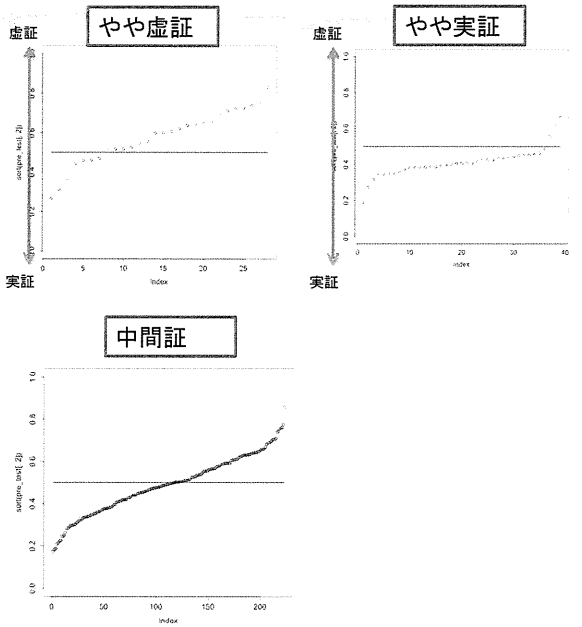
テストデータ　やや虚証　やや実証

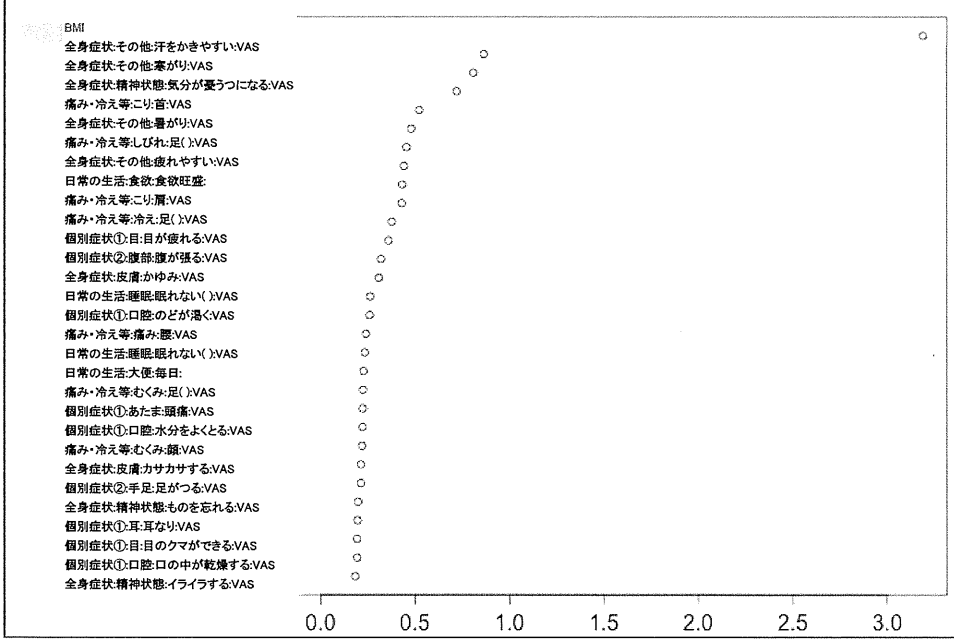
	実証 やや実証	虚証 やや虚証
予測で虚証	21	3
予測で実証	7	36
計	28	39

判別率　85.1%

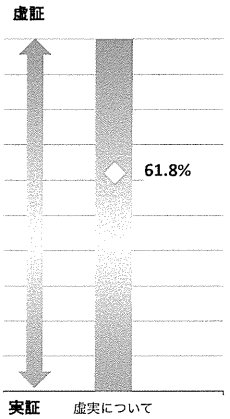
	実証 中間証
予測で虚証	105
予測で実証	118
計	223



BMIを入れたRFにおける変数の重要度



NHK での放送



虚証患者20、実証患者20のデータにより構築した
虚実予測方式により番号3368の患者の虚実を
予測する。

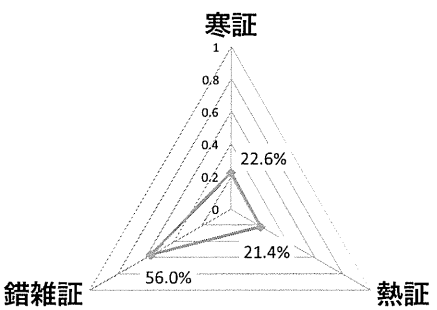
結果
虚証度61.8%

結論
中間証からやや虚証の傾向があると予測される。

渡辺先生の診断

15

NHK での放送

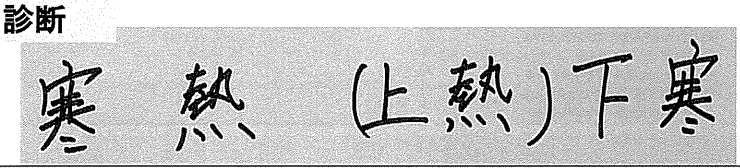


寒証患者70人、熱証患者70人、
錯雑証患者40人のデータにより
構築した寒熱予測方式により
番号3368の患者の寒熱を予測する。

結果
寒証度22.6%
熱証度21.4%
錯雑証度56.0%

結論
錯雑証の傾向があると予測される。

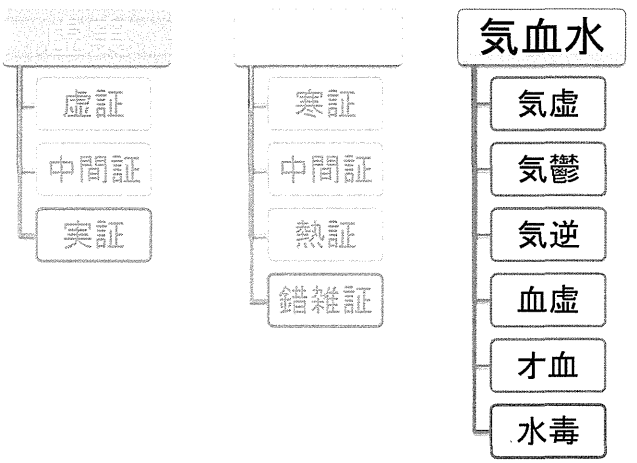
渡辺先生の診断



16

虚実

◆気血水について



17

ランダムフォレストによる気血水予測

	判別率
気虚	49.30%
気鬱	53.20%
気逆	45.00%
血虚	48.50%
お血	49.00%
水毒	52.20%

18

Terasawa Score

1）気虚

気虚スコアー					
症　候	配点	得点	症　候	配点	得点
身体がだるい	10	(　)	眼光・音声に力がない	6	(　)
気力がない	10	(　)	舌が淡白紅・腫大	8	(　)
疲れやすい	10	(　)	脈が弱い	8	(　)
日中の眠気	6	(　)	腹力が軟弱	8	(　)
食欲不振	4	(　)	内臓のアトニー症状 1)	10	(　)
風邪をひき易い	8	(　)	小腹不仁 2)	6	(　)
物事に驚き易い	4	(　)	下痢傾向	4	(　)

