Vol.2017-NL-233 No.6
2017/10/24

将棋解説文へのモダリティ情報アノテーション

松吉 俊[1,a)]　村脇 有吾[2,b)]　亀甲 博貴[3,c)]　森 信介[2,d)]

概要：近年，現実世界の物事を自然言語によって自動的に記述することや検索することに注目が集まっている．我々は，現実世界の具体的な非テキストデータとして将棋に着目している．以前の研究において，将棋の局面とそれに対応する解説文を収集してコーパスを作成し，コーパス整備の第 1 歩として，将棋解説文に単語分割情報，品詞情報，将棋に特有の固有表現をアノテーションした．解説テキストには，断定的な平叙文のみが存在するわけではなく，選ばれなかった戦型や解説者が予想した今後の駒の進行なども言及される．これら否定や推量，仮定などの情報発信者の態度は，モダリティ表現によって表出される．テキストに含まれるモダリティ情報を適切に捉えるため，本研究では，上記のコーパスに対して，3 種類のモダリティ情報 (モダリティ表現，事象クラス，事実性) をアノテーションした．本論文では，提案するアノテーション体系のラベルについて説明するとともに，構築したアノテーション済みコーパスの統計情報を報告する．また，解説文自動生成やシンボルグラウンディングなど，本コーパスの将来の展望についても考察する．

SGC-MEF: A Shogi Commentary Corpus Annotated with Modality Information

Suguru Matsuyoshi[1,a)]　Yugo Murawaki[2,b)]　Hirotaka Kameko[3,c)]　Shinsuke Mori[2,d)]

1. はじめに

近年，現実世界の物事を自然言語によって自動的に記述することや検索することに注目が集まっている．その理由の 1 つは，テキストとそれに紐づけられた非テキスト情報をインターネットを通して大量に入手することが容易になったからであろう．例えば，画像とそのキャプションテキスト，株価チャートとその解説記事などは，比較的容易に収集することが可能である．収集したデータで学習を行うことにより，画像や映像などの非テキストデータから自然言語文を生成する方法についての研究が活発に行われている [1], [15], [22], [25]．非テキスト情報を用いることにより言語モデルの性能を向上させる方法も提案されている [9]．テキストにおいて単なる記号列として記述されるもの (語，句，文) を実世界内の物事と対応づける処理は，**シンボルグラウンディング (記号接地)** と呼ばれる [3]．非テキスト情報を伴った大量のテキストデータが利用可能になったことにより，シンボルグラウンディング問題に取り組めるようになったと言える．

非テキストデータからの自然言語文生成タスクやシンボルグラウンディング問題の研究を遂行するにあたり，我々は，現実世界の具体的な非テキストデータとして以前より将棋局面データに着目している [12]．このデータに着目している主な理由は，次の 3 点である．

i. すべての局面ではないが，プロの試合の多くの局面に対して，その状況を他のプロ棋士 (解説者) が解説した解説文が存在する
ii. 通常の非テキストデータ (画像や映像) と異なり，局面データの内容を曖昧性なく記述できる表記法 (Shogi Forsyth-Edwards notation) が存在する
iii. 今後の良い手を高い精度で自動予測するアルゴリズ

[1] 電気通信大学，The University of Electro-Communications
[2] 京都大学，Kyoto University
[3] 東京大学，The University of Tokyo
[a)] matuyosi@uec.ac.jp
[b)] murawaki@i.kyoto-u.ac.jp
[c)] kameko@logos.t.u-tokyo.ac.jp
[d)] forest@i.kyoto-u.ac.jp

ム [20] が存在する

上の i. は，非テキスト情報 (局面データ) に対してテキストが紐づけられていることを意味しており，上記の研究を遂行するにあたり，最低限満たすべき要件である．

画像や映像内の事物とテキスト内のシンボルを対応させようとすると，物体認識の精度や認識範囲の曖昧性が問題になってくる．上の ii. は，これらの問題を回避して研究を遂行できることを意味する．

2 章で見るように，将棋解説文には，現在の局面についての言及のみでなく，解説者が予想した将来の手筋などについてのコメントも含まれる．それゆえ，解説文自動生成タスクやシンボルグラウンディングにおいては，人間が行っているこの予測を模倣できることが望ましい．上の iii. は，必要ならば，この予測を計算機上で遂行できることを意味する．

我々は，以前の研究 [12] において，将棋の局面とそれに対応する解説コメントを収集してコーパスを作成した．このコーパスの一部に対して，人手で将棋特有の固有表現をアノテーションし，機械学習手法を利用することにより，コーパス全体に対して将棋特有の固有表現のラベルを自動付与した．また，このコーパスを利用して将棋解説文を自動生成する手法を提案した [5], [31], [32]．しかしながら，この生成モデルは，人間と同等の解説文を生成することはできていない．駒の動きを予測することは実現できており，今後の駒の動きについて述べることには成功しているが，自然なテキストとして生成することはできない．

将棋解説文の自動生成や，将棋特有の固有表現のシンボルグラウンディングにおいて，解説文に存在する**モダリティ表現**を認識することは重要である．なぜならば，人間は，現在目に見えていないが，重要である物事・命題を表現するために，モダリティ表現を利用するからである．その命題が予測であることやその命題が成立しないことを伝えたい時や，ある命題を仮定して別の命題を主張したい時に，モダリティ表現が利用される．本研究では，上記の研究などにおいて有効的に利用されることを目的として，将棋解説文にモダリティ情報をアノテーションする．

