形態素・構文タグ付きコーパス作成の作業基準
version 1.8

黒橋 禎夫　居蔵 由衣子　坂口 昌子

平成 12 年 4 月

目 次

1　はじめに

本マニュアルは「京都大学コーパス作成プロジェクト」における作業者 (以下アノテータとよぶ) の作業基準を示したものである．アノテータの作業は，自動解析システム (JUMAN，KNP) の解析結果を，このマニュアルの基準にしたがって正しく，かく均一に修正することである．

このプロジェクトでは，手修正されたコーパスの作成と同時に，我々の解析システム (JUMAN, KNP) を高精度なものに修正していくことをもう一つの目標としている．そのため，種々の解析誤りに対してマニュアルの中で示されているマークを挿入することも重要な作業である (5 章に一覧)．また同様に，形態素解析辞書の更新のために形態素解析結果の修正の際に形態素の仮辞書登録，仮辞書削除も行う (2.4 節，3 章)．

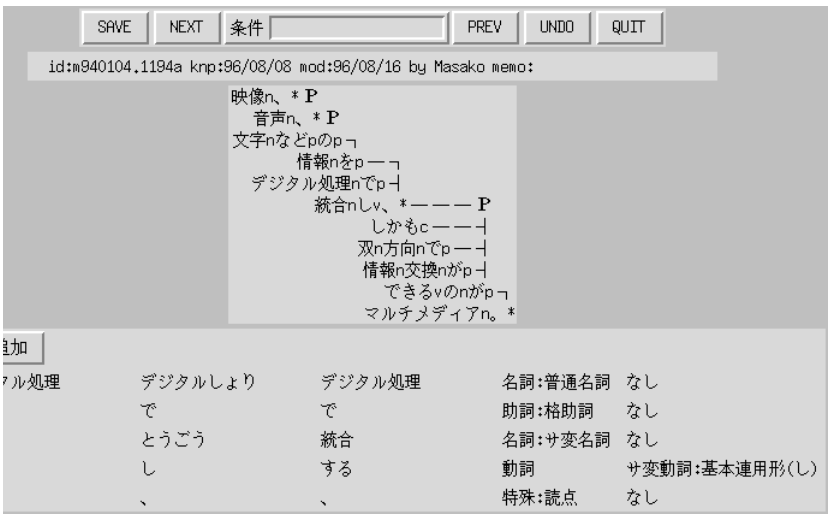


図 1: 解析結果修正用インターフェース

2 ユーザインターフェース

解析結果の修正は図 1 に示すシステムを用いて一文単位で行う.

2.1 コマンド

インタフェースの最上部にはコマンド・ボタンが並べられている．各コマンドは以下の機能を持つ.

SAVE ：それまで一文に対して行った修正を保存する.
NEXT ：次の文に進む (条件指定がある場合は条件を満たす次の文).
条件 ：この部分に文管理情報 (文 ID，メモ欄など) に対する条件を指定すれば，NEXT, PREV がそれにマッチする文だけを対象とする.
PREV ：前の文に戻る (条件指定がある場合は条件を満たす前の文).
UNDO：それまで一文に対して行った修正を無効にし，文解析結果をもとの状態にもどす.
QUIT ：コーパス修正作業を終了する.

2.2 文管理情報

インターフェースの 2 段目には各文の管理情報が示されている．各文には ID が与えられており，解析の日付け，修正の日付け，修正作業者の名前などは自動的に記録される．また，任意のメモを書き込むことができるようになっている (5 章).

2.3 係り受け構造

中央の画面は文の係り受け構造を表示・修正するためのもので，対角線上に順に文節が並び，右上の領域で係り受け関係を持つ文節がリンクされている．この領域はすべてマウスセンシティブになっていて，ここで適当な位置をクリックすることにより，その左側と下側に位置する文節間の係り受け関係を入力できる．マウスボタンの選択によって通常の係り受け関係 (左ボタン，マークなし; 4.1 節),

並列関係 (中ボタン，P または I マーク; 4.2 節)，同格関係 (右ボタン，A マーク; 4.3 節) が選択できる．修正した結果が係り受けの非交差条件を破る場合には，その部分の色が変わり注意を促すようになっている．

2.4 形態素・文節情報

インターフェースの最下部は形態素解析結果と文節区切りを表示・修正するための画面である．この画面は通常は表示されておらず，係り受け構造の表示画面で文節をクリックするとその文節内部の形態素情報が表示される．複数の文節の内容を表示する場合は，前から順に連続する文節を選択 (クリック) していく必要がある (図 1 は「デジタル処理で」と「統合し,」を順にクリックした結果).

各列には形態素の (文中での) 表記，読み，原形，品詞:細分類，活用型:活用形が示されている．表記，読み，原形はテキストとしてのエディットによって修正でき，品詞，活用形などはメニューによって変更できる．活用型は品詞に対応するものだけ，活用形は活用型に対応するものだけがメニュー表示される．また，活用形を修正すると原形は自動修正される．なお，品詞のメニューの末尾には**仮辞書登録**と**仮辞書削除**という項目がある．この項目を選択すると，その形態素情報が辞書登録・削除の候補として別の場所に蓄積されるが，その時点で即座に形態素解析システムの辞書が更新されるわけではない．

各列の先頭のマークは文節の先頭の形態素を示すものである．このマークの変更により文節の区切りを修正できる．また,「追加」ボタンをクリックすると表示されている形態素列の末尾に新たな形態素が追加される．一つの形態素を二つ以上の形態素に分割する場合などに用いる．形態素・文節区切りの修正が終了したら「OK」ボタンを押してこの画面を閉じる．

