『日本語話し言葉コーパス』における係り受け構造付与

(Version 1.0)

内元清貴[1]　丸山岳彦[2]　高梨克也[1]　井佐原均[1]

[1]情報通信研究機構

[2]ＡＴＲ音声言語コミュニケーション研究所／国立国語研究所

【目次】

1. はじめに

日本語文の意味を理解するためには、文節間の依存関係を特定することが最も必要となるが、語順が比較的自由であるため難しい場合が多い。したがって、日本語の処理においては、この問題に焦点を絞り、統語構造として文節間係り受け構造を採用することが多い。コーパスなど言語資源の作成においても同様である。例えば、京大コーパス[1]には、形態素や係り受け構造の情報が付与されている。このコーパスは我々が入手できる代表的な書き言葉のタグ付きコーパスのひとつで、これまで、機械翻訳、情報抽出、要約、質問応答など様々な処理のために利用されてきた。

開放的融合研究「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトでは主に講演などのモノローグを対象とした自発的な話し言葉の大規模コーパス、『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese (CSJ))を作成した[2]。また、このコーパスを利用した話し言葉の自動要約プロトタイプシステムの作成も行なった[3]。書き言葉だけでなく話し言葉の要約にも統語構造の情報が必要であることが指摘されており[4]、このプロジェクトにおいてもコーパスに統語構造の情報を付与することにした。コーパスの対象は日本語であるため、上述の理由で、京大コーパスのような書き言葉のコーパスと同様、統語構造として文節間係り受け構造を採用した。

2. 係り受け構造

CSJ における文節間係り受けは原則として京大コーパスの基準に準拠するものとする。しかし、書き言葉と話し言葉では現象が異なることが多く、この基準だけではすべてを網羅することはできない。したがって、話し言葉特有の現象に対しては新たな基準を設ける。

自発的な話し言葉特有の現象として、まず、文は必ずしも自明な単位ではない、ということがあげられる。CSJ において、いわゆる文を単位とすることには次のような問題点がある。(文献[5]より)

- 書き言葉では書き手自身が句点によって区切りを確定するのに対して、 話し言葉にはこうした情報がない。
- 独話の特徴は一人の話者が続けて話し続けることであるが、文法的に明確な文末形式が頻繁に用いられるとは限らないため、極端に長い文が生じてしまう場合がある。
- 自発的な話し言葉では、言い直し、言い換え、言い差し(言いやめ)などの要因により文の範囲が確定しにくい場合や語や文の断片だけで発話が構成される場合がある。

そこで、我々はいわゆる文に代わる単位として「節」に基づく単位を採用することにした[5]。係り受け構造の情報はこの単位内で付与する。以降で、この単位を便宜上、節単位と呼び、この単位境界の認定を節境界認定と呼ぶ。

自発的な話し言葉特有の現象として、次に、言い差し(言いやめ)、言い直し、言い換え、挿入構造、倒置、ねじれなどの非流暢現象があげられる。これらは話し言葉の次のような特徴により生じる現象である。(文献[6]より)

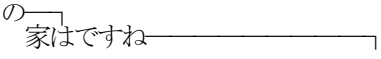
- 話し言葉は、発話という一回的な行動の中で実時間的に組み立てられていく言語様式である点で、書き言葉とは異なる。発話の産出は時系列に沿って線条的に行なわれるため、何らかの理由によって発話の形式が途中でくずれたり、発話内容の間違いに気づいて発話し直したりして、当初想定されていた発話の形式や内容が変容を受けることがある。

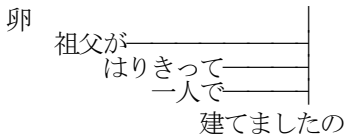
これらの非流暢現象に対しては、基本的に次のように対処している。

- 言い差し(言いやめ)

基本的に節境界認定の作業により別の節として切り出されるが、言い差し部分を越えて係り受けがある場合などは切り出されないことがある。この場合は、言い差しについては係り先なしとする。

例:「卵」が言い差し

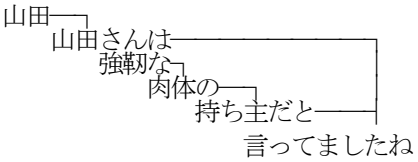




- 言い直し、言い換え

節単位内の言い直し、言い換えは新たに基準を設けて対応する。言い直しや言い換えにも様々な種類のものが考えられるが、CSJ における係り受け構造においては、詳細な種類の分類は行なわず、言い直し、言い換えに関係する範囲を特定することに主眼を置く。

