自 然 言 語 処 理 109－1
（1995．9．14）

日英機械翻訳における読点の前処理

岡崎洋三
大阪大学留学生センター
〒565 吹田市山田丘 2－7

あらまし
日本語の読点の使用法については、指針的なものはあるが、原則的に統一されてはおらず、読解（解析）上適切なものと不適切なものとが混在している。そこで、不適切な読点の発見システムとして改行方式を提案する。これは、文を読点のところで機械的に改行し、分かたれた部分の文要素としての意味的まとまり具合を検討することによって、読点の必要度と適切性を判定するものである。つまり、これは読点制御による前編集である。
キーワード 読点の編集、文要素認知、改行方式

Pre-processing of Punctuation in Japanese-English Machine Translation

Yozo OKAZAKI
International Student Center, Osaka University
2-7, Yamadaoka, Suita-shi, Osaka 565 Japan

Abstract
There are some guidelines for the use of punctuation in Japanese writing. Appropriate use of punctuation marks, especially the comma, is important for enhancing comprehensibility, yet in actual practice there occur many inappropriate commas as well. This paper proposes an automated pre-editing system for finding and deleting unnecessary commas, which works by splitting the text at the commas into separate lines (*versifying*), and examining the semantic coherence of the resulting units.
Key Words editing for punctuation, identifying extended sentential units, versifying method

1　はじめに

日本語表記法では、ひらがな・カタカナ・漢字を使い分けることによって漢字仮名交じり文をつくり、これによって分かち書き効果を生み出している。その結果、仮名が多すぎる文や逆に漢字が多すぎる文は読みにくくなる。日本語の読点は、誤読を避けるためと読みやすさのために使用されるが、具体的使用法については原則的に確立することが出来ないでいる.(文献[5])読点の規範的な使用法としては,文化庁の句読法案とその準拠案があるが、その一方でそれらを批判する形で本多勝一の論考[12]がある。両者の違いをまとめると次のようになる。

本多案では、文の論理的明晰性を高めるために読点の使用法が検討されており、中点で代用できるものは読点を使わないなど、その原則的な役割は構文的なものにしぼられている。「長い修飾語が二つ以上あるときその境界にテンを打つ」「語順が逆になったときにテンを打つ」という２大原則があり、これ以外の用法は書き手の主観によるものとされている。語順との関係で読点の用法が考察されているのが注目される。文化庁案およびその準拠案では、これに比べると語句の分かち書き的な役割として読点が認識されており、語順との関係についての考察も見られない。

読点の使用法が確立されない理由としては次の二点が考えられる。

(1) 日本語表記法では特に必要ではない読点があったとしても読解の障害になるところまでには至らないことも多く、問題視されにくいということ

(2) 社会的な背景として日本の国語教育では表現技術の教育が未発達であり、分かりやすい日本語というものが価値を持つようになったのは近年のことだということ

読点の機械処理の第一の課題は、不適切な読点を発見するシステムの開発である。

2　読点による文要素認知

英語圏での日本語教育において、翻訳せずに日本文を読解する方法が提案されているが、それは拡大文節(Extended Sentential Unit)という概念によって文要素を認知する方法である.([6])

自然言語処理における文要素の認知・確定は、第一に助詞を手がかりにして行なわれ、第二に語彙の共起性・意味的結束性が手がかりになる。これに、第三の手がかりとして読点を使うことが期待される。しかし新聞や雑誌などで使われている読点には適切なものと不適切なものとが混在しているため、不適切なものを前処理時に見つけ出す必要がある。そこで、問題となりそうな文を例にあげて、読点が打たれているところで機械的に改行し、読点で分かたれた文要素間の意味的結束性を検討することを通じて読点の適不適を判定するという作業を行なう。

2－1 短文（単文）について

例１　将来、が見えない。
　⇒将来、
　　が見えない。
例１’将来が、見えない。
　⇒将来が、
　　見えない。
例２　その子は、夢中でしゃべる人形の声を聞いた。
　⇒その子は、
　　夢中でしゃべる人形の声を聞いた。

例2’その子は夢中で、しゃべる人形の声を聞いた。
⇒その子は夢中で、
しゃべる人形の声を聞いた。
例3　私は、英語が話せる人を探している。
⇒私は、
英語が話せる人を探している。

例1のように、助詞がなかったり、英語の前置詞のように助詞が前置されると文要素の認定が出来ない。例2では、文意が「夢中で」「聞いた」であれば2’になる。例3の「私は」の後の読点は、この後に続く「英語が話せる」という文要素が「私は」の陳述部ではなく、その後に続く「人」の修飾部であることを明らかにする。以下、例文番号に’または”がついたものは、人手による処理を行なったことを示す。