3 形態素と文節に関する基準

形態素解析結果を修正した場合は，形態素解析辞書の更新のために，必要に応じて形態素の仮辞書登録，仮辞書削除を行う．辞書登録については，正解コーパス中の語と辞書の差分をとりだし，それらを検討しながら辞書登録していくこともできる．しかし，辞書登録することがほぼまちがいなく必要であると考えられる形態素については形態素解析修正時に仮辞書登録を行うことにする．逆に，不適当な形態素の削除については，その候補を自動収集することはできないので，形態素解析修正時の仮辞書削除が非常に重要である．

3.1 品詞の区別

コーパスの品詞体系は形態素解析システム JUMAN の品詞体系に準じたものである．JUMAN の品詞体系は 2 段階になっており，1 段階目が名詞，動詞など，2 段階目に各品詞の細分類がある．

品詞	細分類
特殊	句点 読点 括弧始 括弧終 記号 空白
動詞	
形容詞	
判定詞	
助動詞	
名詞	普通名詞 サ変名詞 固有名詞 地名 人名 組織名 数詞 形式名詞 副詞的名詞 時相名詞
指示詞	名詞形態指示詞 連体詞形態指示詞 副詞形態指示詞
副詞	
助詞	格助詞 副助詞 接続助詞 終助詞
接続詞	
連体詞	
感動詞	
接頭辞	名詞接頭辞 動詞接頭辞 イ形容詞接頭辞 ナ形容詞接頭辞
接尾辞	名詞性述語接尾辞 名詞性名詞接尾辞 名詞性名詞助数辞 名詞性特殊接尾辞 形容詞性述語接尾辞 形容詞性名詞接尾辞 動詞性接尾辞

インターフェースの係り受け構造表示画面では，各語の品詞は以下の 1 文字の記号で表示されている (本マニュアルの以降の係り受け構造表示でもこれらの記号を用いている)．

特殊	*	名詞	n	接続詞	c
動詞	v	人名	J	連体詞	m
形容詞	j	地名	C	感動詞	!
判定詞	c	固有名詞	N	接頭辞	p
助動詞	x	指示詞	d	接尾辞	s
		副詞	a	未定義語	?
		助詞	p		

語の品詞については，本質的に区別が困難な問題，その問題だけを集中的に作業する方が効率的であると考えられる問題などがあり，現時点では以下に示す基準を考えている．この基準に従って修正の必要がある場合には形態素・文節情報の画面を開いて各語の品詞を修正する．なお，以下の基準でカバーされない不明確な問題があれば随時指摘して下さい．

3.1.1 動詞の連用形と名詞

動詞の連用形と，それが名詞化して用いられている場合とは**区別を行う**.

例) 右に動き (動詞) よけた.
動き (名詞) がにぶい.

システムは，読点の前は動詞，格助詞の前は名詞という程度の区別を行っているが，それが誤りである場合には修正を行う．なお，動詞連用形の名詞化したものとそれに続く名詞が非常に固定化された表現である場合は一語として扱う (例「読み物」).

3.1.2 カタカナ語

カタカナ語が固有名詞，サ変名詞，普通名詞などのいずれであるかは**区別する**．システムは辞書登録されていれば固有名詞を優先し，そうでなければサ変名詞としている．その結果が誤りである場合には修正を行う.

3.1.3 名詞と副詞

名詞と副詞の**区別は行わない**．システムは JUMAN の辞書にしたがって名詞か副詞の区別をしており，この区別は基本的に前後の文脈に関係なく単語に対して固定である．これを修正することはしない．[1]

例) ごはんをみんな (副詞) 食べた.
みんな (副詞; 名詞?) が食べた.

3.1.4 助詞「で」と判定詞「で」

助詞「で」と判定詞「で」は区別を行う．システムは基本的に助詞の解釈を優先するが，それが誤りである場合は修正する．ただし,「〜ではなく」,「〜でもあり」などの「で」はシステムによって判定詞と解釈され，この解釈でよいとする.

3.1.5 名詞＋助詞「に・の」と形容詞

名詞＋助詞「に・の」とナ形容詞・ナノ形容詞の連用形・連体形は**区別を行わない**．システムは基本的にナ形容詞の解釈を優先するが，それが不適当である場合も修正は行わない.

例) 健康に留意する． (本来，名詞＋助詞)
健康に働く． (ナ形容詞)

次の例文は「生活と」の並列からも「健康に」を名詞＋助詞と解釈することが妥当であるが，このような場合も修正はしない．しかし，並列に関する例外的扱いとして**メモ欄に PA** と入力しておく.

[1]「相当」についてはサ変名詞 (「この研究は賞に相当する」) と副詞 (「このような研究は相当した」) を区別をする必要があると考えているが，これはあとからまとめて修正する予定なので，通常の作業では扱わなくてよい．他にこのような区別すべき語があれば随時指摘して下さい.