本研究の主な貢献は次の 3 つである．

1. おそらく世界で初めて，対応する非テキスト情報を伴うテキストに対して明示的にモダリティ情報をラベル付けした (6.5 節参照)
2. 日本語の自然言語処理 (特に，テキスト解析) において利用しやすい形で，モダリティ情報の体系を再整理した (6.2 節参照)
3. 含意関係認識や情報検索などの応用タスクも考慮して事象らしさを定義し，網羅的にラベル付けした (4.2 節参照)

本論文は，以下のように構成される．まず，2 章において，将棋解説文の特徴と，我々が収集した将棋解説文コーパスについて述べる．次に，3 章で，モダリティ情報アノテーションの関連研究について紹介する．4 章で，本研究で提案するモダリティアノテーション体系を説明する．5 章において，アノテーション済みコーパスの統計情報を報告する．6 章では，本コーパスの応用について考察する．7 章はまとめである．

図 1 将棋局面とその解説文

2. 将棋解説文

2.1 将棋

将棋は 2 人で行うボードゲームである．9 × 9 のマスの盤面と，14 種類の駒[*1]を用いる．チェスと異なり，取った相手の駒を自身の持ち駒とすることができ，盤面上の駒を動かす代わりに，持ち駒を盤面上の空いているマスに打つことができる．先手と後手の分を合わせ，盤面上のすべての駒の配置と各自の持ち駒の状態を総称して**局面**と呼ぶ．将棋は完全情報ゲームであり，各局面にはその時点のゲームの状態に関するすべての情報が保存される．Shogi Forsyth-Edwards notation と呼ばれる表記法により，局面を曖昧性なく記述可能である．

将棋にはプロ制度があり，日々，プロの間で対局が行われている．多くの対局において，対局者以外のプロ棋士が将棋ファンのためにその解説を行う．本研究では，このテキストを**将棋解説文**と呼ぶ．図 1 に，インターネット配信されている対局と解説の例を示す．画面中央が将棋の盤面であり，持ち駒と合わせて局面を構成する．画面下部には，現在の局面に対する解説文が掲載されている．

2.2 将棋解説文の特徴

将棋解説文においては，次のような内容が述べられている．

指し手

(1) 羽生は１０分弱で△４四歩を着手．

[*1] 通常の駒と，成った駒を合わせた数．

(2) ▲1五銀に△1四歩.

(3) 飛車取りではなく，5七の地点に香を成った.

指し手の評価

(4) 好手ですね.

(5) ほほー，これは渋い手ですねー.

(6) すさまじく筋の悪い手ですね（笑）.

その指し手を選んだ理由の推測

(7) 7筋を攻められる展開になったときに，▲7三角が王手にならないようにした意味もある.

(8) △2九飛成の桂取りと，△2七飛成の両狙い.

その局面の状況

(9) 端に2手かけているので，生かせるかどうかが序盤のポイントになりそうだ.

(10) 先手の先攻力対後手のスキのない金矢倉.

(11) 早い段階で仕掛けていく展開になりやすく，かなり攻撃的です.

次の指し手の予想

(12) 次は△7六歩▲同銀△6六銀の筋がある.

(13) 検討では，▲4一同馬△同金▲3五歩が示されている.

(14) 封じ手予想は△7五歩が人気を博した.

戦型や囲いの予想とその当たり外れ

(15) 深浦九段は前夜祭で角換わりを予想していた.

(16) 第4局に続く矢倉角対抗の将棋になるとは，誰が予想できただろうか.

解説文のほとんどは局面に対するコメントである．局面に関係のないコメントとして，対局者の出入りや食事，残りの持ち時間などが述べられるが，少量である.

将棋解説文の文体は統制されていない．上の例文 (3), (7), (13) に見られるように，新聞記事に近い書き言葉のものもあれば，例文 (2), (5), (11) に見られるように，くだけた話し言葉のものや丁寧体が混じるものもある．通常のテキストに比べ，体言止めが多いのも特徴である.

将棋解説文には推測や予想が多く含まれるが，それらのすべてが解説者1人によるものであるとは限らない．名人戦の場合，本解説の他に，テレビやインターネット配信の解説が並行して行われており，それらの解説が1つのテキストファイルに混じっている．また，対局の間，棋士室 (関係者控室) では現在の対局についての検討が行われており，解説者は棋士室での検討・予想の様子も伝える．例文 (13), (14), (15) では予想が述べられているが，いずれも解説者自身のものではない.

2.3 将棋解説文コーパス (SGC コーパス)

我々は，インターネット[*2]で配信されている将棋解説文と局面データを対応付け，将棋解説文コーパスを構築した [12]．このコーパスは 6,523 対局に対する 744,327 文，11,083,669 語[*3]のテキストを含む．我々は，このうち，9つの対局を選択し，そのテキストに対して人手で単語分割を行い，品詞タグと将棋特有の固有表現をアノテーションした．この固有表現には，駒，配置，戦型名，囲い名，人名など 21 種類が定義されている．アノテーション済みテキストは，2,041 文，34,184 語である．以下，本論文では，便宜上の理由により，将棋解説文コーパスのうち，固有表現などがアノテーション済みの部分のみを指して，**SGC コーパス**と呼ぶ.

