2004－NL－164（5）
2004／11／4

境界認定の提案: (1) コンセプトと実現法

佐　藤　理　史
京都大学大学院情報学研究科　〒606-8501 京都市左京区吉田本町
sato@i.kyoto-u.ac.jp

形態素解析に代わる新しい日本語文解析の第1ステップとして、「境界認定」という枠組を提案する。境界認定では、語(単位)を認定するのではなく、境界とその種別を認定する。本稿では、その考え方と実現法について述べる。
キーワード: 境界認定、形態素解析、単位認定

Boundary Identification: (1) Concept and Implementation

Satoshi Sato
Graduate School of Informatics, Kyoto University　Sakyo, Kyoto, 606-8501, JAPAN
sato@i.kyoto-u.ac.jp

This paper proposes *boundary identification*, a new framework of the first step of Japanese sentence analysis. Boundary identification identifies boundaries and their types between linguistic units in a given sentence. This paper describes the concept and an implementation of the framework.
Keyword: boundary identification, morphological analysis, word segmentation

1. はじめに

形態素解析システムは、現在、日本語処理の各種応用において、広く用いられている。誰もが利用できるJUMANやChaSenの出現は、日本語処理のハードルを低くするのに、大きく寄与してきた。

しかしながら、前稿[1]でも述べたように、現在の日本語形態素解析は、次のような問題を抱えていると筆者は考える。

(1) **表記ゆれの問題**
表記のゆれに対処していない。同一語でも異表記であれば、まったく別の語として認定される。

(2) **単位の問題**
語の単位の問題に対処していない。出力される語には、形態素、語、複合語などが存在し、単位がばらばらである。

(3) **名称の問題**
実際に処理している内容に対して、「形態素解析」という名称がそぐわない。

このうち、「表記のゆれの問題」に対しては、すでに前稿[1]において、その解決策を示した。

本稿で扱うのは、主に「単位の問題」である。まずは、この問題を復習しておこう。

単位の問題：どの長さを1語と考えるかという問題である。これは、テキストの表記法において、語の区切り記号が存在せず、語境界があいまいであることに起因する。国立国語研究所が行なった各種語彙調査では、複数の単位が採用されている。[1]

この問題に対する標準的かつ正当的アプローチは、(1) なんらかの方法で語の単位を定義し、(2) 辞書の見出し語をすべてその単位で揃え、(3) 解析結果の出力もその単位で揃える、というアプローチであろう。もし、語彙調査のための単語分かち書きを実現するとすれば、おそらく、それが正解なのだろう。しかし、「日本語文解析の第1ステップという文脈では、それが唯一の解ではない」というのが、本稿の主張である。では、どんな別解があるかというと、それが本稿で提案する「境界認定」である。

境界認定とは、「与えられた文字列に対し、語境界となる文字境界を認定する」ものである。もし、語境界を1種類しか考えないのであれば、それは、語認定＝単語分割と同値である。しかし、「複数の語境界を認め、その種別を含めて境界を認定する」ならば、話は

違ってくる。以下に簡単な例を示そう。

(1) ワイン城完成記念パーティに行った
(2) ワイン / 城 / 完成 / 記念 / パーティ / に / 行っ / た
(3) ワイン / 城 / 完成 / 記念 / パーティ / に / 行った
(4) ワイン城 / 完成 / 記念 / パーティ / に / 行った
(5) ワイン城完成記念パーティ / に / 行った
(6) ワイン城完成記念パーティに / 行った
(7) ワイン $|_4$ 城 $|_3$ 完成 $|_3$ 記念 $|_3$ パーティ $|_2$ に $|_1$ 行っ $|_5$ た

正当的アプローチに立った語認定＝単語分かち書きは、その認定単位を定め、(1) に対して、上記の (2)-(6) のうちの<u>いずれか１つ</u>を出力することになるだろう。これに対して、語境界認定は、境界とその種別のみを認定し、(7) のような形式で出力する。もし、(a) 適切な境界種別を設計でき、かつ、(b) それが機械的に認定できるならば、ほしい単位の語の列を、境界認定結果から得ることができる。たとえば、上記の (4) は、(7) において、境界１から境界３までを境界として採用し、その他を捨てた場合に得られる列である。