判定基準　　気虚スコアー（総計）

総計30点以上を気虚とする。いずれも顕著に認められるものに該当するスコアーを全点与え、程度の軽いものには各々の1/2を与える。

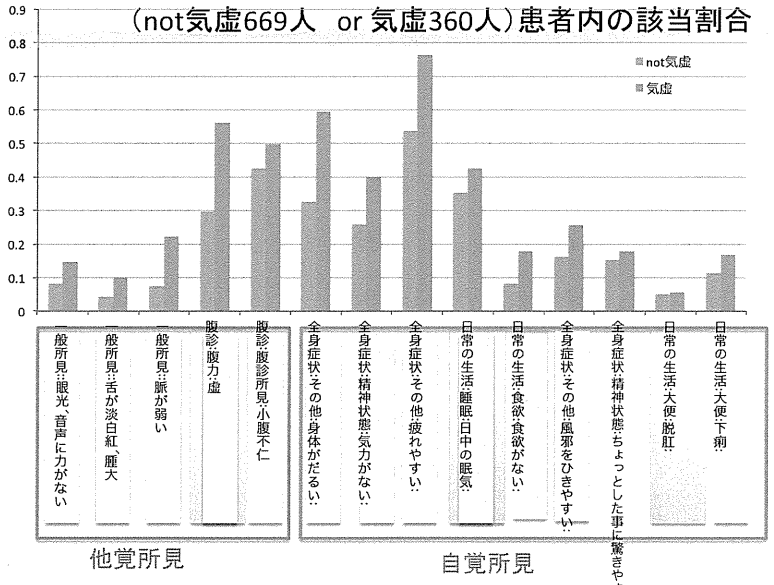
注1）内臓のアトニー症状とは胃下垂、腎下垂、子宮脱、脱肛などをいう。

注2）小腹不仁とは臍下部の腹壁トーヌスの低下である。

19

千葉　気虚項目回答傾向 統合版

（not気虚669人　or 気虚360人）患者内の該当割合



気虚スコアではor条件として扱われる項目をマージしたもの

帰無仮説:気虚患者内での該当割合とnot気虚患者内での該当割合は等しいとしたときに，5%水準で棄却されたもの（カイ二乗検定）

Ⅳ. 学 会 等 発 表 実 績

研究発表

１．論文発表

1. 吉野鉄大, 堀場裕子, 牧田和也, 他. 当院漢方医学センター外来の月経困難症患者の特徴と頻用処方予測モデルの提案. 産婦人科漢方研究のあゆみ. 2014; 04: 80-84.
2. Uematsu, Sumire, Matsuzaki, Takuya, Hanaoka, HirokiMiyao, Yusuke, Mima, Hideki, Integrating Multiple Dependency Corpora for Inducing Wide-Coverage Japanese CCG Resources, ACM Trans. Asian Low-Resour. Lang. Inf. Process. Vol. 14. Num. 1. January 2015.
3. S. Yakubo, M. Ito, Y. Ueda, H. Okamoto, Y. Kimura, Y. Amano, T. Togo, H. Adachi,T. Mitsuma, and K.Watanabe : Pattern Classification in Kampo Medicine. Evidence-Based Complementary and Alternative Medicine Volume 2014, Article ID 535146, 5 pages, http://dx.doi.org/10.1155/2014/535146.

２．学会発表

1. Yoshino T, Katayama K, Munakata K, 他. Kampo Traditional Pattern Diagnosis and the Clustering Analysis of Patients with Cold Sensation. The Journal of Alternative and Complementary Medicine. 2014; 20: A47-A47.
2. Yuko H, Tetsuhiro Y, Kenji W. Kampo Traditional Pattern Diagnosis and the Clustering Analysis of Patients with Insomnia. The Journal of Alternative and Complementary Medicine. 2014; 20: A47-A47.
3. 吉野鉄大, 片山琴絵, 堀場裕子, 他. 漢方自動問診システムを用いた寒熱診断の予測. 第 31 回和漢医薬学会学術集会. 2014.
4. 堀場裕子, 吉野鉄大, 木村容子, 他. 共通のプラットフォームを用いた漢方診断の施設間比較. 第 31 回和漢医薬学会学術集会. 2014.
5. 吉野鉄大, 片山琴絵, 堀場裕子, 他. 冷えを訴える患者に対する処方をロジスティック回帰分析により予測するモデルの提案. 第 64 回日本東洋医学会学術総会. 2014.
6. 堀場裕子, 吉野鉄大, 渡辺賢治. 当院外来における不眠を訴える患者の問診回答状況による分類. 第 64 回日本東洋医学会学術総会. 2014.
7. 吉野鉄大, 牧田和也, 堀場裕子, 他. 月経困難症に対して頻用される当帰芍薬散と桂枝茯苓丸の使い分けを支援する統計的モデルの検証. 第 34 回産婦人科漢方研究会学術集会. 2014.
8. テキスト情報の解析と可視化に向けた高被覆日本語 CCG 構文解析の実現, 植松すみれ,美馬秀樹, 可視化情報学会第 42 回可視化情報シンポジウム 2014 年

H．知的財産権の出願・登録状況

該当なし

V. 研究成果の刊行物・別刷

論文②

当院漢方医学センター外来の月経困難症患者の特徴と頻用処方予測モデルの提案

慶應義塾大学医学部漢方医学センター[1)] 慶應義塾大学医学部産婦人科学教室[2)] 牧田産婦人科医院[3)]
東京大学医科学研究所ヒトゲノム解析センター[4)] 慶應義塾大学SFC研究所[5)]
吉野 鉄大[1)] 堀場 裕子[2)] 牧田 和也[2)3)] 片山 琴絵[4)] 宗形 佳織[5)]
山口 類[4)] 井元 清哉[4)] 宮野 悟[4)] 渡辺 賢治[1)]

はじめに

月経困難症は，月経時あるいはその直前から始まる強い下腹痛や腰痛(いわゆる月経痛)のために，日常生活を営むことが困難な状態を指す．月経痛に対する非ステロイド性抗炎症剤の投与という対症療法以外に，月経困難症の代表的な西洋医学的治療として，低用量ピルと偽閉経療法が挙げられる．しかしながら低用量ピルには，嘔気や頭痛の増悪といった副作用とともに，血栓性素因を有する者，乳癌の治療歴のある者，前兆のある片頭痛患者など慎重投与例ないし投与禁忌例も多い．また，偽閉経療法には，低エストロゲン血症に由来する諸症状のために治療を継続できないことが少なくない．いずれの治療法も排卵を抑制するため，妊娠を希望する性成熟期女性患者には選択できないという大きなデメリットがある．そこで，排卵を抑制せず月経困難症の症状を改善することを目指す漢方治療が選択される．月経困難症は婦人科のみならず漢方クリニックへの受診理由としても頻度が高い病態の1つである[1)]．月経困難症の漢方治療では，三大婦人薬と称される当帰芍薬散，加味逍遙散，桂枝茯苓丸に加えて，桃核承気湯や温経湯など，駆瘀血作用をもつ漢方製剤が用いられることが多いが，その使い分けは漢方を学び始めたばかりの医師にとって難しい．

当院漢方医学センターでは，タッチパネルを使用した自動問診システムを導入し，その結果の解析を進めている．今回われわれは，月経困難症と診断された初診患者を対象とし，患者の自覚症状である問診データ，漢方専門医の証診断，投薬状況の解析を行うことで月経困難症患者の特徴と漢方処方の状況を明らかにした．また漢方を専門としない医師が月経困難症患者に頻用される処方を問診と腹診により使い分けることができるよう，処方を予測するモデルを得ることを目的として，ロジスティック回帰分析によるモデル構築を試みたので報告する．

I 対象と方法

対象は，2008年5月から2013年3月までに慶應義塾大学病院漢方医学センター外来を受診した女性初診患者で，現在月経があり，研究参加に同意が得られたものとした．なお，年齢・性別・月経周期・出血期間の入力に欠損がある患者，60歳以上の患者，症状の入力がない患者は，今回の解析対象から除外した．

問診項目は合計128項目あり，そのうち106項目はVisual Analogue Scale(VAS)でその程度を評価している．なお，VAS以外の問診項目のうち7項目は，例えば経血量について「月経量が多いかどうか」と「月経量が少ないかどうか」を別々に問うているため，実際の問診項目数は135となっている．腹診所見は13項目あり，腹力は虚，中間，実の3段階とし，小腹不仁と小腹拘急は小腹不仁の腹証，瘀血の圧痛と小腹鞕満と少腹急結は瘀血の腹証としてまとめて扱った．漢方診断につ

