

境界認定の提案: (2) 背景と思想

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

境界認定は、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、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述べる。

キーワード: 境界認定、形態素解析、単位認定

Boundary Identification: (2) Background and Philosophy

SATOSHI SATO

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

This paper describes background and philosophy of *boundary identification*, which identifies boundaries and their types between linguistic units in a given sentence. The proposal of boundary identification aims to restructure Japanese sentence analysis method.

Keyword: boundary identification, morphological analysis, word segmentation

1. はじめに

パート1で述べたように、「境界認定」は、次の考え方に基づく。

境界認定の考え方

- (1) 境界とその種別の存在を仮定する。
- (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)。

境界認定の枠組は、これまでの解析系 (形態素解析や構文解析) における単位認定と、どのように異なるのであろうか。単に、考え方が異なるだけで、その本質は同じなのであろうか。それとも何か本質的に異なるところがあるのだろうか。

本稿では、これらの疑問に答えるために、境界認定の考え方が生まれてきた背景についてまとめるとともに、形式文法と境界認定の比較、および、ブロックモデルと境界モデルの比較を行なって、従来の単位認定との差異を明確化することを試み、境界認定の背後にある思想を明らかにする。

2. 背 景

2.1 単位の問題

境界認定の考えは、日本語の語の単位の問題に対する検討から生まれてきた。パート1でも述べたように、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。あるいは、辞書の見出し語としては複数の単位を認め、それぞれの見出し語には、それがどの単位なのか、複合語の場合はそれがどのような構成要素から構成されているのかを記述するというアプローチも考えられる。実際、伝らの UniDic^{2),3)} は、この後者のアプローチを取っている。

当初、私自身、これと同じアプローチで単位の問題を解決することを考えていた。しかしながら、次のような問題に直面し、方向転換を余儀なくされた。

問題: 次の2つの「ワイン城」は、同じ単位と考えてよいのか?

- (1) ワイン城₁ に行った。
- (2) ワイン城₂ 完成記念パーティに行った。

この問題は、結局のところ、単位（語）というものを、絶対的なもの（絶対単位）と考えるのか、文中での位置、あるいは、処理の目的や用途等に依存して決まる相対的なもの（相対単位）と考えるのかという問題である*。「文の構造の基本単位としての語」という立場から考えると、「ワイン城₁」は、助詞「に」と結合して補足語を構成する。これに対して、「ワイン城₂」は、「完成記念パーティ」と結合して複合名詞を構成する。明らかに、2つの「ワイン城」は、文の構造における位置付けが異なり、「文の構造の基本単位としての語」という立場からは、異なる単位と考えるのが妥当である。

絶対単位に基づくアプローチ（絶対単位系）では、「ワイン城」が全体としてどのような単位であるかを文脈ぬきで定義する。これに対して、相対単位系（境界認定）では、『「ワイン」と「城」の間に、境界7400がある』ことだけを定義し、全体がどんな単位であるかは定義しない。それは、文中に現れた時、はじめて定まるものだと考えるのである。

2.2 品詞の問題

単位（語）の問題は、品詞の問題とも関連している。

そもそも品詞とは、どのようなものだと考えるべきなのであろうか。品詞を文中の語の役割と考える立場（構文的品詞）と、語の属性と考える立場（語彙的品詞）があり、それらを両極として、その中間のどこかに位置するものと考えられているのではないかと思う。たとえば、英語の多くの単語は、多品詞語で、文中に現れたときにその品詞が定まると考えるのが普通である。つまり、英語の品詞は、構文的品詞と考えるのが妥当である。一方、日本語では、形容詞は、連体修飾、連用修飾、述語のどの用法で現れても形容詞とみなすのが普通であり、これは、語彙的品詞の考え方に近い。いずれにせよ、ある具体的な語に対して、それがどの品詞であるかを決定する何らかの基準が本来は必要ではなくであるが、それを完全な形で明確化せよということになると、なかなかやっかいなことになる。