以上が、境界認定の骨子である。本稿 (パート 1) では、境界認定の考え方と実現法について述べる。なお、境界認定という考え方が生まれてきた背景と、境界認定の背後にある思想については、稿を改め、パート２で述べる。

2. 境界認定の考え方

2.1 単語分割と基本ユニット

現在の形態素解析が行なっている処理（日本語文解析の第１ステップ）を何と呼ぶのはさておき、そこでのタスクの本質は「語認定」である。しかしながら、日本語において、「語」とは何かを明確に定めることが非常に難しいため、その帰結として「語認定」処理を明確に定義することができない。

影浦[2] が指摘するように、「語」には、

- 語構成／造語能力を考えるときの語、
- 辞書の見出し語としての語、
- 文の構造の基本単位としての語、

など、いくつかの種類の「語」がある。日本語の場合、これらの「語」が、少しずつずれている（一致しない）ところに、問題の根がある。

英語やフランス語など、テキスト表記法に語の区切り記号（スペース）が存在する言語では、語の単位は**形式として明確**であり、それを基本単位とみなすことは、当然かつ妥当である☆。だからと言って、「それに相当する基本単位が、日本語にも当然あってしかるべきである」と考えることは、まったく根拠がない。あってほしいと期待するのは自由であるが、あるという保証はない。

実は、「そのような基本単位は日本語にはないのではないか」というのが筆者の仮説である。より正確にいうならば、「日本語において、そのような基本単位を定義することに労力を費やしても、労力に見合う成果が得られない」ということである。これは、単位の問題を考えることを放棄しようということではない。あらゆるところで基本となる唯一無二の単位（絶対基本単位）を求めることを放棄しようということである。

2.2 なぜ基本単位が必要か

では、絶対基本単位がないとしたら、どう考えていけばよいのだろうか。その解は、日本語文解析において、なぜ基本単位（＝語）が必要なのかを考えることから得られる。

日本語文解析において必要な「語」は、つぎの２種類の「語」である。

(1) **外部知識ベースとのインタフェースとしての語**：解析系に対して与えられる入力は、文字列である。それを解析するためには、各種の語彙的知識が必要である。語彙的知識は、外部の知識ベース（辞書）に格納されており、この知識ベースとのインタフェースとなる単位（＝語）が必要となる。この場合の語は、「辞書の見出し語としての語」である。
(2) **その後の処理へのインタフェースとしての語**：たとえば、その後の処理が文の構造解析であれば、構造解析に適切な基本単位を認定することが必要である。これは、おおよそ「長い単位の語」（文節）であろう。また、その後の処理が情報検索システムのインデックス作成であるならば、インデックスとして適切な基本単位が認定されるべきであろう。

この２種類の区別は重要である。「辞書の見出し語で切る」のは、解析系が必要とする語彙的情報を解析系にインポートするためである。単語分割という処理自身、品詞や接続情報といった語彙的情報を必要とす

☆ 英語では、より小さな単位として、形態素という単位を考えるが、形態素から構成的に語を定義しているのではなく、語から分解的に形態素を取り出している。つまり、最も小さな単位からボトムアップ的に積み上げて、より大きな単位を定義しているのではない点に注意しよう。形態素を定義しなくとも、語を定義することはできるのである。

るのであるから、このような単位（辞書の見出し語）を認定することは不可欠である*。しかし、その一方で、辞書の見出し語が、そのまま、次の処理で必要とされる基本単位となる保証はまったくない。

ここでもう一度確認しておこう。もし、唯一無二の基本単位（絶対基本単位）が存在し、単語分割の結果がその単位となっているのであれば、このような問題は発生しない。なぜならば、定義上、絶対基本単位は、どんな処理に対しても基本単位となるからである。しかしながら、絶対基本単位がないとするならば、辞書の見出し語が、そのまま、次の処理において適切な基本単位となる保証は、どこにもない。