3. 関連研究

この章では，モダリティに関連する情報のアノテーションについての関連研究を述べる.

3.1 モダリティ表現

一般に，文章には命題だけでなく，その命題に対する情報発信者の主観的な態度も記述される [13], [33]．このような態度を**モダリティ**と呼び，それを示唆する，文章中の表現を**モダリティ表現**と呼ぶ.

英語において，主要なモダリティ表現は “must” や “may” に代表される助動詞である．TimeML [18] において，これらのモダリティ表現は，事象を表す <MAKEINSTANCE> タグの@modality 属性に記述される.

日本語においては，多数の文末表現を収録した辞書が編纂されており [26], [28]，必要に応じて，各研究者がこのような辞書を利用してモダリティ表現を自動検出することが多い [4], [19]．Kamioka らは，独自に機能表現集合を定義し，IOB2 フォーマットによりテキストに対して人手でアノテーションを行っている [6]．日本語における先行研究が対象としている主なモダリティ表現は，助動詞や助動詞型機能表現，および，叙実動詞である.

本研究では，品詞にかかわらず，広くモダリティ表現をラベル付けする.

3.2 事象のモダリティ

テキストにおいて述語項構造によって表現される命題のことを**事象**と呼ぶ．TimeML [18] に従い，行動や出来事のみでなく，述語項構造によって表現される状態や状況のことも事象という用語で指すこととする.

事象に対する情報発信者の主観的な態度 (事象のモダリティ) は，テキストにおいて事象の周辺に存在する複数のモダリティ表現により決定される．英語においては，

*2 『名人戦棋譜速報』http://www.meijinsen.jp
*3 自動解析による推測値.

TimeML [18] に見られるように，関連する助動詞や否定辞を直接ラベル付けする．日本語において，松吉らは，いくつかのモダリティクラスを定義し，事象にそのクラスの 1 つを付与している [27]．彼らがそのようにした理由は，日本語が膠着言語であり，英語での方法を直接適用することが難しいことと，日本語には同じような意味を表す複合辞が豊富に存在するからである．

本研究では，松吉らの分類を再整理し，事象のモダリティに含意関係認識や情報検索などの応用タスクも考慮したクラスも追加する．

3.3 事実性

英語において，Sauri らは，事象のモダリティクラスのうち，推測と否定に関する項目を切り出し，**事実性**ラベルの体系を提案した [17]．彼女らは，この体系に従って人手でアノテーションを行い，FactBank を構築した．日本語においてもこの体系が取り入れられ，事実性アノテーションが実施されている [6], [27]．

本研究でも，FactBank の事実性アノテーション体系を利用する．

3.4 態度表明者

事象のモダリティは，情報発信者の主観的なものであるため，それをアノテーションする際には，情報発信者を明記する必要がある．

英語の意見抽出タスクにおいて，Wiebe らは，情報発信者を明記するための「入れ子構造の枠組み」を提案した [24]．FactBank において，この入れ子構造は，モダリティを表現する人物 (**態度表明者**) を明記するために採用されている [17]．日本語においても，モダリティをラベル付けしたコーパスにおいて同じ方法により態度表明者がラベル付けされている [27]．

2.2 節で述べたように，将棋解説文の推測や予想は解説者 1 人によるものであるとは限らないため，モダリティや事実性をアノテーションするに際し，態度表明者をアノテーションし，誰の態度であるのかを明記する必要がある．本研究でも，先行研究と同じ体系を採用し，態度表明者を明記する．態度表明者の層[*4]のみアノテーションが完了していないため，以降の章では態度表明者については言及しない．SGC コーパスにおいては，すでに人名や「擬人化された部屋名」に固有名詞ラベル Hu が付与されている [12] ので，態度表明者のアノテーションは比較的スムーズに遂行できると思われる．

4. アノテーション体系

我々は，モダリティ情報として，次の 3 種類の情報をアノテーションする．

1. モダリティ表現 (ME*): 複合辞やモダリティ副詞など
2. 事象クラス (EV*): 事実性ラベルを付けるべき述語句かどうか
3. 事実性 (FP*, FN*): 事実性の確信度と，否定の有無

表 1 に，提案するアノテーション体系によるラベル付けの例を示す．「品詞」と「固有名詞」の層は，先行研究 [12] によりすでにラベル付与が済んでいる．本研究では，その下の 3 つの層のアノテーションを提案する．

本論文で提案する体系は，汎用的なモダリティ情報アノテーションである．将棋などのゲームや将棋解説文に特化したアノテーションではないことに注意されたい．

以下，この章では，3 種類のアノテーションについて説明する．

4.1 モダリティ表現

モダリティ表現に付与するラベルの一覧を，表 2 の上部に示す．モダリティ表現のラベルは，大きく 2 つのグループに別れる．

事実性関連　確信度や否定に関する表現 (5 種類)
MEy，MEa，ME0，MEm，MEn

時間関連　時間軸に関する表現 (3 種類)
MEp，MEf，MEh

3.1 節で述べたように，本研究では，品詞にかかわらず，これらのラベルをテキスト中の形態素列に付与する．以下，各ラベルのモダリティ表現の例を示す．

MEy　確実な肯定を示唆
(17) 切り返しを狙っていることは間違いない$_{\text{MEy}}$.
(18) 銀の捕獲に成功$_{\text{MEy}}$.

