放送教育開発センター研究紀要第12号（1995年）
Bulletin of the National Institute of Multimedia Education, No. 12, 1995

研究ノート

ニューラル・ネットワーク・モデルによる日本語文節の識別の試み

吉田文彦

Clause Recognition of Written Japanese by Neural Network Model

Fumihiko Yoshida

Abstract

Recognizing that clauses in Japanese tend to end with a relatively limited number of letters, this study attempted to recognize ends of clauses in written Japanese by a neural network model with back propagation.

Focusing on ten letters which are most likely to be used as a last letter in Japanese clauses, this study set up a neural network model to learn to tell whether or not an encountered letter (target letter) was used as an ending letter of a clause.

Although only a partial tests were attempted due to the relatively small sizes of the data sets, the obtained results were quite encouraging. Based on the findings, this study suggests four points of further improvement: (1) an enlargement of volume of data set, (2) an improvement of data generation method, (3) an examination of cases which were not properly recognized, and (4) a possible use of a certain type of dictionary as well as an inclusion of grammatical rules.

キーワード

分かち書き、文節認識、自然言語処理、神経回路網

東海大学文学部広報学科教授

1．はじめに

日本語の文章をコンピュータで処理する際にまず問題になるのは、単語や文節の切れ目をどうやって識別するかということである。この問題に対処するために、最長一致法をはじめとする各種の手法が考案され、その成果は既に日本語ワードプロセッサにも組み込まれ、多くの人々が特に意識することなく日常的にその恩恵にあずかっている。また、日本語の文章を文節により分かち書きするシステムもかなり以前から開発されている[1)]。

本稿で報告する研究の目的は、日本語の文節を識別する方式として、ニューラル・ネットワーク・モデルを用いた手法を開発することにある。なんらかの適切な方法で日本語文章の文節の切れ目をニューラル・ネットワーク・モデルに学習させれば、既存の方式のように構築の困難な文法ルール・システムや膨大な数の語彙からなるコンピュータ化された辞書に頼らずとも、実用に耐える精度で文節の識別（即ち、語の切り出し）が可能になるのではないだろうか、との予想がこの研究の根底となっている。また、システム自体が比較的単純なものですめば、保守は当然容易であろうし、さらには学習済みのいくつかのニューラル・ネットワーク・モデルを組み合わせたり、あるいは新たに追加したりするなど、システムの拡張も容易に行えるはずである。

もしそのような方式が完成すれば、比較的単純なシステムで日本語の文章から語の切り出しが出来ることになり、いろいろな応用が可能となる。例えば、コンピュータによる日本文の内容分析システムを開発する場合には、その基本モジュールのひとつとして利用できることになろう。そのようなシステムの開発は、マスメディアによる報道、文学作品、あるいは教科書などの内容の体系的分析に役立つはずである。この意味で、本稿で扱う文節識別方式が実用化されることの意義は大きいものと考えられる。

以下では、まずこの研究の基本的な考え方とデータ作成の方法を説明し、ついでニューラル・ネットワーク・モデルによる文節識別の結果を報告し、最後に今後の研究へ向けての改善点を論じる。

2．基本的な考え方とデータ作成の方法

日本語の文章から語を切り出す（あるいは分かち書きをする）ためには、語尾または文節末を識別する必要があるが、それには文法のルールを利用するのもひとつの方法であるし、また辞書による識別方法も可能であろう。本稿ではそれらとは違う第三の方式として、いわゆるヒューリスティックスをニューラル・ネットワーク・モデルに組み込むかたちで利用することを試みる[2)]。

例えば、ふたつの漢字の間にひらがなが挟まれている場合は、そのひらがなで文節が終わっている可能性が高いとか、いくつかの漢字が続いている場合はそれらはひとつの文節に含まれる可能性が高いといったヒューリスティックスを、ニューラル・ネットワーク・モデルに学習させるわけである。本稿で試みた方式は、「が」とか「に」などの文節の終わ

りに多用される文字を文節末の候補として選び出し、それらの文字の前後に位置するいくつかの文字に関して、その字種や文字自体に関する情報をニューラル・ネットワーク・モデルに入力し、ヒューリスティックスを明示的なかたちではなく、間接的なかたちで利用することで、文節末を識別する方式をとっている。