例) 人間 n の p ─┐
生活 n と p P
健康に j ─┐
深く j ┤
関わって v いる s. *

3.1.6 助詞の細分類

助詞の細分類については**区別を行わない**．システムは，前後の連接から明らかな場合は区別し，そうでない場合は適当な細分類を選択している．現時点ではこの区別は行わず，修正もしない．

3.2 形態素・文節の区切り

3.2.1 形態素の区切り

原則としては，できるだけ小さな形態素に分割することを基本とする (「戸棚」を「戸」と「棚」にまで分割することはしないが)．JUMAN の辞書には比較的長い単位の語が登録されている場合もあるが，固定化されていてかつ解析の曖昧性をなくすためにも有効と考えられるもの以外は**仮辞書削除**を行い，手修正によって分割する．[2]

三文字または四文字の複合名詞内部の区切りが誤っている場合は，手で修正を行う．区切りに曖昧性がある場合，現システムは**後ろの語が短くなる**区切りを優先する．たとえば「党勢力」は「党勢 (名詞) 力 (名詞)」というように解析されるので，これは「党 (名詞) 勢力 (名詞)」と修正する．

もう一つ，形態素の区切り誤りの可能性が高いのは,「なければならない」のように平仮名が連続する部分の区切りである．このうち，まとまり (平仮名列全体) としてみたときに曖昧性がない表現については，まとまりのまま辞書登録することによって解析誤りを防ぐという機能が用意されている (連語登録機能)．解析誤りのうち，このような連語登録をすることが適当であると考えられる表現があれば，**メモ欄に M** と記入しておく．

3.2.2 文節の区切り

文節の基本的構成は「接頭辞＋自立語＋付属語・接尾辞」というもので，接頭辞，付属語・接尾辞は複数の場合，ない場合もあり，複合名詞などでは自立語が複数の場合もある．形態素解析の誤りなどの原因で文節の区切りが不適切になっている場合には手で修正を行う．

ただし，次のような場合は上記の原則に反するが例外的に一文節として扱っている．この他に慣用的で一文節としても問題がないような表現があれば，**メモ欄に B** と記入する．

- 「しようとする」
- 「〜ざるをえない」,「〜ざるをえません」
- 「〜つつある」
- 「〜かも知れない」,「〜かもわからない」

[2]「〜的だ」という形容動詞が登録されている場合があるが，これは基本的にはそのまま放置する．しかし「自然 (副詞) 科学的な (形容詞)」のようにそのままでは非常に不自然になってしまう場合には「自然 (副詞) 科学 (名詞) 的な (接尾辞)」のように修正する (「自然」について副詞と名詞の区別は行わない; 3.1.3 節).

- 「〜たくありません」
- 「〜のに対し」
- 「〜かどうか (も)」
- 「〜とはいえ」
- 「〜といえど (も)」
- 「〜とともに」,「〜とどうじに」
- 「〜にも関わらず」
- 「〜のみならず」
- 「〜て (も)| ば | たら | たって | と (も) ＋いい | よい」
- 「〜ても | ては | ば | たら | ちゃ| たって | と＋いけない | だめだ | 仕方がない | 仕様がない」
- 「〜ても | ては | ば | たら | ちゃ＋ならない | ならぬ」
- 「〜ても | ては | ば | たら | ちゃ| ずに＋いられない | おられない | すまない | おかない」
- 「〜たくて |〜れて |〜えて＋たまらない | ならない | ならぬ | 仕方がない | 仕様がない」
- 「〜に＋違いない | 相応しい | 過ぎる | 限る | 従う | つれる | する」

3.2.3 複合動詞の区切り

「走りはじめる」の「はじめる」ように接辞的に働く動詞はシステムに登録されており，それらは前の動詞とまとめられて一文節となる．しかし，このような接辞的動詞の登録は完全ではなく，また非常に特定の語にしか結び付かないもの (「建ち並ぶ」の「並ぶ」など) は登録されていない．そのため，本来接辞的であるものが別文節になる場合がある．そのような場合は手で一文節にまとめ，**メモ欄に B** と記入する．

3.2.4 複合名詞の区切り

複合名詞の扱い，特に数詞，接頭辞，接尾辞を中に含むような複合名詞 (3.2.2 節の文節の基本的構成に反するもの) については扱いが非常に難しい．一応，次のような基準を設定する．

- 原則として**アクセントを落とさずに言い切れるもの**を一文節とする．

  例) 七番勝負
  第三回大会
  ４階宴会場
  (↔ 2 文節にするもの：三回会合し，六日告示の)

- 日付け，時間の表現などは各単位ごとに別文節とする．

  例) 九十七年　一月　中旬

- 数詞は，同格的なもの,「につき」の補えるものなどは 2 文節とする．

  例) 戦車　五十両
  一個　十円

  一方，分数，住所，決まり文句的な並列などは一文節とする．

  例) ３分の２
  築地一
  １円２０銭
  ３勝３敗
  ２泊３日

- 肩書は一文節とする．

  例) 鈴木一郎元教授
  クリントン大統領

- 「続柄人名」の部分は別文節とする．

  例) 長男　太
  妻　よしこ

- 「・」,「—」,「＝」などの記号でつながれたものは一文節とする．

  例) メキシコ・新ペソ
  中村祐一郎・三菱自動車社長
  １０時—１９時
  １ドル＝１０１円台

- 括弧は，原則としては，前 (「) は同格的に解釈して別文節，後 (」) は一文節とする．

  例) 本欄　「信念と情熱を持った教師いでよ」
  「グリーン」挺

- 複数文節にすると構造に悩むような場合は例外的に一文節とする[3]．

  例) 国連子どもの権利委員会
  宗派中興の祖・蓮如上人

複合名詞についてはシステムの解析結果に不統一なものや誤りがあるが，今後正解コーパスをもとに精密化する予定である．なお，上記の作業基準で不明確な具体例があれば随時指摘して下さい．

[3]「宗派中興の祖・蓮如上人」を「の」できり「・」でつなぐと「宗派中興の　祖・蓮如上人」となりおかしくなる．

3.3 固有名詞の扱い

固有名詞の問題は非常に複雑で，形態素解析誤りのかなりの部分は固有名詞に関連する誤りである．コーパスは固有名詞処理改善のための基礎データとなるので，以下の基準で注意深く作業する必要がある．

