自動車整備士試験問題に使用される文節の出現頻度について

青木恒夫*

1．は　じ　め　に

中日本自動車短期大学（以下「本学」とする。）では，2005年より自動車分野の知識を日本で学びたいとする留学生向けに日本語を教育する留学生別科を，2009年からは同じく留学生向けに３年間で二級自動車整備士資格の取得を目指す国際自動車工学科を開設している。留学生諸君にとっては，日本語の習得が最も大きなハードルではあるが，とりわけ専門用語の多い自動車工学分野の学習には多くの負担が伴う。参考文献[2)]では，二級自動車整備士試験問題に含まれる専門用語を自動的に抽出する試みをしているが，この際に作成した自動車用語辞書は，自動車専門用語の名詞を中心に拾い出したものであり，コンピューターの日本語変換辞書への応用で，統一した自動車用語の入力などに貢献した。今回，留学生別科教員からの「効果的な日本語教育の基礎資料とするため，付属語を含んだ自動車専門用語の活用について，自動車整備士試験問題に出現する文節ごとの出現数をリストにできないか？」とのご意見から，自動車整備士試験問題に使用される文節について，その出現頻度を求める実験を行ったので，結果を報告する。

2．実　験　の　手　順

2．1　概要

自動車整備士試験問題も通常の日本語文と同じく，複数の文節（１つの自立語と０個以上の付属語で構成される。)が組み合わされてできている。実験は，自動車整備士試験問題のテキスト(以下「原文」とする。）から，あらかじめ用意された文節辞書（以下「辞書」とする。）に含まれる文節を１つずつ検索し，その検索数から文節ごとの出現頻度リストを作成するものである。

辞書は原文中の全種類の文節を含んでいる必要があり，原文を手作業で文節分けする作業から始める。原文には同一文節が複数存在する場合が殆どだが，辞書には同じ文節が２個以上含まれないよう，手作業で文節分けを実施した後，プログラム処理により重複する文節を削除する処置を施した。

原文は2003年７月から2013年３月までの24回分の二級ガソリン自動車整備士試験問題のテキス

*中日本自動車短期大学 MSE学科 教授

ト（7162行, 397387バイト, Shift-JIS）を用いた。テキストは紙ベースの問題文からOCR（Optical Character Recognition：光学文字認識）またはキーボードからの直接入力によりPC上で作成している。2バイト全角文字と改行コードのみで構成されているので，おおよそ19万文字のプレーン・テキストとなる。

原文から辞書に含まれる文節を1つずつ検索し，その数をカウントする作業について，最初は，OS（オペレーティング・システム：実験ではMicrosoft Windows 7）で使えて，無料で配布されているプログラム（wcやawkなどUNIX系のコマンド）を組み合わせて使用することを検討したが，試行錯誤を繰り返しているうち，専用のプログラムの方が効率的であることがわかり，この実験用に新たな文字列検索プログラム（strcount.exe）を開発した。通常の文節分けであれば，その最小単位は「1個の自立語と0個以上の付属語」となるが，自動車整備士試験問題では専門用語の多くは複合語であり（例：「｜かじ取り装置の｜パワー・ステアリング装置の｜油漏れが｜なく｜,」[筆者注] 筆者が入れた文節区切りを縦棒「｜」で示している。），括弧書きを含む場合も多い（例：「｜「自動車点検基準」に｜照らし｜,」や「｜オクタン価（90～95）を｜高めて｜いる｜」など。）。また，文意から判断しないと文節分けが難しいものも多い（例：「大きい｜ときほど｜小さな」[筆者注]「大きいと」という文節が別に存在すると，文節のカウントが正しくできないばかりか，文節分けにヒットされない文字が残される可能性がでてくる。）。このことから，原文から1つの文節を検索した後，検索された文節を原文から削除し，一旦検索された文節は，次の検索では原文中に存在しないという手法で，文節出現数の正確なカウントを実現した。なお，上記のような理由で，原文から抽出した文節内には更に別の小さな文節が存在する可能性があることから，文節の検索・削除は文字長の長いものから順に実施するようにする。