日本語で特に問題となるのが、副詞まわりである。

- (1) ゆっくり 歩く。
- (2) ゆっくりと 歩く。
- (3) その部屋で ゆっくり する。
- (4) ?その部屋で ゆっくりと する。

「ゆっくり」を副詞とするのはまあよろう。しかしながら、「ゆっくりと」はどうであろうか。これを副

詞とするならば、「ゆっくり」と「ゆっくりと」の関係はどうしたらよいのであろうか。あるいは、「-と」を機能語（辞）扱いすべきなのであろうか。さらに、(3)と(4)は、どう考えればよいだろうか。「ゆっくりする」の「-する」を接辞（付属語）扱いするのであれば、「ゆっくりとする」の「-する」も同様にするのであろうか。

(5) 突然 雨が降り出した。

(6) 突然 の雨

「突然」の品詞は何であろうか。この場合は、多品詞語と考えるのであろうか。これ以外にも、副詞、形容動詞語幹、名詞などで、判断に迷うものは多数あり、これらを体系的に整理しようという提案もある⁴⁾。

複合語の解析においても、品詞が問題となる。

(7) ワイン 城

(8) 国際 社会

たとえば、(7)において、「城（じょう）」の品詞は何であろうか。和語の「城（しろ）」は単独で用いられる普通の名詞であるのに対し、「城（じょう）」は通常単独では用いられない⁵⁾。これは、三省堂新明解国語辞典において字音語の造語成分と呼ばれているものである。これを辞と考えるならば、品詞を設定しないという考え方もある。しかし、「国際」は、単独で使われることはない^{5), 6)} からといって、辞と考えるのはつらいであろう⁶⁾。では、名詞とするのであろうか。もし、そうするならば、名詞の条件はどうなるのであろうか。

形態素解析においても、品詞の問題は存在する。特に気になるのは、接続条件を記述するために持ち込まれた詳細分類である。たとえば、「きのう（昨日）」はChaSen(IPADic)では「名詞-副詞可能」と定義されているが、このような品詞細分類を「品詞」として導入するのは適切なものであろうか。つまり、品詞を「接続条件の違いを弁別するためのデバイス」とし、その目的だけから新たな品詞を導入してしまっているものであろうか、という疑問である。

私の基本的な立場は、「品詞は、形式・機能・意味などの色々な側面を考慮した総合的な分類体系である」というものである。それゆえ、接続条件の記述のため詳細分類は、これとは切り離したい。そのため、接続条件を記述するデバイスは別枠で用意しようと考えた。

⁵⁾ 三省堂新明解国語辞典第4版には、「城（じょう）」が見出し語として採録されており、単独用法の用例が記載されている。しかし、現実には、ほとんど使われていないだろう。

⁶⁾ 影山⁶⁾ にならって、語根 (root)、語 (word) 以外に、語幹 (stem) という単位を導入して対処する方法もある。

* この問題に対する私の立場は、すでにパート1で述べたように、「絶対単位を求めることを放棄する」というものである。

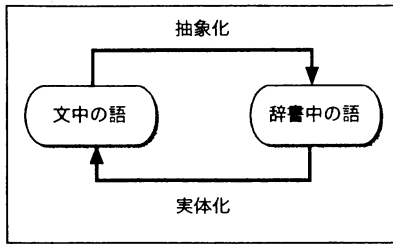


図1 文中の語と辞書の語の関係

パート1の3.2節で述べたように、境界IDは、接続条件を記述するデバイスとして機能する。

2.3 辞書の語と文中の語

先に述べた品詞の問題と関連して、文中の語と辞書中の語の関係をどう考えていけばよいかという問題も浮上してくる。

辞書の語（エントリー）と、実際に文中に現れた語を区別しないという考え方もあるが、辞書の語は、実際に文中に現れる語を抽象化した実体と考えるのが適切であろう。なぜならば、実際に文中に現れる語は、表記、用法、語義などが定まっているのに対し、辞書の語は、それらの可能な値（値域）が示されているのに過ぎないからである。つまり、この2種類の語の関係は、図1に示すような図式として捉えるのがよい。

このように考えるのであれば、語認定の結果は、最終的に、(1) 辞書中の語へのポインタ (ID) と、(2) 辞書中の語から文中の語を実体化 (instantiate) するために必要なパラメータの集合、の組として表現するのが適切であるという帰結となる。

