自然言語処理 51－4
（1985. 9. 13）

日本語の規格化
－係り受け関係の規格化とそれへの変換ルール－

吉田 将・松山 晶子
（九州大学・工学部）

1. まえがき

我々は科学技術文書を対象とした日本語の規格化（コントロール）について研究してきた[1)]。

そこでは、①付属語的表現の規格化、②係り受け・並列構造の規格化、③文章構造の規格化、について第1次案を提示した。

その後、多くの例文について、実際にこの結果を適用しながら問題点を再検討した。また、そこでは未検討であった部分の文法を詳細化・具体化した。本報告は、それらの内で特に係り受け構造の規格化と、一般の文（規格化されていない文）における係り受けからこの規格化された文法に合った構造の係り受けに意味を不変に保って変換する方法、すなわち係り受け構造の規格化とそれへの、意味を保存した変換ルールについて述べる。

規格化文法を考えるとき以下の点が重要である。

① 日本語科学技術文書を記述する。その記述力が充分であること。すなわち表現したい内容が正確に記述できること。

② 文の構造・意味が一意的に定まる文法であること。

③ 人間が読んで分り易いこと。

④ 書き手に大きな負担がかからないで文書が作成できること。

①については、科学技術文書一般を対象としても充分な記述力があることを望みたい。言語をコントロールするという意味では同じであるけれども我々と異なる方式に、対象分野を限定することにより語の意味・用法を明確にするという方式が研究されている。例えば、気象通報文など。この場合、文のパターン、表現内容に強い制限を加えることができ、したがって、語の意味も自立語を含めて限定できる。

我々の場合も、語の意味に制限を加えるが、付属語を中心にコントロールし、自立語については比較的自由に（意味はともかく）使用できるように考えている。これにより、科学技術文書一般に適用範囲を広げたい。

②の、文の構造が一意的に決まるということは、パーザーの構成が容易となり、また、意味のあいまいさも限定できるということである。これにより、機械処理が容易になる。

日本語をコントロールすることの重要性は機械処理の容易化・あいまいさの低減（できれば一意）にあると考える。これにより原言語としての日本語の機械による解析・解釈の高精度化が行える。他言語への翻訳を考えるとき、原言語での解析が充分に行なわれていることは、①相手言語が多い場合、（元がしっかりしていれば、相手言語側での人手の介入が少なくて済む）、②翻訳の質の高さあるいは信頼性の高さが要求される場合、などの場合に重要である。

③の読み易さ（Readability）については以下のことがいえる。

我々は、前年度までの研究で、規格化と読み易さとは同一線上にあることを見出すことができた。すなわち、あいまいさが少なく機械処理の容易な構造をしている文書はまた、人間が読んでも理解し易いということである。文法に制約が多いほど文章が長くなったり、簡潔さが失われるのではないかと思う人が多いかも知れないが、我々の実験では例外なく文章は短かく、簡潔になる。ただし、文書が読み易い、理解し易いということの定量的な測定方法が開発されているわけではないので、今後の研究・評価が必要である。

④の書き易さに関しては以下のような疑問点が挙げられる。

① 出来上がった文書は読み易く、機械処理もし易いに違いない。

② しかし、この文法に従って文書を作成するには大きな負担がかかるのではないか。

③ また、文書作成時に負担が少々あっても、後の機械処理がそれを充分に補償するほど有効・効率的であれば良いが、人手がかかるのではないか。

これらの疑問点に充分答えられるだけの実験結果を我々はまだ、持ち合わせていない。にもかかわらず、我々は次のような理由から規格化日本語の開発

研究は現時点において必要であると考える。

① 出来上がった文書が、我々人間が見て簡潔・明確であることは重要であり、そのことだけでも他に替えがたい。

② 今後、日本には欧米語だけでなくアジア各国語を始めとして、多くの国々の言語で文書を作成することが要求されるだろう。その場合、翻訳の機械化による効率化・省力化は避けられない。この際、原言語のコントロールは有力な手段であるに違いない。さもなければ原言語と相手言語の両方を理解できる人を多数必要とすることになり、このような人手を確保することは不可能になるに違いない。