MEa　肯定の可能性を示唆
(19) このあと居飛車に組む可能性が高$_{\text{MEa}}$そう$_{\text{MEa}}$だ.
(20) 恐らく$_{\text{MEa}}$ △４五角だろう$_{\text{MEa}}$.

ME0　可能性を保留
(21) 相振り飛車にする可能性もある$_{\text{ME0}}$.
(22) その間に先手玉に迫る手段があるかどうか$_{\text{ME0}}$.

MEm　否定の可能性を示唆
(23) 後手の飛車もあまり$_{\text{MEm}}$利いていない.
(24) ▲同金△６七銀が相当受けがたく$_{\text{MEm}}$見える.

MEn　確実な否定を示唆
(25) 銀交換せず$_{\text{MEn}}$に引き揚げる.
(26) ここで△７五歩とするべきでした$_{\text{MEn}}$.

MEp　過去を示唆
(27) ここで銀交換に応じた$_{\text{MEp}}$.
(28) ここまで$_{\text{MEp}}$は谷川好みではないと思われる進行.

MEf　未来を示唆
(29) 先手は将来$_{\text{MEf}}$的に右辺に玉を囲うことになる.
(30) いよいよ$_{\text{MEf}}$戦いが始まる.

MEh　仮定の話であることを示唆

[*4] 第 1 層のモダリティ表現と第 2 層の事象クラスの間に挿入されることとなる．

表 1 将棋解説文に対する 5 層のアノテーション

層																								
テキスト	先手	は	美濃	囲い	が	崩れ	て	い	る	の	で	,	飛車	交換	は	後手	の	得	に	な	り	そう	だ	.
品詞	N	P	N	N	P	V	P	V	Sf	P	Aux	Pnc	N	N	P	N	P	N	P	V	Sf	Adj	Aux	Pnc
固有名詞	Tu-B	O	Ca-B	Ca-I	O	Ao-B	O	O	O	O	O	O	Mn-B	Mn-I	O	Tu-B	O	Ee-B	O	Ao-B	O	O	O	O
モダリティ表現	O	O	O	O	O	MEn-B	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	MEa-B	O	O
事象クラス				EVe		EVe		EVf						EVi						EVe				
事実性				FNc		FPc														FPr				

表 2 モダリティ情報ラベル一覧，および，SGC コーパス内での個数と割合

層	ラベル	説明	記号の由来	FactBank	数	割合
モダリティ表現	MEy	確実な肯定を示唆	yes		49	3%
	MEa	肯定の可能性を示唆	affirmative		224	14%
	ME0	可能性を保留	zero		158	10%
	MEm	否定の可能性を示唆	minus		21	1%
	MEn	確実な否定を示唆	no		269	16%
	MEp	過去を示唆	past		692	43%
	MEf	未来を示唆	future		59	4%
	MEh	仮定の話であることを示唆	hypothesized		150	9%
				合計	1,622	100%
事象クラス	EVe	事象であり，事実性付与が必須	event mention		3,092	62%
	EVc	事象可能	candidate		293	6%
	EVf	機能語につき，事象ではない	functional		761	15%
	EVi	時間軸上にない概念を指示	imaginary		707	14%
	EVs	比況 (明喩，暗喩)	simile		4	0%
	EVa	希望，依頼，命令，勧誘	action		39	1%
	EVq	疑問，問いかけ	question		111	2%
	EVp	許可	permission		7	0%
				合計	5,014	100%
事実性	FPc	肯定 (Positive) かつ確定	certain	CT+	2,646	86%
	FPr	肯定かつ高い確信度	probable	PR+	233	7%
	FPs	肯定かつ低い確信度	possible	PS+	35	1%
	FNc	否定 (Negative) かつ確定	certain	CT−	140	5%
	FNr	否定かつ高い確信度	probable	PR−	34	1%
	FNs	否定かつ低い確信度	possible	PS−	4	0%
				合計	3,092	100%

(31) 歩が入れば$_{\text{MEh}}$ △１六歩から攻め掛かれる．
(32) ▲５五歩は△同歩▲３六角成△８四銀が一例$_{\text{MEh}}$．

モダリティ表現は複数形態素であることがあるため，IOB2 フォーマット [16] によりラベル付けする．それゆえ，実際にテキストに付与するのは，MEy-B，MEy-I，MEa-B，MEa-I のような形式のラベルである．

4.2 事象クラス

事象に付与する事象クラスラベルの一覧を，表 2 の中部に示す．事象クラスラベルは，大きく 2 つのグループに別れる．

事象らしさ関連 事実性ラベルを付けるべき述語句かどうか (5 種類)

EVe，EVc，EVf，EVi，EVs

態度関連 断定と推量を除く，書き手の態度 (3 種類)

EVa，EVq，EVp

これらのクラスを導入する主な理由は，事実性ラベル付与が必須である事象 (EVe) とそうでない事象を事前に区別することにある．

3.2 節で言及したように，本研究では，松吉らの「態度」分類 [27] を上のように再整理した．具体的には，既存のクラス (EVe，EVa，EVq，EVp) に，新しいクラス (EVc，EVf，EVi，EVs) を追加した．松吉らのコーパスでは，EVe，EVa，EVq，EVp のいずれもが付与されていないことにより，EVc，EVf，EVi，EVs のいずれか相当であることが暗示されるのみである．英語において，FactBank のアノテーション指針 [17] には，EVi と EVs の場合を考慮するよう指示があるが，アノテーション結果においてはそのことは陽に反映されないため，真偽が不明の場合と同じラベルが付与される．