現在の形態素解析では、活用する語に対する活用形が、考えられている唯一のパラメータである。しかしながら、前述のように語彙的品詞の立場を取るならば、「用法」もその候補となり得る。たとえば、文中の形容詞に対して、これは連体修飾用法、これは連用修飾（副詞的）用法、といったものを認定するというのである。これは、ある意味で、英語における品詞の曖昧性解消に相当するもので、英語における品詞タグの付与はこのような側面を持つ。

たとえば、先にあげた例文(5)と(6)は、用法という観点からは明確である。

(9) 突然 雨が降り出した。－副詞的用法

(10) 突然の雨－名詞的用法（「の」を伴った連体修飾用法）

一方、例文(1)–(4)も、用法という観点からは、それほど迷いはない。

(11) ゆっくり 歩く。－副詞的用法

(12) ゆっくりと 歩く。－「と」を伴った副詞的用法

(13) その部屋で ゆっくり する。－「する」を伴った動詞述語用法

(14) ?その部屋で ゆっくりと する。－「と-する」を伴った動詞述語用法

このような「用法認定」を境界認定に押し込めるか否かは、用法が形式と局所的な接続からどの程度定まるかに依存する。用法の大多数が局所的に定まるのであれば、それを考慮して境界を設計・認定することによって、用法認定を取り込むことが望ましい。このためには、品詞と用法の明確な分離と、その対応関係の体系的整理が必要である。境界認定では、このような品詞と用法の分離も視野に入れている*。

2.4 形態素解析は何をやっているのか

さて、これまでの日本語の形態素解析 (JUMAN や ChaSen) が実際に何をやっているかをよくよく考えてみると、それは語構成・文節内解析であろうという結論に落ち着く。もちろん、文字列を形態素列に分解するということを行なっているわけであるが、辞書によって規定される可能な形態素列の中から妥当な形態素列を選択するところが肝なわけであり、ここで使われているのが接続条件（連接条件）である。この接続条件は、おおまかに、文節内の接続条件と文節間の接続条件の2種類に分けることができるが、制約として効く（選択力を持つ）のは、文節内の接続条件である**。

上記の理解が正しいとすれば、「なぜ、形態素解析の結果として、複合語や文節の認定結果を出力しないのか」という疑問が湧いてくる。それらの認定はできるはずであるし、形態素解析処理の中で行なうのが自然で、かつ、工学的にも妥当ではないか、ということになる。境界認定は、それを具現化しようとするものである。

2.5 節分割との関係

語境界認定（形態素解析）の次の処理として、文の構造解析を想定した場合、従来の文節係り受け解析以

* 文解析において認定すべきものは「用法」であり、語彙的品詞ではない。語彙的品詞は語に対して定義されているので、語が認定できれば、自動的に品詞が定まる。このように考えるならば、たとえば、「きのう（昨日）」の品詞が何であるかを悩むよりは、文中に現れた「きのう」が名詞的用法なのか、副詞的用法なのかをどうやって決定できるかを考える方が建設的である。このためには、それぞれの語に対して、どのような用法が可能であるかを書き下すことが必要となるが、これを押し進めると、最終的には、語彙的品詞というものは、副次的なものに追いやられることになる。

** 日本語の文において、文節の順序はかなり交換可能である。ということは、文節間の接続条件（制約）はかなり緩いということである。

外に、文節（長い単位の語）より大きな単位を認定するという処理を考えることができる。たとえば、丸山らによって提案される節境界検出⁷⁾は、節 (clause) という単位を認定することを目指している。

おおよそ、ある種の文節境界だけが節境界になる可能性を持つため、語境界認定時に、その可能性判定を行なえるはずである。もし、これが正しいとするならば、その後の節境界認定を想定して、境界を設計・認定していくのが望ましいと考えられる。

3. 形式文法と境界認定

さて、これまでの解析系が立脚してきた数学モデルと、境界認定の数学モデルとを比較検討してみよう。

3.1 形式文法

現在の文解析における単位認定の数学的基盤は、形式文法にあると考える。この認識は、おそらく正しいだろう。以下では、文脈自由文法以下のクラス（文脈自由文法、正規文法）を想定して議論する。