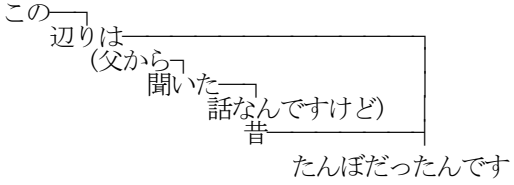
例：「山田」が「山田さん」に言い直されている



- 挿入構造

係り受けは挿入構造内で閉じるものとする。挿入構造は節境界認定作業により特定する。

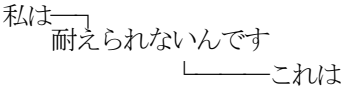
例：「父から聞いた話なんですけど」が挿入節



- 倒置

右から左への係り受けとする。

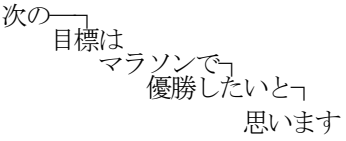
例：「これは」が倒置



- ねじれ

発話プランの変更により、不自然な統語構造となる場合が多いので、基本的に係り先はないものとする。話題導入表現の直後など大きい切れ目においては、節境界認定作業により節境界が認定され、別の単位となっている場合もある。

例：「目標は」の係り先が不自然



本来、係り受け関係と言い直し関係は性質が異なるため、別の作業とするべきであるが、効率上同時に作業している。また、書き言葉における並列、同格も係り受け関係とは異なるが、これらのタグ付け結果から係り受け関係も導出可能であることから、効率上同時に作業している。

係り受け構造を付与する対象は、コア中の独話 177 講演（対話・再朗読以外）とコア以外のテストセット22講演を含む199講演（talk_data.csv参照）である。前節でも述べたように、係り受け構造は節単位内の文節を単位として付与し、基本的に節境界を越える係り受けは付与していない。文節の定義については国立国語研究所で定められた定義[7]に従う。

係り受け構造付与は、一講演あたり、二人の一次作業者と、一人の二次作業者により行なった。一次作業者は基準に従って各文節について係り先を決定し、二次作業者は一次作業者の付与結果を比較しながら最終判定を行なった。テキストだけでなく、音声の情報も参照しながら係り先を決定した。

3. ＣＳＪにおける係り受け構造と係り受け構造付与の基準

基本的に、文節境界に関しては「文節の仕様について」(以下で文節基準と呼ぶ)[7]、節境界に関しては「『日本語話し言葉コーパスＣＳＪ』における節境界認定 」(以下で節境界認定基準と呼ぶ)[8]に従う。係り受けに関しては書き言葉における係り受けと同様の場合には京都大学テキストコーパスの「コーパス作成の作業基準」(以下で京大コーパス基準と呼ぶ)に従い、話し言葉特有の係り受けについては本稿の基準に従うものとする。ただし、本稿に記載の事項と内容が矛盾する場合は本稿の定義を優先する。以下で、CSJ に付与されているラベルやコメント、付与の基準について述べる。

【1】各ラベルやコメントについて

ここでは係り受け関係を示すラベルやコメントについて示す。それぞれ使用にあたっての詳細は後述する。

係り受け関係を示すラベル	P	並列関係
	I	部分並列
	A	同格関係
	A2	具体例と総称の同格関係・具体例と数詞の同格関係
	D	言い直し
	R	倒置
文節境界に関するラベル	B+	後続文節と接続すべき文節
その他	F	フィラー
	C	接続詞
	E	感動詞
	Y	呼び掛け
	N	係り先のない文節
	X	アークのクロス
	K	古文
	S	その他のコメント

1-1　係り受け関係を示すラベル

通常の依存構造にない関係には以下のラベルを付与してアークを結んでいる。

P：並列関係

例)　赤ん坊は―――――――――――――――――+
泣いて―P―+　　|
笑って―P―+　　|
食べて-P+　　|
寝て――|
暮らす

I：部分並列

例)　今日の―+
大会では――I―+
兵庫県が―――――P―+
昨年の―+　　|
試合では――I-|
大阪府が―――――+

開会宣言を――|
それぞれ―|
読み上げた

A: 同格関係

例)　米国大統領――A-+
ジョン・F・ケネディが―+
暗殺された

A2:具体例と総称の同格関係・具体例と数詞の同格関係

例)　みかんとか…P-+
りんごとか……A2-+
そういうものを―+
食べた

例)　英語を――――+
母国語と―|
する―+
話者―A2-+
二名が―+
参加した

D: 言い直し

例)　解かい―D+
解決の-+
糸口が――――+
見付かったよ

R: 倒置

倒置の係り受けは左側に下から上への係り受けとして示される。
しかし、倒置部分に助詞が付いている場合は、普通に係り受けを結んで『R』ラベルを付与している。

例)　{何だろう―R-+
これは}と――――+
思って

1-2　文節境界に関するラベル

転記テキストでは基本的に一行に一文節が記されるが、一文節中に転記基本単位の境界が存在する場合、一つの文節が複数行に跨がって記される（[7]参照)。このような場合には、次のラベルを付与している (p.7【2】参照)。