ニューラル・ネットワーク・モデルにはラメルハートのバックプロパゲーション・モデルを利用しており、3層、4層、5層からなる3種類のモデルを学習データに応じて使い分けている。学習にあたっては、学習パターンごとに重みを逐次修正する方式をとった。また、収束を速める目的で、1ステップ前の修正の影響を加味する、慣性項を用いた。なお、いわゆる学習係数と慣性係数に関しては、全てのモデルにおいて前者には0.6、後者には0.8の各数値を一貫して用いた。

文節識別のための学習用データならびに学習結果による識別テストに用いるデータは、新聞記事をもとに作成した。使用したのは1992年1月から3月末までの朝日新聞の在日外国人に関する60の記事で、CD-ROMにおさめられたものを利用した[4]。

実際のデータ作成に当たっては次の手順に従った。まず、CD-ROMの検索ソフトを利用して記事を選び出し、ハード・ディスクにコピーした。いずれの記事も見出しや日付などを取り除き本文のみのかたちにした。ついで人名の後の年齢などのように括弧でくくられたかたちで本文中に挿入されている数字やそれに類する語句を取り除いた。さらに、(1)句読点、(2)かぎかっこ、(3)ひらがなの「を」、を目印に、どの文章もそれらが出現するたびにその位置の前後でさらに短い単位に分割した。すなわち、どのセンテンスも文の途中であっても明かに文節が終わっていると機械的に判断できる箇所でいくつかに分割されたことになる。以上のプロセスから得られたテキストファイルが数値データ作成のベースとなった。

テキスト・ファイルの日本語の文章から文節識別のための学習データを生成するにあたっては、まず文節末の候補としてとりあげる文字をひらがなの10文字に限定する方式をとった。それらは、「が」、「た」、「て」、「で」、「と」、「に」、「の」、「は」、「も」、「る」の10文字である。さらにこれら10文字の各々について、個別のデータセットを作成する方式をとった。これらの文字のみに限定したのは、少なくとも選び出した新聞記事の文章を見る限り、これらの文字が文節の終わりとなっている場合がきわめて多かったからである。

これらの10文字の各々に関して学習用データと教師データとを作成した。学習用データには表1に示されている94項目の1あるいは0からなるデータを作成した。教師データは文節末であるかどうかを1あるいは0で示すかたちとした。教師データの作成にあたっては、各候補文字に関してそれが文節末なのかどうかを判定する必要があるが、基本的には常識的な判断を繰り返すだけの単純な作業であり、多少の手間はかかっても、ニューラル・ネットワーク・モデルによる分かち書き方式の利点を損なうほどのものではない。

なお、いずれの候補文字の場合も、句読点、かぎかっこなどの記号、ひらがなの「を」の直前に位置する場合は、データには取り入れられなかった。というのは、このような位置に当該文字が存在する限り、次の文字あるいは次の次の文字で文節が終わっていることは明かであり、わざわざそのようなデータをニューラル・ネットワーク・モデルに学習させることは不要だからである。また、膨大な量のデータを正確かつ短時間で作成する必要

表1　ニューラル・ネットワークの入力データ

	測定項目
1	3文字手前の文字タイプはその他か？
2	3文字手前の文字タイプは記号か？
3	3文字手前の文字タイプは数字か？
4	3文字手前の文字タイプは英字か？
5	3文字手前の文字タイプはひらがなか？
6	3文字手前の文字タイプはカタカナか？
7	3文字手前の文字タイプは漢字か？
8	2文字手前の文字タイプはその他か？
9	2文字手前の文字タイプは記号か？
10	2文字手前の文字タイプは数字か？
11	2文字手前の文字タイプは英字か？
12	2文字手前の文字タイプはひらがなか？
13	2文字手前の文字タイプはカタカナか？
14	2文字手前の文字タイプは漢字か？
15	1文字手前の文字タイプはその他か？
16	1文字手前の文字タイプは記号か？
17	1文字手前の文字タイプは数字か？
18	1文字手前の文字タイプは英字か？
19	1文字手前の文字タイプはひらがなか？
20	1文字手前の文字タイプはカタカナか？
21	1文字手前の文字タイプは漢字か？
22	1文字後ろの文字タイプはその他か？
23	1文字後ろの文字タイプは記号か？
24	1文字後ろの文字タイプは数字か？
25	1文字後ろの文字タイプは英字か？
26	1文字後ろの文字タイプはひらがなか？
27	1文字後ろの文字タイプはカタカナか？
28	1文字後ろの文字タイプは漢字か？
29	2文字後ろの文字タイプはその他か？
30	2文字後ろの文字タイプは記号か？
31	2文字後ろの文字タイプは数字か？
32	2文字後ろの文字タイプは英字か？
33	2文字後ろの文字タイプはひらがなか？
34	2文字後ろの文字タイプはカタカナか？
35	2文字後ろの文字タイプは漢字か？
36	3文字後ろの文字タイプはその他か？
37	3文字後ろの文字タイプは記号か？
38	3文字後ろの文字タイプは数字か？
39	3文字後ろの文字タイプは英字か？
40	3文字後ろの文字タイプはひらがなか？
41	3文字後ろの文字タイプはカタカナか？
42	3文字後ろの文字タイプは漢字か？
43	2字手前はひらがなの「が」か？
44	2字手前はひらがなの「た」か？
45	2字手前はひらがなの「て」か？
46	2字手前はひらがなの「で」か？
47	2字手前はひらがなの「と」か？
48	2字手前はひらがなの「に」か？
49	2字手前はひらがなの「の」か？
50	2字手前はひらがなの「は」か？
51	2字手前はひらがなの「も」か？
52	2字手前はひらがなの「る」か？

