構文情報付き法律文コーパスの設計と構築

山田 将之　小川 泰弘　外山 勝彦
名古屋大学大学院情報科学研究科
{myamada,yasuhiro,toyama}@kl.i.is.nagoya-u.ac.jp

1　はじめに

裁判員制度の導入や市民運動の高まりにより，法令はますます身近な存在になる一方で，社会の国際化により，外国の人々が日本法令に接する機会が増えている．それに伴い，法令の内容を素早く正確に伝え，理解し，運用することへの要求が高まっている．そのような要求に対する有効な対処として，文章の作成支援や読解支援といった法令文書の高度処理に基づく支援が考えられる．

ここで，自然言語文書の高度処理において，構文情報は重要な役割を果たす．例えば，統計的な構文情報に基づく構文解析や係り受けの曖昧性解消，用例を利用した文章作成支援，当該分野における語彙知識の獲得などにおいては，構文情報が広く利用されている．これらの言語処理は法分野においても有用と考えられる．しかし，法律文の構文解析は容易でないことから，それらの言語処理の実現には構文情報が付与された法律文コーパスが必要になる．

そこで，本稿では法律文に構文情報を付与した法律文コーパスを構築する．その際に，文中に頻繁に挿入される括弧や，階層的並列構造を作る等位接続詞など，法律文特有の表現が問題となる．これらの表現は，既存のコーパスに含まれていないか，または既存のコーパスにおける表現とは用法が異なっているため，タグ付け基準を新たに定める必要がある．

本稿では，法律文コーパスの基本設計について述べた上で，新たに定める必要のあるタグ付け基準について議論し，法律文に付与するタグセットとタグ付け基準を設計する．さらに，法律文コーパスの構築手順，構築したコーパスの概要について述べる．

2　法律文コーパスの基本設計

形態素情報と構文情報は様々な言語処理において有用な言語情報であり，既存の解析ツールを利用することにより比較的低コストでタグ付けできる．そこで，法律文コーパスにおいては下記の情報を付与する．

形態素情報

- 形態素の区切り
- 形態素の品詞，読み，原形，活用型，活用形

構文情報

- 文節の区切り
- 文節間の係り受け関係
  - 通常の係り受け関係
    主語・述語の関係，修飾・被修飾の関係，接続・被接続の関係
  - 並列関係
    意味上同等な語，句，節間の関係
  - 同格関係
    単に併置され，互いに意味を限定し合うような語と語の関係

日本語の構文解析器において現在広く用いられている形態素辞書として，IPADIC[1] や JUMAN 辞書などがある．本稿では，法律文コーパスのエンドユーザは学校文法に慣れ親しんだ法曹関係者や一般人であると想定して，形態素情報については，品詞体系が学校文法により近い IPADIC に準拠する．また，構文情報については，統一的なタグ付け基準を既に定めている京大コーパス作成基準 [4] に準拠し，通常の係り受け関係，並列関係，同格関係をそれぞれ記号 D,P,A を用いて表す．

本稿では，以上を法律文コーパスの基本設計とする．ただし，基本設計をそのまま適用できない法律文特有の表現などについては，語の意味や用法に基づき，タグ付け基準を新たに定める．

3　タグ付け基準の設計

法律文には，京大コーパスにおけるタグ付け基準をそのまま適用できない特有の表現があり，基本設計に加えて，タグ付け基準を新たに定める必要がある．本節では，そのような表現のうち特徴的な 3 つの表現について特に述べる．

3.1　括弧書き

法律文では，次の例のように括弧付きの文が多用される．

(1) 金融庁の長は、金融庁長官（以下「長官」という。）とする。[1]

(2) 警察庁の課に、課長（室にあつては、室長）を置く。[2]

(3) 民事訴訟法第百五十四条（通訳人の立会い等）の規定は、審判に準用する。[3]