B+: 後続文節と接続すべき文節

例)　必要な―+
『B+』書類-+
が――――+
まだ――|
来ない

1-3　節境界に関するラベル

節境界に関して、節境界検出プログラムＣＢＡＰ[6]により付与されている情報は以下の４９種類である。これらの情報は表層的な情報をもとに自動挿入されている。

［文末］［文末候補］［と文末］ /並列節ガ/ /並列節ケド/ /並列節ケドモ/ /並列節ケレド/ /並列節ケ

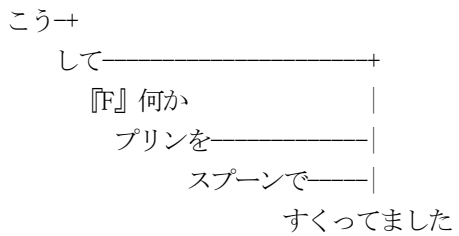
レドモ/ /並列節シ/ 〈タリ節〉 〈タリ節-助詞〉 〈テカラ節〉 〈テカラ節-助詞〉 〈テハ節〉 〈テモ節〉 〈テ節〉 〈テ節-助詞〉 〈トイウ節〉 〈トカ節〉 〈トカ節-助詞〉 〈ノニ節〉 〈フィラー文〉 〈ヨウニ節〉 〈引用節〉 〈引用節-助詞〉 〈引用節トノ〉 〈感動詞〉 〈間接疑問節〉 〈間接疑問節-助詞〉 〈条件節タラ〉 〈条件節タラバ〉 〈条件節ト〉 〈条件節ナラ〉 〈条件節ナラバ〉 〈条件節レバ〉 〈接続詞〉 〈接続詞C〉 〈接続詞CL〉 〈接続詞L〉 〈接続詞M〉 〈並列節ダノ〉 〈並列節デ〉 〈並列節ナリ〉 〈理由節カラ〉 〈理由節カラ-助詞〉 〈理由節カラニハ〉 〈理由節ノデ〉 〈連体節テノ〉 〈連用節〉

1-4 その他のラベル

F: フィラーに付与

＊既に節境界の情報として〈フィラー文〉などの情報が与えられている場合を除く。

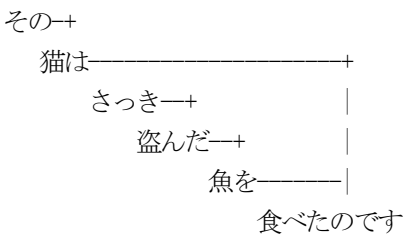
例) 『F』何か



C: 接続詞に付与

＊既に節境界の情報として〈接続詞〉の情報が与えられている場合を除く。

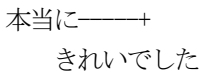
例) 『C』そして



E:感動詞に付与

＊既に節境界の情報として〈感動詞〉の情報が与えられている場合を除く。

例) 『E』もう



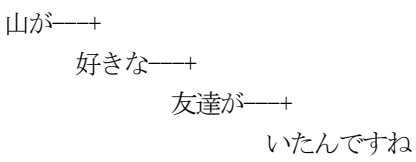
Y:呼び掛けに付与

例) 『Y』菅原君

期待してるからね

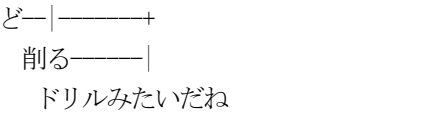
N : 係り先が消失している場合に付与

例) 『N』中学校を



X: 係り受けのアークがクロスする場合に付与

例) 地面を



K:古文に付与

K:S1（古文の開始点）

K:E1（古文の終了点）

S: その他のコメント

S:格表示誤り「修正候補の助詞」（助詞の言い誤りがあったとき）

S:複数文節言い直し（ラベルＤを付与する文節が１対多対応になるとき）

S:複数文節言い直し:S1（ラベルＤを付与する文節が多対１／多対多対応になるとき）

S:複数文節言い直し:E1（ラベルＤを付与する文節が多対１／多対多対応になるとき）

1-5　複数ラベルの併記

付与するラベルが複数個になる場合は、ラベルとラベルの間に半角アンダーバー「_」を挿入している。

```
例）　『C_B+』て─+
          『C』言うか
             これが─+
                原因で──────────────────────+
                   会社の──+                   |
                       机に────+               |
                         しまっておいた──+     |
                                   書類が────|
                                        見つかったの
```

【２】文節境界について

2-1　基準

係り受け作業では、文節基準（[7]参照）をもとに分割された文節を基本単位として利用する。同じ基準で認定された文節が、転記テキストにおいて一文節一行という形式で既に記されているため、本作業ではこの一行単位を文節として利用する。ただし、転記テキストにおいて、一文節中に転記基本単位の境界（基本的に 200ms 以上のポーズがある場合）が存在する場合、一つの文節が複数行に跨がって記されることになる（[7]参照）。このような場合、『B+』を付与している。

```
例）          授業───+
                 受ける──+
              『B+』態度に関して────+
                                 は─────────────+
                                  向こうが───────|
                                     やっぱり────|
                                           積極的です
```

フィラー(F)や言い淀み(D)が独立した文節となる場合は、係り受け関係は付与しない.

```
例）　(F えー)
          (F えー)
             (D い)
               以上で──────────────────────+
                  (F あー)                 |
                     留学の──+             |
                        頃の────+          |
                          (D お)  |        |
```

