不適切な読点配置の自動検出とその推敲支援への応用

伊藤俊一
愛知教育大学情報教育講座

1.　はじめに

伊藤らは，文中に存在する読点配置の適切性が，文節間の係り受け構造上の読点の位置によって決まることを示した（伊藤・上野(2008)，伊藤(2010)，伊藤(2011)）。他にも，村田・大野・松原(2010)，岩畑(2004)などが，読点配置の適切性を文構造との関係によって論じている。

本研究では，それらの研究で得られた知見に基づいて，不適切な読点配置を文中から自動的に検出するための方法を提案する。さらに，それらの方法を応用した読点推敲支援ツールを試作し，実際の推敲作業におけるその効果を検証する。

2.　不適切な読点配置の自動検出方法

本研究で提案する，入力文から不適切な読点配置を自動的に検出するための方法について，本章で説明する。

2.1.入力文の係り受け構造解析

最初に，入力文の文節間の係り受け構造を解析する。ただし，不適切な読点配置によって係り受け構造を誤って解析してしまうことがないように，解析自体は，入力文から読点をすべて取り去った文を用いて行なう。すなわち，ここでは読点を手がかりとせずに係り受け構造解析を行なうことになる。解析には「日本語係り受け解析器 CaboCha」（工藤・松本(2002)）を用いる。

2.2.不適切な読点配置を検出するための規則

入力文の文節間の係り受け構造を手がかりとして不適切な読点配置を検出するための規則を，次の通り，2種類設ける。

[規則 1]：
隣接接点は，読点を配置すべき優先度が極めて低い。

村田・大野・松原(2010)は，社説記事を対象とした調査を行ない，「係り受け関係にある隣接文節間192,540箇所に対して，読点が挿入されたのは5,866箇所で，挿入率は3.04%に過ぎなかった。一方，係り受け関係にない隣接文節間への挿入率は36.99%であった。」と報告している。

本研究では，係り受け関係にある隣接文節間の接点を「隣接接点」，係り受け関係にない隣接文節間の接点を「遠隔接点」と呼ぶことにする。

村田らの報告は，読点が配置される可能性が，隣接接点においては極めて低いことを示している。

[規則 2]：
上流に位置する遠隔接点は，その下流に位置する接点より，読点を配置すべき優先度が高い。

伊藤(2011)は，係り受け構造において，上流に読点を配置することが可能な接点，すなわち，まだ読点が配置されていない遠隔接点が残留しているにもかかわらず，それらを " 差し置いて " 下流の接点に読点を配置することが，読点配置の適切性を損なわせる大きな原因となることを実験的に示している。

2.3.検出結果の出力

[規則 1]あるいは[規則 2]に抵触する読点配置を持つ文節間の接点を文中から自動的に検出し，それぞれ次のように色によってマークした状態で出力する。

(a) 自身に読点が配置されている隣接接点を赤色（＼）でマークする。([規則 1]による)
(b) 上流に読点が配置されていない遠隔接点があり，かつ，自身には読点が配置されている遠隔接点を黄色（＼）でマークする。([規則 2]による)
(c) 下流に読点が配置されている遠隔接点があり，かつ，自身には読点が配置されていない遠隔接点を緑色（＿）でマークする。([規則 2]による)

2.4.動作例

文が入力されてから不適切な読点配置が検出されて出力されるまでの動作の流れを，以下に例を用いて示す。

入力文（例）：
国土交通省が、熊本県の球磨川水系に建設を進めている川辺川ダムに漁民がノーを、出している。

Cabochaによる係り受け構造の解析結果：

```
国土交通省が＼ ───┐          ← 読点Ａ
      熊本県の ┐   │
    球磨川水系に  ┤
          建設を ┤
        進めている ┐          ← 接点b
        川辺川ダムに ───┐     ← 接点c
              漁民が ──┤
              ノーを＼ ┤     ← 読点D
              出している。
```