なお，固有名詞辞書を整備するために，辞書登録を行う必要があると考えられる固有名詞の**仮辞書登録**，頻繁に解析誤りの原因となるような固有名詞の**仮辞書削除**をできるだけ行う．

3.3.1 品詞分類

固有名詞は再分類は以下のように行う．

人名 姓，名

地名 国，都市，地域，山，川，湖など

組織名 会社名，大学名，省庁名，政党名，寺，神社など

その他 商品名 (アコード，八ッ橋)，品種 (ゴールデンデリシャス)，列車名 (のぞみ)，大会名 (インタコンチネンタル)，法律名，年号など (この「その他」の JUMAN における細分類名が「固有名詞」)

3.3.2 扱いの原則

「京都」や「太郎」のように一語で固有名詞の場合は問題ないが，複合語で固有名詞となる場合，特に組織名の扱いが難しい．原則的には語の単位に分割して，それぞれに適当な品詞細分類を与えることにする．

- 市町村など — 前を地名,「市，区，町，村」を接尾辞とする．

  例) 京都 (地名) 市 (名詞性接尾辞) 左京 (地名) 区 (名詞性接尾辞)
  北 (普通) かつらぎ (地名) 町 (名詞性接尾辞) ??

- 山，川，湖など — 原則は前が地名,「山，川，湖」は普通名詞とする．ただし,「琵琶湖」,「高野山」のように，分割すると「琵琶」,「高野」が何だかよくわからなくなるような場合は「琵琶湖」,「高野山」を一語を地名として扱ってもよい．が，この区別はあまり厳密には行わないことにする．

  例) 三笠 (地名) 山 (普通)
  大阪 (地名) 湾 (普通)
  ひょうたん (普通) 山 (普通) ??
  琵琶湖 (地名)
  高野山 (地名)

- 組織名一般 — 多くの場合 2 語からなるが，分割してそれぞれに適当な品詞を与えることを原則とする．ただし，上の場合と同じく切ると変になるものは一語の組織名とするが，厳密な区別は行なわない．

例) 京都 (地名) 大学 (普通名詞)
山梨 (地名) 学院 (普通名詞)
慶応 (組織名) 大学 (普通名詞)

松下 (人名) 電機 (普通名詞)
竹中 (人名) 工務店 (普通名詞)
トヨタ (組織名) 自動車 (普通名詞)
豊田 (人名?) 商事 (普通名詞)
富士 (地名?) 重工業 (普通名詞)

自民党 (組織名)
運輸省 (組織名)

清水 (地名) 寺 (普通名詞)
今宮 (地名) 戎 (普通名詞) 神社 (普通名詞)
伊勢 (地名) 神宮 (普通名詞)
長岡 (地名) 天満宮 (普通名詞)
銀閣寺 (組織名)
金閣寺 (組織名)
法隆寺 (組織名)
東寺 (組織名)
東大寺 (組織名)
本能寺 (組織名)

(※ 現在「〜寺」は地名でしかはいっていないので，見直し・修正が必要)

3.3.3 その他の問題

1. 省略形の扱い

   一文字の国名のような省略形の固有名詞は，基本的に一文字で固有名詞として扱う．ただし,「訪日」のようにまとまりとして一般的になっているものは一語として扱う．

   例) 対 (普通) 日 (地名)
   日 (地名) 米 (地名)
   日 (地名) 独 (地名) 伊 (地名)
   訪日 (サ変名詞)

   また，省略の組合わせで，分割してもほとんど意味がないと考えられるものは全体として一語の固有名詞とする．

   例) 京産大 (京都産業大学) → 固有名詞
   日経 (日本経済新聞) → 固有名詞
   南ア (南アフリカ，南アメリカ?) → 地名
   セ大阪 (セレッソ大阪) → 固有名詞

2. 句の固有名詞の扱い

句の単位で固有名詞となっているものは，分割して個々の語に適当な品詞を与えることを原則とする．分割すると構成要素に固有名詞がなくなったり，複数の文節になる場合もあるが，仕方がないとする．

例) 鳥 (名詞) よ (助詞) はばたけ (動詞)
スターズ (名詞)・アンド (名詞)・ストライプス (名詞)

3. 読みの扱い

読みについては基本的に修正は行わないことにする．ただし，固有名詞関係などで，形態素区切り，品詞などの修正の「ついで」の場合には修正してもらってもよい．その際の基準は一応次のようにしておく．

- システムの与える読みが誤りであることは明らかであるが，正しい読みはわからないという場合には読みを？とする．
- 「ＮＨＫ」のように記号列の読みがその各記号の連続でしかない場合はそのままもとの記号 (ＮＨＫ) を入力する.「ＯＰＥＣ (おぺっく)」のように別の読みがある場合はその読みを入力する．

4　係り受けに関する基準

文節間の係り受け関係については次の 4 種類の区別を行う.

並列関係

「太郎と花子が」,「〜食べ，〜飲んだ.」などの関係.「太郎と」と「花子が」,「食べ,」と「飲んだ.」を P マークでリンクする (P の位置をマウス真中ボタンでクリック).

例)　チーズ n を p ┐
　　食べ v，*━P
　　ビール n を p ┫
　　飲んだ v. *

部分並列内の関係

並列する部分に述語がなく，係り先のない文節がある場合の扱い．次の文の場合,「本を」と「太郎に」,「ノートを」と「次郎に」を I マークでリンクする (I の位置をコントロールキーを押しながらマウス真中ボタンでクリック).