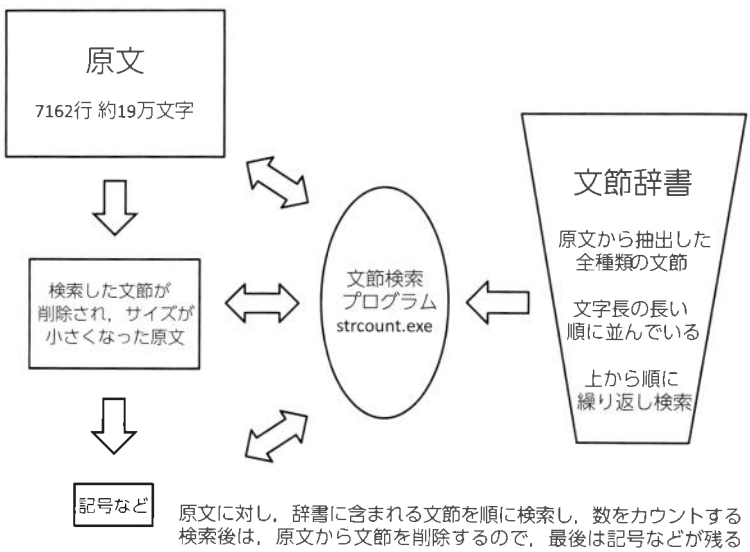

図2.1　各文節の出現数を調査する方法

辞書に含まれる文節の検索が終わると，原文には辞書に含まれていない文字（この場合，文節とはならない問題番号や記号なども含まれる。）が残されることになり，すべての文節の出現数をカウントできるとともに，文節分けできなかった文節が残っていないかも確認できる。これら一連の流れを図2.1に示す。

2．2　辞書の作成

辞書は，１文節１行とするテキスト・ファイルで，原文に含まれるすべての文節が網羅されている必要がある。このため，原文をテキスト・エディタで開き，先頭から目視で文節を判断して手作業で残していく「文節分け作業」を先に行った。この場合，各試験問題の先頭に書かれている試験タイトルや問題番号，選択肢番号などは削除していく。なお，文節ではないが，原文中に含まれる各問の正解情報（「(正解４)」などと，各問に挿入されている。）については，参考文献[3]に関して興味があり，残すこととした。

文節分けの基本方針については，次項に詳しく述べるが，この辞書作成には非常に多くの時間が費やされた。原文7162行を１回分の試験に平均すると約300行となるが，この中に数千個の文節が含まれる。作業の初期段階では，文節分けが遅々として進まない状況がしばらく続いた。同じ文節が複数出現するので，これをすべて残していく作業は大変効率が悪い。上記の作業が，試験問題７回分を終了した時点で，試しに原文から抽出した文節を削除する作業を行うことにした。これは，これまで抽出した文節の多くは，かなりの割合で残りの原文にも存在するだろうという期待からで，少しでも文節抽出作業を効率化したいという目的である。

この時点で，原文から文節のみ仕分けた作業ファイルは10147行（10147文節）あり，これには重複した文節が多数含まれている。これを以下の手法で，重複した文節を含まない中間辞書（[筆者注] 最終的な辞書ではないので，「中間辞書」と呼ぶことにする。）を作成した。コマンドラインより，

> sort 作業ファイル > ソートファイル　　……　単純ソート，uniqの前処理（10147行）
> uniq ソートファイル > 中間辞書ファイル ……　重複行の削除（3883行）

この処理を行った結果，重複しない3883個の文節が抽出できた。作成された中間辞書ファイルを表計算ソフトExcelに読込み，関数「Len」で得た文字長を使って，長さ降順で並べ替える。

次に原文に対して，図2.2のバッチ・ファイルを適用した。

バッチ・ファイルは単純なコマンドの繰り返しであるので，中間辞書を元にテキスト・エディタのマクロを利用して作成した。「strcount」は，今回の実験で作成した検索プログラムで，「-d」オプションを添えると，原文から指定された文節を検索後に削除し，残りを標準出力へ出力する。すなわち，原文から指定した文節を削除する機能がある。通常，標準出力は画面なので，これを中間ファイルにリダイレクトして，次々と文節の削除を繰り返していく。原文に対して複数の文節削除を実行すると，最後に残された原文（図2.2では，「結果.txt」）は，中間辞書に含まれるすべての文節が削除されているので，残った原文に対して引き続き文節分けを行えば，かなり良い

```
echo off
copy 原文.txt temp1.txt                    ..... 原文を作業ファイル1へコピー
set /A LINE = 0

set /A LINE = LINE + 1
strcount -d temp1.txt 文節0001 > temp2.txt  ..... 「-d」オプションは，中間辞書の文節を検索して，
copy /Y temp2.txt temp1.txt                       原文から削除する。結果は作業ファイル2へ
echo  %LINE%                                      続いて作業ファイル2を作業ファイル1にコピー

set /A LINE = LINE + 1
strcount -d temp1.txt 文節0002 > temp2.txt  ..... 同様の文節削除作業を全文節に対して実行
copy /Y temp2.txt temp1.txt
echo  %LINE%
      :
   以下繰り返し
      :
set /A LINE = LINE + 1
strcount -d temp1.txt 文節n > temp2.txt
copy /Y temp2.txt 結果.txt                 ..... 最終結果は「結果.txt」に残る
echo  %LINE%                                      原文から中間辞書の文節が削除されているので
                                                  続く文節分け作業の効率が向上する
echo Finished
```