お話しを――|
終わります

2-2　転記基本単位の境界で分割された１文節の復元

本来１文節であるべき文節が転記基本単位の境界（基本的に200ms以上のポーズがある場合）によって分割されている場合、それらを結び、ラベル『B+』を付与している。この場合、主辞の文節を決め、係り文節は全て主辞に係るとする。（主辞については後述）

例)　ディズニーランドには―+
いつか―――|
『B+』行き―――+
たいと―――+
思っていましたが

2-3　言い直しが絡む場合の文節

言い直しが絡む場合は、次の基準でラベル『B+』を付与している。

1. 『B+』でつなぐことにより文節基準を満たす。
2. 『B+』でつなぐことによりその中に文節の入れ子構造ができない。

○入れ子にならない例
『B+』海岸――――+
を―D―|
に―+
近い

×入れ子になる例
海岸―――――――――――+
『S:複数文節言い直し』近い―D-+
に―|
近い

「入れ子にならない例」では「海岸」と問題部：「を」／「海岸」と訂正部：「に」の２組が『B+』関係にある。「海岸」を訂正部に『B+』で結んでもその間には文節未満の「を」があるだけなので「入れ子構造」にはならない。一方、「入れ子になる例」では、『B+』関係にある「海岸」「に」の間に「近い」という文節の入れ子構造ができている。この場合、『B+』は付与しない。

【３】係り受けについて

文節間の係り受けについては原則として京大コーパス基準に準拠する。以下では、主に話し言葉特有の現象について記述する。

3-1　他の要素への係り受けを結ばないもの

「フィラー」「接続詞」「感動詞」「非言語音」は基本的に他の要素への係り受け関係を結ばない。

3-1-1　フィラー

転記タグ(F　)が付与されているもの、あるいは節境界認定においてコメント「フィラー文」が付与されているものがある。それ以外に「フィラー」と認められるものには『F』を付与している。

例)　(F えーと)
乗り物は――――――――+
ほとんど――――|
(F あの)　|
車ですね

ただし転記タグ(F)は、フィラーの他に感情表出系感動詞にも付与されている。後者が引用等の形式で出現する場合、文の要素として無視できないことがある。その場合に限り、係り受けを結ぶ。

例)　娘が―――――――――――――――――――――――+
　　『B+』(F きゃー)―+　　|
　　って―――――――――|
　　悲鳴を―――|
　　上げたんです

3-1-2　接続詞

自動節ラベル付与によりコメント〈接続詞〉が付与されているものがある。それ以外に、「接続詞」と認められるものには『C』を付与している。

例)　んで〈接続詞〉
　　私が―――――――――――+
　　父と―+　　|
　　一緒に―――|
　　イランに―|
　　行く―+
　　ことに―――+
　　なりました

例)　パラ言語情報―――P―+
　　『C』それから　|
　　非言語情報が―――|
　　伝わっております

3-1-3　感動詞

自動節ラベル付与によりコメント〈感動詞〉が付与されているものがある。それ以外に、「感動詞」と認められるものには『E』を付与している。

例)　うん〈感動詞〉
　　みんなが―――+
　　付いていけない

例)　『E』　そう
　　『E』　そう
　　絶対にさ―――+
　　いいねとか―――|
　　言ってくれなくて

3-1-4　非言語音

〈雑音〉〈笑〉〈ベル〉などの非言語音は係り受けを結ばず、コメントも付与しない。

例)　〈雑音〉
　　鮫と―+
　　泳ぐので―+
　　〈笑〉　　|
　　怖かったですけども

3-2　ＣＳＪに特有の現象

3-2-1　係り先がないもの

話し言葉特有の「係り先が存在しない例」として、文の途中で発話を中止しているもの、発話のプランを変更しているものなどが挙げられる。このような場合、どの部分にも係り受け関係を付与せず、『N』を付与している。以下に具体例を示す。

a. 発話が途中で中断されているために係り先を消失している場合

例)　では
『N』結論か
(F あ)
その―+
前に――――――+
資料について――――――|
ご説明しましょう

b. 途中で発話のプランを変更してしまっているために係り先を消失している場合

プラン変更後の文脈を優先する。

例)　将来の―+
夢は『N』
作家に―――+
なって―――+
頑張ります

例)　一番―+
最近―+
『N』できたのが
センター北っていう―+
町に―+
『N』できた
Aっていう―――+
大きな―――|
デパートが―――+
できたんですね

c. 照応の関係にあって係り受け関係を結べない場合

例)　金曜日の―+
夜の―+
終電―+
中央林間行きとか―+
乗っちゃうと―+
凄い―|
大変『N』
それぐらい―+
たくさんの―+
人が

3-2-2　挿入

既に挿入(「挿入節」「挿入文」)と認定されている部分は(　)で囲まれている(「節境界認定基準」[8]参照)。

挿入内部から外部への関係は削除するが、内部での係り受け関係は結ぶ。

例)　この―+
　　辺りは――――――――――――――――――――+
　　(父から-+　　|
　　聞いた―――+　　|
　　話なんですけど) +　　|
　　昔―――|
　　たんぼだったんです