表1 （続き）

	測定項目
53	1字手前はひらがなの「が」か？
54	1字手前はひらがなの「た」か？
55	1字手前はひらがなの「て」か？
56	1字手前はひらがなの「で」か？
57	1字手前はひらがなの「と」か？
58	1字手前はひらがなの「に」か？
59	1字手前はひらがなの「の」か？
60	1字手前はひらがなの「は」か？
61	1字手前はひらがなの「も」か？
62	1字手前はひらがなの「る」か？
63	手前にひらがな文字列があるなら、それは1字か？
64	手前にひらがな文字列があるなら、それは2字か？
65	手前にひらがな文字列があるなら、それは3字か？
66	手前にひらがな文字列があるなら、それは4字か？
67	手前にひらがな文字列があるなら、それは5字以上か？
68	後ろにひらがな文字列があるなら、それは1字か？
69	後ろにひらがな文字列があるなら、それは2字か？
70	後ろにひらがな文字列があるなら、それは3字か？
71	後ろにひらがな文字列があるなら、それは4字か？
72	後ろにひらがな文字列があるなら、それは5字以上か？
73	1字後ろはひらがなの「が」か？
74	1字後ろはひらがなの「た」か？
75	1字後ろはひらがなの「て」か？
76	1字後ろはひらがなの「で」か？
77	1字後ろはひらがなの「と」か？
78	1字後ろはひらがなの「に」か？
79	1字後ろはひらがなの「の」か？
80	1字後ろはひらがなの「は」か？
81	1字後ろはひらがなの「も」か？
82	1字後ろはひらがなの「る」か？
83	1字後ろはひらがなの「を」か？
84	2字後ろはひらがなの「が」か？
85	2字後ろはひらがなの「た」か？
86	2字後ろはひらがなの「て」か？
87	2字後ろはひらがなの「で」か？
88	2字後ろはひらがなの「と」か？
89	2字後ろはひらがなの「に」か？
90	2字後ろはひらがなの「の」か？
91	2字後ろはひらがなの「は」か？
92	2字後ろはひらがなの「も」か？
93	2字後ろはひらがなの「る」か？
94	2字後ろはひらがなの「を」か？

から、すべてのデータはコンピュータで作成した。

いずれの候補文字に関しても実際に文節末となる場合と、文節末ではない場合とがあるが、学習用のデータの作成にあたっては、原則としてその比率が3対1になるようにした。実際の文中での比率はこれよりも大きいが、これは識別が困難な文節末ではないケースがデータ中になるべく多く含まれるように配慮した結果である。なお、この二つのタイプに属するケースはデータセットのなかでできるだけ規則正しく交互に出現するように配置し

た。

3．学習ならびに識別の結果

以下では分析の結果を報告するが、まず文節末の候補文字ごとに作成されたデータの学習結果に触れ、ついで文節の識別結果を報告する。前述したように、ほとんどの候補文字に関しては4層のモデルを用いたが、「に」と「る」の場合は3層のモデルを用い、「が」に関しては5層のモデルを用いている。また、中間層のユニット数も「に」、「る」、「は」に関しては、他のものと異なっている。これは、試行錯誤の結果を反映するものである。「に」と「る」の場合は、最初に試みた3層構造（中間ユニット数94）のモデルでの学習結果がきわめて良好だったため、それをそのまま用いている。また、「が」と「は」については、恐らく全体のデータ量が少なかったことと、文節末とはならないケースが特に少なかったことが主な原因と思われるが、学習の結果が思わしくなく、中間層の数ならびに中間層のユニット数を増やすなどの対応策をとった。その他の候補文字に関しては、中間層を1層から2層に増やすことで、学習の速度ならびに精度がかなり改善された。