いては，虚実は虚証，やや虚証，中間証，やや実証，実証の5カテゴリ，寒熱は寒証，中間証，熱証，錯雑症の4カテゴリとし，互いに排他なものとした．また，気血水は気虚，気うつ気滞，気逆，血虚，瘀血，水毒，亡津液の7カテゴリとし，重複を許した．

解析対象データを，月経困難症を病名に持つ患者(D群)と月経困難症を病名にもたない有経患者(ND群)で分類し，統計解析を行い比較検討した．検討項目は，年齢，併存疾患，問診項目への回答状況，虚実・寒熱・気血水などの漢方診断，漢方薬の処方頻度とした．処方についてはD群全体において10人以上に処方されているものを対象とし，エキス剤と煎じ薬は同一のものとし，抑肝散加陳皮半夏は抑肝散として，また桂枝茯苓丸加薏苡仁は桂枝茯苓丸として扱った．

さらに処方については，処方頻度が高かった当帰芍薬散か桂枝茯苓丸のいずれかが処方された患者を対象に，どちらの薬が処方されるのかを予測する回帰式を構築した．当帰芍薬散と桂枝茯苓丸が併用された患者は除外した．説明変数として問診項目と腹診所見の有無を採用し，桂枝茯苓丸が処方されたのか当帰芍薬散が処方されたのかを目的変数として，ロジスティック回帰分析を適用した．説明変数として採用した問診項目は，桂枝茯苓丸が処方された患者の回答率を当帰芍薬散が処方された患者の回答率で除することで得た陽性尤度比で0.8以下か1.2以上の30項目であった．次に，回帰式の赤池情報量規準が最小となるように変数選択を行い，最良回帰式を推定した．

統計解析ソフトウェアとしては，R version 2.15.2(2012年10月26日版)を使用した．年齢，月経周期，出血期間，症状数，症状のVASの比較についてはWilcoxonの順位和検定を用い，それ以外に行った月経困難症の有無，もしくは当帰芍薬散が処方された患者と桂枝茯苓丸が処方された患者での2群間の頻度の比較については比率の差の検定を用いた．有意水準は，5%を採用した．

Ⅱ 結 果

1. 基礎情報

研究参加に同意が得られた女性初診患者4,057例のうち，解析対象患者は1,142例であった．月経困難症と診断された患者(D群)は，238例(平均年齢33.7±1.0歳，平均月経周期28.9±0.6日，平均出血期間5.8±0.2日，平均症状数26.3±1.5個)で，月経困難症の診断のない患者(ND群)は904例(平均年齢35.6±0.6歳，平均月経周期29.3±0.4日，平均出血期間5.6±0.1日，平均症状数24.6±0.9個)であった．両群間で年齢のみ有意差を認めた．また，D群の8.0%，ND群の3.3%に子宮内膜症が合併し，有意差を認めた．

2. 症状

月経困難症患者の症状についてD群とND群を比較し，D群に有意に多かった項目を**図1**に示す．反対に，「ものを忘れる」「分娩」「経血量が少ない」「げっぷ」「手のしびれ」「胸痛」「足に力が入らない」はD群に有意に少なかった．このなかで患者の入力した症状のVASについて，D群の方が有意に高値であった項目は，「月経痛」と「膝の痛み」であり，「頭痛」は有意に低値であった．

3. 漢方医学的診断

月経困難症患者の漢方医学的診断についてD群とND群を比較すると，D群に「寒証」「瘀血」「水毒」が有意に多く，「気虚」「血虚」が有意に少なかった．

4. 処方の頻度と予測

処方の解析では，併用を含めて407処方を解析した．59.2%の患者で漢方薬が併用された．併用も含めた処方の頻度を**図2**に示すが，当帰芍薬散33.6%，桂枝茯苓丸29.0%，加味逍遙散11.3%で，以下，五苓散，抑肝散と続いた．一般に鎮痛目的で投与されることが多い安中散は8.0%，芍薬甘草湯は1.7%に投与されていた．

次いで，頻用処方である桂枝茯苓丸と当帰芍薬散のいずれかが処方された患者について問診，腹

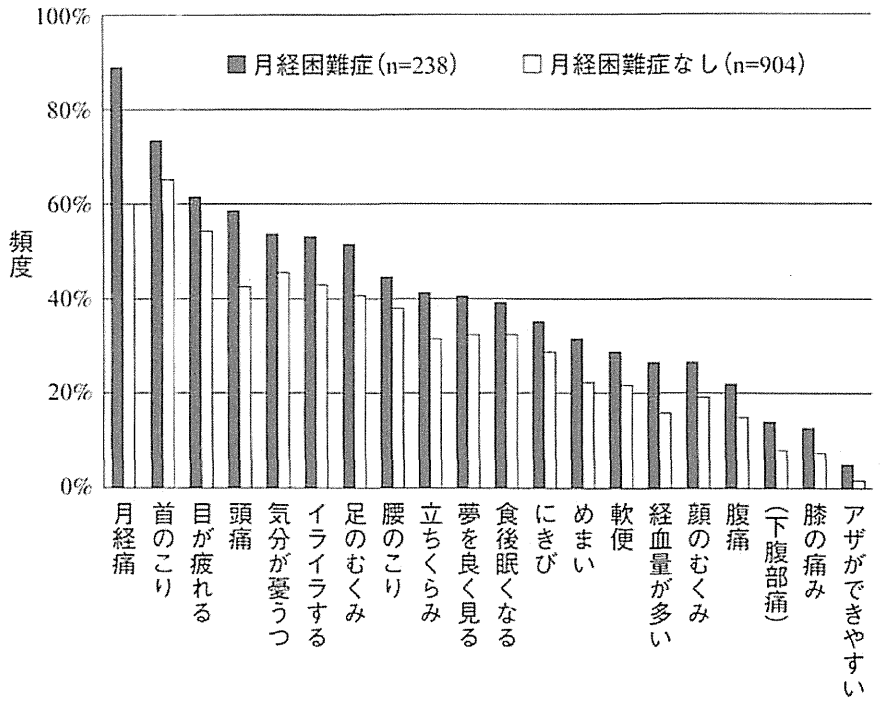
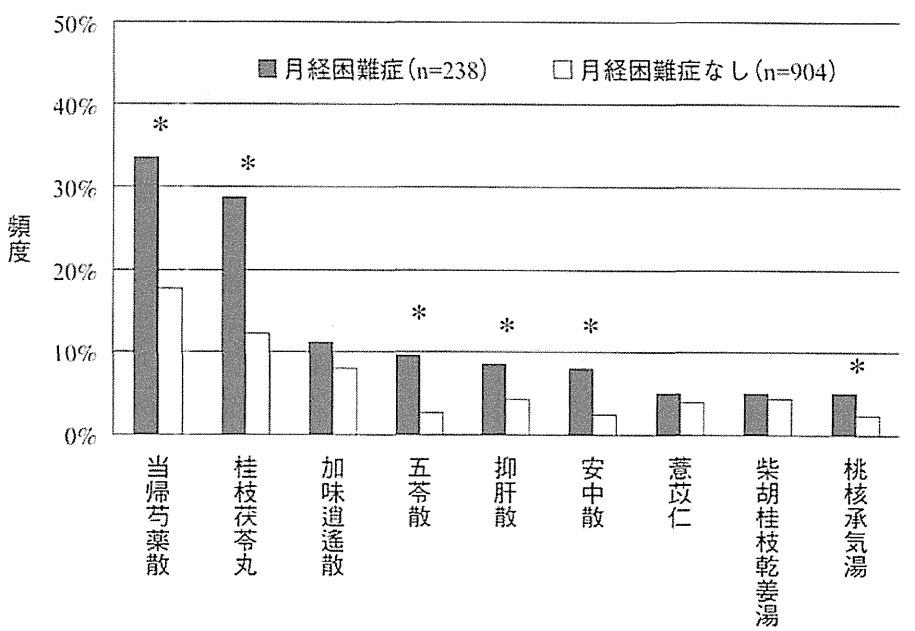