③ 規格化文法による文書作成の負担を軽減するための支援システムの開発は可能に違いない。

上のような理由により、我々は、規格化日本語の開発に際して、①規格化日本語を開発することと、②その文法に従った文書を作成すること、とは一応切り離して平行に作業を進めることにした。もちろん両者を念頭に置きながら作業を進めてはいるが。

2. 係り受け関係の規格化

日本語文は文節の並びからなると考えることができる。文中の各文節は、（最後の文節を除いて）それよりも後方にある適当な文節を一定の意味関係で修飾する。この修飾・被修飾の関係（文節間相互の役割を示す関係）を係り受けと呼ぶ。

日本語文の構造を知るためには係り受け構造を知る必要がある。逆に、日本語文を書く場合には、係り受け構造が明確な文を書く必要がある。一般には、係り受け構造を知るためには、語の意味、文脈の情報などを知る必要がある。

日本語文の係り受けの構造を明確にするために、係り受けに制限をおく。この際、語の意味に関係なく係り受けの構造が決まるように係り受けを制限する。もちろん、文の構造（係り受けの構造）が決まっても文の意味が決まるとは限らない。

係り受けの構造は第ｉ番目の文節が第ｊ番目（ｉ＜ｊ）の文節にどのような意味的関係で係っているかを示すものであるから、２次元の関係である。これを文という文節の１次元的並びで表現しなければならないのであるから、左右括弧の対のように任意の位置に明確な対応関係を付けることができる記号を用いない限り、複雑な文の係り受け構造を一意的に表示することには限度がある。

そこで、我々は次のような基本方針で係り受け構造の一意化を行うことにした。

(1) 係り受けの関係に“結合の強さ”という概念を導入する。

たとえば、‘野の花’のように‘名詞＋の＋名詞’の結合よりも‘花および草’のように‘名詞＋接続詞＋名詞’の方が結合が強いと決めれば‘野の花および草’は‘（野の（花および草））’のような構造と解釈される。

(2) 結合の強さが同順位の場合は左から右（文頭から文末）の順に係り受け関係を決めていく。

(3) 文中における並列構造（結合の強さが同じものが並列している）は構造の決定が困難で解決すべき多くの課題を残している。上述の(1)の方針に従って結合の強さを利用しても並列構造が決められない場合には、

① 並列句に①、②のように番号を付けて明確に表示する、

② 箇条書きにする、

などの形式で表現する。

(4) 文が長い場合には係り受け構造が複雑となり、上のような方法では一意に決められない場合が多い。我々は一文を短かく書くことを基本方針とし、その方法を示す。

(5) 規格化文法に従っていない文を上のルールによって規格化された文に書き替えるための変換ルールを設定した。

本文では、以上の基本方針のうち、(1)・(2)・(3)について述べ、(4)・(5)の文の分割については次の機会に発表することにしたい。文を短かく書くことは係り受け構造を明確にすることの他に、文を読み易く理解し易くするためにも重要である。

2．1 係り受け規則

一般的な係り受け規則は以下の５つである。

(1) 連体修飾語は後方の体言のうち最も近い体言に係る。

(2) 連用修飾語は後方の用言のうち最も近い用言に係る。

(3) 「連体修飾語＋読点」は名詞句の最後の体言に係る。

(4) 「連用修飾語＋読点」は文の最後の体言に係る。

(5) 「は」は「PR_{PP}＋読点」・「Ｐ(連用)＋読

点」・「P(終止)＋句点」のうちの最も近いものに係る。

２．２　連体修飾語間の係り受けに関する変換

連体修飾語に関する変換規則を以下のように決める。

(1) 係り受けの優先順位

① n・n・…　、　nCn

② Tn

③ $nR_{NN}n$　、　$pR_{PN}n$　、　p(連体)n

④ nとnと　、　nかnか　、　nとかnとか

①～④が優先順位を表わす。但し、④における最後の「と・か・とか」を、並列の範囲を示すマーカーとみなす。このマーカーが現れたならばそこまでを一塊とみなす。したがって、「n_1とn_2とのn_3」は「(n_1とn_2と)のn_3」となる。