3-2-3　引用節構造と連体節構造

〈引用節〉や〈トイウ節〉などの引用節の内部や連体修飾節の内部に絶対境界か強境界(「節境界認定基準」[8]参照)が含まれている部分はそれぞれ引用節構造、連体節構造とされ、データ上ではこれらの部分は{　}で囲まれ、その内部の境界は:で示されている。係り受け構造は{　}で囲まれた部分の内部についてのみ付与されている。

1.「:」で区切られた文間の関係　→　係り受け関係を結ばない
2.「:」で区切られた文内部の関係　→　係り受け関係を結ぶ

例)　{この世には――――――+
　　変わらないものが――|
　　二つ―|
　　ある[文末]:
　　それは――+
　　『B+』自然である}――――+
　　というような〈トイウ節〉―+
　　ことが――+
　　言われます

3-2-4　助詞などの省略

助詞などが省略されている場合は省略されている語を補って判断している。

例)　長調-P-+
　　短調の―+
　　違いは――――+
　　誰でも――|
　　わかります

3-2-5　倒置

倒置は基本的に左係りの係り受けとしている。倒置のうち、節境界に関わる倒置には既に<<　>>が付与されている場合がある。

例)　私は――――+
　　耐えられないんです
　　+――<<これは>>

例)　君か
　　+―<<やはり>>

例)　『F』{何か
　　『N』あんまり
　　勘違いしてるな――――――――+

+――――――みんな　　|
って

倒置されている文節に、引用の「と」など不要な助詞が付いている場合は右係りの係り受けとし『R』ラベルを付与している。

例)　{何だろう――R-+
これは}と――――+
思って

3-2-6　係り受けにおける主辞

ポーズの影響で本来１文節であるべきものが分割されている場合がある。このような場合には『B+』を付与して本来の文節を復元し、「この文節に係る文節」がある場合は「主辞」に結んでいる。また、この文節から他の文節に係るときは文節を構成する要素のうち最後尾のものから結んでいる。「主辞」は基本的に、文節を構成する要素のうち用言あるいは体言を含むものとする。

例)　きれいな――+
『B+』花――+
が――+
咲いている

例)　彼は――――――――――+
よく――――――|
『B+』勉強――|
した

3-2-7　比較

「ＡよりもＢのほうが～だ」という形の文においては「Ａよりも」と「Ｂのほうが」を並列にするのではなく、「Ａよりも」「Ｂのほうが」それぞれが述部に係るものとしている。

例)　音楽が――+
流れている―+
時の―+
方が――――――――――――――――――+
静かな―+　　|
時よりも――――――――――|
疲れが――――|
癒えるようです

3-2-8　体言止め

「体言＋する」における「する」や、「体言＋判定詞」での「だ」「です」などが省略されている場合は、その体言を述語として係り受け関係を結んでいる。

例)　そして
こちら側が――――+
ホルマント空間

例)　長男は―+
勉強――――――――――P-+
やんちゃな-+　　|
次男は――|
運動して――――+

日曜を―|
過ごした

3-2-9　格表示誤り

そのままでは主格が目的格になってしまうなどといった格要素の変更を伴うような誤りについては『S:格表示誤り（修正候補の助詞）』を付与している。ただし、単なる言い誤りやある話者特有の「くせ」による言い回しによって起こる多少不自然と思われるような格助詞の誤りは対象外とする。

＊前後関係から明らかに「私がしかった」のではなく「私を兄がしかった」ことが分かる場合。

例）　兄は―――――――+
『S:格表示誤り「を」』　私が―――|
しかった

3-2-10 並列

複数個の文節が対等な関係にあり、それらが異なる対象を示している場合、それらは並列関係にある。並列関係にある二つの文節は『P』ラベルでリンクしている。

例）　太郎と―P
花子という―+
名は

並列関係の認定方法を「文節士の並列関係」「節間の並列関係」「決まり文句的な言い回しにおける並列関係」について示す。

A. 文節の並列

基本的に助詞を伴った文節同士の関係は並列としない。ただし、接続助詞「と」は並列関係とする。用言は、意味的には並列であっても、連用形の並列要素のみを並列関係とし、連体形は並列関係に含めない。