モダリティ解析器・事実性解析器を作成する場合，まずは，入力されたテキストから述語項構造をすべて抽出し，

続いて，それぞれの述語項構造が事実性ラベル付与が必須である事象かどうか判定する必要がある．EVf，EVi，EVsは，この見通しをよくするために導入したものである．

以下，例とともに各ラベルについて説明する．

EVe　断定や推量の事象．事実性ラベルの付与が必須

(33)　歩を成り捨て$_{\mathrm{EVe}}$ た．

(34)　銀交換せずに引き揚げる$_{\mathrm{EVe}}$．

EVc　対象述語が複合名詞の一部になっているもの．または，対象述語が別の述語句の修飾要素になっているもの．本動詞としても解釈できる機能動詞や複合辞内の述語．これらは，必要に応じて，事象と解釈することが可能

(35)　遊び$_{\mathrm{EVc}}$ 駒を竜にぶつけた．

(36)　第４局は途中までかなり優勢$_{\mathrm{EVc}}$ に進めている．

(37)　少し後手が勝っていそうな気がし$_{\mathrm{EVc}}$ ます．

EVf　機能語，もしくは，複合辞の一部である．事象としての解釈が全くないもの

(38)　封じ手につい$_{\mathrm{EVf}}$ て検討が進んでいる．

(39)　この試合では居飛車を採用するかもしれ$_{\mathrm{EVf}}$ ない．

EVi　時間軸上に接地できない概念を指す．主なものは次の 2 種類である．仮定された事象．(事象トークンではなく) 事象タイプを指すもの．

(40)　ここで△１四歩と受け$_{\mathrm{EVi}}$ れば先手はつらい$_{\mathrm{EVi}}$．

(41)　と金の攻め$_{\mathrm{EVi}}$ はコストが低い．

EVs　明喩や暗喩

(42)　３四の銀をあざ笑$_{\mathrm{EVs}}$ うかのように玉を進行させる．

EVa　希望，依頼，命令，勧誘など．対象述語の主語により EVa の事象は細分可能である．例えば，主語が態度表明者に等しいならば，「希望」である．主語が態度表明者でないならば，他者への働きかけ (「依頼」や「命令」) である

(43)　多くのファンに楽し$_{\mathrm{EVa}}$ んでもらいたい．

EVq　疑問や問いかけ

(44)　先手は桂を取る$_{\mathrm{EVq}}$ か．

EVp　許可

(45)　サイン会だけの参加$_{\mathrm{EVp}}$ も可能です．

含意関係認識タスクでは，あるテキスト A の内容から別のテキスト B の内容が正しいことが推測できるかどうかが問われる．通常，複合名詞の構成要素である述語や複合辞内の述語は，事象であるとみなされない．しかしながら，テキスト B にそれらの述語が本動詞として出現する場合，テキスト A 内においてそのような述語を事象に格上げすることは有用である．例えば，テキスト B が「山田は対戦している．」の場合，テキスト A「対戦者の山田は ～．」において，複合名詞内の「対戦」を事象とみなすことは含意関係認識のために必須である．応用を考慮して，必要に応じて事象集合の範囲を増やすことができるように，EVc を導入した．

情報検索においては，目的に応じて適合率や再現率が重要視される．事象を検索するにあたり，再現率を優先するならば，EVf を除くすべての事象が抽出されることが望ましい．一方，適合率を優先するならば，EVf，EVc，EVs，EVq などを抽出する必要はないと思われる．本論文で提案する事象クラスを利用することにより，必要に応じて抽出すべき事象集合を制御できるようになると思われる．

先行研究 [17], [27] に従い，事象クラスのラベルは，事象を構成する述語 1 形態素のみに付与する．2.2 節の例でも見られるように，サ変名詞直後の「する」や名詞・形状詞直後の「だ」は省略されることが少なくないため，「する」「だ」の有無にかかわらず，サ変名詞や名詞・形状詞を事象クラスラベル付与の対象形態素とする．

4.3　事実性

前節においてラベル EVe が付与された事象に対してのみ，事実性のラベルを付与する．付与する事実性ラベルの一覧を，表 2 の下部に示す．

3.3 節で述べたように，本研究では，FactBank の事実性アノテーション体系 [17] を利用する．ただし，Uu ラベル (事実性は不確定) だけは利用しない．その理由は，Uu ラベルが付与されうる事象に対しては，事象クラスのアノテーションにおいて EVq や EVi などが付与されるからである．EVq や EVi は，対象事象の事実性が不確定であることを示唆するため，Uu を利用しなくても，FactBank との互換性は保たれる．