形式文法は、次のような4つ組として定義される。

$$G = \langle V_N, V_T, P, \sigma \rangle \quad (1)$$

ここで、 V_N は非終端記号の集合、 V_T は終端記号の集合、 P は生成規則の集合、 $\sigma (\in V_N)$ は開始記号である。

ここでの議論では、次の点が重要である。

- (1) 形式文法は、本質的には、生成モデルである。すなわち、開始記号 σ から生成（導出）される記号列集合 $L(G)$ を定義するものである。（なお、この集合を $D(\sigma)$ とも書くことにしよう。）
- (2) 形式文法の興味を中心は、ある記号列 s が、 $L(G)$ に含まれるか否かを決定することにある。
- (3) s が $L(G)$ に含まれているどうかを決定することを、 σ からの導出が可能であるかどうかによって調べる。これは、概念的には、 σ から導出できる記号列をすべて列挙し、その中に、 s と等しいものがあるかどうかを調べているに等しい。
- (4) 形式文法において、非終端記号が、結果的に認定される単位に相当する。 s が $L(G)$ に含まれる場合、 s を構成する単位が認定される。
- (5) ある非終端記号 v_n から導出される記号列集合 $D(v_n)$ は、生成規則の集合によって、完全に規定されている。つまり、あらかじめ完全にわかっている。その基盤となるのが、終端記号の集合 V_T である。 V_T の要素は、絶対基本単位であり、そこからボトムアップ的に、 $D(v_n)$ が定義される。

上記のポイントを、以下に示す文法を使って具体的

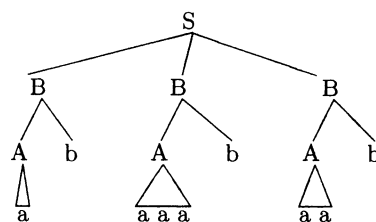


図2 abaabaab の解析木

に説明しよう。（簡単化のため、正規表現の + 記号を使用する。）

$$G = \langle V_N, V_T, P, \sigma \rangle \quad (2)$$

$$V_N = \{A, B, S\} \quad (3)$$

$$V_T = \{a, b\} \quad (4)$$

$$P = \{S \rightarrow B^+, B \rightarrow Ab, A \rightarrow a^+\} \quad (5)$$

$$\sigma = S \quad (6)$$

この文法で、記号列 *abaabaab* を解析すると、図2のような解析木（導出木）が得られる。

この結果、 B という単位は、次のように認定されることになる。

$$ab \ aaab \ aab$$

上記のポイント5で述べたように、非終端記号 B から生成される記号列集合 $D(B)$ は、次のような集合であり、これは、あらかじめわかっている。

$$D(B) = \{ab, aab, aaab, \dots\} \quad (7)$$

これを単位認定という立場から見れば、次のようになる。

どのような記号列がその単位として認定されるかは、あらかじめ完全にわかっている

非終端記号から生成される記号列集合は、生成規則（集合）によって定義される。たとえば、以下の規則を例にとろう。

$$B \rightarrow Ab \quad (8)$$

この規則は、単位認定という立場からは、次のように解釈される。

B という単位は、（より小さい） A という単位と b から構成される

つまり、

それぞれの単位は、最小基本単位 (V_T の要素) から構成的に (ボトムアップ的に) 定義されている

のである。この最小基本単位は、前節の絶対基本単位に相当する。

自然言語の解析系に、形式文法（に基づく方法）を使用する場合、与えられる文（記号列）が文法から生成されるかどうかを調べることに、目的があるわけではない。与えられる文が文法から生成されることは

前提として仮定されており、そのときの導出木（文の構造、あるいは、構成単位）を得ることに目的がある。このとき、

解析系に入力されうる記号列集合は $L(G)$ として完全に想定されている。

以上の議論からわかるように、形式文法は、いわば「全智モデル」なのである。これを日本語の場合に当てはめると、次のようになる。

- (a) 日本語の形態素は、すべてわかっている。
- (b) 文を構成するそれぞれの単位（語、文節、節）がどのような構成をとるかは、すべてわかっている。
- (c) 日本語の文は、すべてわかっている。

3.2 境界認定の枠組

一方、境界認定に対しては、どのようなモデルを考えればよいであろうか。形式文法と境界認定ではタスクが異なるので、単純には比較できない。しかしながら、できるだけ形式文法に対応する形でモデルを構成してみよう。