例)　本 n を p ― I
　　兄 n の p ┤
　　太郎 N に p，*━━P
　　ノート n を p ― I
　　弟 n の p ┫
　　三郎 N に p ┐
　　かして v いる s. *

同格関係

「指輪など，高級品を」,「会社員，太郎が」などの関係．それぞれを A マークでリンクする (A の位置をマウス右ボタンでクリック).

例)　泥棒 n は p ┌─────┐
　　指輪 n など p，*―A　|
　　多数 n の p ┤　|
　　高級 n 品 n を p ┤
　　盗んだ v. *

通常の係り受け関係

他の関係，すなわち,

- 「僕が書いた」(格要素と述語),
- 「僕の本が」(体言と体言),
- 「書いた本が」(連体修飾節の述語と被修飾語),
- 「書けば，売れる」(従属節の述語と主節の述語)

などは，それぞれの間にマークのない線を引く (線の折れ曲がる部分をマウス左ボタンでクリック).

以下に，問題となる係り受け関係について基準を示す.

4.1 通常の係り受け関係

基本的に，活用形や付属語の文法的働きに従った係り受け関係を考える．同じ述語に係っている格要素間には強い意味的関係が認められることもあるが，それらは構文的な関係ではないと考え，係り受け関係とはとらえない．たとえば「分割が大きな問題になっている」という文では「分割が」と「問題に」の間に意味的関係が認められるが，それらは「なっている」を通した関係ととらえる．

通常の係り受け関係に対するシステムの解析誤りは，多くの場合，特別な働きをする文節に対してその働きが (正確に) 記述されていないことに原因がある．たとえば「できる (動詞) だけ (助詞)」というのは用言の直前では副詞的に働き，この文節が格要素の係り先になることはない (例:「本をできるだけ読みなさい」)．また，たとえば「(〜を) めぐって」というのは助詞相当句として働き，これも直前のヲ格以外の格要素や連用形の用言の係り先となることは通常ない．このように文節に対する情報不足に原因すると考えられるような係り受け誤りがあった場合は**メモ欄に B** を入力しておく．

4.1.1 格要素と複数の述語の関係

文中のある格要素が意味的には複数の述語と関係を持つことは少なくない．そのような場合は以下の基準で係り先を決定する．

1. 並列する複数の述語との関係
   並列する述語の中の最後の述語に係るかたちで扱い，このかたちで並列する全ての述語と関係を持つことを表すことにする．

   例) 問題 n を p ┐
   先生 n が p ┐ │
   与え v，* ─ P
   生徒 n が p ┤
   解いた v．*

2. 主語と，従属節・主節の述語との関係
   主節の述語に係るかたちで扱う．(主節の主語は (文脈中に) 別に存在し，従属節の主語でしかない場合はもちろん従属節に係るとする．)

   例) 私 n が p ┐
   東京 N へ p ┐ │
   いく v ┐ │
   途中 n で p ┤
   考えた v．*

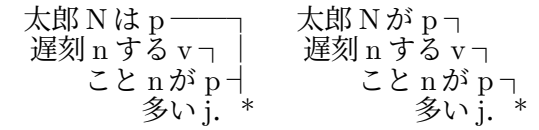
3. 主語以外の格要素と，従属節・主節の述語との関係
   従属節の述語に係るかたちで扱う．このような場合は偶然同一の表層格となっただけで，実際には主節の方の格要素が省略されているとみなす．

   例) 東京 N に p ┐
   引っ越した v が p，* ┐
   まだ a ┤
   馴染め v ない s．*

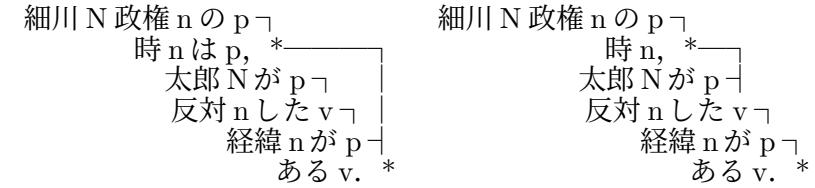
4. 文末に補助的述語がある場合
   システムの解析は，以下の原則にそっているが，それらが明らかに不自然な場合は修正する．なお，判断に迷う文の場合は**メモ欄に DA** と入力しておく．

- 「〜ことが多い」,「〜経緯がある」,「〜する予定」など
  原則としては,「〜は」(「〜では」,「〜には」などでなく単独の「〜は」) は外側 (文末), それ以外は内側 (埋め込み文の述語) に係るとみなす.

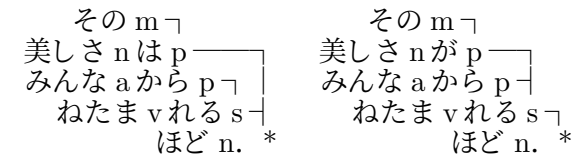
例)



例)

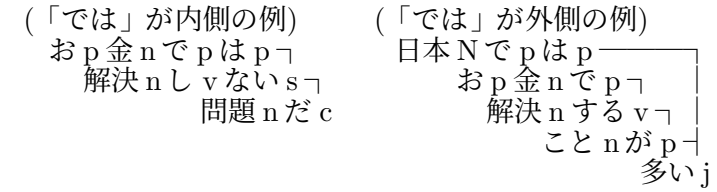


例)



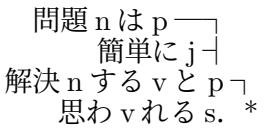
「〜では」,「〜には」などの判断は特に難しいが, どちらともいえない場合は内側とする.