図2.2　文節を削除するバッチ・ファイル

効率で文節分け作業が行える。ただし，この文節削除作業については，文字長の長い順に適用する必要がある。

なお，このような中間辞書を使った文節削除は，残された原文に含まれるタイプ・ミスを発見することに役だった。本来なら統一された用語からできている原文が，紙ベースの試験問題からテキストに起こされる段階で，タイプ・ミスが生じている場合があり，肉眼ではなかなか見分けのつかない違い（「一」（数字の），「ー」（長音記号），「－」（マイナス記号）など。）が検索では区別されるため，予想外のタイプ・ミスに気づくことが多々あった。

手作業による文節分け作業と中間辞書を使った原文からの文節削除を数回行った結果，最終的に7414個の文節が抽出できた。この7414個の文節を原文に適用して削除すると，原文には１つの文節も残らないことになる。別の言い方をすれば，原文は7414個の文節の組み合わせであるとも言える。

２．３　文節分け

前項の辞書作成作業は，手作業による「文節分け」の作業を前提にしている。一般的な文節は，「１個の自立語と０個以上の付属語」を最小単位としており，参考文献[6]では「言語単位の一つ。文を読む際，自然な発音によって区切られる最小の単位。」とある。初等教育では，「ネ」「サ」「ヨ」などを挿入して，意味の通じるところで区切ると教えられる。

文節分けを始めた最初の段階では，原文に対して一般的な文節分け（分かち書き）の手法で作業を進めた。ところが，前述したように，自動車整備士試験問題には，「複合語」，「但し書き」，「括弧書き」といった，コンピューターで一括処理が難しい文が多数存在し，正しい文節出現頻度を計測するのが難しく思えた。そこで，できる限り問題文の「文意」を損なわないよう，あまり細かい単位の文節には拘らないよう，拡大した文節の解釈を用いることとした。以下に，今回用い

た主な文節分けの方針と，それにより分けられた１文節の例を示す。

・名詞の複合語，中点「・」区切りの外来語は，１個の名詞として扱う。

例)「リング・ギヤが」「ポペット先端の」「ドライブ回路に」「無負荷運転して」

・括弧書き，但し書きは，括弧やカギ括弧を文節に含める。

例)「比重（20℃）が」「変速点（車速）は」「「輪荷重」とは」「「自動車点検基準」に」

・細かく文節分けをすると，文字が取り残される可能性がある場合は，１つにまとめる。

例)「大きいときには」(「大きい」「大きいと」「大きいとき」などを含むので。)

・単位を含む数字，数字を含む量は，ひとまとめに名詞として扱う。

例)「５．６Ａで」「５分の１で」「４／８から」「３人以上で」「「１．５」は」「“０”となり」

・意味上は１つと考えられるものは，１つに扱う。

例)「A，B間の」「低温・低圧の」「点検・修正に」「円周率（π）＝」

・虫食い問題の括弧付き記号（空白も含む）も文字として扱う。

例)「(　) 以下で」「(イ) なので」「(ハ) レベルと」

3．文節数カウント・プログラム

3．1　文節数カウント･プログラムの概要

原文に散らばっている文節を検索して，数を数えるプログラムについて，当初は一般的なUNIXコマンドを組み合わせて実現することを考えた。ただ，これまでの辞書作成手法を考えたとき，専用のプログラムを開発した方が，効率的であることがわかった。そこで，コマンドラインで動く簡単な文字列検索プログラム（strcount.exe）を新たに作成した。開発はWindows 7で動くMicrosoft Visual Studio 2010の環境上で行い，プログラミング言語はC#を用いた。