- (1) ここでは、1種類の境界を認定するモデルを考えよう。この境界によって区切られる単位を X とするとき、この境界を認定するモデルを $B(X)$ と呼ぶことにしよう。
- (2) 境界認定規則は、ある部分記号列を条件として境界を認定する形としよう。たとえば、記号列 ba に対して、 b と a の間に境界を認定する場合、次のような形式で書くことにしよう。

$$b a \rightarrow b | a \quad (9)$$

- (3) 境界認定モデルの最も重要な部分は、上記のような境界認定規則の集合である。これを I と書こう。
- (4) 境界認定規則の集合を定義するためには、終端記号の集合のある部分集合が既知であることが必要である。これを V_K と書くことにしよう。上記の例では、

$$V_K = \{a, b\} \quad (10)$$

である。

- (5) 境界認定規則に現れない終端記号は、不明でよい（定義する必要がない）。すなわち、終端記号の集合 V_T は未知である。

以上をまとめると、単位 X を決定するための境界を認定するモデル $B(X)$ は、次の2つ組として定義できることになる。

$$B(X) = \langle V_K, I \rangle \quad (11)$$

さて、ここで、先ほどと同じ記号列 $abaaabaab$ が与えられたとしよう。上記の境界認定規則集合により、

表 1 与えられる記号列集合と認定される記号列集合	
与えられる記号列集合 S	X として認定される記号列集合
$(a^+b)^+$	a^+b
$(a^+b^+)^+$	a^+b^+
$(a^+c^+b)^+$	a^+c^+b

次のような境界が認定される。

$$ab | aab | aab$$

この結果、認定される単位 X は、先ほどの形式文法と全く同じものとなる。

では、形式文法と、どこが異なるのであろうか。

- (1) 形式文法においては、単位 X として認定される記号列は、あらかじめ集合として定義されている。これに対して、境界認定モデルでは、定義されていない。
- (2) 境界認定モデルでは、このモデルに与えられる記号列集合が定まった時、はじめて、単位 X として認定される記号列集合が定まる。

境界認定モデルでは、そのモデルに与えられる記号列集合は想定されていない。どのような記号列集合が与えられるかによって、認定される単位の集合が異なることになる。表 1 に例を示す。

以上の議論からわかるように、境界認定モデルは、全智モデルではない。境界認定モデルは、ある局所的手がかりから、境界を認定する能力を持っているに過ぎず、どのような記号列がモデルに与えられるのか、最終的にどのような記号列が単位として認定されるか、については「知らない」のである。

このように、形式文法と境界認定モデルは、「全智モデル」対「非全智モデル」という形で区別されうる。

4. ブロックモデルと境界モデル

「境界認定」の考え方は、もう少し一般化して議論することができるように思う。ここでは、ブロックモデル (block model) と境界モデル (boundary model) という新しい2つの用語を導入して、議論してみよう。どちらのモデルも、全体と部分の関係に対するモデルであるが、その考え方が根本的に異なる。

4.1 ブロックモデル

ブロックモデルとは、「全体は部品から構成されている」と考えるモデルである。いわゆる「構成的定義」は、この考え方に基づいている。

ブロックモデルの典型例は、レゴ (Lego) である。レゴは、基本的な部品が用意され、これを組み合わせて、色々なもの（オブジェクト）を作成する。図 3 に、レゴで作ったさかなを示す。

ここで、重要なのは、次のことである。

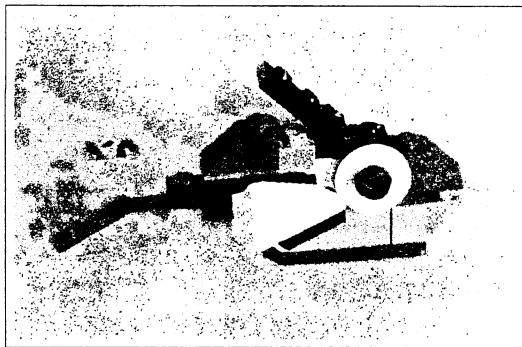


図 3 ブロックモデル: レゴで作ったさかな

- (1) ひと目見ただけで、部品から構成されていることがわかる。
- (2) 部品の単位は明確である。だれが分解しても、まったく同じ部品群に分解される。つまり、ここでの部品とは、絶対基本単位である。
- (3) この「さかな」の構造を知るということは、どのような部品がどのように組み合わせられて、この「さかな」ができていないか、を知ることである。