となると、どんな帰結となるだろうか。素直な帰結は、上記の２種類の基本単位（＝語）を考えざろう得ないということである。後者の単位は、後続する処理に依存するので、以下の議論では、後続する処理として、文の構造解析を仮定する。

2.3 単位が先か、境界が先か

とにかく、基本単位（語）が必要なのであるから、これを構成的に定義しようと考えるのは素直な考え方である。しかし、現時点では、**私はこれをしないことにする**。なぜならば、

(1) 辞書の見出し語としての語：
辞書において、何を見出し語とするのかは、その語の構成（内部構造）によって決まるのではなく、どのような情報を解析系にインポートしたいのかによって決まる。たとえば、「自然言語処理」が「学問分野」であるという情報を解析系にインポートしたいのであれば、「自然言語処理」という見出し語を立てるしかない。ゆえに、見出し語という単位を構成的に定義するのはナンセンスである。何を見出し語とするかは、外部要因によって定まる。

(2) 文解析のための基本単位としての語：
文解析のための基本単位は、ある具体的な文においてのみ存在する。具体的な文がなければ、その単位を認定できない。ということは、構成的に定義できない。

前者は外部要因によって定まるので、これ以上議論する必要はない。後者は、構成的に定義できないとしても、依然として、文が与えられたならば、その単位を認定しなければならないことには変わりがない。つまり、なんとかしなければならない。

そこで、発想を 180 度転換して、単位を認定するのではなく、単位の境界を認定することを考える。その心は、

(1) 単位の実在を仮定する。
(2) 具体的な文においてのみ、単位を考える。
(3) 単位を構成的に定義することはしない。
(4) その代わりに、単位間の境界と、その判定法を定義する。

さらに、もう一歩進めて、境界がまず実在し、それによって単位が定まると考えよう。

境界認定の考え方
(1) 境界とその種別の実在を仮定する。
(2) 具体的な文が与えられた時、そこに含まれる境界とその種別を認定する方法を与える。
(3) その結果に基づき、文中の単位が定まる（を定める）。

この考え方に基づくならば、境界の認定と単位の認定は、一応、別の処理として切り離すことができることになる**。境界の種別をうまく設計できれば、境界認定結果から、複数種類の単位を認定できる可能性も生まれてくる。本論文で提案する語境界認定は、このような考え方に基づく。

境界認定における次の大きな課題は、実際にどのような境界を認定すべきかという問題である。その問題に進む前に、まず、境界認定がどのような方法で実現できるかを検討しよう。

3. 境界認定システムの実現法

境界認定システムの実現法には、いくつかの方法が考えられる。ここでは、現在の形態素解析システムをベースとした実現法を考える。

3.1 形態素解析システムをベースとした実現法

現在の形態素解析システムは、主に、次の２つの知識から構成されている。

(1) 形態素辞書：
形態素として認定する対象を規定した辞書。各見出し語（表記）に対して、品詞と活用型を定義する。

(2) 連接規則：
２つの形態素が連接可能かどうかを規定した規

* JUMAN の現在の開発者である黒橋は、『自然言語処理』において、次のように書いている。「工学的に見たときの形態素解析の目的は、入力文をとにかく辞書中の項目の組合せに分解することである (文献 3); p120)」。この言明は、「形態素解析は基本的に辞書の見出し語の認定のみを目的としている」と解釈できる。

** どのような単位を最終的に認定したいかに基づいて、境界の種別を設計しなければならないので、完全に切り離せるわけではない。しかし、処理としては、切り離せる。

	...	する	できる	可能	...
構築	...	1	1	1	...
分析	...	1	1	1	...

図 1 連接行列

	...	する	できる	可能	...
名詞–サ変名詞	...	1	1	1	...