3．2　strcountの仕様

以下に文字列検索プログラム「strcount.exe」の仕様を示す。

［概要］

ファイルから指定の文字列を検索し，結果を標準出力へ出力する。

［使用法］

Strcount オプション ファイル名 検索文字列［return］

［オプション］

-c：検索数を出力

-i：検索位置を出力

-d：検索文字を削除

［出力の書式］

-c → 一致する検索文字数の情報を，以下の書式で標準出力に出力する。

［出力例］

ドライブ，29（検索文字列，検索数）

※ファイル中に29個の「ドライブ」が存在することを意味する。

-i → 検索文字列に一致したすべての文字について，「検索文字列，連番，先頭からの行数，先頭からの文字数」を標準出力へ出力する。

［出力例］

検索文字，連番，行，文字　　…… 項目名（１行目）

ドライブ，１，72，2055　　…… 検索された文字列について，先頭から位置情報を出力

ドライブ，２，75，2326

ドライブ，３，740，20391

：　　…… 検索されただけ，繰り返し出力される

※ファイル中に存在する検索文字のすべての位置情報を出力する。

-d → 検索文字列以外のすべての文字を標準出力へ出力する。

※検索文字を除いた残りのテキストを出力する。結果的に，検索文字列を削除することになる。

４．文節出現数のカウント

４．１　文節出現数のカウント方法

これまでの手法で作成された辞書および文字列検索プログラム「strcount.exe」を使って，原文に含まれるすべての文節について，出現数のカウントとリストの作成を試みた。

既に辞書作成で，原文からの文節削除を行っているが，これに加えて文節の出現数をカウント

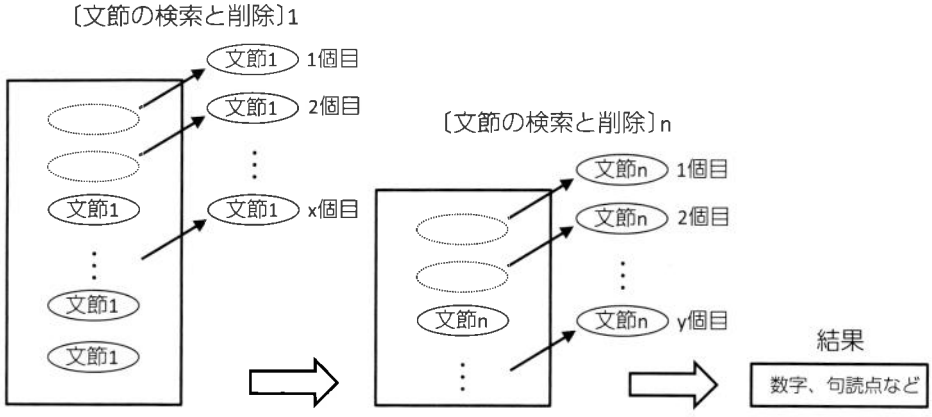

図4.1　すべての文節について出現頻度を調査する方法

```
echo off
copy 原文.txt temp1.txt                    ..... 原文を作業ファイル１へコピー
set /A LINE = 0

set /A LINE = LINE + 1
strcount -c temp1.txt 文節1 >> result.txt  ..... 文節を検索し，結果をresult.txtに 追加保存
echo. >> result.txt                        ..... result.txt に改行出力
strcount -d temp1.txt 文節1 > temp2.txt    ..... 文節を削除して作業ファイル２へ出力
copy /Y temp2.txt temp1.txt                ..... 作業ファイル２を作業ファイル１へコピー
echo  %LINE%
set /A LINE = LINE + 1
      :
   以下繰り返し
      :
strcount -c temp1.txt 文節n >> result.txt  ..... 結果が result.txt に集約される
echo. >> result.txt
strcount -d temp1.txt 文節n > temp2.txt
copy /Y temp2.txt 残り.txt                 ..... 残り.txtには，辞書に含まれない
                                                 問題番号や句読点が残る
echo  %LINE%
echo Finished!
```

図4.2　すべての文節について出現頻度を調査するバッチ・ファイル

する。作業の流れは概ね 図4.1のようになる。辞書（文字長が長い順に並んでいる。）に含まれる文節を原文から検索し（strcount -c），結果をリストに出力する。直後に同じ文節を原文から削除（strcount -d）する。引き続き，少し小さくなった原文に対し次の文節を検索し，同じく原文から削除する。この作業を辞書の末尾まで繰り返す。結果的に，原文には辞書に存在しない文字のみ残り，通常は問題番号や句読点，数字のみ残ることが期待できる。