3.a. 学習結果

学習結果は表2にまとめられている。学習の回数と収束精度は、二乗誤差0.05を一応の目安としたが、「が」と「は」に関しては、表2に示されているように他の候補文字に比べ学習回数が多い割りには収束の精度が低くなっている。文節末の総合識別率については、「が」と「は」を除けば、いずれも94%以上が正しく識別できるところまで学習させることができた。

表2には、この他、各候補文字に関して、文節末となる場合と、文節末とはならない場合とに分けて、各々の学習データによる文節識別の結果が示されている。文節末の場合は、きわめて高精度で識別されており、「る」の場合が最も低い値となっているが、それでも97.69%はある。これに対し、文節末とはならない場合の学習精度はかなり低くなっている。

表2　学習用データによる学習の結果

文節末候補	モデル		学習状況		識別率（%）		
	層数	ユニット数	収束回数	二乗誤差	総合識別率	文節末	非文節末
「が」	5	94/30/30/30/1	300	0.0696	91.00	98.67	68.00
「た」	4	94/30/30/1	50	0.0302	96.55	99.20	89.58
「て」	4	94/30/30/1	100	0.0456	95.16	98.92	83.87
「で」	4	94/30/30/1	50	0.0527	94.72	99.39	85.37
「と」	4	94/30/30/1	50	0.0469	94.36	99.59	77.03
「に」	3	94/94/1	16	0.0002	100.00	100.00	100.00
「の」	4	94/30/30/1	50	0.0228	97.70	100.00	90.80
「は」	4	94/40/40/1	350	0.0525	92.50	100.00	70.00
「も」	4	94/30/30/1	9	0.0008	100.00	100.00	100.00
「る」	3	94/94/1	8	0.0074	98.19	97.69	100.00

ただし、「に」、「も」、「る」の場合は、100%学習できており、候補文字による差異は大きい。

3.b. 未学習データによる検証の手順

学習した結果を用いると、未学習の文章の文節がどの程度識別できるかを検証することがこの研究では最も重要な側面であるが、表2に示されているように、今回の試行では用意したデータ量が少なく、本格的なテストを行うまでには至らなかった。しかし、未学習データでの文節識別テストを省略するわけにはいかないので、次のように可能な範囲内でのテストを行った。

まず、今回とりあげた10の候補文字の内、比較的にデータ量の多かった「た」、「で」、「と」、「の」に関しては、用意したデータを半分ずつに分け、そのひとつを学習用データに、残り半分をテスト用の未学習データとすることで、識別テストを行った。残りの6つの候補文字に関しては、「が」、「て」、「に」、「は」の場合は、学習用のデータには含まれなかったデータを使用してテストを行った。ただし、これらのデータは、いずれの候補文字に関しても文節末となる場合とならない場合との二つのケースの内、文節末となる場合のデータしかなく、やむを得ずそれらだけによるテストしか行えなかった[5]。以下では、この二つの方式による文節の識別テストの結果を報告する。

3.c. 「た」、「で」、「と」、「の」の識別結果

表3は文節末の候補文字「た」、「で」、「と」、「の」について、各々全データを二分し、半分のデータを用いて学習を行った結果を示している。学習結果はかなり良好で、二乗誤差、総合識別率とも全データによる学習結果（表2参照）と優劣をつけ難い。文節末の場合とそれ以外の場合とに分けてみても、前者に関してはいずれの候補文字も99%か100%の識別率であり、文節末ではない場合も「の」の84.09%から「た」の92.31%の間におさまっている。

学習の結果得られたシナプス荷重値を用いて未学習データによる文節末の識別テストを行った結果は表4に示されている。まず、総合識別率であるが、最も精度が高かったのは「の」の91.95%、逆に最も低かったのは「た」の82.76%であった。文節末の場合と、そうではない場合とに分けてみると、前者に関しては、いずれの候補文字の場合も95%程度の識別率が得られた。学習したデータ量が少ない割りにはまずまずの結果といえようが、いずれの候補文字も、文節末となる場合の方が圧倒的に多いことでもあり、当然といえば当