以下，例とともに各ラベルについて説明する．FactBank のオリジナルラベルとの対応を表 2 内に示す．

FPc　事象の成立を断定する

(46)　歩を成り捨て$_{\mathrm{FPc}}$ た．

FPr　事象の成立を高い確信度で推測する

(47)　おそらく△１四香が良好$_{\mathrm{FPr}}$．

FPs　確信度は低いが，事象の成立を推測する

(48)　この試合では居飛車を採用$_{\mathrm{FPs}}$ するかもしれない．

FNc　事象の不成立を断定する

(49)　角交換$_{\mathrm{FNc}}$ には応じなかった．

FNr　事象の不成立を高い確信度で推測する

(50)　穴熊に組む$_{\mathrm{FNr}}$ つもりはないだろう．

FNs　確信度は低いが，事象の不成立を推測する

(51)　△９四歩は指し$_{\mathrm{FNs}}$ づらいかもしれません．

5.　アノテーションコーパス

この章では，実際のモダリティ情報アノテーション作業について説明し，コーパスの統計情報を報告する．

5.1　アノテーション作業

4 章において説明したアノテーション体系に従い，2.3 節

で紹介した SGC コーパスにモダリティ情報を人手でアノテーションした．アノテーション対象は，将棋解説文 2,041 文である．

アノテーションは，モダリティ表現，事象クラス，事実性の順に行った．作業者は 1 名であり，各層のアノテーションにかかった時間は，モダリティ表現 430 分，事象クラス 750 分，事実性 250 分であった．

5.2 統計情報

SGC コーパス内のラベルの分布を表 2 の右側に示す．

将棋解説文 2,041 文に，モダリティ表現は 1,622 個存在した．一番多く出現したのは，MEp が付与された，完了・過去の助動詞「た」である．否定辞「ない」や「ず」など，MEn が付与された表現は，269 個存在した．推量を表す MEa の表現は 224 個出現した．事前に予期していたことではあるが，MEa の表現は，他のラベルに比べ，表現の多様性が高いことが確認された．MEh のほぼすべては，接続助詞「ば」，「たら」，「と」と，「一例」や「順がある」など，将棋分野に特有の用語であった．

事象クラスラベルは，合計 5,014 個あった．これは，1 文あたり平均 2.5 個の事象クラスラベルが存在することを意味する．断定や推量を表す EVe の事象は 3,092 個あり，これは，ラベル全体の 62%を占める．機能語相当である EVf は 761 個存在した．これはラベル全体の 15%であり，日本語において事象 (述語項構造) を処理する際に，機能語相当表現を軽視せず適切に処理することが必要であることを示唆している．本コーパスの事象クラスラベル分布と先行研究のコーパス [27] の「態度」ラベルの分布を比較する．4.2 節で述べたように先行研究は EVc，EVf，EVi，EVs を扱っていないため，直接的な比較はできないが，これら以外のラベルの割合を比較すると，本コーパスは，新聞や書籍に近い分布を持っていることが確認できる．

事実性ラベルの総数は，EVe の事象の数と等しいため，3,092 個である．「肯定かつ断定」の FPc は 2,646 個であり，全体の 86%を占める．「否定かつ断定」の FNc は 140 個であり，全体の 5%であった．「肯定かつ推量」の FPr は 233 個であり，全体の 7%であった．将棋解説文には解説者の予想が多く含まれる印象が強いが，FPr の数は予期していたほど多くはなかった．FPr の数が少ない理由は，解説者の予想は，例文 (31)，(40) に見られるように仮定の構文や，例文 (12)，(13)，(32) に見られるように特有の用語を伴った断定の構文においても表現されうるからである．これらの構文を利用した予測・推量は，汎用的な事実性解析ではうまく捉えられない．将棋に特有の固有表現の認識器とモダリティ表現認識器を用いて，このような予測・推量に特別な配慮をする必要がある．

本コーパスの事実性ラベル分布と先行研究のコーパス [27] の「真偽判断」ラベルの分布を比較する．4.3 節で述べたように本研究では事実性の層において Uu を扱っていないため，直接的な比較はできないが，これら以外のラベルの割合を比較すると，本コーパスは，新聞や Yahoo!知恵袋に近い分布を持っていることが確認できる．

5.3 コーパスの配布

本研究で作成したアノテーション済みコーパス[*5]は，希望者に無償で配布する予定である．配布ファイルの詳細については，我々のウェブサイト[*6]にて確認いただきたい．

6. 本コーパスの応用

この章では，本研究で作成したコーパスの応用について考察する．

6.1 モダリティ表現認識

4.1 節にて例示したようなモダリティ表現を認識するシステムの構築に本コーパスは直接応用可能である．

本コーパスの元となった将棋解説文コーパスは 6,523 の対局を収録し，744,327 文の各々に対して対応する局面データが存在する．局面データとともにテキストデータを学習することにより，アノテーション済みコーパスには出現しないようなモダリティ表現を認識できる可能性がある．

6.2 事象クラス解析および事実性解析

4.2 節で述べた事象クラス，および，4.3 節の事実性を解析するシステムの構築に本コーパスは直接応用可能である．日本語における先行研究である拡張モダリティの体系 [27] は複雑であり，6 つの項目間の依存関係を考慮しながらその解析器を設計するのは非常に困難である．一方，本研究におけるアノテーション体系は，解析器構築のことも考慮して設計されている．例えば，事実性解析器を構築する場合，上の 2 つの層は無視し，3 つ目の層の事実性ラベルのみを利用することが可能である．同様に，事象クラス解析器を構築する場合は，2 つ目の層の事象クラスのラベルのみ利用すればよい．