上記の作業を連続的に行うため，図4.2のようなバッチ・ファイルを作成した。バッチ・ファイルは単純な繰り返しであるので，辞書を元にテキスト･エディタのマクロを使用して作成した。実行後にリスト「result.txt」が作成されるが，「文節,出現数」というCSV形式になっており，表計算ソフトなどでも利用しやすい形式である。

４．２　実験の結果

上記のバッチ・ファイルを実行した結果，原文（図4.2では「残り.txt」）には，期待通り「タイトル」「番号」「選択肢番号」「句読点」「図番号」「単位を伴わない数値」のみ残され，すべての文節が検索できた。作成された「result.txt」を表計算ソフトに読み込み，次項に結果を概観する。

４．２．１　文節の出現頻度

表4.1に文節の出現頻度上位60を示す。

上位に自動車関連用語を含む文節は比較的少なく，22位「タイヤの」，24位「自動車の」，41位「エンジンの」，表には一部しか表示されていないが，59位「「道路運送車両の保安基準」」，60位「「道路運送車両の保安基準の細目を定める告示」に」（59位，60位はいずれもカギ括弧を含む）の５種類であった。

１位から５位は，四択問題の解答方法を指示する「適切なものは次のうちどれか。」を構成す

表4.1　原文に使われている文節の出現頻度（上位60）

No.	文節	個数	No.	文節	個数	No.	文節	個数
1	次の	980	21	ときの	135	41	エンジンの	76
2	うち	961	22	タイヤの	132	42	一般に	76
3	どれか	960	23	示す	129	43	ほど	76
4	ものは	673	24	自動車の	123	44	約	76
5	適切な	659	25	とき	120	45	行う	74
6	いる	562	26	大きい	112	46	発生する	71
7	関する	524	27	文章の	112	47	こと	69
8	記述として	492	28	（イ）	110	48	高く	68
9	ある	429	29	（ロ）	108	49	もので	65
10	不適切な	296	30	照らし	101	50	又は	64
11	なる	279	31	ため	99	51	方が	63
12	（正解３）	271	32	図に	99	52	ときに	61
13	及び	265	33	当てはまる	98	53	よって	60
14	（正解２）	264	34	できる	94	54	ことを	59
15	（正解４）	256	35	おいて	92	55	小さい	59
16	（正解１）	169	36	比べて	87	56	低く	57
17	いう	157	37	小さく	85	57	なると	53
18	する	145	38	下の	84	58	その	52
19	大きく	144	39	組み合わせの	82	59	「道路運送車両の保	51
20	ものとして	136	40	（　）に	82	60	「道路運送車両の保	49

るすべての文節である。二級自動車整備士試験問題は１回分が40問題から構成される。したがって，24回分960問のほとんどに上記の「決まり文句」が使われていることになる。この指示とは別に「不適切なものは次のうちどれか。」という設問もあり，10位に「不適切」が出現している。

12位と14位から16位は四択問題の正解であり，二級ガソリン自動車整備士試験問題では，「正解３」（28.2%）が最も多く，以下順に「正解２」（27.5%），「正解４」（26.7%），「正解１」（17.6%）と並ぶ。

４．２．２　同じ自立語からなる文節の変化

表4.2は，同じ自立語（複合語も含む）からなる文節の変化の例を示している。日本語は付属

表4.2　同じ自立語からなる文節の変化例

文節	個数	文節	個数	文節	個数
コントロール・ユニット	4	ターボ・チャージャで	2	取り付け	3
コントロール・ユニットが	4	ターボ・チャージャと	1	取り付けた	6
コントロール・ユニットからの	8	ターボ・チャージャに	14	取り付けられ	1
コントロール・ユニットで	3	ターボ・チャージャの	6	取り付けられた	6
コントロール・ユニットに	10	ターボ・チャージャは	7	取り付けられて	11
コントロール・ユニットにより	1			取り付けられなければ	1
コントロール・ユニットの	5	自己診断コード表と	1	取り付けると	1
コントロール・ユニットは	3	自己診断コード表の	1	取り付けるときは	2
コントロール・ユニットへ	2	自己診断システムが	3	取り付け位置の	3
		自己診断システムにより	1	取り付け高さが	1
シリンダが	2	自己診断システムの	1	取り付け面の	2
シリンダとして	14	自己診断システムは	1		
シリンダに	3	自己診断する	1	大きい	112
シリンダの	17	自己診断を	1	大きいが	2
シリンダへ	5	自己診断用コネクタとして	1	大きいと	2
				大きいときに	4
トルク・コンバータで	2	周速が	1	大きいときには	4
トルク・コンバータと	1	周速で	3	大きいときの	1
トルク・コンバータに	9	周速と	2	大きいときは	3
トルク・コンバータの	12	周速の	3	大きいので	13