[1]金融庁設置法第二条第二項
[2]警察法第二十六条第二項
[3]特許法第百四十六条

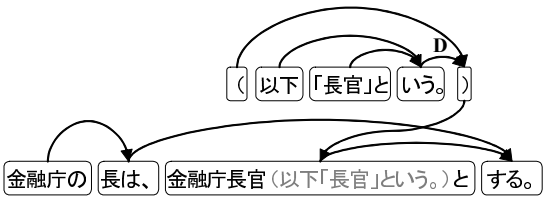

図 1: 括弧書き-被括弧書き文間の通常の係り受け関係

一般に，括弧「( )」を含む文は，形態素解析や構文解析における対処が容易ではない．加えて，多くの場合，括弧内の文字列 (以下，括弧書きという) を削除しても，元の文の形態素情報や構文情報は変化しないことから，京大コーパスにおいては括弧書きをあらかじめ除去してタグ付けしている．一方，法律文における括弧はその用法が定まっており，多くの場合，文の意味上省略不可能な字句が括弧内に頻繁に挿入される．そのため，法律文コーパスにおいては括弧書きを除去せずにタグ付けする必要がある．

括弧を含む文のタグ付けにあたっては，括弧書きの挿入によって分割された語句の扱いが問題となる．それに対し本稿では，**括弧の内外の文字列は各々独立して文を成す**と仮定して解決を図る．この仮定により，字面の上では文節間の係り受けが複雑な括弧を含む文についても，括弧内の文字列 (すなわち，括弧書き) と括弧外の文字列 (以下，被括弧書き文という) に対するタグ付けを各々独立に考えることができる．

ここで，括弧書きと被括弧書き文は，もともと 1 文中に記述されていること，一般的に 1 文中のすべての文節間には何らかの係り受け関係が成り立つことを考慮すると，括弧書きと被括弧書き文の間にも何らかの係り受け関係を定めるべきである．そこで以下では，括弧書きの被括弧書き文に対する係り受け関係を定め，従来，法律文中において用法別に分類されてきた [2] 括弧書きを係り受け関係の種類別に再分類する．その際，括弧自体を括弧書きと被括弧書き文を繋ぐ特別な文節として捉え，括弧書き-被括弧書き文間の係り受け関係は，閉括弧を介して成り立たせる．

直前の字句に通常の関係で係る括弧書き

直前の字句に対する定義・略称，字句の意味の縮小・拡張に用いられる括弧書きは，(1) のように「述語＋句点」で終わる形で記述される．このような括弧書きの述語は被括弧書き文の字句を修飾していると考え，括弧書きの文末の文節は括弧書きの直前の字句を含む文節に対して通常の関係で係るとする．

例えば，(1) の係り受け構造は図 1 のように表される．なお，係り受け関係の非交差条件を満たすために，「いう。」と「金融庁長官と」の係り受け関係を，「いう。」と閉括弧「)」の間の係り受け関係で代替して表す．図 1 の場合，「いう。」と「金融庁長官と」が通常の係り受け関係にあることが，「いう。」と閉括弧「)」の間にある記号 D が付された関係により表されている．

直前の字句と並列関係にある括弧書き

条件を満たす場合に直前の字句を置き換えることを規定する括弧書きは，(2) のように「条件を記した節＋置き換える字句」という形で記述され，置き換えられる字句と置き換える字句は同じ品詞をとる．このような置き換える字句と置き換えられる字句は意味上同等であると考え，括弧書きの文末の文節は，括弧書きの直前の字句を含む文節に対して並列関係にあるとする．

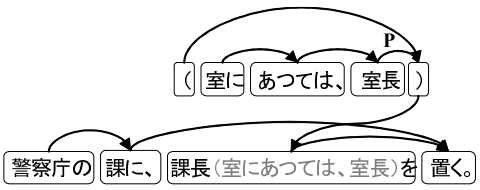

図 2: 括弧書き-被括弧書き文間の並列関係

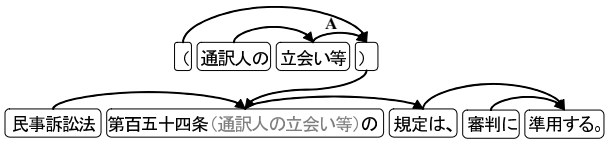

図 3: 括弧書き-被括弧書き文間の同格関係

例えば，(2) の係り受け構造は図 2 のように表される．前述の通常の係り受け関係の場合と同様に，「室長」と閉括弧「)」の間にある記号 P が付された関係は，「室長」と「課長を」が並列関係にあることを表している．