SGC コーパスには，将棋特有の固有表現のラベルも付与されている．それゆえ，事象クラスや事実性の解析において，固有表現の情報も利用可能である．局面データも利用可能であり，必要ならば，現在の手までの局面履歴データも利用可能である．局面データ利用のもと，事象クラス・事実性と固有表現を同時に解析することにより，これらの解析精度をお互いに高め合うことができる可能性がある．

[*5] 我々が配布するのは，文字列位置情報が付いたラベル列のみである．将棋解説文のテキストおよび局面データは，別途 http://www.meijinsen.jp より有償で入手する必要がある．このサイトのアカウント情報を利用して上記データのダウンロードを支援するツールを https://github.com/hkmk/shogi-comment-tools にて公開している．

[*6] http://plata.ar.media.kyoto-u.ac.jp/data/game/

6.3 将棋解説文の自動生成

モダリティ情報が付与された本コーパスを利用することにより，先行研究の解説文自動生成手法 [5], [7] を改善できると思われる．先行研究では，その局面に特徴的な語を特定することにより，解説文を自動生成していた．より良い解説文を生成するために，単純な語ではなく，固有表現やモダリティ表現が利用可能であり，同時に事象の事実性解析結果も参照可能である．この応用にあたり，自動的に生成されたテンプレートに基づくテキスト生成 [11], [14] や，対話行為の代わりに固有表現を使用した深層学習に基づくテキスト生成 [23] を利用することができる．

局面に関する解説生成・質問生成の先行研究 [29], [30] において，人間らしい解説を行う上で，最善手順を予測して述べるだけでなく，実際には指されることはないが，特別な性質を持った手への言及も必要であることが示唆されている．事実性ラベルと局面データを合わせて学習することにより，このような言及を自動的に特定できる可能性がある．

6.4 将棋局面検索

キーワードではなく，自然言語文を用いて将棋局面を検索するシステムの構築に本コーパスは応用可能である．先行研究では，駒の配置や戦型などのキーワードによる局面検索が提案されている [2], [21]．局面データと事象の事実性の間の関連性が学習されれば，次のような自然言語文を局面検索の入力として受け付けることができる可能性がある．

- 今後 銀が活躍する局面
- 穴熊を控えるべき局面

前者の検索では，銀が活躍できると予測できた局面を返すことが求められる．後者の検索では，穴熊が選択可能であるが，実際には選択されなかった局面を返すことが求められる．これらの検索には，推量や否定といった事実性が強く関連している．

6.5 シンボルグラウンディング

本コーパスの応用において最も興味深いものは，シンボルグラウンディングである．これまで本章で述べてきた種々の応用は，モダリティ表現を局面データの中にグラウンディングすることを間接的に含んでいる．具体物を表す名詞や具体的な動作を表す動詞を画像や映像の中にグラウンディングする (対応物を見つける) ことは，直感的で分かりやすいシンボルグラウンディングの例である．一方，画像や映像，その他の非テキストデータの中に，機能語であるモダリティ表現をグラウンディングする方法を見つけることは，挑戦的で未解決の問題である．対応する局面データを持ったテキストに対して，表 1 のように 5 層のアノテーションを施したコーパスを構築することにより，おそらく世界で初めてこの問題に挑戦できる基盤ができたと言ってもよいかもしれない．

将棋のゲーム木を可能世界の集合とみなすと，様相論理 [10] の応用により，いくつかの典型的なモダリティ表現をシンボルグラウンディングできる見込みがある．

各言語におけるモダリティ表現のグラウンディングが可能になれば，先行研究の手法 [8] を用いて，モダリティ表現に関して 2 言語間の対訳辞書を自動構築できる可能性がある．

7. おわりに

本論文では，日本語モダリティ情報のアノテーション体系を提案し，その体系に従って，将棋解説文コーパスに対して，モダリティ表現，事象クラス，事実性のラベルを付与した．構築したコーパスは，1,622 のモダリティ表現，5,014 の事象クラスラベル，3,092 の事実性ラベルを含む．

今後の課題は大きく 2 つある．1 つは，6 章で述べたように，固有表現や局面データを利用しつつ，モダリティ表現や事実性を解析するシステムを構築することである．

もう 1 つは，先行研究 [27] と同じ対象である『現代日本語書き言葉均衡コーパス』[*7] に対して，本論文で提案するアノテーションを適用することである．自動変換することにより，すでに付与されているラベルの大部分を利用できるので，他の生テキストを対象とするよりも早く大規模なアノテーション済みコーパスが構築できることが期待される．

参考文献

[1] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J. and Forsyth, D.: Every Picture Tells a Story: Generating Sentences from Images, *Proc. of the ECCV10*, pp. 15–29 (2010).
[2] Ganguly, D., Leveling, J. and Jones, G. J.: Retrieval of Similar Chess Positions, *Proc. of the SIGIR14*, ACM, pp. 687–696 (2014).
[3] Harnad, S.: The Symbol Grounding Problem, *Physica D*, Vol. 42, pp. 335–346 (1990).
[4] Izumi, T., Imamura, K., Asami, T., Saito, K., Kikui, G. and Sato, S.: Normalizing Complex Functional Expressions in Japanese Predicates: Linguistically-Directed Rule-Based Paraphrasing and Its Application, *ACM Transactions on Asian Language Information Processing*, Vol. 12, No. 3, pp. 1–20 (2013).
[5] Kameko, H., Mori, S. and Tsuruoka, Y.: Learning a Game Commentary Generator with Grounded Move Expressions, *Proc. of the CIG15* (2015).
[6] Kamioka, Y., Narita, K., Mizuno, J., Kanno, M. and Inui, K.: Semantic Annotation of Japanese Functional Expressions and its Impact on Factuality Analysis, *Proceedings of The 9th Linguistic Annotation Workshop*, pp. 52–61 (2015).
[7] Kaneko, T.: Real Time Commentary System for Shogi, *First Workshop on Games and NLP* (2012).
[8] Kiela, D., Vulić, I. and Clark, S.: Visual Bilingual Lexi-