4.2 境界モデル

一方、図 4 に示すような、木製のさかなの模型では、話はまったく違ってくる。

この「さかな」は、部品から構成されていない。「まず『さかな』ありき」である。しかし、この「さかな」の上に、適当な境界線を引けば、尾、ひれ、胴体などの部分に分けることができる。尾やひれや胴体などは、最初から部品として存在するわけではなく、境界を設定することで、はじめて見えてくるものである。そして、一旦、このような境界が設定されれば、尾やひれや胴体は、さかなを構成する部分とみなすことができるようになり、「さかなは、胴体と尾とひれから構成されている」と違和感なく言うことができるようになる。

このように、境界を設定してはじめて見えてくるような、全体と部分の関係を境界モデルと呼ぼう。境界モデルとは、全体と部分の関係を、「まず、全体があり、境界を設定したとき、はじめてその部分が見えてくる」と捉えるモデルである。

このような境界モデルの例は、他にもある。たとえば、マンションは境界モデルの一つの例である。私が住んでいるマンションは、128 戸から構成されているが、各戸をブロックのように積み上げて、マンションが作られているわけではない。まず、マンション全体があり、そこに床や壁などの境界を作って、初めて各戸を実体化させているのである。稼働棚板を持った書庫も、境界モデルの例である。書庫のそれぞれの段は、

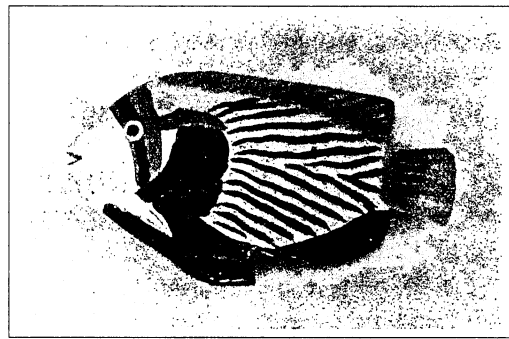


図 4 境界モデル: 木製のさかな

棚板（境界）を置くことによって、初めて実体化する。

4.3 日本語文解析のモデル

さて、本題に進もう。

- (1) 文の生成のためのモデルには、ブロックモデルが妥当であろう。

文の生成では、（おそらく、）言いたいことが部品として存在し、それを組み合わせることに主眼があるのであるから、これは、ブロックモデルで考えるのが適切であろう。

では、解析はどうか。

- (2) 従来の文解析のモデルは、ブロックモデルであった。

これは、すでに前節で考察した。形式文法の本質は、文の生成モデルである。形式文法の最大の興味は、正しい文と非文の区別である。これを解析系として使用する場合、与えられる単語列は正しい文であることがあらかじめ想定されており、文の構造（部分要素）を知ることが目的となる。

- (3) 英語文の解析のモデルは、ブロックモデルでもよい。

英語の文の場合、ひと目見ただけで、部品（語）の存在は明らかである。ゆえに、英語文解析を、ブロックモデルで捉えようという方向性は、なんら不思議ではない。特に反対する理由もない。

- (4) 日本語文の解析のモデルは、ブロックモデルよりも境界モデルの方が適切。

しかし、日本語文の場合は、境界線を引かない限り、部分は見えてこない。とすれば、これは、境界モデルとして考えるのが適切であろう。

ここまで来ると、境界認定の主張点ははっきりしてくる。

- (A) まず、文ありき*

* それが正しい文であることは大前提。文と非文を区別すること

- (B) 文を理解するために、(理解する人が) その文内に境界を(能動的に) 設定する。
 (C) その結果として、文が部分に分解される。
 (D) そのような部分を、単位(文節、語)と呼ぼう。

4.4 構成的定義至上主義からの脱却

語を定義するために、(1) それより小さい単位を定義し、(2) それがどのように組み合わせあった場合に語となるかを定義する、という構成的定義が、本当に必要なのであろうか。それが語を定義する唯一の科学的方法なのであろうか。

文を定義するために、(1) それより小さい単位を定義し、(2) それがどのように組み合わせあった場合に文となるかを定義する、という構成的定義が、本当に必要なのであろうか。それが文を定義する唯一の科学的方法なのであろうか。

