節境界と係り受け解析

柏岡 秀紀, 丸山 岳彦, 田中 英輝
ATR音声言語コミュニケーション研究所

1 はじめに

近年の自然言語処理技術および、音声認識、音声合成技術の向上により、実用的な音声翻訳が身近なものとなりつつある。音声翻訳の実用化の一形態として講演などの同時通訳を考えることは自然であり、その技術は重要である。同時通訳を実現するには、時間的な制約から漸進的な処理が必要とされる。

漸進的な解析処理に関する研究は[1, 2]などがこれまでに報告されている。その多くは、一語ずつの入力に対して、解析候補の効率的な絞り込みに焦点を当てたもの、あるいは、全体の処理時間について議論しているものである。漸進的な解析による翻訳処理を考えた場合、意味的あるいは構造的にまとまりのある翻訳単位での解析効率が重要なポイントになる。従来の翻訳処理は、文を処理単位としている。しかし、話し言葉、特に講演などの独話では、文の単位の認定が困難であり、また、1文が比較的長くなるため、文は適切な単位とはいえない[3]。一方、音声の特徴であるポーズにより分割した発話の単位は、比較的短く、時間的な制約は満たすと思われる。しかし、ポーズによる単位は、1語のみの単純な場合から句、節のような構造を持つ場合までさまざまな言語的な単位が混ざることになり、翻訳処理に向いているとはいえない。そこで、述部を中心としたまとまりである「節」を、処理単位として検討する。

本稿では、まず、我々が翻訳単位に想定している「節」について述べ、係り受け構造との整合性を調べる。「節」への分割は、局所的な(品詞情報を含む)単語列により自動的に行った[4]。さらに、話し言葉、書き言葉による特徴が、「節」と係り受け構造の関係に現れるか否かを検討する。対象データには、「あすを読む」、「NHKニュース原稿」、「日経新聞記事」という3種類のデータを取り上げた。それぞれ、実際に話された話し言葉、ある程度読み上げを想定して作成された書き言葉、読まれることを目的とした書き言葉のデータである。

2 分析対象のデータ

本稿では、分析対象として以下の3つのテキストを利用した。ただし、以下に示すデータのうち、テキストをクリーニングする段階で、一部のデータを対象から除外している。

1. 「あすを読む」
   NHKの解説委員が1番組10分間で話題になっているトピックについて解説を行っている番組の書き起こしテキスト。1999年末からの300番組分を利用。
2. 「NHKニュース原稿」
   NHKのニュース原稿。1999年1月、3月、7月の3ヶ月分のデータを利用。
3. 「日経新聞記事」
   日本経済新聞、日経産業新聞、日系流通新聞、日系金融新聞の記事データベース。1995年1月の1ヶ月分のデータを利用。

これら3つのデータは、それぞれ、1. 実際に話された話し言葉(の書き起こし)、2. 話すことを前提に書かれた言葉、3. 読むことを前提に書かれた書き言葉、を代表するものである。これにより、同時通訳を目的とした翻訳単位としての節相当語句と係り受け構造との関係を調べる。また、様々なデータで節相当語句が処理単位として有効であることを検証する。

3 「節」境界の判定と係り受け構造

本節では、まず、分析に利用した節への分割処理について簡単に紹介する。次に、分割処理から得られる節相当語句と係り受け構造との整合性について述べる。

3.1 節相当語句

従来から多くの機械翻訳では翻訳の単位を文としている。文が翻訳などの処理単位として適切であるのは、主に以下の二点のためと考えられる。

- 書かれたテキストの場合、句点により明確に単位を判定することができる。
- 係り受け構造を文単位で独立に扱える.

ところが、音声翻訳が扱う話し言葉においては、「文」の境界を明確に判定することが困難である。旅行会話などの対話では話者の交代に伴う発話の単位が比較的短いことから、発話を「文」とみなし、処理単位としている。一方、講演などの独話では、1人の話者が話し続けることから、発話が長くなり、対話のように発話を処理単位と見なす事は困難である。さらに、実際に書き起こされたテキストにおいても、1文は比較的長く、「文」の判定にも揺れが生じている。

「述語を中心としたまとまり」である「節」の係り受け構造は、その内部では単文の構造に類似していると考えられる。我々は、独話等の話し言葉においても明確に判定でき、また、係り受け構造としてもある程度独立して扱える処理単位として、「節」が適切であると考えた。「節」の単位を検出するために、漸進的な処理との相性を考慮し、局所的な形態素の連鎖により分割点を検出する手法を用いている。形態素に関する情報は得られていると仮定している。分析にあたって形態素の情報は、茶筌[5]による解析結果を利用した。「節」の境界を検出する際に、主題を示す句や接続詞などを含む談話標識、などの境界の検出も行っている。以後、本稿で扱っている「節」の単位より少し細かい単位を、節相当語句と記す。また、分割点の検出に伴い、各節相当語句にラベルを付与している。ラベル は、全部で144種類あり、大まかな分類では、表1に示すタイプにまとめることができる。