例）　りんご―P―+
みかんと――P+
『C』それから　|
ぶどう―P+
なしが―+
大好きだ

例）　肉を―――――――――――+
豆腐と―P―+　　|
野菜を―――|
一緒に―|
いためて

例）　明るい―――――――――+
きれいで――P―+　　|
静かな―|
町であるという―――+
ことが

例）　部屋の―+
掃除や―――――+
『C』それから　|
ショッピングしたり

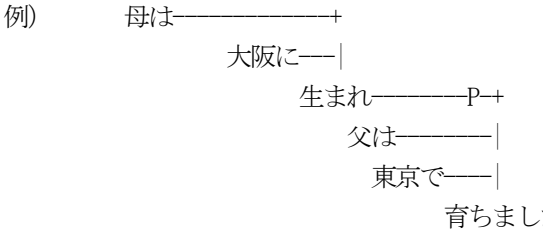
B.　節の並列

節同士の関係においては原則として以下の条件により並列を認定する。

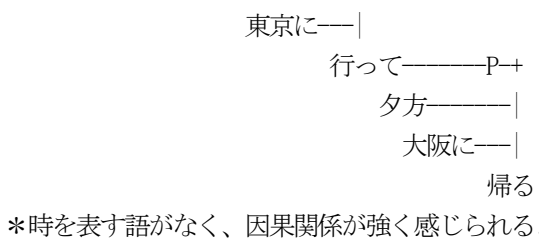
・表層的に and/or の関係である
・表層的な形の類似性が高い
したがって、意味的な並列関係が感じられても表層的な形が大きく違えば並列としない。ただし、「～たり」「～とか」「～やら」「～だの」など、並列関係を表す語があれば、形に類似性がなくても並列とする。
以下、具体例をいくつか示す。

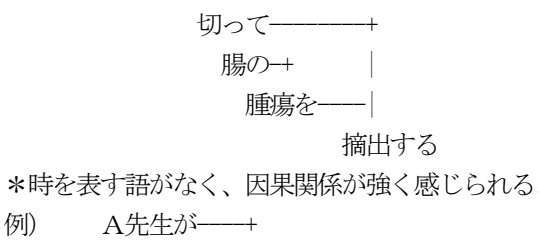
a. 時間経過が感じられず因果関係もない場合で、表層的な形が類似している場合

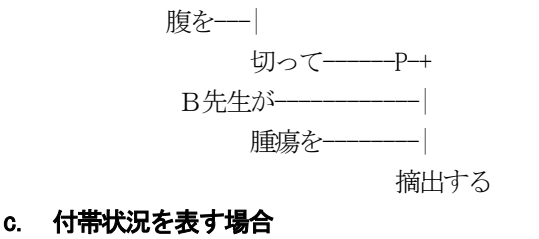
並列関係とする。

例）


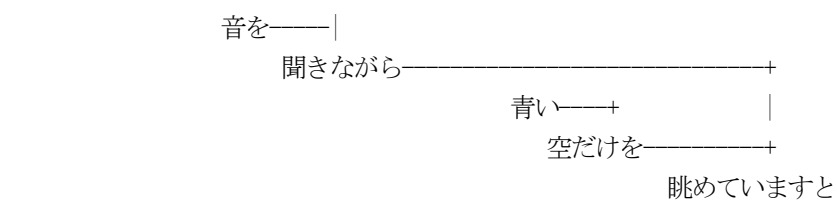
b. 時間経過、因果関係があっても、表層的な形の類似性が高い場合

前後を入れ替えることが可能であれば並列関係とする。
＊時を表す語があれば並列関係をとれる場合が多い。
例）


＊時を表す語がなく、因果関係が強く感じられるような場合は並列関係としない。
例）


＊時を表す語がなく、因果関係が強く感じられるような場合でも、並列関係とする場合もある。
例）


c. 付帯状況を表す場合

「～ながら」がついて付帯状況を表す場合は並列関係としない。
例）


d. ～たり～たり、～か～か

「～たり～たり」、「～か～か」という並列関係を強く示す語句がある場合には、表層的な形の類似をみなく

てもよいものとする。

例)　　凧を―+
　　あげたり――――P―+
　　庭で――――|
　　みんなと――|
　　食事したり――+
　　しました。

例)　　食べ物なのか―P―+
　　食べ物とは――|
　　全く――――|
　　違うものなのか――+
　　知りませんが

C.　決まり文句的な言い回し

「～かどうか」「～か何か」などの言い回しも並列関係とする。

a.　～かどうか

例)　暖かい―+
　　土地の+
　　人―+
　　特有なのか―P+
　　どうか――+
　　分かりません

b.　～か何か

例)　　知恵熱だか―P+
　　何か――+
　　知らないんですけど

3-2-11　同格

複数個の文節が対等な関係にあり、それらが同一の対象を示している場合、それらは同格関係にある。同格関係にある二つの文節は『A』または『A2』ラベルでリンクしている。同格関係のうち、「他にも同格の要素がありそうなとき」「他にも同格の要素があるかどうか断定できないとき」に『A2』を用いている。