表3　文節識別テスト用データの学習結果

文節末候補	モデル		学習状況		識別率（%）		
	層数	ユニット数	収束回数	二乗誤差	総合識別率	文節末	非文節末
「た」	4	94/30/30/1	50	0.0220	97.70	99.26	92.31
「で」	4	94/30/30/1	50	0.0386	95.94	98.78	90.24
「と」	4	94/30/30/1	50	0.0241	96.88	100.00	87.50
「の」	4	94/30/30/1	50	0.0402	95.98	100.00	84.09

表4　文節識別テスト用データによる識別結果

文節末候補	二乗誤差	識別率（%）		
		総合識別率	文節末	非文節末
「た」	0.1550	82.76	94.87	57.89
「で」	0.1000	89.43	95.12	78.05
「と」	0.1093	87.42	96.00	55.88
「の」	0.0754	91.95	95.42	81.40

然の結果であろう。

同じ候補文字でも文節末ではない場合の識別率は、文節末の場合に比べて大幅に低く、最高でも「の」の81.4%にとどまり、「た」と「と」の場合には、それぞれ57.89%と55.88%の識別率しか得られなかった。識別率の低かった「た」と「と」に関していえば、およそ40数パーセントの場合には、本来は文節の終わりではないものも文節の終わりであると誤認されてしまったことになる。

このように、確かに誤認が多く、もちろんこのままでは実用にはならないが、これだけでこの研究の試みが期待の持てないものであることを意味するわけではない。というのは、「で」と「の」に関しては、文節末ではない場合の識別率はおよそ80%前後はあり、大いに期待の持てる数字だからである。学習に用いたデータの量が非常に少なかった点を勘案すると、全体的にみて、さらなる研究に期待をつなぐことのできる結果であるといえよう。また、後述するように、データ自体の作成方法にも改善の余地があるし、コンピュータ用の辞書や文法ルールを併用することも可能であり、多少の手直しとデータ量の面での対処のみで識別率が大幅に向上する可能性もある。

3.d.「が」、「て」、「に」、「は」の識別結果

既に触れたように、「が」、「て」、「に」、「は」の四つの候補文字に関しては、文節末の場合のデータのみではあるが、同じく未学習データによる識別テストを行った。テストにあたっては、初めに報告した学習によって得られたシナプス荷重値を用いた。識別の結果は表5に示されている。文節末の候補文字「て」、「に」に関しては、それぞれ97.14%と100%と良好な識別結果となっているが、「が」、「は」に関しては、もともと学習精度が低かったこともあり、それぞれ92.04%と88.28%と、かなり低い値になっている。表4に示されているように、「た」、「で」、「と」、「の」の場合は、文節末のデータに関しては未学習でもほぼ95%前後の識別率が得られており、これらと比べるとかなり見劣りする結果になっている。

なお、「も」と「る」に関しては、データ量が少なく、学習結果を用いたテストは全く行

表5　未学習データによる文節末の識別結果

文節末候補	「が」	「て」	「に」	「は」
識別率（%）	92.04	97.14	100.00	88.28

えなかったが、いずれの場合も学習データでは、表2に示されているように、特に識別の困難な文節末以外のケースの識別率が100%となっており、未学習データであってもかなり良好な結果が期待できよう。

4．考　察

以上の報告のように、一部の文字に関してはその前後の文字の種類を入力データに用いることで、ニューラル・ネットワーク・モデルによって、かなりの精度で文節末であるかどうかが識別できることが判明したが、この実験の結果を活かして実用的な、日本語の文章の分かち書きシステム（あるいは文節識別システム）を開発するためには、まだ多くの作業が残されている。以下では、今回の試行の結果を振り返りつつ、今後、研究をさらに進めるための指針ならびに改善点を検討してみたい。

第一に何よりも最優先すべきなのは大量のデータを整えることである。今回は、データ不足のため一部の候補文字についてしか未学習データによるテストが行えなかったが、とりあえず、テスト用のデータだけでも作成して、未テストの6つの候補文字に関しても、文節識別の結果を得ることが必要である。また、今回の試行では、新聞記事にして60記事、総字数19、931字の日本文よりデータを得たが、今後の研究では少なくともその5～10倍程度のデータによる再分析が必要であろう。

第二に、研究の対象となる、文節末の候補文字の範囲を拡大することも必要である。今回は特に出現頻度の高い10文字を対象としたが、それ以外の文字に関してもデータを整えるべきであろう[6]。