表 1: 節のタイプ

並列節	並列節
連用節	条件節, 譲歩節, 連用節 (その他) 時間節, 理由節
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

3.2 係り受け構造

係り受け構造には、茶筌、CaboCha[6]による解析結果を利用した。「あすを読む」の1番組のデータで、茶筌、CaboChaによる解析結果の精度を調べたところ、20個弱の係り受け関係に誤りが含まれていた。文単位では、52文中12文の誤りが含まれていることになる。話し言葉において大量のデータでの比較を行うには、比較的良好なデータであると思われる。各データの節相当語句数、文節数、および、一つの節相当語句あたりに含まれる文節数を表2に示す。

表 2: 分析対象テキストのデータ量

	あすを読む	NHKニュース原稿	日経新聞記事
節相当語句数	70002	420362	648852
文節数	175764	1236135	1799580
文節／節	2.51	2.94	2.77

4 節相当語句と係り受け構造の関係

ここでは、以下の2点について検討を加える。

1. 節相当語句内部での係り受け構造
2. 節相当語句の係り先の特徴

4.1 節相当語句内部の係り受け構造

まず、節相当語句内部での係り受け構造について、例を図1に示す。図1に示す円の内部で係り受け構造が閉じていれば、最後の文節の係り先を残して、優先的に解析結果を構成することができる。「裁判で／問われましたのは」や「どのような／場合に／死刑を／適用するのかという」は、節相当語句内で係り受け構造が閉じており、最終文節である「問われましたのは」と「適用するのかという」の部分の係り先のみが、節外にある。この様な構造は、漸進的な処理に都合が良い。表3に3つのデータにおける節相当語句内から節外への係り受け関係を持つ文節数の頻度を示す。

表3に示されている節外に係る文節数が0のものは、全て文末である。他の所に文末があらわれることはなかった。また、1であるものは、節相当語句内の最終文節が他の節相当語句に係るもので、漸進的な処理に適している節相当語句である。文末を除く節相当語句に対して、漸進的処理に適している節相当語句の割合は「あすを読む」で、91.2%、「NHKニュース原稿」で87.7%、「日経新聞記事」で、89.0%であった。

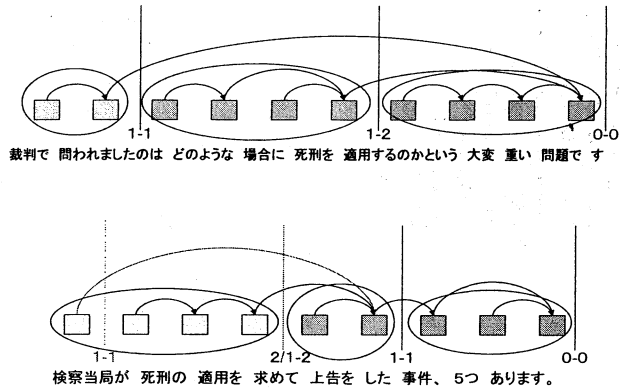

図 1: 節分割点と係り受け構造の関係

表 3: 節相当語句の節外へ係る文節数

節外に係る文節数	あすを読む	NHK ニュース原稿	日経新聞 記事
0	16393	67676	181690
1	48911	309100	414434
2	3878	35690	44316
3	686	6414	6015
4	113	1067	905
5	16	224	133
6	4	57	16
≥7	1	57	7
total	70002	420285	647516

最終文節以外に係り先が節外にある節相当語句は、「あすを読む」のデータでは、“主題ハ”のつけられた節相当語句が最も多く、次いで、“連体節”、“連用節 (その他)”が続き、“並列節”、“体言止”、“補足節”の順になっていた。各節タイプの出現頻度との割合で見ると、“体言止”が 25% で最も高く、“主題ハ”、“連用節 (その他)”、“間投句”、“並列節”が 10% を上回っていた。

節相当語句の平均文節数は、2.5 から 3 である。節相当語句の内部での係り受け関係は、比較的容易であり、長い係り受け関係を考慮しないですむため、効率的な処理が可能と思われる。また、節タイプの違いに応じて最終文節以外にも節外への係り先を考慮するなどの処理を導入することにより、適切かつ効率的な処理が可能と思われる。

さらに、節相当語句に含まれる文節数に注目すると、「あすを読む」において係り受け構造が閉じていない節相当語句は、平均の文節長に比べて約 1 文節、長くなっている。節タイプ毎の文節長を見ても、特に節タイプによる差は見られない。