[規則 1]および[規則 2]の適用：

読点Ａの上流に位置する接点は接点ｂと接点ｃである。これらのうち，接点ｂは隣接接点であり，[規則 1]により読点を配置すべき優先度は極めて低い。したがって，接点ｂに読点が配置されていないのは，[規則 1]に適合するものと判断される。

接点ｃは遠隔接点であり，[規則 2]により読点Ａよりも読点を配置すべき優先度が高いにも関わらず，読点が配置されていない。

読点Ｄが配置されている接点は隣接接点であり，[規則 1]により読点を配置すべき優先度が極めて低いにも関わらず，読点が配置されてしまっている。

出力文（例）：
国土交通省が＼熊本県の球磨川水系に建設を進めている川辺川ダムに＿漁民がノーを＼出している。

3. 不適切な読点配置の検出状況調査

書き手の層がそれぞれ異なる 4 種類のコーパスに含まれる文に対して，2 章で提案した「不適切な読点配置の自動検出方法」を適用してみることで，その検出状況を調査する。

本調査の目的は，大きく分けて 2 つある。1 つは，本研究で検出しようと企てているタイプの不適切な読点配置が，一般的な文章において，どのくらいの頻度で出現するものなのかを見積もることである。一般的な文章において出現する可能性がほとんど存在しないような読点配置については，そもそも，それを検出するための方法論を確立することの意義は低いと言わざるを得ないわけである。逆に，それなりの頻度で出現する不適切な読点配置については，それを検出し，問題が生じていることを指摘し，警告することには意義があると考えられる。

本調査のもう 1 つの目的は，本研究で検出しようと企てているタイプの不適切な読点配置が出現する頻度が書き手の層によってどのように異なるのかを明らかにすることである。これらの結果は，2 章で提案した検出方法を応用する読点推敲支援ツールのターゲット（利用者）を選定する際に参考にするべきものであると考えられる。（読点推敲支援ツールを用いた推敲作業については，4 章で取り上げる。）

方法：

材料　書き手の層がそれぞれ異なる 4 種類の文章（小学生の卒業文集・中学生の卒業文集・授業「レポートライティング」で提出された大学生の初稿レポート・新聞社説）を，それぞれ 60 編ずつ用意した。

手続き　2 章で設けた [規則 1]あるいは[規則 2]に抵触する文節間接点，および，それらに抵触しない接点を自動的に検出し，それぞれの出現頻度を 4 種類の文章ごとに集計した。

結果：

本調査で検出された，[規則 1]あるいは[規則 2]に抵触する文節間接点，および，それらに抵触しない接点の出現頻度を Table 1 に示す。X^2 検定の結果，出現頻度の偏りは有意であった（$X^2(12)= 678.27, p<.01$）。残差分析の結果を Table 1 内に示す。

Table 1 に示した出現頻度をもとに，4 種類の文章（小学生卒業文集・中学生卒業文集・大学生初稿レポート・新聞社説）ごとに，[規則 1]あるいは[規則 2]に抵触する文節間接点の出現率を算出した。結果を Fig.1 に示す。

小学生・中学生の卒業文集においては，いずれも，[規則 1]に抵触する読点，すなわち，本来は読点を配置すべき優先度が極めて低い隣接読点に配置されてしまった読点が有意に多く認められた。その一方で，[規則 2]に抵触する読点，すなわち，上流に読点が配置されていない遠隔接点が残留しているにも関わらず，それらを " 差し置いて " 下流の遠隔接点に配置されてしまった読点は，他と比べて特に多いという傾向は認められなかった。

大学生初稿レポートにおいては，小学生・中学生の卒業文集とは逆に，[規則 1]に抵触する読点が有意に少なかった一方で，[規則 2]に抵触する読点が有意に多く認められた。

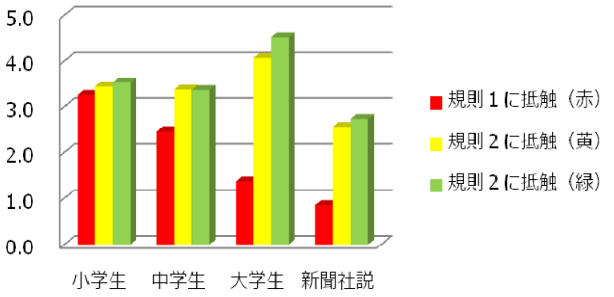
新聞社説においては，[規則 1]と[規則 2]のいずれに抵触する読点も，他と比べて有意に少なかった。