図 1　月経困難症の診断の有無による症状の頻度の違い

月経困難症患者の症状について月経困難症群と月経困難症なし群を比較し，比率の差の検定により月経困難症群に有意に多かった項目を抽出した．このなかで患者の入力した症状の VAS について，Wilcoxon の順位和検定により比較すると，「月経痛」と「膝の痛み」は月経困難症群が有意に高値であり，「頭痛」は有意に低値であった．



＊p-value<0.05(比率の差の検定)

図 2　月経困難症の診断の有無による処方の頻度の違い

月経困難症群において，59.2% の患者に漢方薬が併用された．当帰芍薬散が 33.6%，桂枝茯苓丸が 29.0%，加味逍遙散が 11.3% の患者に投与され，以下，五苓散，抑肝散と続いた．一般に鎮痛目的で投与されることが多い安中散は 8.0%，芍薬甘草湯は 1.7% の患者に投与されていた

表1　月経困難症患者に対する当帰芍薬散と桂枝茯苓丸の，ロジスティックモデルによる最良回帰式における係数と信頼区間

	係数	2.5%	97.5%
瘀血の腹証	2.6	1.3	4.2
腹力実	2.2	0.6	4.4
足の痛み	2.2	0.6	3.9
背中のこり	1.7	0.7	3.0
腰の冷え	1.4	-0.3	3.4
顔のほてり	1.4	0.1	2.8
のどが渇く	1.3	0.0	2.7
中途覚醒	0.9	-0.2	2.0
皮膚のかゆみ	-1.3	-2.5	-0.2
口の中が乾燥する	-1.6	-3.3	-0.1
切片	-1.8	-3.3	-0.6
立ちくらみ	-1.9	-3.3	-0.7
胸やけ	-2.0	-3.7	-0.5
腹力虚	-2.6	-4.1	-1.4

桂枝茯苓丸と当帰芍薬散のいずれかが処方された患者について，処方を目的変数とする回帰式を検討した．赤池情報量規準による変数選択により得た最良回帰式の係数と，95% 信頼区間を表に示す．最良回帰式について，leave-one-out 法による交差検定での医師の処方との一致率は 82.2% であった．

診所見を比較すると，「汗をかきやすい」「腰が冷える」「顔のほてり」「腹力実」「瘀血の腹証」が桂枝茯苓丸を処方された患者で有意に多く，「腹力虚」が当帰芍薬散を処方された患者で有意に多かった．漢方診断は，「やや実証」「実証」「錯雑証」「瘀血」が桂枝茯苓丸を処方された患者で有意に多く，「やや虚証」「虚証」「寒証」「水毒」が当帰芍薬散を処方された患者で有意に多かった．さらに処方が桂枝茯苓丸であるか当帰芍薬散であるかを予測するロジスティック回帰分析を構築した．問診 30 項目と腹診 13 項目を使用した変数選択前の回帰式の赤池情報量規準は 171.8 で，最良回帰式の赤池情報量規準は 128.4 となった．また，leave-one-out 法による交差検定での医師の処方との一致率は 82.2% であった．最良回帰式に採用された項目の係数と，その 95% 信頼区間を**表 1** に示す．

Ⅲ 考 察

当院漢方医学センター外来初診患者全体における病名の頻度をみると，冷え症，アトピー性皮膚炎，不眠症，便秘症に次いで月経困難症が 5 番目に多い．さらに月経のある女性に限れば，その 20.8% が月経困難症と診断されていた．過去の当研究会でも，漢方薬が適応とされる月経異常として，月経困難症がもっとも頻度の高い病名であることは報告されている[2)3)]．

今回の検討から，月経困難症と診断された患者において有意に頻度の高い，もしくは低い症状が抽出された．しかしながら，この検定は問診項目全てについてそれぞれに比率の差の検定を行っているため，問診項目全体に対する検定の α エラーが 5% を大幅に上回ってしまう．そこで，各々の検定について棄却水準を 5 / 135 = 0.037% とする Bonferroni の補正を行うと，「月経痛」「頭痛」「経血量が多い」が月経困難症と診断された患者において有意に頻度の高い症状となり，頻度の低い症状はなくなる．症状の程度の比較でも同様に有意水準として 5/106 = 0.047% を採用すると，「月経痛」のみが有意となった．これらの結果として，月経困難症と診断されている患者は，月経痛と頭痛と過多月経の頻度が高く，月経痛の程度が強いという，医師の一般常識に照らしても妥当な結果を得ることができると思われた．

次に，漢方診断についてみると，月経困難症と診断されている患者は，月経困難症と診断されていない患者と比較して寒証，瘀血，水毒と診断されている頻度が高かった．これは以前から指摘されている月経困難症の漢方医学的病態に矛盾しない[4)]．一方で，月経困難症と診断されている患者は気虚，血虚が少ないという結果になったが，月経困難症と診断されていない有経女性において，気虚や血虚がより重要となるアトピー性皮膚炎などの病名の頻度が高いためであり，月経困難症患者における気虚や血虚の重要性を否定するものではないと考える．

そして，月経困難症に対する処方の頻度は，当帰芍薬散と桂枝茯苓丸が特に高く，それに加味逍遙散が続いた．当帰芍薬散と桂枝茯苓丸は主に月

経痛を目標に使用される処方であるのに対して，加味逍遙散は月経前症候群を目標に使用されることが多い処方であるとされており[5]，われわれの結果も過去の報告に矛盾しないと考えられた．当帰芍薬散と桂枝茯苓丸の使い分けは虚実によるところが多いと考えられ，当研究会においても小山の質問表[6]を用いた虚実の判定をもとに検討した報告がなされている[7]．われわれの最良回帰式でも腹力の虚実は信頼できる説明変数であったが，この2処方に限らず，より様々な処方の予測に応用可能な方法を検討する必要があると考える．そこで，今回は様々な説明変数を導入でき，結果を直感的に理解しやすいロジスティックモデルを採用した．説明変数に腹診を入れず，問診だけで回帰式が作成できれば，対象を医師に限らないことも想定できるが，われわれのデータでは一致率が60%程度から上昇せず，今後の課題とした．

漢方専門医の判断には，今回検討していない舌診を含む望診や，脈診などの情報も重要であるが，漢方を学んだことがなく，また近くにその師を求めることができない環境で漢方治療を取り入れたいと考えている医師にとっては，今回提案したモデルは2つの効能があると考えられる．1つ目の効能は，取り組みやすい問診と腹診により月経困難症の代表処方である桂枝茯苓丸と当帰芍薬散を使い分けることを通して，虚実や気血水の概念を理解する一助になりうることである．桂枝茯苓丸は実証かつ錯雑証かつ瘀血証である患者に，当帰芍薬散は虚証かつ寒証かつ水毒証である患者に使用されることが多い処方であることが今回の結果からも確認されたが，それだけでは漢方医学の理論体系が体得できていない場合には理解が難しい．そこで，西洋医学でも行う問診と腹部診察から漢方独特の診察を導入したうえで，漢方医学独特の概念の理解や診察方法の修練に進めばスムーズに漢方医学を習得できる可能性がある．2つ目の効能は，漢方導入当初より，有効率を高めながら，安全に処方を運用しうることである．今回のモデルは漢方専門医の月経困難症症例に対する頻用二処方の使い分けについて高い一致率を得ており，「がっちりタイプの月経困難症の第一選択はこの処方」というような使用法よりも臨床に適用した場合の有効率や安全性が高い可能性がある．ただし，今回のモデルを用いた投薬後の患者の経過については，今後の慎重な評価が必要である．