例)



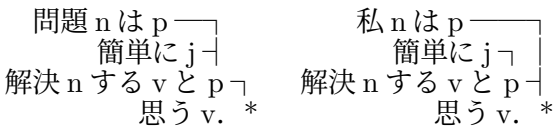
- 「〜は〜すると (ように) 思われる」(いわれる, される, みられる, 考えられる)
  これらは「書き手は『〜は〜する』と思う」の変形であると考え,「〜は」は内側の「〜する」に係ると考える.

例)



- 「〜は〜すると (ように) 思う」(思える, 見える, 聞く, いう, いえる, いってよい, いわざるをえない, など)
  これらは能動態でも「書き手は『〜は〜する』と思う」あるいは「噂では『〜は〜する』らしい」の意味と考え,「〜は」は内側の「〜する」に係ると考える. ただし「私 (ぼく, われわれ) は〜すると思う」の場合の「私は」の係り先は「思う」である. (もちろん, これらは原則であって, 文脈に応じて個別に判断する必要がある.)

例)



- 「〜しかない」,「〜ほかない」などは一文節とする方向で検討中.

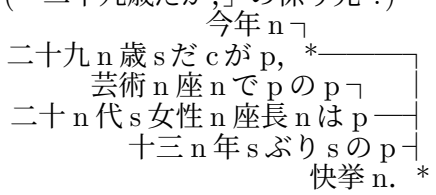
4.1.2 係り先が非常に曖昧な場合

以下のそれぞれの基準で対処する．

従属節，文頭の接続詞など

従属節 (「〜するが」)，文頭の接続詞 (「また」) などについて係り先が非常に曖昧な場合は，文末を係り先とすることを原則とする．

例) (「二十九歳だが,」の係り先 ?)



連体修飾の連続

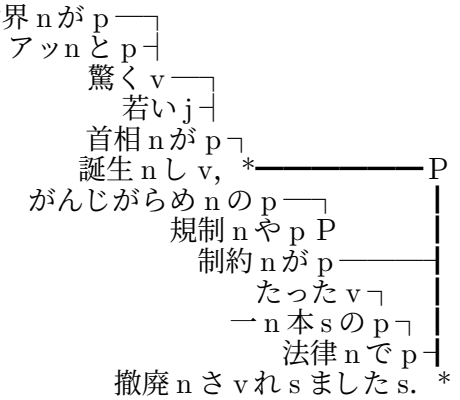
「A の B の C」,「A した B の C」などで係り先が非常に曖昧な場合は近い係り受けを優先することを原則とする．つまり「A の」が「B」に係る構造を優先する．なお，この場合は**メモ欄に DA** と入力しておく．(同格が関係する場合も同様，4.3.8 節参照)

4.2 並列

文中に並列構造がある場合は，並列する部分の主辞間を P マークでリンクする．3 つ以上の部分が並列する場合には，各部分の主辞からその右隣の並列部分の主辞にリンクをはる．

例) 走ったり v ── P
激しい j
運動 n を p
したり v する s
こと n が p
難しく j なった s. *

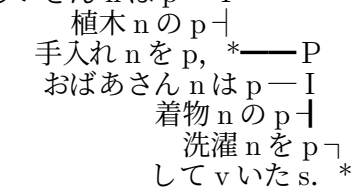
例) 走ったり v P
投げたり v P
飛んだり v と p
いう v
激しい j
運動 n は p
難しく j なった s. *

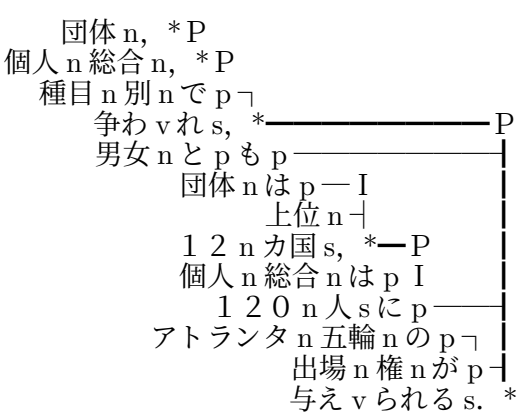
例)


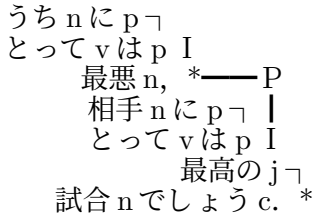
以下，並列の扱いの複雑な問題，特別の問題について基準を示す.

4.2.1 部分並列

「おじいさんは植木の手入れを，おばあさんは着物の洗濯を していた.」という文では下線部分が並列しているが，この各部分は一つの係り受け構造にまとめることができない．このような場合には，係り先のない文節を各並列部分の主辞に I というマークでリンクする．このような文節は，最終的には最後の並列部分の主辞の係り先に係るとみなす (上の例の場合「していた.」).