図 2 連接行列: 行の縮退

則。概念的には、連接行列と考えて良い。

一方、境界認定システムでは、上記の2つに加えて、次のような境界認定規則が必要である。

(3) 境界認定規則:
形態素の並びに対して、その間の境界の種別を決定するための規則。左右の形態素のみから、境界が定まると仮定するのであれば、現在の形態素解析システムの連接行列の各要素に、その境界の種別を記述すればよい。

最後の境界認定規則について、もう少し補足しよう。形態素解析システムでは、2つの形態素が接続するかどうかを、連接行列によって表現する。ここでは、連接しない場合は0、連接する場合は1で表現する[⋆]。たとえば、「構築」や「分析」の後に、「する」、「できる」、「可能」などが連接することは、**図 1** のように表現される。

このような行列のサイズは、語彙の数を n とすると、$n \times n$ となる。これは大きすぎるので、通常は、品詞および品詞細分類を用いて抽象化し、行列を縮退化する。たとえば、名詞の下にサ変名詞という品詞細分類を導入すると、「構築」と「分析」の行がマージできる (**図 2**)。

さて、ここで、境界認定のための拡張を考えよう。たとえば、サ変名詞と「する／できる」の間の境界 ID を 89、サ変名詞と「可能」の間の境界 ID を 90 とすることにしよう。そして、これを、**図 3** のように連接行列の要素として書くことにしよう。

形態素解析システムは、連接行列を見て、連接できるかどうかを判断する。このチェックを行なった際、連接可能の場合はその境界 ID を（最終的な出力時に出力できるように）記憶しておくことにしよう。もし、

⋆ この他に、連接する場合には、その連接コストを定義する。

	...	する	できる	可能	...
名詞–サ変名詞	...	89	89	90	...

図 3 連接行列: 境界 ID を記述

	...	する	できる	可能	...
構築	...	89	89	90	...
分析	...	89	89	90	...

図 4 連接行列: 境界 ID を記述

	右接続
構築	89, 90
分析	89, 90

	左接続
する	89
できる	89
可能	90

図 5 連接行列を 2 つの表へ分割

ある形態素の並びが、連接行列によって複数の境界によって連接できると定義されている場合は、複数の境界 ID を記憶しておけばよい[⋆⋆]。こうして記憶しておいた境界 ID を、最終的な形態素列を出力する際に、同時に出力するようにすれば、境界の認定結果が得られることになる。

以上のような方法で、境界認定システムを実現できる。

3.2 連接行列の分解

前節の説明では、サ変名詞という品詞細分類を導入して連接行列を縮退した後に、境界認定のための拡張を行なった。これに対して、境界認定のための拡張を、$n \times n$ の連接行列から出発することにすると、**図 4** のような連接行列が得られる。

この連接行列は、境界 ID の導入により、2つの表に分割できる (**図 5**)。そして、これら2つの表は、再度、1つの表にまとめることができる (**図 6**)。

この表のサイズは、n となる。それゆえ、左接続、右接続の情報を、辞書に個々の見出し語に書くことにしてもよい[⋆⋆⋆]。

ここで重要なことは、境界 ID を持ち込むことにより、連接条件を記述するためだけの品詞細分類を導入

⋆⋆ つまり、境界 ID が一意に定まらない場合も考えられる。

⋆⋆⋆ 少なくとも、例外的な接続をとりうるものに関しては、そのような方法が適切であるように思える。

	左接続	右接続
⋮	⋮	⋮
構築		89, 90
分析		89, 90
⋮	⋮	⋮
する	89	
できる	89	
可能だ	89	
⋮	⋮	⋮

図 6　接続情報の記述

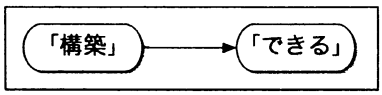

図 7　語＝ノード、接続＝リンク

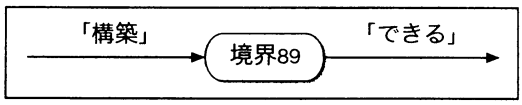

図 8　境界＝ノード、語＝リンク

することが不要となるという点である。品詞細分類に相当する抽象化は、「境界 89」、「境界 90」のような境界種別として導入される。

以上のような、連接行列の分解を行なうと、境界認定に必要な道具立ては、次のような辞書だけとなる☆。

(1)　形態素／語の辞書：
語彙情報を供給する単位を規定した辞書。各見出し語に対して、品詞（と活用型）の情報を定義する他に、左接続可能な境界 ID と、右接続可能な境界 ID を定義する。