おわりに

当科を受診する月経困難症患者の臨床像の一部を明らかにするとともに，頻用処方である桂枝茯苓丸と当帰芍薬散の使い分けについて問診と腹診により予測するモデルを提示した．今後はさらに対象処方を広げ，3つ以上の処方にも対応できるモデルを構築した上で，慎重に経過を評価し，モデルの妥当性について検討していきたいと考えている．

◆文献

1) 中村恵子，他：漢方薬服用患者における西洋薬の使用実態及び併用治療の意義に関する検討．薬事新報 2009；**1**：25-32
2) 田中栄一，他：月経異常に対する漢方治療について　有効例と無効例の検討．産婦漢方研のあゆみ 2002；**19**：35-37
3) 柴田健雄，他：当科における婦人科外来患者に対する漢方薬処方の実態調査．産婦漢方研のあゆみ 2012；**4**：44-47
4) 後山尚久：女性を悩ます痛みとそのケア　婦人科医に必要な最新情報．痛みと漢方．産婦人科治療 2010；**101**：119-122
5) 森裕紀子，他：ワンランク上の漢方診療．漢方処方の実際 月経困難症．臨床婦産 2012；**66**：62-66
6) 小山嵩夫：若年卵巣欠落に伴う不定愁訴と漢方療法．漢方と最新治療 1992；**1**：245-249
7) 武市和之，他：月経困難症に対する漢方薬(当帰芍薬散および桂枝茯苓丸と芍薬甘草湯併用)の有効性について．産婦漢方研のあゆみ 2007；**5**：24-28

Integrating Multiple Dependency Corpora for Inducing Wide-Coverage Japanese CCG Resources

SUMIRE UEMATSU, The University of Tokyo
TAKUYA MATSUZAKI, National Institute of Informatics
HIROKI HANAOKA, The University of Tokyo
YUSUKE MIYAO, National Institute of Informatics
HIDEKI MIMA, The University of Tokyo

A novel method to induce wide-coverage Combinatory Categorial Grammar (CCG) resources for Japanese is proposed in this article. For some languages including English, the availability of large annotated corpora and the development of data-based induction of lexicalized grammar have enabled deep parsing, i.e., parsing based on lexicalized grammars. However, deep parsing for Japanese has not been widely studied. This is mainly because most Japanese syntactic resources are represented in chunk-based dependency structures, while previous methods for inducing grammars are dependent on tree corpora. To translate syntactic information presented in chunk-based dependencies to phrase structures as accurately as possible, integration of annotation from multiple dependency-based corpora is proposed. Our method first integrates dependency structures and predicate-argument information and converts them into phrase structure trees. The trees are then transformed into CCG derivations in a similar way to previously proposed methods. The quality of the conversion is empirically evaluated in terms of the coverage of the obtained CCG lexicon and the accuracy of the parsing with the grammar. While the transforming process used in this study is specialized for Japanese, the framework of our method would be applicable to other languages for which dependency-based analysis has been regarded as more appropriate than phrase structure-based analysis due to morphosyntactic features.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—Language parsing and understanding

General Terms: Algorithms, Experimentation, Languages, Theory

Additional Key Words and Phrases: Combinatory Categorial Grammar, dependency annotation, grammar development, Japanese parsing

ACM Reference Format:
Uematsu, S., Matsuzaki, T., Hanaoka, H., Miyao, Y., and Mima, H. 2015. Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 14, 1, Article 1 (January 2015), 24 pages.
DOI: http://dx.doi.org/10.1145/2658997

1. INTRODUCTION

Syntactic parsing for Japanese has been dominated by a dependency-based pipeline architecture in which chunk-based dependency parsing is applied and then semantic role labeling is done on the dependencies [Hayashibe et al. 2011; Iida and Poesio 2011;

This article is an updated and extended version of Uematsu et al. [2013] published in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).
Authors' addresses: S. Uematsu and H. Mima, Center for Knowledge Structuring, The University of Tokyo; email: uematsu@cks.u-tokyo.ac.jp; T. Matsuzaki, Research Center for Community Knowledge, National Institute of Informatics; H. Hanaoka, The University of Tokyo; Y. Miyao, Digital Content and Media Sciences Research Division, National Institute of Informatics.

DOI: http://dx.doi.org/10.1145/2658997

Kawahara and Kurohashi 2011; Kudo and Matsumoto 2002; Sasano and Kurohashi 2011]. This dominance is mainly because chunk-based dependency analysis apparently seems most appropriate for Japanese syntax due to its morphosyntactic typology, which includes agglutination and scrambling [Bekki 2010]. However, it is also true that this type of analysis has prevented us from deeper analysis such as deep parsing [Clark and Curran 2007; Miyao and Tsujii 2008] and logical inference [Bos 2007; Bos et al. 2004].

In this article, we present our work on inducing wide-coverage Japanese resources based on Combinatory Categorial Grammar (CCG) [Steedman 2001]. Our work is basically an extension of a seminal work on CCGbank [Hockenmaier and Steedman 2007], in which the phrase structure trees of the Penn Treebank (PTB) [Marcus et al. 1993] are converted into CCG derivations and a wide-coverage CCG lexicon is extracted from these derivations. Since CCGbank has enabled a variety of outstanding studies on wide-coverage deep parsing for English, our resources are expected to significantly contribute to Japanese deep parsing.

The application of the CCGbank method to Japanese, however, is not trivial, as resources like PTB are not available in Japanese. In Japan, the widely used resources for parsing research are the Kyoto corpus [Kawahara et al. 2002; Kurohashi and Nagao 2003] and the NAIST text corpus [Iida et al. 2007], both of which are corpora with annotations of dependency structures of chunks. An internal structure of a chunk is dependent on the words comprising the chunk. Moreover, a dependency between two chunks can be interpreted as a relation of one chunk to a certain part of the other, as well as between two complete chunks. Therefore, the relation between a chunk-based dependency structure and a CCG derivation is not obvious.

We propose a method for integrating multiple dependency-based corpora into phrase structure trees augmented with predicate argument relations. We can then convert the phrase structure trees into CCG derivations. In the following, we describe the details of the integration method as well as Japanese-specific issues in the conversion. We empirically evaluate the method in terms of the quality of the corpus conversion, the coverage of the obtained lexicon, and the accuracy of parsing with the obtained grammar. Additionally, we discuss problems that remain in Japanese resources from the viewpoint of developing CCG derivations.

There are three primary contributions of this article: (1) we provide the first comprehensive results for Japanese CCG parsing, (2) we present a methodology for integrating multiple dependency-based resources to induce CCG derivations, and (3) we investigate the possibility of further improving CCG analysis by using additional resources.

Our proposed method is not limited to Japanese. It should be possible to apply a similar method to other languages for which chunk-based dependency analysis is widely regarded as more appropriate than word-based phrase structure analysis due to morphosyntactic features similar to Japanese.

In Section 2, we introduce a CCG-based theory of Japanese syntax and related works on the induction of CCG resources. In Section 3, we explain the details of our method for developing Japanese CCG resources. In Section 4, we present the experimental results on evaluating the obtained resources and conclude the article in Section 5 with a brief summary and a mention of future work.