2-2 長文（複文）について

2-2-1　主部の認知

物理的に文が長くなると主部・修飾部などの文要素を読解過程において認知・確認する必要性が出て来る。一般的には、３０字を越える文に読みやすさのための読点を使うことは妥当である。

例4　私が英語を話せる人を探しているというのは来月イギリスから来るお客さんを案内する人がいないからである。(50字)
例4’私が英語を話せる人を探しているというのは、来月イギリスから来るお客さんを案内する人がいないからである。

2-2-2　修飾部の認知

例5　この章ではもうひとつの重要な要素である、辞書についてご紹介します。
⇒この章ではもうひとつの重要な要素である、
辞書についてご紹介します。
例5’この章では、もうひとつの重要な要素である辞書についてご紹介します。
⇒この章では、
もうひとつの重要な要素である辞書についてご紹介します。

「この章では」は「ご紹介します」にかかる。また、「もうひとつの重要な要素である」は「辞書」の修飾部として意味的まとまりを持つ。「辞書」の前の読点は、これが話し言葉であればここにポーズがあると解釈することも出来るが、実用文においては修飾部と被修飾語の間の読点は変則的である。

2-2-3　補部の認知

例6　あれは実は深い意味があったのだ、と折に触れて思い出す度に反省させられる。
⇒あれは実は深い意味があったのだ、
と折に触れて思い出す度に反省させられる。
例6’あれは実は深い意味があったのだと、折に触れて思い出す度に反省させられる。
⇒あれは実は深い意味があったのだと、

折に触れて思い出す度に反省させられる。

例7　Cross/36では、スクリーンまたは、rpgc や autoc のコマンドで指定した出力デバイスに、コンパイル時に出されるすべてのメッセージが出力されます。

⇒Cross/36では、
スクリーンまたは、
rpgc や autoc のコマンドで指定した出力デバイスに、
コンパイル時に出されるすべてのメッセージが出力されます。

例7'　Cross/36では、
⇒スクリーンに、
または、
rpgc や autoc のコマンドで指定した出力デバイスに、
コンパイル時に出されるすべてのメッセージが出力されます。

例8　依頼事項とは別次元のこんな抗議を公表する「原稿」として加えるべきではないと思います。

例8'　依頼事項とは別次元のこんな抗議を、公表する「原稿」として加えるべきではないと思います。

⇒依頼事項とは別次元のこんな抗議を、
公表する「原稿」として加えるべきではないと思います。

例6は「あれは実は深い意味があったのだと」「反省させられる」という係り受け関係であるから、6'の方が良い。例7の「スクリーン」の係り先は、7'のように助詞と読点を補えば明確になる。例8は「抗議を」が「公表する」の目的語ではないことを確認する必要がある。文要素認知第二の手がかりである語彙の意味的結束性がむしろ障害になる場合は、第三の手がかりとしての読点が不可欠である。

2-2-4　否定辞作用域の決定

読点が否定辞の作用域を決定する場合がある。

例9　A方式を試用したので解析に失敗したとは思わない。

例9'　A方式を試用したので、解析に失敗したとは思わない。

⇒A方式を試用したので、
解析に失敗したとは思わない。 cf.解析に失敗したとは思わない。A方式を試用したので。

例9"　A方式を試用したので解析に失敗した、とは思わない。

⇒A方式を試用したので解析に失敗した、
とは思わない。

2-2-5　従属節依存関係の認知

2-2-5-1　従属節が従属節の一部になる場合

南[1]によれば例10・11のように従属節が従属節の一部になって意味的まとまりをつくる場合があるが、これに読点を打てば10' 11'になる。

例10　月末になれば新製品が入荷しますのでそれまでお待ちください。

例10’ 月末になれば新製品が入荷しますので、それまでお待ちください。
⇒月末になれば新製品が入荷しますので、
それまでお待ちください。
例11 デパートは午後こむから午前中行って来たがそれでもけっこうお客は多かった。
例11’ デパートは午後こむから午前中行って来たが、それでもけっこうお客は多かった。
⇒デパートは午後こむから午前中行って来たが、
それでもけっこうお客は多かった。

2-2-5-2 従属節が従属節の一部になれない場合

また、従属節が従属節の一部になれない場合([1])については次の通りである。

例12 危ないところへさしかかったら手をつないで歩きなさい。
例12’ 危ないところへさしかかったら、手をつないで歩きなさい。
⇒危ないところへさしかかったら、
手をつないで歩きなさい。
例13 焦げるおそれがありますからかきまぜながら煮ましょう。
例13’ 焦げるおそれがありますから、かきまぜながら煮ましょう。
⇒焦げるおそれがありますから、
かきまぜながら煮ましょう。