例)


例)


例)


(この場合，名詞とナノ形容詞の並列となってしまうので**メモ欄に PA** と記入; 3.1.5 節)

4.2.2 括弧内の複数文

括弧の内部に句点によって区切られている文が複数ある場合は，各文の末尾の述語を並列関係として扱う．

例) 太郎 N は p ┐
「*少し a ┐ │
疲れた v．*──P │
明日 n ─┤ │
また a ┤ │
来る v」*と p ┐ │
言って v ┤
帰った v．*

現システムの解析結果もこの基準に従っているが，誤りがある場合は修正する．

4.2.3 テ形

「用いて」,「読んで」のようなテ形の連用形については，並列とみなすか連用修飾とみなすかが難しい場合が少なくない．以下の基準で扱う．

1. 単なる修飾の場合 → 並列としない．

   例) 魚介類 n を p ┐
   用いて v ┐
   作った v．*

2. (弱い) 時間経過の場合 → 並列としない．

   例) 扉 n を p ┐
   開けて v ┐
   入った v．*

3. 非常に類似性が高く，等位接続とみなせる場合 → 並列とする．

   例) 本 n を p ┐
   読んで v，*─P
   レコード n も p ┤
   聞いた v．*

   例) 今日 n ─┐
   東京 N に p ┤
   いって v，*──P
   明日 n ─┤
   京都 N に p ┤
   帰る v．*

現システムでは，テ形の前後の文節列が非常に類似する場合は並列関係，そうでなければ通常の係り受け関係 (後ろの適当な用言を修飾) として扱っている．この扱いで上記の基準と矛盾する解析結果となっていれば手修正する．

4.2.4 連体形

連体形が続く場合は並列とはせず，それぞれが体言に係っているとみなす.

例) 緊迫 n した v ─┐
重大な j ┤
状況 n. *

4.2.5 「〜から〜まで」

「〜から〜まで」という表現は次のように扱う.

1. 「AからBまで扱う」の場合,「Aから」と「Bまで」を並列,「Bまで」と「扱う」を通常の係り受けとして扱う.

   例) ミサイル n から p P
   ゆりかご n まで p ┐
   扱う v. *

2. 「AからBまで多くのものを扱う」の場合,「Aから」と「Bまで」を並列,「Bまで」と「ものを」を同格として扱う.

   例) ミサイル n から p P
   ゆりかご n まで p ─A
   多く a の p ┤
   商品 n を p ┐
   扱う v. *

3. 「AからBまで行く」の場合,「Aから」,「Bまで」がともに「行く」に通常の関係で係ると扱う.

   例) 東京 N から p ─┐
   大阪 N まで p ┤
   行く v. *

現システムでは,「〜から」と「〜まで」が類似する文節列の場合は並列関係，そうでなければ通常の係り受け関係 (「〜から」も「〜まで」も後ろの適当な用言を修飾) として扱っている．この扱いで上記の基準と矛盾する解析結果となっていれば手修正する.

4.3 同格

同格関係にある 2 つの文節は A マークでリンクする．同格については，システムの解析は全く不十分なので，手作業が中心となる．また，表現のバリエーションについても調査が十分ではなく，これまでに目についた表現だけを扱っている状況である．以下に示す表現以外に同格表現と考えられるものがあれば，**メモ欄に AA** と記入し，随時指摘して下さい.

4.3.1 住所，職業，続柄と人名

新聞記事によく見られる「住所，職業，人名」「住所の職業，人名」などの表現では，住所は人名に係り，職業は人名と同格として扱う．また，続柄 (「父」,「母」,「兄」,「長男」など) と人名も同格とみなす.

例) 住所 n 不定 n，*┐
無職 n，*A
山田 N 太郎 N．*

例) 東京 N 杉並区 N の p ┐
会社員 n，*A
山田 N 太郎 N．*

例) 容疑者 n の p ┐
長女 n，*A
花子 N．*

4.3.2 「〜から〜まで」

次のような文では「ゆりかごまで」と「商品を」を同格とみなす．この同格表現はシステムではまったく解析できないので，すべて手作業による修正となる (4.2.5 節).

例) ミサイル n から p P
ゆりかご n まで p ―A
多く a の p ┤
商品 n を p ┐
扱う v．*

4.3.3 「体言＋ら | たち | その他 | など | と | すなわち | つまり | とりわけ | 特に」

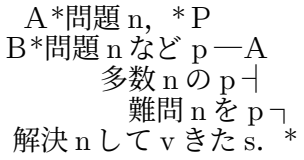
これらの表現は後の適当な体言と同格とみなす．ただし

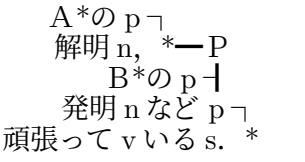
- 同格関係をみとめられるような体言がない場合は，適当な用言に通常の関係で係るとして扱う.
- 「この問題など簡単だ」のような場合は「など」は取り立てを示しているだけなので，通常の関係で「簡単だ」に係るとする.
- 「A と B」の「と」は並列である.
- 「とりわけ」,「特に」は厳密には同格ではないが，当面同格として扱う.
- 「すなわち」,「つまり」,「とりわけ」,「特に」は (読点がなくても) 別文節,「その他」,「など」,「と」は (読点があっても) 一文節として扱う.

例) プログラマー n ら s A
四十 n 人 s の p ┐
ストライキ n．*

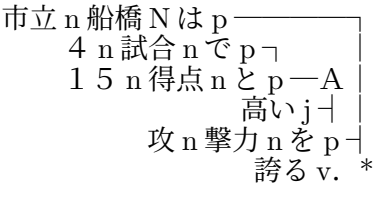
例) 行政 n 改革 n，*――――A
とりわけ a ――――┤
規制 n 緩和 n，*―P │
特殊 n 法人 n の p ┤ │
見直し n，*P
地方 n 分権 n など p ―A
大きな m ┤
課題 n が p ┐
ある v．*

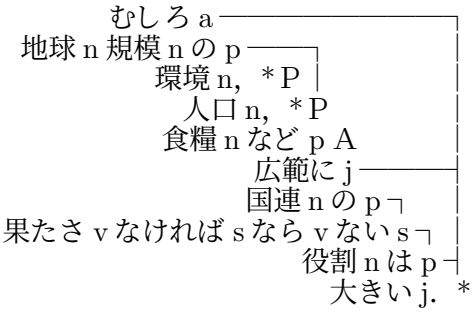
(後に体言の並列がある場合は全体と同格とみなす)