Table 1.　規則に抵触する文節間接点の出現頻度（個）

	規則 1 に抵触		規則 2 に抵触		規則に抵触しない		
	＼	＼		＼		計
小学生	415 ▲	439 n.s.	451 n.s.	1407 ▲	10054 ▽	12766
中学生	224 ▲	306 n.s.	305 n.s.	899 ▲	7341 ▽	9075
大学生	195 ▽	576 ▲	638 ▲	912 ▽	11830 ▲	14151
新聞	121 ▽	364 ▽	389 ▽	922 ▽	12427 ▲	14223
計	955	1685	1783	4140	41652	50215

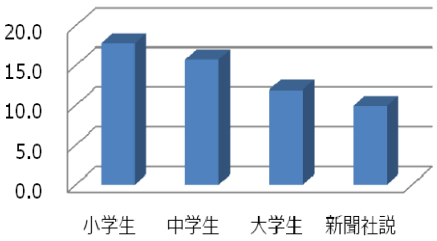
（▲有意に多い，▽有意に少ない，n.s.有意でない，p<.05）

Fig.1.　規則に抵触する文節間接点の出現率（%）



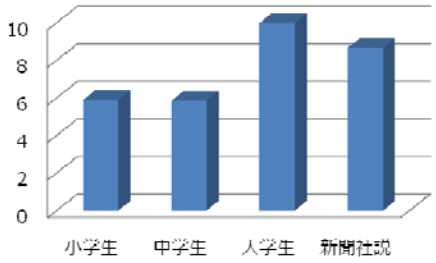
なお，4 種類の文章ごとに見た文節間接点あたりの読点の出現率は Fig.2 に示す通りであった。小学生卒業文集＞中学生卒業文集＞大学生初稿レポート＞新聞社説の順に，読点が配置される割合が高かった。

Fig.2. 文節間接点あたりの読点出現率（%）



また，4 種類の文章ごとに見た一文あたりの文節間接点数は Fig.3 に示す通りであった。小学生・中学生の卒業文集では，一文一文が少ない文節数から成り立っているのに対して，大学生初稿レポートは，そのおよそ倍の文節数から成る長文が多かった。新聞社説は，それらの中間程度の文節数から成っていた。

Fig.3. 文あたりの平均文節間接点数（個）



考察：

書き手の層ごとに，それぞれが作成した文章から検出された読点配置の特徴をまとめると次のことが言える。

小学生・中学生の卒業文集　1 文 1 文が短く，係り受け構造が単純であるため，上流の接点を " 差し置いて " 下流の遠隔接点に読点を配置してしまう問題は回避される傾向にある，ただ，必要以上に多くの文節間に読点を配置する傾向が強く，本来は優先度が極めて低いと考えられる隣接接点間にも読点を配置してしまう問題が生じやすい。

大学生初稿レポート　1 文 1 文が長く，複雑な係り受け構造を持った文が多い。そのため，上流の接点を " 差し置いて " 下流の遠隔接点に読点を配置してしまう問題が頻出する。一方で，隣接接点間に読点を配置してしまう問題は，ほぼ回避されている。

新聞社説　いずれのタイプの不適切な読点配置も出現率が低く，全般的に問題が回避されていると言える。

4.　読点推敲支援ツールを用いた推敲作業の分析

3 章では，4 種類の文章（小学生卒業文集・中学生卒業文集・大学生初稿レポート・新聞社説）の中で，特に大学生初稿レポートにおいて，上流に読点が配置されていない遠隔接点が残留しているにも関わらず，それらを " 差し置いて " 下流の遠隔接点に読点を配置してしまう問題が多く認められた。