3.3　有限オートマトンへの拡張

形態素解析の過程を説明する場合、形態素をノード、接続をリンクとするグラフを書くのが普通である（**図 7**）。

これに対して、境界認定では、その逆を考えるのがわかりやすい。すなわち、境界をノード、語をリンクとするグラフである（図 8）。

このように考えると、語は、ある境界（ノード）から別の境界（ノード）に対する遷移に対応する。前節では、語に対して、左接続可能な境界 ID と右接続可能な境界 ID を指定することにしたわけであるが、このような指定方法にとらわれる必要はない。たとえば、各語に対して、遷移元境界 ID と遷移先境界 ID の組を指定することにしてもよい。この場合、この指定は、有限オートマトンの状態遷移規則となり、その処理系として、有限オートマトンを準備することになる☆☆。

3.4　境界認定システムの試作版

最終的には、有限オートマトンへの拡張が望ましいが、当面は、既存の形態素解析システムを改造することで、境界認定システムを実現する方針を取る。なぜならば、実際にどんな境界を設定すべきかを明らかにすることが当面の懸案であり、その間、システムは、それをテストできる範囲で「動けばよい」からである。具体的には、JUMAN-4.0☆☆☆ を改造して、境界認定システムを実現する。

JUMAN-4.0 では、連接規則は、`JUMAN.connect.c` というファイルに記述されている。個々の連接規則は、左右の形態素集合とコストの 3 つ組によって定義される。たとえば、次の規則は、「数詞」の後に、「名詞性名詞助数辞」が連接可能であり、そのコストが 5 であることを定義している。

```
(((名詞 数詞))
 ((接尾辞 名詞性名詞助数辞))
 5)
```

ここでは、連接可能を定義しているわけであるが、これと同時に、その連接の境界の種別を定義することにしよう。つまり、3 つ組の連接規則を、4 つ組の境界認定規則に拡張する。上記の連接の境界の ID を 7500 とするならば、これを次のように記述する。

```
(((名詞 数詞))
 7500
 ((接尾辞 名詞性名詞助数辞))
 5)
```

これで、知識ベース（境界認定規則）の方は、準備が整ったので、あとは、JUMAN-4.0 のシステム（プログラム）を改造して、使用した境界認定規則（連接規則）の ID を出力するようにすればよい。具体的には、連接行列に ID を埋め込み、連接チェック時に、その ID を取り出して記録しておく方法を取る。**図 9** に、改造したプログラムの実行例を示す。この例では、「三」と「本」の間の境界が、上記で説明した 7500 境界として認定されている。

4.　境界の設計

さて、いよいよ境界の設計に進もう。

☆ 連接コストは、ここでは考えていない。実際には、どこかで定義する必要がある。色々な選択肢がありえるが、境界 ID に対して定義するのが順当だろう。

☆☆ この場合、現在の形態素解析システムを、機能拡張することになる。但し、ChaSen の内部は、実際には、有限オートマトンとして動いているとのことである[4]。

☆☆☆ http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

```
- 1100 -
バラ ばら バラ 名詞 普通名詞 * * (JC1:cA243200W!A0)
- 3100 -
を を を 助詞 格助詞 * * (JC1:jX000020W!A0)
- 2300 -
三 さん 三 名詞 数詞 * * (JC1:nX000180K!A1)
- 7500 -
本 ほん 本 接尾辞 名詞性名詞助数辞 * * (JC1:sX000150K!A1)
- 2522 -
買った かった 買う 動詞 * 子音動詞ワ行 タ形 (JC1:cA104070W!A1)
- 1200 -
EOS
```