直前の字句と同格関係にある括弧書き

直前の名詞句の法令番号，要旨の記述に用いられる括弧書きは，(3) のように名詞句の形で記述される．直前の名詞句とこのような括弧書きは互いに意味を限定し合っていると考え，括弧書きはその直前の字句を含む文節に対して同格関係にあるとする．

例えば，(3) の係り受け構造は図 3 のように表される．前述の通常の係り受け関係の場合と同様に，「立ち会い等」と閉括弧「)」の間にある記号 A が付された関係は，「立会い等」と「第百五十四条の」が同格関係にあることを表している．

3.2 等位接続詞

法律文に頻出する等位接続詞として，「又は」「若しくは」「及び」「並びに」「かつ」がある．「又は」「若しくは」は，ある事物と他の事物を選択的に結び付ける語であり，「及び」「並びに」「かつ」は，ある事物と他の事物を併合的に結び付ける語である．法律文では，並列構造を階層的に表すために，これらの語の使い分けが明確に定められている [2].

形態素の品詞，文節の区切り

京大コーパスにおいては，等位接続詞は，(4) のように直前の語に後続して記述される場合は接続助詞，(5) のように読点に後続して記述される場合は接続詞としている [4].

(4) 外国の政府又は地方公共団体の公務に従事する者[4]

(5) 競争関係にある他人の営業上の信用を害する虚偽の事実を告知し、又は流布する行為[5]

さらに，どちらの場合においても，等位接続詞は直前の語と合わせて一文節を成すとしている．しかし，「接頭辞＋自立語＋付属語・接尾辞」という文節の基本構成，「文を意味上・発音上不自然でない範囲で分解した

[4]不正競争防止法第十八条第二項第一号
[5]不正競争防止法第二条第一項第十四号

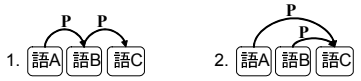

図 4: 京大コーパス基準に基づく並列構造表現

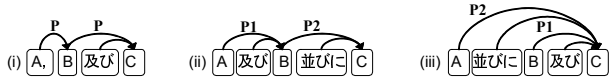

図 5: 法律文コーパスにおける並列構造表現

場合の最小単位」という文節の一般的な定義を考えた場合,「自立語＋読点＋自立語」を 1 文節とするのは不自然である.

法律文における等位接続詞は，原則として，併記される語が動詞，形容詞，副詞のいずれかの場合にのみ読点に後続して記述される．すなわち，読点の有無は用字上決められるものであり，等位接続詞の統語的働きは読点の有無に関わらず本質的に同じであると考えられる.

したがって，法律文における等位接続詞の品詞は一貫して接続詞とし，1 語で 1 文節を成すとする.

階層的並列構造

法律文では,「又は」と「並びに」,「及び」と「若しくは」の使い分けにより並列関係の強弱を表す．例えば，次のように併記された語 A,B,C はそれぞれ括弧で括った順で結びつきが強い.

(i) A，B 及び C → (A，B 及び C)
(ii) A 及び B 並びに C → ((A 及び B) 並びに C)
(iii) A 並びに B 及び C → (A 並びに (B 及び C))

京大コーパスにおいては，並列関係を記号 P のみを用いて表している．しかしその場合，併記された語 A,B,C に対しては図 4 に示す 2 通りの並列構造しか表せず，上記の 3 文の構文的差異をタグ付けにより表すことができない.

そこで，並列関係については，並列関係であることを示す記号 P と並列関係の結びつきの強さを示す数値 (小さいほど結びつきが強い) を用いて，階層的な並列構造を表すことにする．これにより，前述の (i)〜(ii) の係り受け構造は図 5 のように表される．ここで，P1 が付された「及び」の並列関係は，P2 が付された「並びに」の並列関係より結びつきが強いことを表している．なお，(i) のように並列関係の結びつきの強さがどれも対等な場合，単に記号 P を付すことにする.