例)　　山田太郎さんの――+
　　長男――A―+
　　一郎くんが

以下のみを同格の関係とする。基本的に助詞を伴った文節同士の関係は同格としない。

a.　同格を表す助詞「の」が省略された関係（上位概念と下位概念）

同格『A』ラベルでリンクしている。

例)　　アメリカ大統領――A―+
　　クリントンは

例)　　日本で――+
　　大評判の―+
　　外車――A―+
　　ベンツを

＊同格の助詞「の」が省略されていない場合、単なる修飾関係とし同格関係とはしない。

例)　　ビールメーカーの――+
　　Y社が

＊同格の助詞「の」で結ばれる関係は「～であるところの」であり、「言い換えれば」のような関係は「同格」とはせず、「言い直し（後述）」としている。

例）　水―――D―+
　　即ち　　|
　　Ｈ２Ｏの―+
　　痕跡が

＊それぞれの文節が助詞を伴っていれば同格関係としていない。

例）　内閣総理大臣が―――――――+
　　小泉純一郎が―――――――|
　　事実上の―――――――|
　　責任者だ

b.　助詞がない文節を指示詞や人称代名詞などの照応関係にある語句で受けている場合

指示詞、人称代名詞などを省略して言い換えても自然な場合、同格『A』ラベルでリンクしている。

例）　次の―-+
　　問題-A―+
　　これについて―――+
　　検討します

例）　歌手の―-+
　　山口百恵―A-+
　　彼女を―-+
　　選びました

例）　歌手の―-+
　　山口百恵―-A-+
　　この―-|
　　女性を

c. 具体例に対する総称の関係

同格『A2』ラベルでリンクしている。

＊「そういう週刊誌」と「ヴォーグ」を同格関係としたいが、「そういう週刊誌」と同格の要素は「ヴォーグ」の他にもありそうだという意味で『A2』とする。

例）　ヴォーグとか-―-A2-+
　　そういう―――|
　　週刊誌の

＊「他にも同格の要素があるか断定はできないが可能性としてありそう」なら『A2』を使用する。

例）　東大寺とか-―P-+
　　興福寺―-P-+
　　春日大社―-A2―+
　　いくつかの―-|
　　奈良市内の―――|
　　社寺では

d. 具体例に対する数詞の関係

同格『A』または『A2』ラベルでリンクしている。

例）　日本語を―――――+
　　母国語と―|
　　する―+
　　話者-A2-+

二名を

3-2-12　言い直し

(D　)(D2　)の転記タグが付与されているものは対象外とし、それ以外の文節を対象とする。「言い直し」と認定された場合、「問題部」から「訂正部」へのリンクを付与しラベル『D』を付与している。

例)　山田――D-+
　　山田さんは――――――――――――――――+
　　強靭な-+　　|
　　肉体の――+　　|
　　持ち主だと――――|
　　言ってましたね

「言い直し」の認定には「並列」「同格」との識別が必要になるが、ここでは「5-2-10　並列」「5-2-11　同格」で示した関係のみを「並列」「同格」とし、それ以外を「言い直し」とする。以下で「体言」の例について示す。

a.　問題部の途中で言い直している場合

「言い直し」とする。

例)　欽ちゃんの――+
　　週刊(D き)――D――+
　　(F ん)　　|
　　週刊欽曜日とかいうような―+
　　番組が―+
　　あって

例)　混合数は―+
　　六十四時間
　　学習データー量は――+
　　五十――D-+
　　五十九――D-+
　　五十九時間分です

b.　問題部が訂正部に「訂正」されている場合

話者本人が「言い誤った」と認識し訂正している場合のみ、「言い直し」とする。この場合、訂正部が助詞を伴うか否かによって「言い直し」の認定に違いがある。

b-1　問題部が助詞を伴わない場合

問題部が助詞を伴わない場合は「言い直し」とする。

例)　彼は――――――――――+
　　大阪――D―+　　|
　　神戸の――|
　　出身だ

b-2　問題部が助詞を伴う場合

問題部が助詞を伴う場合は基本的に「言い直し」としない。話者本人が「言い誤り」を認識し訂正している場合のみ、つまり

・判断材料となる接続詞・フィラーなどがある場合
・音声チェックにより「言い直し」と認められる場合

など明らかな「言い直し」のみに限定する。

＊フィラー「いや」によって判断、「言い直し」と認定する。

例)　『S:複数文節言い直し』みかんを――D―+
　　『F』いや　|
　　甘い――|
　　みかんを

例)　私は――――――――――――――――――+
　　大きい――+　|
　　看板が――D―+　|
　　看板に――――|
　　気が―|
　　ついた

＊話者本人の「言い誤り」の認識が感じられない場合、「言い直し」とはしない。音声で確認しても、特に「言い直し」ていると感じられない。「太郎が」「次郎が」の関係は単なる追加と判断、「言い直し」とはしない。

例)　太郎が――――――――――――――――+
　　次郎が――――――――――――――|
　　みかんを――――――――――――|
　　食べた

c.　訂正部で不足分を補っている場合

話者本人に「言い誤った」という認識はないが、何か「言い足りない」と感じている。そのようなとき、類語で言い直したり、何か言葉を補って言い直す場合があり、問題部が助詞を伴うか否かによって「言い直し」の認定に違いがある。

c-1　問題部が助詞を伴わない場合

問題部が助詞を伴わない場合は「言い直し」とする。

類語での言い直し

＊「アップル」と「りんご」の関係は「同格」との識別が必要である。この場合、「言い換えれば」の関係であるので「同格」とはしない。「つまり」「即ち」で結ばれている場合もある。