図 9 境界認定の実行例

4.1 境界の全体設計

実際にどのような境界を設定すべきかは、この「境界認定」という枠組の中で、どのような問題を解きたいかに依存する。当面、想定している問題は、次のようなものである。

(1) いわゆる文節の境界を認定する
(2) 文節内の内容語と付属語の境界を認定する
(3) 複合語の構造解析に有用な情報を提供する

このうち、(1) と (2) が実現できれば、いわゆる「長い単位の語」の認定が可能となる。(3) を想定するのは、日本語文解析において、複合名詞解析（特に、動詞性名詞を含むもの）が必須と考えるからである。

これらの問題を解くために、大分類で 8 種類の境界を設定し、その詳細設計を行なっている。現在設定している境界の概要を**表 1** に示す。

実際の作業は、前述の JUMAN.connect.c で定義されている連接規則 235 個に対して、まず、すべて異なる ID を付与し、それを順次、種類別に分類していくという方法を取った。この過程で、必要に応じて、規則を分割したり、併合したりした。現在の境界認定規則の数は、302 個である。

4.2 文節境界と問題点

8 種類の大分類の境界のうち、2200 番台を除く b 境界が文節境界に対応する。大多数を占める典型的な文節に対しては、境界設計および境界認定のいずれも、大きな問題はない。問題となるのは、次に述べるような点である。

4.2.1 境界設計に関わる問題

内容語と機能語（付属語）の中間的な表現をどう扱うか、意味的な単位と形式的な単位に不一致が見られるときにどちらをとるか、といった問題は、どのように境界を設計するかに関わる問題となる。具体的には、次のような点が問題となる。

(1) 形式名詞
形式名詞を内容語と考え、その直前を文節境界とするか。あるいは、付属語と考え、その直前を文節境界としないか。

(2) 助動詞相当の複合辞
どの範囲の表現を助動詞相当の複合辞（付属語相当）とするか。現在のところ、複合辞はその構成に関わらず、付属語扱いとする方針。

(3) 助詞相当の複合辞
どの範囲の表現を助詞相当の複合辞とするか。その下位分類をどうするか。本来の表現か複合辞かを形式的に決定できないもの（曖昧なもの）をどう扱うか。

(4) その他
たとえば、以下のような例。(| が問題となる境界。(b)-(f) は影浦峡による。(g)-(j) は文献 5) による。)
- (a) プログラムの開発 | 支援環境
- (b) ロミオと | ジュリエット | 饅頭
- (c) 性格 | 穏和な彼でも怒ることがある
- (d) 漱石全集を全文 | 入力する。
- (e) 板チョコ | 半分をわけてあげる。
- (f) 彼は綺麗に | なる。
- (g) 現 | 会長と | 副 | 会長
- (h) 大きい | 婦人服 | 専門店
- (i) 不精な | 中年男性 | 用
- (j) 幻の | 著者 | 探し

4.2.2 境界認定に関わる問題

文節境界と文節内境界のどちらにすべきかは明らかであるが、機械処理による認定が難しいものには、次のようなものがある。

(1) 複合動詞
後項の動詞を認定して☆、前項／後項間の境界を特別扱いする必要ある。どのくらいの精度でできるか。(100%正しく認定することは難しい。)