また、N・Pは修飾語を含む句を表わし、また、n・pは単語を表わす。

例)

{(①)①(③)}　→ ① ④ ③ ◇

$\{(n\cdot n)\cdot(nR_{NN}n)\}$ → $n\cdot n$と$nR_{NN}n$と

外側の括弧の内の数字（優先順位を表わす数字）が内側の括弧の内の数字よりも大きくなるように変換する。◇は並列の範囲を示すマーカーである。

(2) 読点の使用

$n_1R_{NN}n_2\cdots n_{m-1}R_{NN}n_m$ においてn_1がn_mに係るとき、n_1R_{NN}の後に読点を付ける。但し、「連体修飾語＋読点」は名詞句の最後の体言に係るので以下の場合には適用できない。

(1) $n_1R_{NN}n_2R_{NN}n_3\cdots n_{m-1}R_{NN}n_m$ において、n_iが後方のnのうちn_{i+1}・n_m以外のnに係る。

(2) $PR_{PN}n$を含む名詞句において、Pの中の名詞句が読点を有する。

また、連体修飾語中で読点を使用することは文としてはあまり望ましくない。したがって、他に方法があればそちらを優先的に採用する。方法がなければ読点を使用することに決める。そこで、ある特定の条件を満たした場合の変換方法を求める。

(3) 並列構造における箇条書き

N_1＜並列＞N_2＜並列＞---＜並列＞N_mにおいて、以下の条件を満たせば箇条書きにする。

$m\geqq 3$であり、かつN_iのいずれかが名詞句である。または、N_iの中に「nとnと」が含まれる。

N_1＜並列＞N_2＜並列＞---＜並列＞N_m

→ ① N_1、② N_2、--- ⓜ N_m、

このパターンにおける最後の読点も並列の範囲のマーカーとみなす。

この規則は文の分割規則につながる。

(4) 分配：x(y＋z)＝xy＋xz

(x＋y)z＝xz＋yz

※　＋；並列関係　　×；修飾関係

但し、nR_{NN}(n・n)のような単純な句に対しては適用しない。読点を必要とする句（例えば、$nR_{NN}n$、n・nとnと）に対して適用する。また、この方法は文の意味の点から適用できない場合がある。例えば、「xとyとの関係」においては、適用できない。

(5) 位置の移動

TとPR_{PN}とは移動可能である。

$\{T(nR_{NN}n)\}$ → $nR_{NN}Tn$

$[\{PR_{PN}(nR_{NN}n)\}\cdot n]$

→ $nR_{NN}PR_{PN}n$とnと

(6) 「nのn」の「の」を省略する。

$n_1R_{NN}(n_2のn_3)\ R_{NP}$

→ $n_1R_{NN}n_2n_3R_{NP}$

$n_1の(n_2R_{NN}n_3)\ R_{NP}$

→ $n_2R_{NN}n_1n_3R_{NP}$

(7) サ変名詞を動詞化する。

$n_1R_{NN}(n_2のn_3＜サ変＞)\ R_{NP}$

→ $n_1R_{NP}n_2をn_3すること R_{NP}$

(但し、R_{NP}はR_{NN}と同じ意味である)

$n_1の(n_2R_{NN}n_3＜サ変＞)\ R_{NP}$

→ $n_2R_{NP}n_1をn_3すること R_{NP}$

(8) 受動態を能動態に変える。

P＜受身＞($n_1のn_2$)R_{NP}

→ n_1がP＜能動＞n_2R_{NP}

(9) $n_1R_{NN}(n_2R_{NN}n_3)$

→ $n_1R_{NN}n_3$(C、$n_2R_{NN}n_3$)