2. BACKGROUND

2.1. Combinatory Categorial Grammar

Combinatory Categorial Grammar is a syntactic theory widely accepted in the NLP field [Steedman 2001]. A grammar based on CCG theory consists of *categories*, which

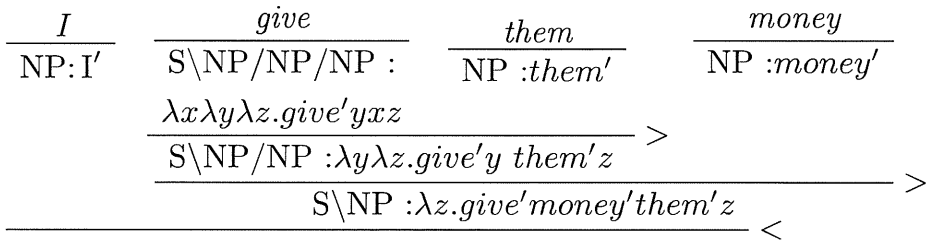


Fig. 1. A CCG derivation. An example from Hockenmaier and Steedman [2007].

$$
\begin{array}{llll}
\mathrm{X/Y}:f & \mathrm{Y}:a & \rightarrow & \mathrm{X}:fa \quad (>) \\
\mathrm{Y}:a & \mathrm{X\backslash Y}:a & \rightarrow & \mathrm{X}:fa \quad (<) \\
\mathrm{X/Y}:f & \mathrm{Y/Z}:g & \rightarrow & \mathrm{X/Z}:\lambda x.f(gx) \quad (>\mathrm{B}) \\
\mathrm{Y\backslash Z}:g & \mathrm{X\backslash Y}:f & \rightarrow & \mathrm{X\backslash Z}:\lambda x.f(gx) \quad (<\mathrm{B})
\end{array}
$$

Fig. 2. Combinatory rules (used in the current implementation).

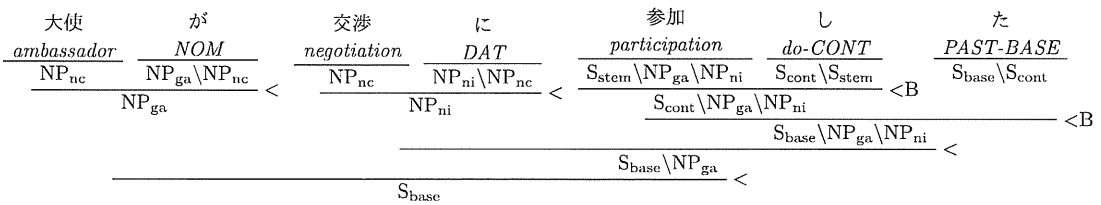


Fig. 3. A simplified CCG analysis of the sentence "The ambassador participated in the negotiation".

represent syntactic categories of words and phrases, and *combinatory rules*, which are rules to combine the categories. Categories are either *ground categories* such as S and NP or *complex categories* in the form of X/Y or X\Y, where X and Y are the categories. Category X/Y intuitively means that it becomes category X when it is combined with another category Y to its right, and X\Y means it takes a category Y to its left. Categories are combined by applying combinatory rules (Figure 2) to form categories for larger phrases. Figure 1 shows a CCG analysis of a simple English sentence, which is called a *derivation*. The verb *give* is assigned category S\NP/NP/NP, which indicates that it takes two NPs to its right, one NP to its left, and finally becomes S. Starting from lexical categories assigned to words, we can obtain categories for phrases by applying the rules recursively.

An important property of CCG is a clear interface between syntax and semantics. As shown in Figure 1, each category is associated with a lambda term of semantic representations, and each combinatory rule is associated with rules for semantic composition. For example, the first rule in Figure 2 states that the left phrase has semantic representation f, the right phrase has a, and the result of rule application creates semantic representation fa. Since the rules in the figure are universal, we can obtain different semantic representations by switching the semantic representations of lexical categories. This means that we can plug in a variety of semantic theories with CCG-based syntactic parsing [Bos et al. 2004]. For example, Bos et al. [2004] proposed a system that computes formal semantic representations based on Discourse Representation Theory (DRT) [Kamp and Reyle 1993] by replacing semantic representations of the original CCG parser with discourse representation structures.

Table I. Features for Japanese Syntax (Those Used in the Examples in this Article)

Category	Feature	Value	Interpretation
NP	case	ga	nominative
		o	accusative
		ni	dative
		to	comitative, complementizer, etc.
		nc	none
S	form	stem	stem
		base	base
		neg	imperfect or negative
		cont	continuative
		vo_s	causative

Table II. Typical Categories for Japanese Syntax. $ stands for Zero or More of NP with a Backslash Between Them

Sentence	S	Predicate	S\$ (e.g. $S\backslash NP_{ga}$)
Noun phrase	NP	Post position	$NP_{ga\|o\|ni\|to}\backslash NP_{nc}$
Auxiliary	S\S		

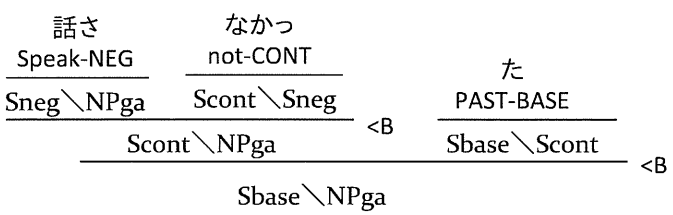


Fig. 4. Agglutination of auxiliaries.

2.2. CCG-Based Syntactic Theory for Japanese

Bekki [2010] proposed a comprehensive theory for Japanese syntax based on CCG. While the theory is based on Steedman [2001], it provides concrete explanations for a variety of morphological and syntactic constructions of Japanese, such as agglutination, scrambling, and long-distance dependencies (Figure 3).

The ground categories in his theory are S, NP, and CONJ (for conjunctions). Syntactic features are assigned to categories NP and S (Table I). The feature *case* represents a syntactic case of a noun phrase. The feature *form* denotes an inflection form. Table II lists typical lexical categories. Predicates, *i.e.*, verbs, adjectives, and verbal nouns, are represented as $S\backslash NP_{ga}$, $S\backslash NP_{ni}\backslash NP_{ga}$ or $S\backslash NP_{to}\backslash NP_{ga}\backslash NP_{o}$, etc., depending on their arguments. For example, $S\backslash NP_{ga}$ denotes intransitive predicates and $S\backslash NP_{ga}\backslash NP_{o}$ is a transitive one. Postpositions are $NP_{ga}\backslash NP_{nc}$ and $NP_{ni}\backslash NP_{nc}$, etc. For example "が NOM ga" is represented as $NP_{ga}\backslash NP_{nc}$ as it takes the left NP to form a nominative NP. Categories for auxiliary verbs require an explanation. In Japanese, auxiliary verbs are extensively used to express semantic information such as tense and modality. The auxiliaries and the main verb are combined in sequential order. For example, a verb "話さ/speak-NEG" and auxiliaries "なかっ/not-CONT" and "た/PAST-BASE" form a VP "話さなかった", which means "did not speak". This is explained in Bekki's theory by the category S\S, and the category is combined with a main verb via the function composition rule (<B in Figure 2) as shown in Figure 4. As stated above, the feature *form* of category S denotes an inflection form. Since the agglutination of auxiliary verbs is restricted by inflection forms, this form feature is necessary for constraining the grammaticality of agglutination.

$$S \rightarrow NP/NP \quad \text{(RelExt)}$$
$$S\backslash NP_1 \rightarrow NP_1/NP_1 \quad \text{(RelIn)}$$
$$S \rightarrow S_1/S_1 \quad \text{(Con)}$$
$$S\backslash\$_1\backslash NP_1 \rightarrow (S_1\backslash\$_1\backslash NP_1)/(S_1\backslash\$_1\backslash NP_1) \quad \text{(ConCoord)}$$

Fig. 5. Type changing rules. The upper two are for relative clauses and the others are for continuous clauses.