例)


例)


(同格関係の体言がない例)

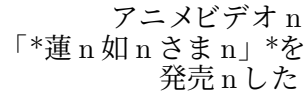
例)


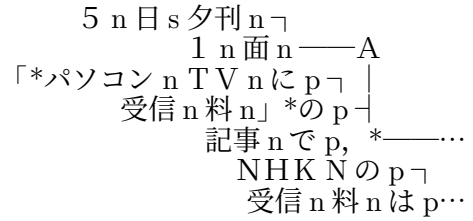
例)


(「広範に」を名詞＋助詞と見なすべきかもしれないが (3.1.5 節)，現在は「広範に」のまま放置)

4.3.4 「体言 (＋，) ＋「〜」」

体言 (＋，) が直接，鍵括弧に続く場合は同格の可能性を考慮する．同格関係の相手先は鍵括弧の最後の文節である場合と，そうでない場合がありえる (鍵括弧内の途中の文節ということはないはず)．

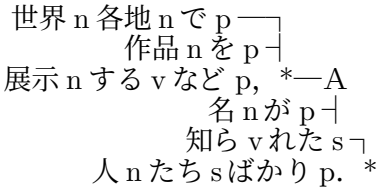
例)


例)


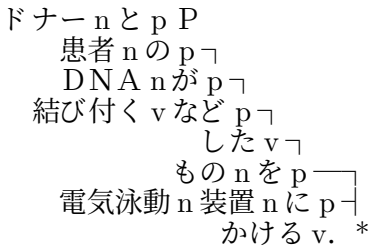
(この例は厳密な同格とはいえないが，当面「同格」を広く解釈することにする)

4.3.5 「用言＋など」

「用言＋など」はその後ろの用言と同格であるとみなす.

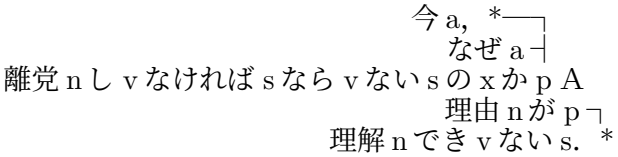
例)


ただし,「〜など＋する」の場合は通常の係り受けとする.

例)


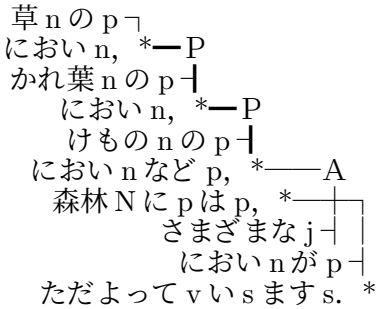
4.3.6 節とそれをまとめる名詞

次のような場合の節と「理由」の関係も同格とみなす.

例)


4.3.7 同格と非交差条件の問題

同格関係については係り受けの非交差条件を破るものも認めることにする．ただし，その場合は**メモ欄に DT** と記入する.

例)


4.3.8 同格に関連する係り受けの曖昧性

「A の B(同格表現)C」で「A の」の係り先が曖昧な場合，一般の名詞句「A の B の C」の場合と同様に基本的には「B」に係る構造を優先する.

ただし，一般の名詞句の場合よりも「C」が係り先であると感じられる場合が多く，最終的判断は作業者にまかせる．

近くに係る例 (優先される構造)

米国内の 電波望遠鏡 十基
村山内閣の 防衛費 〇・八五五%増
米国第二位の ビールメーカー 「〜」社
未完の オペラ 「〜」
自民党の 総裁 「〜」
アルバムの中の 一曲 「〜」
香港の 中立系紙 「〜」
遊眠社の 若手人気俳優、 羽場裕一ら
健三郎さんの 長男 光さん

遠くに係る例

子供 二人の 家族 四人
自民、社会、さきがけの 与党 三党
同社の 創立 ３０周年記念
一日付の 機関紙 「赤旗」
俳優座劇場で 公演の 俳優座 「〜」
同乗の 妻 晴子さん

5　メモ記号一覧

各文のメモ欄には必要に応じて以下のメモ記号を入力する (複数ある場合はスペースで区切る).

?	保留／相談を要する文
M	連語の候補を含む文
B	特別な文節区切り，特別な働きの文節 (複合辞，弱い用言など) を含む文
DA	通常の係り受けで判断の迷う文
DT	非交差条件を破るような係り受けがある文
PA	並列に関して判断の迷う文
AA	同格に関して判断の迷う文
G	放棄 (文語体，アンケート結果など，コーパスとして意味のないもの)

6　その他

- 丸括弧の扱い
  丸括弧は任意の場所に挿入され，形態素・構文解析での対処が非常に困難である．逆に，文から丸括弧を削除しても，もとの文の形態素・構文構造はほとんどの場合変化しない．これらのことから本プロジェクトでは丸括弧はあらかじめテキストから取り除いて処理することとした.

  例)　解（わか）る
  　　‥‥（社会面に関連記事）

- 品詞の判定ができないもの
  単語として認められず，品詞を与えるとこができない場合は，品詞メニューで「未定義語:その他」を選ぶ.

  例)　そ，そんなあ… (「そ」→ 未定義語:その他)
  　　“ 親切 ”な… (「な」→ 未定義語:その他)

  上の 2 例はいずれも一文節として扱う.「未定義語:その他」のバリエーション，扱いについては，コーパスでの出現のつど検討する.