n_1R_{NN}（n_2R_{NN}＜同列＞n_3）

→ $n_1R_{NN}n_3$（C、n_2）

＊＜同列＞：の・のような・という

C：すなわち・例えば・φ

(10) 連体修飾における分割

(10)-1 名詞句がｐを含む場合

名詞句を［N＝$n_1R_{NP}p_1\cdots n_2R_{NP}p_mn_m$］と置く。

(a) p_mの直後で切る。

→ $n_1R_{NP}p_1\cdots n_2R_{NP}p_m$。

そのようなn_m---

(b) 名詞句全体を文の外に出す。

(c) p_iのいずれかの直後で切ってみる。

(10)-2 名詞句がｐを含まない場合

名詞句を［N＝$n_1R_{NN}n_2R_{NN}\cdots n_{m-1}R_{NN}n_m$］と置く。

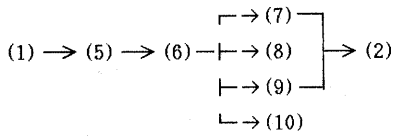
(a) 名詞句全体を文の外に出す。

(10) に関しては次回の文の分割で詳しく述べる。

以上の(1)～(10)の方法の適用順位は以下のとおりである。

並列関係　(1) →(3)、(4)

修飾関係



２．３　連用修飾語間の係り受けに関する変換

連用修飾語に関する変換においては以下の規則を用いる。

① 文の分割
② 読点の使用
③ 位置の移動
④ 分配
⑤ 「は」の削除

(1) NR_{NP}---P_1---P_m。

(a) Nがｐを含まない場合

→ ---P_1---$NR_{NP}P_m$。

（位置の移動）

(b) Nがｐを含む場合

→ NR_{NP}、---P_1---P_m。

（読点の使用）

(c) 文が60字以上である場合

→ ---P_{1o} C、NR_{NP}---P_m。

（文の分割）

(2) NR_{NP}---P_1---P_m。

(a) 文が60字以上である場合

→ NR_{NP}---P_{1o} --- C、NR_{NP}---P_m。

（分配）

(b) 文が60字未満である場合

→ NR_{NP}、---P_1---P_m。

（読点の使用）

(3) NR_{NP}は（または「、」）---P_ℓ---P_m。

(a) 文が60字以上である場合

→ NR_{NP}は---P_ℓ。C、---P_m。

（文の分割）

(b) 文が60字未満である場合

→ NR_{NP}は---P_ℓ、---P_m。

（読点の使用）

→ NR_{NP}---P_ℓ---P_m。

（「は」の削除）

→ ---$NR_{NP}P_\ell$---P_m。

（位置の移動）

(4) NR_{NP}---P_1---P_ℓ---P_m。

(a) → NR_{NP}、---P_1---P_ℓ。

C、---P_m。　（文の分割）

［参考文献］

１）吉田将：日本語の規格化に関する基礎的研究、昭和５８年度科学研究費補助金、一般研究（Ｂ）研究成果報告書、１９８４．

［連体修飾語に関する変換例］

{文脈依存文法や{タイプ０の文法}}の解析手法については	→ (1)	文脈依存文法とかタイプ０の文法とかの解析手法については
[{標準形の句構造文法}、{ブラケット文法}、{ブラケット言語}、{{{標準形の句構造文法}と{ブラケット文法}}の対応}]を定義する	→ (3)	①標準形の句構造文法、②ブラケット文法、③ブラケット言語、④標準形の句構造文法とブラケット文法との対応、を定義する
各々のΣの要素の間に現れる{括弧の列}は	→ (6)	各々のΣの要素の間に現れる括弧列は
Ｇの{ＣＦ規則の適用}には	→ (6)	ＧのＣＦ規則適用には
	(7)	ＧにおいてＣＦ規則を適用することには
文脈自由文法の構文解析法である{Earleyの方法}の拡張になっている	→ (9)	Earleyの方法（文脈自由文法の構文解析法）の拡張になっている
辞書の{使われ方、内容、構成}は	→ (1)	辞書の使われ方・内容・構成は