私には、境界モデルが、別の方法を提供しているように思えるが、読者諸氏はいかに*。

5. 関連研究

文解析において、境界を考えるというアイデア自身は、とりたてて新しいものではない。これまでに次のような研究が存在する。

5.1 80年代の形態素解析

JUMAN 以前の形態素解析システムでは、連接行列の値が2値(接続不可と接続可能)以外のものが存在した**。たとえば、文献9)に記述されている荻野の方式やMuシステムの形態素解析¹⁰⁾では、3値(0: 接続不可、1: 接続可だが切れ目は入れない、2: 接続可で切れ目を入れる)を採用している。これは、概念的に2種類の境界を考えていることとみなすことができる。また、文献11)のp142には、文献9)を紹介したのち、「接続表の作り方により、文節の終わりに切れ目を付けることもできるということである」との記述がある***。しかしながら、これらの文献を読む限り、

には興味がない(もはや問題ではない)。

* 国立国語研究所のこれまでの語彙調査における語の認定基準の記述⁸⁾を読むと、実際の認定マニュアル(ガイドライン)は、「どのような場合に、どこで切るか/切らないか」を規定している。これは、構成的に語を定義しているのではなく、事実上、境界を定義していると見なすのが適切であろう。とすれば、すでに、構成的定義から逸脱している。

** 5.1節で参照している文献は、颯々野学氏に教えていただいた。深く感謝する。

*** なぜ、このような方向に研究が進まなかったのか、よく考えてみる必要があろう。私の一つの仮説は、英語の文解析の枠組をそのまま持ち込むことが前提となっていたため、というものである。つまり、英語文は、あらかじめ語に分割されているので、日本語文もそれと同じような形式にすればよい、という考え方が支配的だったのではないだろうか。

多数の境界種別を導入し、それらの認定結果を境界情報として実体化させて出力することは、想定されていなかったようである。その点において、ここで提案した境界認定は、新しい要素を含んでいると言える。

5.2 日本語の文節まとめ上げ

ルールによる日本語文節のまとめ上げの処理には、明示的ではないにせよ、境界を認定するという側面を持つのが普通である。たとえば、JUMAN/KNPでは、KNPのなかで文節まとめ上げの処理を行なっているが、knp2.0b6のマニュアルおよびソースコードを読むと、まず、形態素に対して、自立語、付属語、複合語の先頭、複合語の末尾、複合語内、などのラベル(属性)を付与した後、文節の範囲を決定している。これらのラベルを付与する処理は、局所的な文脈(2~3個の形態素の並び)に基づいており、3.1節で述べた形式文法に基づく方法というよりは、3.2節で述べた境界認定の枠組に近い。

これは、次のように理解できる。文節は、近似的に、

文節 = 自立語 + 付属語 *

と書けるが、このようなモデルでは、連続する自立語間が、文節境界なのか、文節内境界なのか定まらない。典型的には、以下のような例である。

(15) 昨日 | 講習会へ行った。

(16) 夏季 | 講習会へ行った。

このような場合、文節がどのような構造をとりうるかという構成的観点で文節を定めるのではなく、より積極的にその境界が文節境界か否かを決定しなければならない。となると、必然的に、構成的定義に基づく認定から逸脱しなければならない。

5.3 統計的手法によるチャンキング

与えられた列に含まれる部分列をひとまとまりのチャンクと認定するチャンキングタスクを統計的手法で実現する場合、注目する要素の前後に、ある幅の窓(window)を設定し、その窓に含まれる範囲の情報から、注目する要素に付与すべきラベルを決定する方法がとられるのが普通である。このような方法で、たとえば、base NPを認定することを行なった場合、そのチャンカーは、base NPがどのように構成的に定義されるかを知らないまま、チャンキングタスクを実行していることになる。これは、非全智モデルであり、その点で境界認定と類似点がある。

チャンキングで使用されるラベル体系に、Inside/Outside と Start/End がある。文献12)の表1(表2として引用)に、これらの比較が示されている。

(1) Inside/Outside

表 2 ラベル体系の比較 (文献 12))