第三に、データ作成の方式に関しても改善の余地がある。今回の試行では、表1の94項目のデータを作成し、ニューラル・ネットワーク・モデルの入力データとして用いたが、データ作成に際し一部の情報が欠落してしまったことも事実である。今回は、「対象文字の1字後ろは『が』か？」という項目に関しては、そうであるかどうかを1あるいは0で示すかたちのデータを作成したが、別の方式をとることも可能である。

例えば、対象文字の前後のいくつかの文字をとりあげ、それぞれのJISコードを二進数に直しそのまま1と0のデータとしてニューラル・ネットワーク・モデルの入力データとして用いることも可能である。この方式をとれば、文節末の候補文字の前後の文字が漢字か、記号か、ひらがなか、カタカナか、数字か、という字種に関する情報のみならず、それらがどの文字なのかという情報も同時にデータに含まれることになる。全角文字ひとつにつき16ビットの情報が必要であり、全部で96ビットあれば、候補文字の前後3文字ずつ、合計6文字分の情報を表わすことができる。このように、96個の1と0からなる測定項目で、今回の94個の項目からなるデータよりもはるかに多くの情報をデータに含めることが出来るようになる。ただし、実際に使用してみない限りは、この方式で作成したデータが文節末の識別にどれほどの効果を持つかは分からないが、試みるだけの価値は十分にあるといえよう。

第四に、誤った識別結果の内容の検討ならびにその対処の検討が必要であろう。誤りに

一定の傾向があるかどうかを調べるとともに、具体的な対策を講じる必要がある。例えば、ある程度まで文節末の識別精度が高まれば、その結果を利用してコンピュータ用の文節あるいは単語の辞書を構築することで、さらに精度を高めることもできよう。また、入力ユニットのいくつかに文法ルールの条件部を割り当てることにより、何らかの文法ルールをシステムに追加することで、文節識別の精度を高めることも可能かもしれない。今後の研究では、これらの点も具体的に検討すべきであろう。特に、辞書の構築の方は、漢字のみからなる文字列をいくつかの語に正しく分割する上でも必要となろう。ただし、この場合は、本稿で報告した手法ではなく、最長一致法など既存の方式も含め、その他の方式を用いる必要があろう。

以上では、(1)データの量的な面での改善、(2)文節末の候補となる文字を増やすこと、(3)データ作成面での改善、(4)誤った識別結果の分析と、辞書や文法ルールの利用の検討、などを骨子とする改善方針を示した。今回の試行の結果からみる限り、以上のような改善が行われれば、ニューラル・ネットワーク・モデルを用いた、実用に足る、日本文の分かち書きシステム（ないしは文節識別システム）の構築は十分に可能性のあるものだといえよう。

〈注〉

1）田中（1986）は、(1)字種切り、(2)最長一致の原則による辞書引き、(3)接続表の利用、を統合した方式で、98%程度の精度で自動分かち書きが達成できた事例を説明している。漢字の分布を出発点として語の認定を行うシステムの報告としては坂本（1980）がある。
坂本義行「日本語処理システムのソフトウェアにおける特徴」、『日本語情報処理』（マーケティング・レポート・サービス、1980）。
田中穂積「2　構文解析と意味解析」、高橋延匹編『日本語情報処理』（近代科学社、1986）。

2）自然言語処理においてのヒューリスティックスの利用については、以下の文献を参照せよ。
野村浩郷『自然言語処理の基礎技術』、電気情報通信学会、1988年。
野村雅昭「漢字かなまじり文字連続」、『電子計算機による国語研究IV』国語研報告46、1972。

3）ラメルハート、マクレランド、PDP リサーチグループ、『PDP モデル：認知科学とニューロン回路網の探索』、甘利俊一監訳（産業図書、1989）。

4）1992年版『朝日新聞全文記事情報 CD-HIASK』（紀伊国屋書店・日外アソシエーツ）を用いた。在日外国人関係の記事を用いたのは、特定の理論的根拠があってのことではない。本稿で報告する文節識別システムを組み込んだコンピュータによる内容分析システムが完成すれば是非、在日外国人関係の記事を分析したいとかねがね考えており、個人的な関心からこのような選択となった。

5）これらの候補文字の場合は、ほとんどが文節の終わりとして用いられており、もともと文節末とはならないケースのデータが少なく、しかもそれらは全て学習用データに使用してしまったため、このような方式をとらざるを得なかった。

6）例えば、「な」、「い」、「や」、「へ」、「く」、「し」などがその候補となろう。