3.3 「その他」,「その他の」

法律文において事物を列挙する場合，通常は等位接続詞が用いられるが，事物を例示的に列挙する場合やグループ的に全体をまとめるような場合には，次のように，列挙する事物を読点で結び，最後に「その他」「その他の」を用いて括る.

(6) 法律の規定に基づき施設、区間、地域その他これらに類するものを指定する命令又は規則[6]
(7) 公務員の給与、勤務時間その他の勤務条件について定める命令等[7]

[6]行政手続法第三条第二項第四号
[7]行政手続法第三条第二項第五号

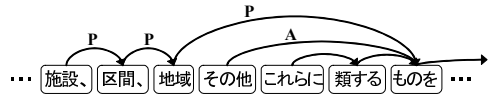

図 6: 「その他」を含む法律文の係り受け関係

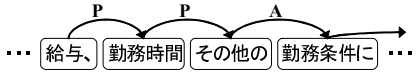

図 7: 「その他の」を含む法律文の係り受け関係

「その他」と「その他の」は，一般の文では区別されることなく用いられるが，法律文ではこれらは明確に区別される [2].

「その他」

京大コーパスでは,「〜その他」については「〜など」と同様に，後続する適当な体言と同格関係にあるとしている.

一方，法律文における「その他」は，事物を並列的に例示する際に用いられ,「〜など」と用法が異なる．例えば (6) における「施設」「区間」「地域」「これらに類するもの」は意味上同等であり,「これらに類するもの」には「施設」「区間」「地域」が含まれない．すなわち，この場合の「その他」は，ある事物の集合から「施設」「区間」「地域」を除いたものを指し示しており,「これらに類するもの」と同じ事物を指し示し，意味を限定し合っている.

したがって,「その他」の品詞は代名詞であり，後続する適当な語と同格関係にあるとする．また,「その他」の前後に列挙された語は並列関係にあるとする．(6) のうち「…施設、区間、地域その他これらに類するものを…」の部分の係り受け構造は図 6 のように表される.

「その他の」

京大コーパスでは,「〜その他の」は，それに後続する適当な語に通常の関係で係るとしている.

一方，法律文における「その他の」は，前に挙げられた事物が「その他の」の後ろに挙げられた事物の下位概念であることを示す際に用いられる．例えば，(7) における「給与」「勤務時間」は意味上同等であり，これらは「勤務条件」の例示であり下位概念である．すなわち，この場合の「その他」は「勤務条件」から「給与」「勤務時間」を除いたものを指し示しており,「給与」「勤務時間」と意味上同等であると考えられる．さらに,「勤務条件」には「給与」「勤務時間」以外の事物も含まれ,「給与、勤務時間その他」と「勤務条件」は同じ事物を指し示し，意味を限定し合っていることから,「〜その他の」中の「の」は同格の「の」であると考えられる.

したがって,「その他の」中の「その他」の品詞は代名詞であり，前の適当な語と並列関係にあるとする．また,「その他の」中の「の」は同格の「の」であることから,「〜その他の」とそれに後続する適当な語は同格関係にあるとする．(7) のうち「…給与、勤務時間その他の勤務条件に…」の部分の係り受け構造は図 7 のように表される.

4　法律文コーパスの構築

本節では，2 節，3 節において設計したタグセットとタグ付け基準に基づき，法律文コーパスを構築する手順について述べ，法律文コーパスを構築する．

4.1　手順

タグ付けする法律文書の収集

現在，日本政府によって日本法令の英訳事業が進められている [7]．英訳文が対応付けられている日本語文に対して構文情報を付与した場合，単言語内の用例検索や文章作成支援だけでなく，言語横断用例検索や翻訳支援の実現に寄与できると考えられる．そこで，当事業により英訳が作成されている日本語法律をタグ付けの対象とし，収集する．

法律文の抽出

収集した法律文書を，法令文書の XML 文書型定義 [6] に基づき XML 化する．この XML 文書型定義では，法律文書において 1 文にほぼ相当する「段」という構造要素を定めている．本稿では各段をそれぞれ 1 文とみなして抽出する．

前処理

法律文においては，「…にあつては」のように促音「っ」が大書きされている場合がある．このような記述は形態素解析の誤りの原因となるため，事前に大書きされた促音を小書きに変換する．

