ないと思われるものや，代名詞化されたために，節の対応付けで誤ったものなどであると考えられる．また，訳出としては，動詞句 (VP)，名詞句 (NP) の形式になりやすい．

以上に述べた訳出の形式は，対訳英語の Charniak parser の解析結果から，連体節に対応する部分取り出した際の先頭カテゴリに着目したものである．図 3に，その分布を示す．

5　考察

連体節は，その節の長さ，節ラベルの種類により特定の訳出がなされるという強い相関は，現在のところ見受けられていない．それよりも，訳出に関連しているのは，連体節の受け側の文節による処がある．連体節の受け側の体言が形式名詞である場合には，その受け側の節が訳出されず，動詞句，あるいは，名詞句として訳出される傾向が強いことがわかる．また，連体節の受け側文節を含む節が分割されるか否かも，受け側文節がその文節を含む節内でどのような格要素になっているかに関係していると考えられる．

6　まとめ

本稿では，講演など独話の音声翻訳システムを構築するために，節単位で翻訳を行うことを考え， NHK の解説番組「あすを読む」の節対応の取れた対訳データに対して，連体節の対訳関係を調べた結果を示した．

量的には，連体節は後方からの修飾として，受け側の文節を含む節の対訳を分割せずに関係節を利用した訳出が多く，また，受け側の体言が形式名詞の場合に受け側の文節を含む節は訳出せず，連体節を名詞句あるいは動詞句として訳出されることが多いことが分った．今後は，受け側文節の格関係などに注目して，適切な訳出のための特徴をみつけたい．さらに，連体節だけでなく，連用節 (特に条件節や理由節)，補足節など他の節ラベルについても訳出の傾向と適切な訳出のための特徴分析を行うつもりである．これらの分析より得られる特徴を翻訳システムに組み込み，節単位での対訳システムの構築を行う予定である．

謝辞

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

参考文献

[1] 柏岡．講演の同時通訳データの作成と分析．思考と言語研究会 TL2000-33, 電子情報通信学会, 2000.

[2] H. Kashioka, T. Maruyama, and H. Tanaka. Building a parallel corpus for monologue with clause aligmnemt. In *MT Summit IX*, 2003.

[3] 工藤, 松本．チャンキングの段階適用による係り受け解析．情報処理学会論文誌, 43(6):1834–1842, 2002.

[4] 丸山, 熊野, 柏岡. 日本語における独話の特徴と文分割. 言語処理学会 第 7 回年次大会発表論文集, pp. 429–432, 2001.

[5] 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.

[6] 大野, 松原, 丸山, 柏岡, 田中, 稲垣．節境界に基づく独話文の係り受け解析とその評価．言語処理学会 第１０回年次大会発表論文集, 2004.