(2) 動詞の連用形と名詞の区別
難しい。JUMAN では放棄されている。

(3) 副詞と副詞的に働く名詞の扱い
これらの語の整理が不十分のため、正しく境界を認定することが困難。副詞および副詞的に働く名詞の体系的分類が必要。

(4) 助詞の下位分類
JUMAN/京大コーパスでは正しく認定することを放棄。判定基準の明確化が必須。(文節境界認定のみならば不要だが、節境界認定を視野

☆ ChaSen では、「非自立」。JUMAN/KNP では、KNP の中で処理。

表 1　境界の種別

大分類	中分類	ID	小分類の数	例
a 境界	（いわゆる文境界となりうるもの）			
	文頭	1100 ～ 1150	6	| 学校に行った。
	文末	1200 ～ 1290	6	学校に行った。|
	句点の直後	1300 ～ 1340	4	「がんばってね。|」と言った。
	句点の直前	1400 ～ 1410	2	学校に行った | 。
b 境界	（おおよそ文節境界に相当するもの）			
	読点の直後	2100 ～ 2130	4	彼は、| 学校に行った。
	読点の前	2200 ～ 2240	5	彼は |、学校に行った。
	助詞→助詞以外	2300 ～ 2360	9	学校に | 行った。
	助詞の省略	2390	1	ご飯 | 食べる？
	連体修飾の直後	2400 ～ 2410	3	美しい | 日本に生まれた。
	連用修飾の直後	2500 ～ 2522	6	確実に | 実施しよう。
	接続詞の直後	2600 ～ 2610	2	しかし | 彼は頑固だった。
	感動詞の直後	2700	1	ええ | そうです。
	副詞の直後	2800 ～ 2820	3	ゆっくり | 食事をしよう。
c 境界	（文節内の内容部から機能部への境界）			
	格助詞の前	3100 ～ 3150	6	学校 | に行った。
	副助詞の前	3200 ～ 3340	15	彼 | は学校に行った。
	接続助詞の前	3400 ～ 3630	24	疲れた | が休まなかった。
	終助詞の前	3700 ～ 3820	13	楽しもう | ね。
	連体助詞の前	3900 ～ 3920	3	彼 | の本
d 境界	（助詞から助詞へ。機能部内の境界）			
	助詞から格助詞	4100 ～ 4200	11	明日まで | が最初の締切だ。
	助詞から副助詞	4300 ～ 4480	19	明日から | は気をつけよう。
	助詞から接続助詞	4500 ～ 4610	12	
	助詞から終助詞	4700 ～ 4760	7	話しながら | ね。
	助詞から副詞的名詞	4800	1	子どもの | ころから知っている。
e 境界	（助動詞相当への境界。）			
	助動詞の前	5100 ～ 5240	15	知っている | はずだ。
	判定詞の前	5500 ～ 5550	7	机 | だ。
	形式名詞・副詞的名詞の前	5700 ～ 5750	6	勉強した | のは本当だ。
f 境界	（内容部内の境界。複合名詞内）			
	その他	6000 ～ 6030	4	
	後ろが名詞	6100 ～ 6130	4	ワイン城完成記念 | パーティ
	後ろが形容詞性名詞	6200 ～ 6230	4	当選 | 確実
	後ろが動詞性名詞	6300 ～ 6330	4	ワイン城 | 完成 | 記念パーティ
	～的→形容詞	6500	1	幾何学的 | 美しさ
g 境界	（語構成。形態素境界）			
	サ変→する	7000	1	勉強 | する
	* → 動詞性接尾辞	7100 ～ 7280	19	強 | がる
	* → 形容詞性接尾辞	7300 ～ 7350	8	食べ | にくい
	* → 名詞性接尾辞/造語成分	7400 ～ 7440	5	ワイン | 城
	* → 名詞性名詞助数辞	7500 ～ 7520	3	三 | 本
	* → 名詞性特殊接尾辞	7600 ～ 7660	7	三回 | 以内
	* → 名詞性述語接尾辞	7700 ～ 7750	6	寒 | さもつらい
	数詞関係	7800 ～ 7810	2	
	接頭辞	7900 ～ 7980	9	お | 食べになる
i 境界	（特殊な境界）			
	括弧関係	9100 ～ 9120	3	
	鉤括弧関係	9200 ～ 9220	3	
	記号関係	9300 ～ 9360	7	
	解析誤りを防ぐ	9800 ～ 9810	2	
	ダミー等	9050 ～ 9054	2	
未整理	連語関係	8100 ～ 8236	17	
計			302	