例)　アップル―D―+
　　りんごを

言葉を補う言い直し

例)　『S:複数文節言い直し』りんご―D―+
　　完熟の――|
　　りんごを――+
　　ください

c-2　問題部が助詞を伴う場合

問題部が助詞を伴う場合は基本的に「言い直し」としない。

類語での言い直し

例)　アップルを――+
　　りんごを――|
　　ください

言葉を補う言い直し

例)　私は――――――――――――――――――+
　　みかんを――――+　|
　　みかんの―+　|　|
　　皮を――――|　|
　　むくのが――|
　　下手だ

d. 倒置の言い直し

「訂正部」が倒置部分にある場合、「言い直し」としない。

例)　私は――――――+　　間違い例)　私は――――+
　　海へ――|　　海へ――|――D―+
　　行った　　行った　|
　　|　(F あ)　　|　|　(F あ)
　　+――――川へ　　+――――川へ

＊この例では問題部：「海へ」 訂正部：「川へ」であるが、「川へ」は「行った」に倒置で係る。このように訂正部が倒置部分にある場合、『D』によるリンク付けは行わない。

＊「用言」の「言い直し」についても上記の例に準ずる。

例)　人口は――――――+
　　だんだん――――|
　　減ってく―D―+
　　減ってきた

e. 複数文節の言い直し

「言い直し」の「問題部」と「訂正部」が「１文節：１文節」対応でない場合、『S:複数文節言い直し』を付与している。

「「１：１」対応ではないことを示す」ために、「１：多」「多：１」「多：多」対応の言い直しに次の基準で付与する

・問題部が１文節の時（『B+』で結ばれた文節も１文節）→『S:複数文節言い直し』

・問題部が複数文節の時のみ→『S:複数文節言い直し:S1』『S:複数文節言い直し:E1』

例)　そこに――――――――――――――――――――――――――――――+
　　『S:複数文節言い直し』遠因―D―+　　|
　　元々の――――|　　|
　　原因というものが――|
　　あるようです

＊問題部が複数文節である場合、開始点と終了点を明示する
　開始点『S:複数文節言い直し:S1』
　終了点『S:複数文節言い直し:E1』

例)　『S:複数文節言い直し:S1』 色んな―+
　　紛争――+
　　『S:複数文節言い直し:E1』 起こってまし―D――――+
　　色んな――――+　　|
　　国内紛争が――|
　　起こってますが

3-2-13　古文の処理

学会発表などで古文が引用されている場合がある。このような場合、係り受けはすべて隣の文節に係るものとし、開始点『K:S1』／終了点『K:E1』を付与している。

例)　『K:S1』きどくに――+
　　夜回りを――+
　　『K:E1』するよ

ただし、現代文中に古文が引用されていれば外への係り受けは付与している。

4. XML における記述について

以上の情報はXML ファイルの一部として提供される。SUW 要素の属性のうち"Dep_"で始まる名称のものが係り受け関係の情報である（[9]参照)。詳細は次の通り。

(a) Dep_BunsetsuUnitID
節単位内で文節に一意に付与された ID（0 スタートの連番）を記述する。

(b) Dep_ModifieeBunsetsuUnitID
同一単位内で当該文節が係っていく先の文節の ID(Dep_BunsetsuUnitID)(ない場合がある)を記述する。

(c) Dep_Label
係り受け関係を示すラベルを記述する。この属性の値が取り得るのは、p.4 の【1】の表にある「係り受け関係を示すラベル」である。

(d) Dep_ObligateComment
義務的コメントを記述する。この属性の値が取り得るのは、p.4 の【1】の表にある「文節境界に関するラベル」と「その他」のコメントである。

参考文献

[1] 黒橋, 長尾：“京都大学テキストコーパス・プロジェクト”, 言語処理学会大３回年次大会発表論文集, pp.115-118（1997）

[2] 前川：“『日本語話し言葉コーパス』の概観”,『日本語話し言葉コーパス』の解説文書（overview.pdf）

[3] 古井:“話し言葉の音声認識と自動要約”,第３回話し言葉の科学と工学ワークショップ講演予稿集,pp.1-6（2004）

[4] 堀、古井：“講演録作成を目的とした講演音声自動要約”, 日本音響学会秋季講演論文集, 第１巻, pp.67-68（2001）

[5] 高梨, 丸山, 内元, 井佐原：“『日本語話し言葉コーパス』における節境界認定”, 平成１５年度国立国語研究所公開研究発表会予稿集, pp.33-34（2003）

[6] 丸山, 柏岡, 熊野, 田中：“節境界自動検出ルールの作成と評価”, 言語処理学会第９回年次大会発表論文集, pp.517-520（2003）

[7] 西川, 小椋, 相馬, 小磯, 間淵, 土屋, 斎藤：“文節の仕様について”,『日本語話し言葉コーパス』の解説文書（bunsetsu.pdf）

[8] 高梨, 内元, 丸山：“『日本語話し言葉コーパス』における節境界認定”,『日本語話し言葉コーパス』の解説文書（clause.pdf）

[9] 菊池, 塚原, 小町, 山田, 高梨：“『日本語話し言葉コーパス』XML 文書について”,『日本語話し言葉コーパス』の解説文書（xml.pdf）