本章では，大学生をターゲット（利用者）として想定し，2 章で提案した「不適切な読点配置の自動検出方法」を応用する読点推敲支援ツールを試作する。そして，本支援ツールを利用した文章の作成において，推敲時に初稿がどのように書き直されて最終稿に至るのかを調べる。

方法：

被験者　大学生 20 名であった。

手続き　被験者は，与えられたテーマにしたがって，30 文以上からなる文章を作成する。

文章を作成する際，文の入力は，本研究で試作した読点推敲支援ツールの入力部（エディタ）を使って行なう。文が入力される度に 2 章で設けた [規則 1]あるいは[規則 2]に抵触する文節間接点が自動的に検出され，その結果が色によってマークされた状態で出力部に表示される。

被験者は，表示される色の意味を次の文言によって予め教示される。

赤色（　）：

「そこに打たれた読点は必要ない可能性が高い。」

黄色（　）および緑色（　）：

「黄色の位置に打たれた読点が必要であるなら，
緑色の位置にも読点が必要である可能性が高い。」

被験者は，出力部に色表示された結果を参照しながら，文の推敲を行なう。

文章の作成が完了した後，被験者は，本支援ツールによる支援の有効性を問うためのアンケートに答える。アンケートは，出力部に表示されたそれぞれの色（赤色・黄色・緑色）が，推敲の際に，どの程度参考になったかを 5 段階（1:まったく参考にならなかった ～ 5:とても参考になった）で問う設問，出力部に表示された色についての評価を自由記述で問う設問，本支援ツール全般についての評価を自由記述で問う設問，から構成されている。

結果：

本実験において，被験者によって作成された文は計 661 文であった。これら 661 文のうち，推敲によって読点以外の文字列の変更が生じた 163 文は，今回の分析の対象からは除外した。すなわち，分析の対象となった文は，初稿と最終稿を比較したときに，読点以外は全て同一の文字列から成る文 498 文とした。これら 498 文のうち，初稿と最終稿で読点配置にも変更が生じなかった文は 408 文，読点配置に変更が生じた文は 90 文であった。

本読点推敲支援ツールの出力部に色表示された，[規則 1]あるいは[規則 2]に抵触する文節間接点の検出結果を受けて，その後，被験者が読点配置に関してどのような推敲行動を取ったのかを調べるために，同一の文節間接点ごとに，初稿と最終稿に対する本支援ツールの検出結果を比較し，クロス集計を行なった。その結果を，Table 2 に示す。

Table 2.　規則に抵触する文節間接点の初稿×最終稿のクロス集計（個）

		最終稿					
		赤	黄	緑	黄		計
初稿	規則 1 に抵触（赤）	28	-	-	-	32	60
	規則 2 に抵触（黄）	-	47	1	39	30	117
	規則 2 に抵触（緑）	-	1	46	52	21	120
	規則に抵触しない（黄）	-	0	1	514	0	515
	規則に抵触しない	0	0	0	1	3327	3328
	計	28	48	48	606	3410	4140

Table 2 より，規則に抵触することが検出された文節間接点のうち，推敲によって最終稿でその問題が解消された接点の比率は，それぞれ，

[規則 1]（■）：53.3%
[規則 2]（■）：59.0%
[規則 2]（■）：60.8%

であった。

[規則 2] に抵触する文節間接点において，その問題を解消するためには，次の 2 通りの方法のいずれかを取る必要がある。

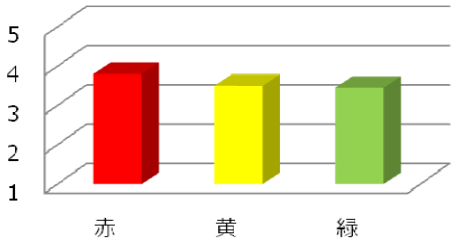
(a) 読点が配置されずに残留している上流の遠隔接点（■）に新たに読点を追加する。
(b) 残留している上流の遠隔接点を " 差し置いて " 下流の遠隔接点に配置されている読点（■）を削除する。