入力した原稿（テキスト）の編集（誤り修正、推敲、文章構成、印刷形式など）を	→ (9)	入力した原稿（テキスト）の編集（例えば、誤り修正・推敲・文章構成・印刷形式等）を
このような{日本語を使う側の特性}に	→ (5)	日本語を使う側のこのような特性に
[{自立語と付属語}、{付属語と付属語}の接続]には	→ (4)	自立語・付属語の接続と付属語・付属語の接続とには
[自動インデクシング機能、すなわち{{文中に現れる単語（句）}と{その現れる場所}}の一覧表の自動作成機能]なども望まれる	→ (1)(9)	自動インデクシング機能（すなわち、文中に現れる単語（または、句）とその現れる場所との一覧表の自動作成機能）なども望まれる
不要語テーブルとのマッチングで生じる{切断の誤り}を	→ (6)	不要語テーブルとのマッチングで生じる切断誤りを
[{連体詞、接続詞、副詞、形容詞など}{{専門用語や{その構成要素}}に成り得ない８１４単語}]が	→ (9)	専門用語とかその構成要素とかに成り得ない８１４単語（例えば、連体詞・接続詞・副詞・形容詞など）が
図－３に示すような{漢字と平仮名で混ぜ書きされたもの}である	→ (2)	図－３に示すような、漢字と平仮名とで混ぜ書きされたものである
[{その意味内容}と{記述法}]の間には	→ (1)	その意味内容と記述法との間には

［｛ＤＷの抽出｝および｛ＤＷ－ＥＷの関係付け｝］の精度をあげる → (4) ＤＷの抽出の精度とＤＷ－ＥＷの関係付けの精度とをあげる

意味処理・推論処理を伴う｛自然言語の知的処理｝を行うためには、 → (2) 意味処理・推論処理を伴う、自然言語の知的処理を行うためには、

漢字表記の正書法になっていない場合（ひらがな書きなど） → (9) 漢字表記の正書法になっていない場合（例えば、ひらがな書きなど）

｛(1)漢字の読みの除去｝と、｛(2)｛括弧とドット｝の処理（括弧の削除）｝からなる → (1)(3) (1)漢字の読みの除去、(2)括弧・ドットの処理（括弧の削除）、からなる

［これに関連した、［｛語釈義文の｛標準文への変換処理｝、および｛語釈義文（の文末表現）の構造的特徴の調査｝］］についても考察した → (3)(7) これに関連した、①語釈義文を標準文へ変換処理すること、②語釈義文（または、語釈義文の文末表現）の構造的特徴の調査、についても考察した

一回で引き起こされる｛二次記憶から一次記憶へのデータ転送回数（又は、木の高さ）｝は → (2) 一回で引き起こされる、二次記憶から一次記憶へのデータ転送回数（又は、木の高さ）は

｛挿入アルゴリズム｝又は｛４節の削除アルゴリズム｝で → (1) 挿入アルゴリズムか４節の削除アルゴリズムかで

｛α、β、γ、δ｝は［｛ポインター｝と｛第一種のキー｝］の列 → (1) α・β・γ・δはポインターと第一種のキーとの列

｛綴り、読み、及び文法情報（品詞、活用）｝を → (1) 綴り・読み・文法情報（品詞・活用）を

｛１０種類のモータ｝、｛３０種類のドラム｝、｛４種類のギア｝及び｛フック｝から成り立っており → (3) ①１０種類のモータ、②３０種類のドラム、③４種類のギア、④フック、から成り立っており

｛パワーと変速｝にそれぞれ二つづつの選択が可能な｛ミキサという製品｝の例である → (1)(9) パワーと変速とにそれぞれ二つづつの選択が可能な製品（ミキサ）の例である

モジュール作成中になされた｛設計者の意志決定｝ → (8) 設計者がモジュール作成中に行った意志決定

本論文で我々は、［｛与えられたモジュール外部仕様に基づいて｝｛モジュール内部論理の設計・文書化｝を実行する｝］過程における計算機支援の一方式を考察した → (10) 与えられたモジュール外部仕様に基づいてモジュール内部論理の設計・文書化を実行する。その過程における計算機支援の一方式を本論文で我々は考察した