	IOB1	IOB2	IOE1	IOE2	Start/End
In	O	O	O	O	O
early	I	B	I	I	B
trading	I	I	I	E	E
in	O	O	O	O	O
busy	I	B	I	I	B
Hong	I	I	I	I	I
Kong	I	I	E	E	E
Monday	B	B	I	E	S
,	O	O	O	O	O
gold	I	B	I	E	S
was	O	O	O	O	O

I チャンク内

O チャンク外

B チャンクの先頭 (一つ前は別のチャンク)

(2) Start/End

B チャンクの先頭

E チャンクの末尾

I チャンク内

S その要素だけで一つのチャンク

O チャンク外

このような各種の表現が存在するのは、本来、境界が持つべき情報を、要素 (トークン) に付加しようとすることに起因すると考えられる。境界に情報を付加する場合、次のようなラベル体系を考えればよい。

B その境界からチャンクが始まる

E その境界でチャンクが始まる

EB 2つのチャンクの境界

X それ以外

このようなラベル体系を用いると、表 2 のチャンキングの結果は、次のように表現されることになる。(ラベル X は省略した。)

(17) In_B early trading_E in_B busy Hong Kong_E Monday_E,_B gold_E was ...

このようなラベル体系を使用したチャンキングは、境界認定の一つの実現法となる。

5.4 Constituent Boundary Parsing

文解析において、境界を積極的に利用した手法に、Furuse らの Constituent Boundary Parsing (CBP) がある¹³⁾。CBP では、主に、機能語 (functional word) を境界として使用する他に、ある種の POS bigram を挿入して、境界として使用する。例を以下に示す。

(18) The bus leaves Kyoto at eleven a.m.

(19) The bus *noun-verb* leav_s *verb-propn* Kyoto at eleven a.m.

日本語においては、機能語が頻出するため、POS bigram の境界を挿入するのは稀である。しかし、つぎ

のように、助詞が省略された場合は、POS bigram 境界が挿入される。

(15) Kochira jimukyoku

(16) kochira *pron-noun* jimukyoku

この手法は、境界を実体化し、明示的に表現する点において、境界認定と類似点がある。

謝 辞

本稿の内容をまとめるに際して、影浦峯、竹内孔一、丸山岳彦、柏野和佳子、颯々野学、菊井玄一郎、宇津呂武仁の諸氏に、大変有益なコメントをいただいた。深く感謝する。また、表の引用を快諾して下さいた工藤拓氏に感謝する。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集. 情報処理学会自然言語処理研究会, 2004-NL-161, pp. 97-104, 2004.
- 2) 伝康晴, 宇津呂武仁, 山田篤, 浅原正幸, 松本裕治. 話し言葉研究に適した電子化辞書の設計. 第 2 回話し言葉の科学と工学ワークショップ講演予稿集, pp. 39-46, 2002.
- 3) 伝康晴, 宇津呂武仁, 山田篤. UniDic version 1.1.2 ユーザーズマニュアル. 千葉大学文学部行動科学科・音声対話技術コンソーシアム, 2004.
- 4) 水谷静夫, 星野和子. 名詞から副詞まで - 語類の新しい枠づけ. 計量言語学, Vol.19, No.7, pp331-340, 1994.
- 5) 内山清子, 竹内孔一, 吉岡真治, 影浦峯, 小山照夫. 専門分野における複合名詞解析のための名詞文法属性の分類について. 計量言語学, Vol.23, No.1, 2001.
- 6) 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.
- 7) 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界検出プログラム CBAP の開発と評価. 自然言語処理, Vol.11, No.3, pp. 39-68, 2004.
- 8) 島村直己, 鶴岡昭夫, 正保 勇. 国立国語研究所の語彙調査—資料編—. 国立国語研究所, 2004.
- 9) 高橋延匡 編. 日本語情報処理. 近代科学社, 1986.
- 10) 電子技術総合研究所, 京都大学. 日英科学技術文献の速報システムに関する研究, 言語処理システムの開発に関する報告書. 1986.
- 11) 田中穂積. 自然言語解析の基礎. 産業図書, 1989.
- 12) Taku Kudo and Yuji Matsumoto. Chunking with Support Vector Machine. Proc. of NAACL-01, pp. 192-199, 2001.
- 13) Osamu Furuse and Hitoshi Iida. Constituent Boundary Parsing for Example-Based Machine Translation. Proc. of COLING-94, pp. 105-111, 1994.