[規則 2]（■）に抵触することが検出された接点のうち，56.5%の接点においては(a)の方法が，残りの 43.5%においては(b)の方法が講じられた。同様に，[規則 2]（■）に抵触することが検出された接点のうち，71.2%の接点においては(a)の方法が，残りの 28.8%においては(b)の方法が講じられた。

[規則 1]と[規則 2]のいずれにも抵触しなかった接点，すなわち，不適切な読点配置が検出されなかった接点について見ると，推敲を経て最終稿で新たな問題が発生し検出された接点は 1 例（0.03%）に過ぎなかった。

文章作成が完了した後に，本支援ツールが出力する色表示について被験者がその有効性を 5 段階で評定した結果を Fig.4 に示す。赤色，黄色，緑色，いずれの色表示についても，平均評定値は，3:どちらともいえない ～ 4:やや参考になった，の間の値であった。特に，赤色の評価が，黄色，緑色に比べて高かった。

Fig.4. 色表示の有効性の平均評定値



考察：

本支援ツールで問題が検出された文節間接点のうち半数以上は，推敲時に，それらの問題が検出されなくなるように最終稿で書き直しが行なわれていた。逆に，推敲によって最終稿で新たな問題が検出された例は，ほぼ皆無に等しかった。

文章作成完了後のアンケートでは，本支援ツールが出力する色表示について，推敲時に「参考になった」という側に評価が傾いた。

これらのことから，本支援ツールによる読点配置上の問題の検出および利用者に対するその指摘は，推敲支援において一定の成果を上げていたと言える。

5. 今後の課題

4 章の実験で被験者によって作成された文のうち，推敲時に読点配置に変更が生じた 90 文については，改めて，読み手の立場から初稿と最終稿の読点配置の適切性を比較評価するための実験を行なった（伊藤，準備中）。その結果，本支援ツールによって初稿で問題が検出された文節間接点に対して，それらの問題が検出されなくなるように最終稿で書き直しが行なわれた文の中には，読み手の立場から見た評価がかえって低下してしまったものも含まれることがわかった。その 1 つの原因としては，本研究で不適切な読点配置を検出する際に拠り所とした，読点を手がかりとしない係り受け構造解析が，必ずしも適切な係り受け構造を出力しなかったことが挙げられる。そのために，本支援ツールが誤った検出結果を被験者に伝えてしまい，被験者は，かえって不適切な推敲を促されてしまった可能性がある。

また，2 章で示した[規則 1]および[規則 2]に抵触しないことだけでは解消できない読点配置上の問題が，本支援ツールにおいては検出結果として被験者に伝えられなかったため，かえってそれらの問題が推敲時に軽視あるいは無視されてしまった可能性がある。さらには，[規則 1]および[規則 2]に抵触しないことに特化して行われた書き直しが別の新たな問題を発生させた可能性も考えられる。本研究では扱えなかったそれらの問題の性質，それらを検出するための方法，解消するための方法については，さらに研究を進める必要がある。

引用文献

伊藤俊一・上野慎之介 (2008) 文推敲者による読点打ち行動の分析 言語処理学会第 14 回年次大会発表論文集，1101-1104.

伊藤俊一 (2010) 打点方略が読点配置の適切性に及ぼす影響 言語処理学会第 16 回年次大会発表論文集，407-410.

伊藤俊一 (2011) 読点配置の適切性を規定する文構造上の要因について 言語処理学会第 17 回年次大会発表論文集，695-698.

岩畑貴弘 (2004) 読点の使用とその決定要素について 神奈川大学「人文研究」，154，51-81.

工藤拓・松本裕治 (2002) チャンキングの段階適用による日本語係り受け解析 情報処理学会論文誌，43(6)，1834-1842.

村田匡輝・大野誠寛・松原茂樹 (2010) 日本語テキストにおける読点位置の検出 言語処理学会第 16 回年次大会発表論文集，812-815.

注

1) 本研究における実験の実施，および，データの分析にあたり，森部大悟氏（愛知教育大学大学院教育学研究科発達教育科学専攻情報教育領域）の協力を得た