表 2　複合名詞解析のための境界

前接	後接			
	その他	普通	形容詞性	動詞性
その他	6000 連体修飾 (の/φ)	6100 連体修飾 (の/φ)	6200 格 (が)	6300 格 (が)
普通名詞	6010 連体修飾 (の/φ)	6110 連体修飾 (の/φ)	6210 格 (が)	6310 格 (を/が)
形容詞性名詞	6020 連体修飾	6120 連体修飾	6220 連用修飾/並列・付帯	6320 連用修飾
動詞性名詞	6030 連体修飾 (する)	6130 連体修飾 (する)	6230 格 (することが)	6330 格 (することを)/並列・付帯

に入れると、助詞の下位分類に基づく境界認定が必要。)

4.3　複合名詞の解析と問題点

現在の境界認定システムは、冒頭にあげた「ワイン城完成記念パーティに行った」に対して、次のような境界を認定する。

$|_{1100}$ ワイン $|_{7400}$ 城 $|_{6300}$ 完成 $|_{6330}$ 記念 $|_{6130}$ パーティ $|_{3100}$ に $|_{2302}$ 行った $|_{1200}$

現在、複合名詞解析のために設定している境界 (f 境界) を**表 2** に示す。この表のそれぞれの欄の下段は、その接続の典型的な関係を表している。この表に基づくと

ワイン城 を/が 完成 することを 記念 する パーティ

のように分析する（言い換える）ことができる。複合名詞内において例外的に働く語彙は多数あり、これをリストアップしていく必要はあるが、境界を区別することによって、複合名詞内の内部構造をある程度解析することが可能であると考えている。

そこでの問題は、次のような点である。

- 複数の単位で「形容詞性名詞」や「動詞性名詞」を構成する場合、その単位を認定しないと適切な境界の認定ができない。(前後の形態素から境界を決定するのではなく、オートマトンを構成して、形態素列から境界を決定できるようにすれば、問題は解決する)
- いわゆる造語成分の体系的整理が必要。(「ワイン $|_{7400}$ 城」と認定された場合は、読みは、「しろ」ではなく「じょう」となる。)

5. 論　　点

現時点において、「境界認定」の枠組は、十分に固まっているわけではない。これから、多くの検討が必要である。次のような点が論点になると考えている。

- 実装の立場から
  情報を単位に記述するのではなく、境界に記述することにどんなメリットがあるか。どれだけの処理を境界認定に押し込むことができるか。
- 文法記述の立場から
  日本語の語構成に、正面から向き合うことになる。整理が進むか。
- 文法論の立場から
  単位に情報を記述するのではなく、境界に情報を記述するとすれば、どういう変革がありえるか。
- 辞書の見出し語中の境界
  どのように記述しておくか。文中の境界と区別するか否か。
- その他
  辞書の見出し語としての語と、文中の語の関係。クラスとインスタンスという関係として整理できるか。(パート 2 を参照のこと)

謝　　辞

本稿の内容をまとめるに際して、7 月 14 日の北大での議論、8 月 4 日の国立国語研究所での議論、9 月 15 日の国立情報学研究所での議論、および、これに関連したメールでの議論が大変有益であった。これらに参加された方々に感謝する。本研究の一部は、次の研究費による；基盤研究 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009)、特定領域研究「実世界の関連性を投影した語彙空間の構築」(課題番号 16016249)、21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」。

参　考　文　献

1) 佐藤理史. 異表記同語認定のための辞書編纂. 情報処理学会自然言語処理研究会, 2004-NL-161, pp. 97–104, 2004.
2) 影浦峡. パーソナルコミュニケーション.
3) 長尾真 編著. 自然言語処理. 岩波書店, 1996.
4) 松本裕治. パーソナルコミュニケーション.
5) 影山太郎. 文法と形態論. In 松本裕治, 影山太郎, 永田昌明, 齊藤洋典, 徳永健伸. 単語と意味. 岩波書店, pp1–51, 1997.