語の種類が多く，様々な結びつきで文節が多様化しているが，自立語を中心に眺めてみると変化の違いがよく分かり，日本語初学者の理解を助けるかもしれない。

4．2．3　文節長さの分布傾向

今回の実験では，7414種類の文節について，その出現頻度を調べたが，最も長い文節長が47文字（「道路運送車両の保安基準第２章及び第３章の規定の適用関係の整理のため必要な事項を定める告示」に）であった。カギ括弧を取れば，10個以上の文節に分けられることから，本実験の文節分け方針による特例となる。最も短い文節長は１文字で，「÷（16），音（７），後（19），誤（13），際（５），時（５），錫（１），正（11），又（１），約（76）」（カッコ内は出現数）の10種類であった。文節長さの平均は5.8文字で，その分布を図4.3に示す。３文字(1442個)，４文字(1392文字)，５文字（1298文字）が他に比べて圧倒的に多かった。

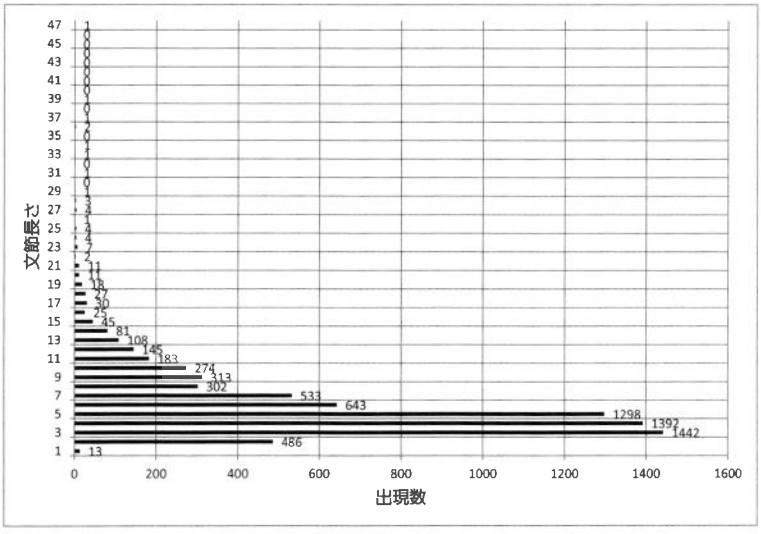

図4.3　文節長さの分布

5．ま　と　め

初めての試みとして，二級ガソリン自動車整備士試験問題に含まれる文節の出現頻度を調べてみた。大半の時間を文節辞書の作成に費やした。機械的に調査する関係で，通常の文節分けとは幾分違ったルールを設定したが，日本語文に含まれる文節の多様さ，自動車専門用語の多さに驚かされた。今回は，文節辞書の作成と出現頻度を調べることにテーマをおいたため，実験結果に対する十分な考察が出来ていない。今後，結果を精査し，更に精度の高い文節辞書の作成や二級ガソリン自動車整備士試験問題以外の整備士試験問題での調査を進める計画である。また，出現頻度データは，学内で共有し，自動車技術を学ぼうとする留学生諸君の日本語学習に供されることを期待する。

最後に，この実験を実施する機会を与えていただいた留学生別科の先生方に謝意を表します。

参　考　文　献

1）青木恒夫，２級自動車整備士試験問題における過去20年間の出題状況（工学一般分野），中日本自動車短期大学 論叢,94（1991）
2）青木恒夫, 二級自動車整備士試験問題からのキーワード自動抽出の試み, 中日本自動車短期大学 論叢,57（1999）
3）青木恒夫，自動車整備士登録試験における正解群の偏りについて，中日本自動車短期大学 論叢,61（2010）
4）益岡隆志・田窪行則，基礎日本語文法―改訂版―，くろしお出版,（1996）
5）日本点字委員会，日本点字表記法 2001年版，日本点字委員会,29（2008）
6）新村　出，広辞苑第六版 DVD-ROM版，株式会社岩波書店,（2008）