*7 http://pj.ninjal.ac.jp/corpus_center/bccwj/

con Induction with Transferred ConvNet Features, *Proc. of the 2015 EMNLP*, pp. 148–158 (2015).
[9] Kiros, R., Salakhutdinov, R. and Zemel, R.: Multimodal Neural Language Models, *Proceedings of the 31st International Conference on Machine Learn ing*, pp. 595–603 (2014).
[10] Kripke, S. A.: Semantical Considerations on Modal Logic, *Acta Philosophica Fennica*, Vol. 16, pp. 83–94 (1963).
[11] Mori, S., Maeta, H., Sasada, T., Yoshino, K., Hashimoto, A., Funatomi, T. and Yamakata, Y.: FlowGraph2Text: Automatic Sentence Skeleton Compilation for Procedural Text Generation, *Proc. of the INLG14*, pp. 118–122 (2014).
[12] Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H. and Tsuruoka, Y.: A Japanese Chess Commentary Corpus, *Proc. of the LREC16* (2016).
[13] Palmer, F.: *Mood and Modality Second edition*, Cambridge University Press (2001).
[14] Reiter, E.: NLG vs. Templates, *Proc. of the EWNLG95*, pp. 147–151 (1995).
[15] Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M. and Schiele, B.: Translating Video Content to Natural Language Descriptions, *Proc. of the ICCV13* (2013).
[16] Sang, E. F. T. K. and Meulder, F. D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proc. of the CoNLL2003*, pp. 142–147 (2003).
[17] Sauri, R.: *FactBank 1.0 Annotation Guidelines* (2008).
[18] Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J.: *TimeML Annotation Guidelines Version 1.2.1* (2006).
[19] Suzuki, T., Abe, Y., Toyota, I., Utsuro, T., Matsuyoshi, S. and Tsuchiya, M.: Detecting Japanese Compound Functional Expressions using Canonical/Derivational Relation, *International Conference on Language Resources and Evaluation* (2012).
[20] Tsuruoka, Y., Yokoyama, D. and Chikayama, T.: Game-Tree Search Algorithm Based On Realization Probability, *ICGA Journal*, Vol. 25, No. 3, pp. 145–152 (2002).
[21] Ushiku, A., Mori, S., Kameko, H. and Tsuruoka, Y.: Game State Retrieval with Keyword Queries, *SIGIR* (2017).
[22] Ushiku, Y., Harada, T. and Kuniyoshi, Y.: Automatic Sentence Generation from Images, *Proc. of the ACMMM11*, pp. 1533–1536 (2011).
[23] Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D. and Young, S.: Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems, *Proc. of the 2015 EMNLP*, pp. 207–213 (2015).
[24] Wiebe, J., Wilson, T. and Cardie, C.: Annotating Expressions of Opinions and Emotions in Language, *Language resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210 (2005).
[25] Yang, Y., Teo, C. L., III, H. D. and Aloimonos, Y.: Corpus-Guided Sentence Generation of Natural Images, *Proc. of the 2011 EMNLP* (2011).
[26] 首藤公昭，田辺利文：日本語複単語表現辞書：JDMWE，自然言語処理， Vol. 17, No. 5, pp. 51–74 (2010).
[27] 松吉　俊，江口　萌，佐尾ちとせ，村上浩司，乾健太郎，松本裕治：テキスト情報分析のための判断情報アノテーション，電子情報通信学会論文誌 D 情報・システム，Vol. 93, No. 6, pp. 705–713 (2010).
[28] 松吉　俊，佐藤理史，宇津呂武仁：日本語機能表現辞書の編纂，自然言語処理，Vol. 14, No. 5, pp. 123–146 (2007).
[29] 金子知適：コンピュータ将棋を用いた棋譜の自動解説と評価，情報処理， Vol. 53, No. 11, pp. 2525–2532 (2012).
[30] 小川直希，石脇滉己，荒川達也：詰将棋大盤解説聞き手エージェントのための質問自動生成の提案，ゲームプログラミングワークショップ 2015 論文集，pp. 40–45 (2015).
[31] 亀甲博貴，森　信介，鶴岡慶雅：実現確率に基づく解説すべき指し手の推定，第 21 回ゲームプログラミングワークショップ，pp. 28–35 (2016).
[32] 亀甲博貴，三輪　誠，鶴岡慶雅，森　信介，近山　隆：対数線形言語モデルを用いた将棋解説文の自動生成，情処論，Vol. 55, No. 11, pp. 2431–2440 (2014).
[33] 益岡隆志：モダリティの文法，くろしお出版 (1991).