2-2-5-3 依存関係の認知か個々の節の認知か

節と節との依存関係認知の第一の手がかりは接続助詞・接続表現であるが、次のような場合の読点はこれでよいかどうか。

例14 おじいさんは山へ行き、おばあさんは川へ行った。
例15 おじいさんは山へ行って、おばあさんは川へ行った。
例16 おじいさんは山へ行ったが、おばあさんは川へ行った。
例17 おじいさんは山へ行ったので、おばあさんは川へ行った。
例18 おじいさんは山へ行ったから、おばあさんは川へ行った。
例19 おじいさんは山へ行ったのに、おばあさんは川へ行った。
例20 おじいさんは山へ行ったけれども、おばあさんは川へ行った。
例21 おじいさんは山へ行ったにもかかわらず、おばあさんは川へ行った。

例文14～16(連用中止形「行き」・て形「行って」・接続助詞「が」)では節同士が並列の関係を示しており、いわゆる重文と見なされるが、接続助詞「ので」「から」「のに」「けれども」「にもかかわらず」を使えば、主節と従属節の関係を持つ複文になると考えられる。

この場合の節の依存関係と読点との関係について言えば、例文14～16の読点は後に係ると考えなくてもよい読点である。そして、例17～21の読点は、節と節との係り受け関係を示すというよりも、節そのものを言わば節の分かち書きとして示していると解釈される。これについてさらに発展させると、例えば例文19は次のようになる。

例22　おじいさんは山へ行ったのに、おばあさんは川へ行った。私はどちらについて行ったらよいか悩んだ。
例23　おじいさんは山へ行ったのに、おばあさんは川へ行ったので、私はどちらについて行ったらよいか悩んだ。
例23'　おじいさんは山へ行ったのにおばあさんは川へ行ったので、私はどちらについて行ったらよいか悩んだ。(47字)

上記例は語彙が平易だが、文が長くなって語彙が難解になると、節の依存関係認知のための読点に、個々の節そのものを認知して読みやすくするための読点の問題が加わり、どちらを優先するか判定が難しくなる。読みやすさか論理的結束性かという選択と、前処理時の問題か後処理時の問題かという選択がある。

2-2-6　埋め込み文の認知

例24のような構造の埋め込み文([4])に読点を打つと、理論的には24'になる。しかし、これは機械が理解できたとしても、人間にとって悪文的である。そこで語順操作を行なって24"のようにする。

例24　あの論文は僕は皆が君が真っ先に読んだと信じていると思う。
例24'　あの論文は、僕は、皆が、君が真っ先に読んだと信じていると思う。
⇒あの論文は、
僕は、
皆が、
君が真っ先に読んだと信じていると思う。
例24"　あの論文は君が真っ先に読んだと皆が信じていると僕は思う。

3　マニュアル文における読点の前処理と後処理－「～すると、～する」の場合について－

あるメーカーのＣＤラジオカセットレコーダーのマニュアル文の読点使用について調べてみたところ、「～すると～する」という構文が繰り返し使われており、数えてみると全部で76文あった。そしてそのうち「～すると、～する」と読点が打たれているものが48文あった。読点使用率は約63%である。構文・語彙・字数が似ていて読点がある文とない文とを比較したところ、読点使用の規則性があると解釈されるものと、規則性がないと見られるものとがあった。

例25　選局がずれていますと留守録音ができません。
例26　選局がずれていますと、正常な録音ができません。
例27　分解したり、改造したりして内部に触れると、感電や故障の原因になります。
例28　内部に金属類（針、ヘアピンなど）を差し込むと感電や故障の原因になります。
例29　ＣＤが一時停止状態のとき、録音ボタンを押すとＣＤが演奏状態になり録音がスタートします。
例30　ＣＤを演奏中、お好みの曲のところにきたとき、録音ボタンを押すと、そこから録音することができます。
例31　ほこりの多い場所で使用すると故障の原因になります。
例32　そのまま使用すると、火災や感電のおそれがあり、非常に危険です。

例33　これでタイマーＯＮ時刻がくると電源が入り音が出ます。
例34　これでタイマーＯＮ時刻がくると、自動的に設定した内容で電源が入り、録音を始めます。
例35　標準プラグ付マイクの場合は、プラグアダプターＴ-27（別売）をご使用になると接続できます。
例36　一時停止状態でサーチすると、ボタンから指を離すと一時停止状態になります。
例37　早送りで最終曲の終端にくると停止状態になり、早戻しで最初の曲の頭にくると一時停止状態になります。
例38　単１形乾電池が消耗すると、「カセットホルダーが開かない」、「テープが動作しない」、「ＣＤ演奏中に動作しなくなる」、「動作中に電源が自動的に切れる」などの状態になることがあります。