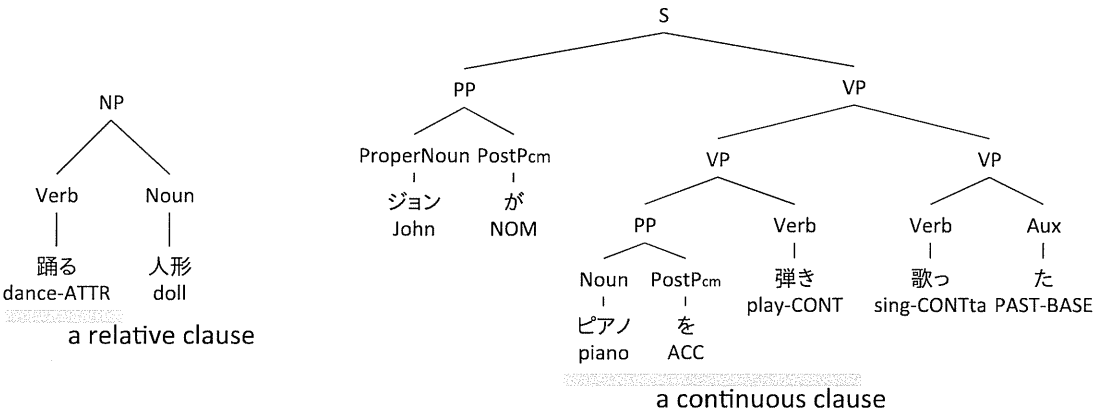


Fig. 6. A noun phrase with a relative clause (left) and a sentence with a continuous clause (right). In the left NP "a doll that dances," the underlined phrase is the relative clause. For the right tree "John played the piano and sang," the underlined phrase is the continuous clause with a shared argument, which means "played the piano".

Our implementation of the grammar basically follows Bekki's theory [Bekki 2010]. However, as a first step in implementing a wide-coverage Japanese parser, we focused on the frequent syntactic constructions that are necessary for computing predicate argument relations, including agglutination, inflection, scrambling, and case alternation. Other details of the theory are largely simplified (Figure 3), coordination and semantic representation in particular. The current implementation recognizes coordinated verbs in continuous clauses (see Figure 6), but the treatment of other types of coordination is largely simplified. For semantic representation, we define *predicate argument structures* (PASs) rather than the theory's formal representation based on dynamic logic. A PAS consists of a predicate word, set of argument types, and argument phrases. An argument type is represented by a deep syntactic case, *ga* for nominative, *o* for accusative, *ni* for dative, and *to* for comitative, *etc.*, all of which are from Japanese postpositions used as case markers. For the example in Figure 3, the PAS for the predicate "参加/participation" has arguments of ga and ni, where the ga-argument refers to the phrase "大使が/ambassador-NOM" and the ni-argument is "交渉に/negotiation-DAT". As a result of defining simple semantics, some of the grammatical distinctions that are required for semantic representations in the theory of Bekki [2010] are excluded from our implementation. Sophisticating our semantic representation is left for future work.

For parsing efficiency, we modified the treatment of some constructions so that empty elements are excluded from the implementation. First, we define type changing rules to produce relative and continuous clauses (shown in Figure 5). A relative clause in Japanese does not have a relativizer. Figure 7 exemplifies a relative and a continuous clause in Japanese. In the original theory, the two types of clauses are explained by using empty elements: *pro*, operators *rel* and ϕ (see Figure 7). The newly defined rules produce almost the same results as the theory's treatment but without

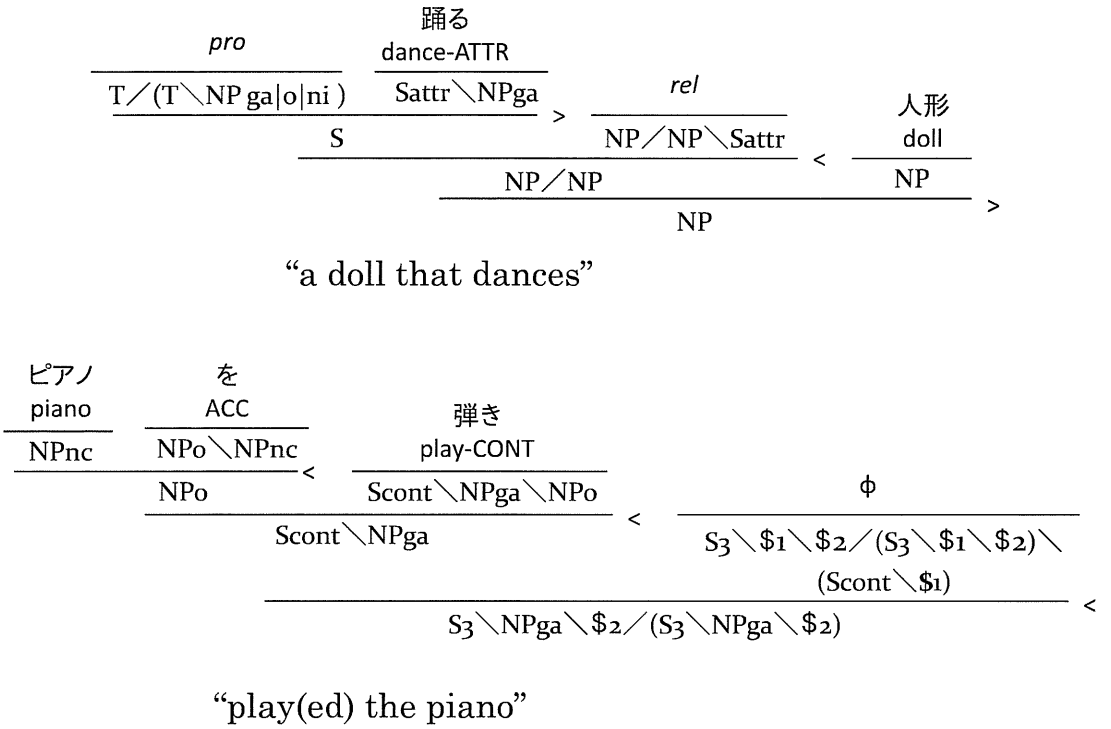


Fig. 7. Derivation for a relative clause and a continuous clause in the original theory.

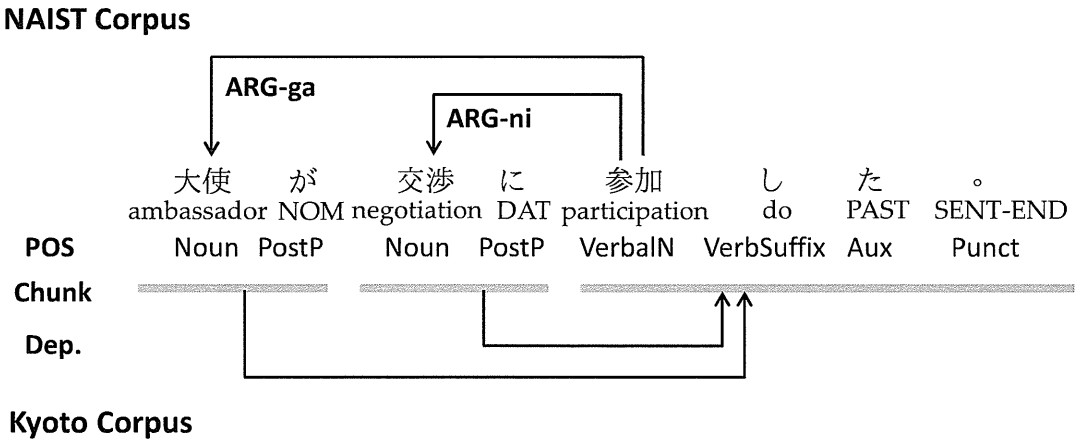


Fig. 8. The Kyoto and NAIST annotations for “the ambassador participated in the negotiation”.

using the empty elements. Second, we treat pro-drop and scrambling by simply adding lexical entries to the lexicon. For the sentence in Figure 3, the deletion of the nominative (大使が/ambassador-NOM), the dative (交渉に/negotiation-DAT), or both, results in valid sentences, and shuffling the two phrases does so as well. Lexical entries with the scrambled or dropped arguments are produced using simple methods, which permutate or delete arguments in categories.