節相当語句から 2 つ以上の係り先を節外に持つデータを幾つか調べてみると、a) 明らかに、係り先を節外に 2 つ以上持つ場合と、b) 文全体での係り受け構造としては、節相当語句外に係り先があるが、節相当語句内にも係り先の候補を持つ場合、c) 茶筌、CaboCha の誤りによる場合があった。a) の場合、ガ格の重複など既存の解析上の制約を利用して節外に係る文節を見つけることがある程度可能と思わる。b) の場合、節相当語句内で係り受け構造を構築することで、より近い文節間での係り受け構造を出力する解析が可能となる。図 1内の「検察当局が／死刑の／適用を／求めて」という節が、その例である。「検察当局が」は、節内にある「求めて」に係るとしても誤りではない。また、節ごとの翻訳という立場からは、このような解析の方が有用と思われる。

4.2 節相当語句の係り先

従来より、話し言葉では複雑な係り受け構造は現れにくく、比較的理解しやすい構造が現れるといわれている。会話における発話は比較的短く、係り受け構造として遠くに係るものは特徴的な表現が多いため、ある程度

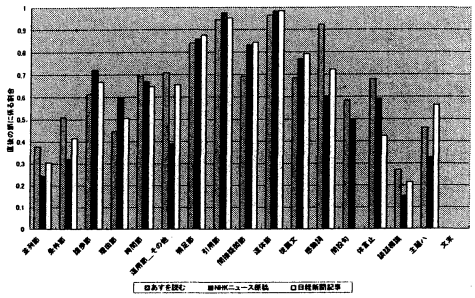

図 2: 直後の節に係る節相当語句の割合

予測できると考えられる。

図 2に直後の節に係る節相当語句の節タイプ毎の割合を示す。連体節は、どのデータにおいても、そのほとんどが直後の節に係っている。引用節、補足節もほぼ同様の傾向を示しているといえる。それぞれの節の統語的な特性から、これらはごく自然な傾向といえる。感動詞、間投句は、元々、出現頻度がほとんどない上に、形態素解析誤りによるものが比較的多く見受けられるため、これらについての考察は控たい。並列節、条件節、連用節 (その他) で、3 つのデータを比較すると、「あすを読む」が比較的多く直後に係っており、次いで「日経新聞記事」、「NHK ニュース記事」となっている。

また、節相当語句の最終文節がどの程度離れた文節と係り受けを構成するかを調べた。「あすを読む」の結果を図 3に示す。他の「NHK ニュース原稿」、「日経新聞記事」でも、ほぼ同じ傾向が見られた。

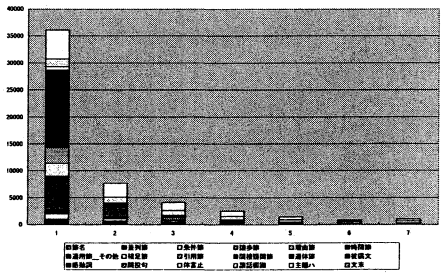

図 3: 最終文節の係り先までの距離

5 まとめ

本稿では、「あすを読む」、「NHK ニュース原稿」、「日経新聞記事」に対して、同時通訳を考慮した翻訳単位である節相当語句の係り受け構造について分析、検討を行った。節相当語句は、局所的な形態素情報を利用した節分割の結果であり、係り受け構造は、茶筌、Cabocha による係り受け解析結果による。その結果、今回利用している節相当語句は、ほぼ 90% 前後のものが、内部で閉じた係り受け構造をもつことから、漸進的な解析に有効な単位であることが判った。また、今回調べた 3 つのデータでは、節構造と係り受け構造の明確な差異は見られなかったが、細かな点で、直後の節へ係る文節の割合いなど節タイプにより、データ毎の差が見られた。今後は、節分割を利用した節相当語句内の係り受け解析の実現を目指すとともに、節の間の解析の効率化のための分析を進めたい。

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] Mima, H., Iida, H. and Furuse, O.: “Simultaneous Interpretation Utilizing Example-based Incremental Transfer”, Proc. of COLING-ACL ’98, Vol.2, pp.855–861, (1998).

[2] 渡辺, 松原, 外山, 稲垣 “英日同時翻訳のための漸進的日本語生成” 言語処理学会 第 6 回年次大会発表論文集, pp.272–275, (2000).

[3] 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).

[4] 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 弟 9 回年次大会発表論文集,(2003).

[5] 日本語形態素解析システム ChaSen「茶筌」(奈良先端科学技術大学院大学 松本研究室), http://chasen.aist-nara.ac.jp/

[6] 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、 Vol.43, No06,(2002).