例25と26、27と28、29と30、31と32、33と34を比べると、読点の規則性を見つけることが出来ない。そして、例29･32･33は節の従属関係を示すために読点が使われるのがより良い。

例29'　ＣＤが一時停止状態のとき、録音ボタンを押すとＣＤが演奏状態になり、録音がスタートします。
⇒ＣＤが一時停止状態のとき、
録音ボタンを押すとＣＤが演奏状態になり、
録音がスタートします。
例32'　そのまま使用すると火災や感電のおそれがあり、非常に危険です。
⇒そのまま使用すると火災や感電のおそれがあり、
非常に危険です。

一方、規則性が見られると解釈できるものは例35～38である。例35のように文末位置に来る場合は全て読点が打たれていなかった。例36～38では文要素を認知するのに適切に読点が打たれている。以上の検討から、読点の後処理として次のように言うことが出来る。

第一に、文要素が明確になるように読点を打つ　例29'32'
第二に、長い文には読みやすさのための読点が許容される　例27
第三に、自由な読点は障害にならない限り許容される　例26･30

また、前処理という観点からでは次のように言うことが出来る。

長文(30字以上)で読点がないものには人間が読点を打つ　例28･29'
係り受け関係を明らかにする場合に人間が読点を打つ（打ち直す）例29'32'

4　おわりに

最後に、悪文的長文([6])を例にまとめてみる。

例39　これは、もともとは床に１メートル四方の小学生の子供がしゃがんですっぽり入るくらいの深さの穴を掘って、その穴の真ん中で炭をおこし、穴の上に正方形の机を乗せ、さらにその上に、こたつぶとんという正方形のふとんをかけたものであった。（111字）

⇒これは、
もともとは床に１メートル四方の小学生の子供がしゃがんですっぽり入るくらいの深さの穴を掘って、
その穴の真ん中で炭をおこし、
穴の上に正方形の机を乗せ、
さらにその上に、
こたつぶとんという正方形のふとんをかけたものであった。

例39' これは、
もともとは、
１メートル四方の、
小学生の子供がしゃがんですっぽり入るくらいの深さの穴を床に掘って、
その穴の真ん中で炭をおこし、
穴の上に正方形の机を乗せ、
さらにその上にこたつぶとんという正方形のふとんをかけたものであった。

文要素を認知する際に構文が複雑であったり語彙の共起性が障害になるというような場合には、第三の手がかりとしての読点の役割が相対的に重くなる。本稿で試みた改行方式は、長文短文化システム、悪文推敲支援システムのひとつ、そして文の座り具合の検討にも応用することが出来る。

謝辞 本稿については、大阪大学言語文化部成田一先生に貴重な示唆をいただきました。
なお、本稿で検討した例文の一部は、下記参考文献から借用させていただきました。

主な参考文献

[１]南不二男(1974)：現代日本語の構造　大修館書店
[２]木下是雄(1981)：理科系の作文技術　中央公論社
[３]長尾眞(1983)　：制限言語の試み　自然言語処理技術シンポジウム報告集
[４]郡司隆男(1988)：言語科学への招待　丸善
[５]岡崎洋三(1988)：日本語とテンの打ち方　晩聲社
[６]牧野成一・畑佐由紀子(1989)：読解－拡大文節の認知－　荒竹出版
[７]吉村、平川、天野(1989)：自然な文章生成のための規範　情報処理学会自然言語処理研究会資料 89-NL-74
[８]熊野、吉村、野上(1990)：自然な日本語生成のための指針　情報処理学会第41回全国大会講演論文集
[９]成田一(1990)　：機械翻訳における言語処理の問題点　Kansai Linguistic Soceity 10
[10]成田一(1990)　：システム編集部の言語処理　情報処理学会研究報告90-NL-76
[11]益岡隆志・田窪行則(1992)：基礎日本語文法－改訂版－　くろしお出版
[12]本多勝一(1994)：実戦・日本語の作文技術　朝日新聞社
[13]白井、横尾、池原、木村、小見(1994)：日本語従属節の依存構造に着目した係り受け解析
情報処理学会研究報告94-NL-102
[14]成田一編著(1994)：こうすれば使える機械翻訳　バベル・プレス
[15]長尾眞・牧野武則編著(1995)：コンピュータで翻訳する　共立出版