また，3 節で述べたように，括弧の内外の文は各々独立して文を成すと考えてタグ付けするため，括弧書きと被括弧書き文をそれぞれ別の文とみなして分割する．

法律文の形態素解析，構文解析

ChaSen[5] と CaboCha[3] を用いて法律文を解析し，形態素情報，構文情報を付与する．

後処理

新たに設計したタグ付け基準のうち，「その他」「その他の」のように字面を手掛かりにして構文的位置付けを比較的容易に特定できるものについて，CaboCha により付与された構文情報を機械的に修正する．また，各々独立して解析した括弧書きと被括弧書き文の解析結果を統合し，括弧書きと被括弧書き文の間の係り受け関係を付与する．さらに，前処理で小書きに変換した促音を大書きに再変換する．

人手による言語情報の修正

機械的な処理によっては正しく付与できなかった形態素情報，構文情報を人手で修正する．

4.2　結果

前節で述べた手順に従い，行政手続法と不正競争防止法から抽出した 414 文に対して形態素情報，構文情報を付与した．タグ付けした法律文の一部[8]を図 8 に示す．ここでは CaboCha のデフォルトの出力形式と同様のデータ形式で出力している．図 8 では，3.2 で前述した記号 P1,P2,P3 を用いて階層的並列構造を表している．これにより，「学生」，「生徒」，「児童」，「幼児」，「保護者」，「講習生」の結びつきが次のような順で強いことが容易に読み取れる．

[8]行政手続法第三条第一項第七号

```
：：：
：：：
* 14 15P1
学生 ガクセイ 学生 名詞-一般
、 、 、 記号-読点
* 15 16P1
生徒 セイト 生徒 名詞-一般
、 、 、 記号-読点
* 16 18P1
児童 ジドウ 児童 名詞-一般
* 17 18D
若しくは モシクハ 若しくは 接続詞
* 18 21P2
幼児 ヨウジ 幼児 名詞-一般
* 19 21D
若しくは モシクハ 若しくは 接続詞
* 20 21D
これら コレラ これら 名詞-代名詞-一般
の ノ の 助詞-連体化
* 21 22P3
保護 ホゴ 保護 名詞-サ変接続
者 シャ 者 名詞-接尾-一般
、 、 、 記号-読点
* 22 23P3
講習 コウシュウ 講習 名詞-サ変接続
生 セイ 生 名詞-接尾-一般
、 、 、 記号-読点
：：：
：：：
```

図 8: 形態素情報，構文情報を付与した法律文

- (((学生、生徒、児童若しくは幼児) 若しくはこれらの保護者、) 講習生、…)

5　おわりに

本稿では，計算機による法律文書の効率的かつ高度な利用の実現において重要な役割を果たす構文情報付き法律文コーパスを設計し，構築した．その過程で，既存のコーパス構築手法を適用できないような法律文特有の表現に対するタグ付け基準を定めた．本コーパスを利用することにより，法律文に特化した構文解析器の精度向上，法律文の作成支援，翻訳支援などの実現が期待される．今後，引き続き，コーパス収録文の増加に努める．

参考文献

[1] 浅原正幸, 松本裕治: ipadic version 2.7.0 ユーザーズマニュアル, 2003.

[2] 石毛正純: 自治立法実務のための法制執務詳解《四訂版》, ぎょうせい, 2004.

[3] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol.43, No.6, pp.1834–1842, 2002.

[4] 黒橋禎夫, 居蔵由衣子, 坂口昌子: 形態素・構文タグ付きコーパス作成の作業基準 version 1.8, 2000.

[5] 松本裕治: 形態素解析システム「茶筌」, 情報処理学会誌, Vol.41, No.11, pp.1208–1214, 2000.

[6] 村瀬裕城, 小川泰弘, 外山勝彦: 法令文書の XML 文書型定義, 平成 18 年度電気関係学会東海支部連合大会, O-395, 2006.

[7] 内閣官房: 法令翻訳データ集, http://www.cas.go.jp/jp/seisaku/hourei/data1.html, 2008/1/21 アクセス.