2.3. Linguistic Resources for Japanese Parsing

As described in Section 1, chunk-based dependency analysis has been considered as a standard in Japanese syntactic parsing. Research on Japanese parsing also relies on dependency-based corpora. We used the following resources in this work as they have annotations for the same set of texts and can be used in a complementary way.

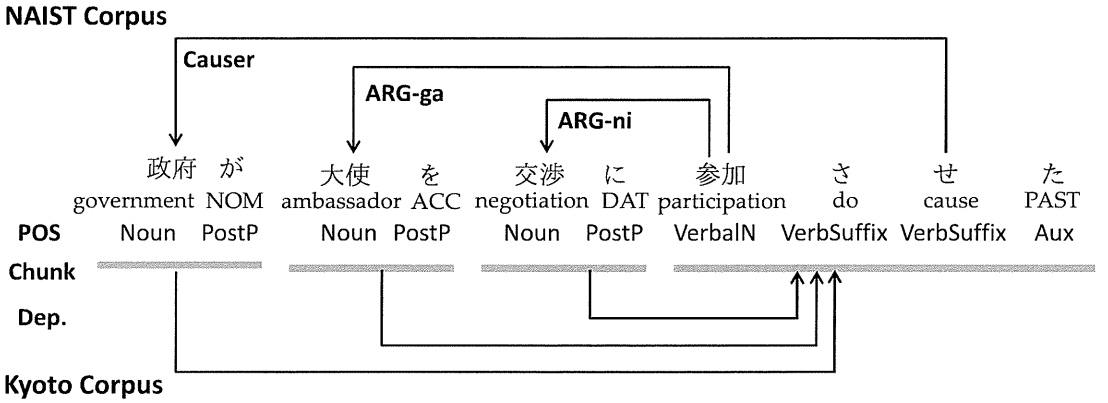


Fig. 9. The Kyoto and NAIST annotations for "The government had the ambassador participate in the negotiation." Punctuation is omitted in this figure. Accusatives are labeled as ARG-ga in causative sentences (see Section 3.2).

Kyoto corpus. This is a news text corpus annotated with morphological information, chunk boundaries, and dependency relations among chunks (Figure 8). The text consists of approximately 40,000 sentences from articles appearing in *The Mainichi Newspaper*: all news articles from January 1st to 17th 1995 and all editorials from January to December 1995 [Kurohashi and Nagao 2003; Kawahara et al. 2002]. The morphological information is annotated basically based on a Japanese grammar book [Masuoka and Takubo 1989]. The dependencies are classified into four types: Para (coordination), A (apposition), I (argument cluster), and Dep (default). Most of the dependencies are annotated as Dep, so the dependencies in the corpus can be regarded as unlabeled relations.

NAIST text corpus. This is a corpus annotated with predicate argument relations for verbs, adjectives, and nouns referring to events [Iida et al. 2007]. It also contains voice information of the predicates and annotations of anaphora and coreference relations. The same set as the Kyoto corpus is annotated.[1] The labels used for predicate argument annotation are deep syntactic cases. In Japanese, arguments of a predicate are typically marked by a postposition, which functions as a case marker. As a result, the label for an argument is basically the same as the postposition following the argument if the sentence has no case alternation. In the example shown in Figure 8, "大使/ ambassador" and "交渉/ negotiation" are arguments of the predicate "参加// participate" and are followed by postpositions "が / NOM / ga" and "に / DAT / ni", respectively. Therefore, their labels are *ARG-ga* and *ARG-ni*. Figure 9 shows an example involving case alternations. Since the sentence "The government had the ambassador participate in the negotiation." is a causative sentence, arguments of the event noun 参加 *participation* have altered case markers. As a result, "大使 / ambassador", superficially with the postposition "を / ACC / o", is labeled with *ARG-ga*. Note that the label is the same as that in Figure 8. The corpus now focuses on three cases: "ga" (subject), "o" (direct object), and "ni" (indirect object).

Japanese particle corpus (JP). This is a corpus annotated with distinct grammatical functions of the Japanese postposition "と to" [Hanaoka et al. 2010]. The same set as the Kyoto corpus is annotated. In Japanese, the postposition "to" has many functions, including a complementizer (similar to "that"), a subordinate conjunction (similar to "then"), a coordinating conjunction (similar to "and"), and a case marker (similar to "with"). In our work, the case marker information is used to construct *to*-labeled

[1]In fact, the NAIST text corpus includes additional texts, but in this work we only use the news text section.

arguments for predicate argument relations, while other annotations are used for the detection of constructions such as coordination.

2.4. Related Work

Research on Japanese deep parsing is fairly limited. Formal theories of Japanese syntax were presented by Gunji [1987] based on Head-Driven Phrase Structure Grammar (HPSG) [Sag et al. 2003] and by Komagata [1999] based on CCG. Komagata [1999] has also presented implemented work, although the implementation has not been very successful in terms of parsing real-world texts. JACY [Siegel and Bender 2002] is a large-scale Japanese grammar based on HPSG that has been used for wide-coverage deep parsing of Japanese. While JACY has been successful in producing precise and detailed semantic representations for realistic sentences, our focus is more on developing deep parsing systems capable of processing a large amount of real-word text. Yoshida [2005], who led the previous studies most similar to our present work, proposed methods for extracting a wide-coverage lexicon based on HPSG from a phrase structure treebank of Japanese. We largely extended his work by exploiting the standard chunk-based Japanese corpora based on dependency structures and obtained the first results for Japanese deep parsing with grammar induced from large corpora. In addition, we demonstrated the first results for Japanese CCG parsing of real-world news texts.

Corpus-based acquisition of wide-coverage CCG resources has enjoyed great success for English [Hockenmaier and Steedman 2007]. In that method, PTB is converted into CCG-based derivations from which a wide-coverage CCG lexicon is extracted. CCGbank has been used for the development of wide-coverage CCG parsers [Clark and Curran 2007]. The same methodology has been applied to German [Hockenmaier 2006], Italian [Bos et al. 2009], Turkish [Çakıcı 2005], and Hindi [Ambati et al. 2013]. These works also suffered from a lack of PTB-like resources and used dependency treebanks as source resources. Their treebanks are annotated with dependencies of *words*, the conversion of which into phrase structures is not a big concern. A notable contribution of the present work is to propose a method for inducing CCG grammars from chunk-based dependency structures, which is not obvious, as we discuss later in this article.

CCG parsing provides not only predicate argument relations but also CCG derivations, which can be used for various semantic processing tasks. Bos et al. [2004], Bos [2007] proposed a method for computing DRT-based semantic representations from CCG derivations for English, and their system has been applied to NLP tasks such as textual entailment recognition [Bos 2007]. Our work constitutes a starting point for such deep linguistic processing for languages similar to Japanese.

3. CORPUS INTEGRATION AND CONVERSION

For wide-coverage CCG parsing, we need, (a) a wide-coverage CCG lexicon, (b) combinatory rules, (c) training data for parse disambiguation, and (d) a parser (*e.g.*, a CKY parser). Since d) is grammar- and language-independent, all we have to develop for a new language is (a)–(c).

We adopt the methodology of Hockenmaier and Steedman [2007]. In that work, existing linguistic resources (Penn Treebank) are converted into CCG derivations (CCGbank), which are then used for extracting a wide-coverage CCG lexicon, as well as for training parse disambiguation models [Clark and Curran 2007]. Combinatory rules can be hand-coded because the number of rules and their language dependence are limited. In our case, the number of combinatory rules is nine: four general rules in Figure 2, four type changing rules in Figure 5, and a type changing rule to handle a