文内に出現する談話関係を認定するための接続表現の調査

飯田 龍　　徳永 健伸
東京工業大学 大学院情報理工学研究科
{ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

因果関係や逆接関係などの文章中の談話関係を同定する問題は，談話の理解を計算機で実現し，それを応用した自動要約や情報抽出などの自然言語処理の応用処理を構築するための重要な課題である．近年，談話関係の解析に関しても形態素・構文解析などの基盤処理と同様に，文章に談話単位の範囲とその間の関係をアノテーションしたコーパスを作成し，それを基礎データとして自動解析の研究が進められている [2, 8, 10]．特に，Penn Discourse TreeBank [8] を解析対象とした研究が進められているが，それらの研究の多く [1, 3, 4, 7] が文章中に出現する接続表現や談話単位中に出現する語彙的な情報に依拠した解析を行っており，そのような情報だけでは解析精度を向上させることができないという状況に陥っている．つまり，どのような観点から談話解析に取り組めばよいかを未だに模索中という状況にある．

このような状況を打破するために，本研究では，文章中の広範な談話関係を対象として，談話関係タグ付きコーパスを構築する研究とは異なり，まずは特定の現象に着目し，そこに出現する談話関係の特徴を明確に捉えることを目的とする．このため，研究の対象を文内の談話単位間の関係，特に特定の出現パタンに限定して分析を行う．分析対象とするパタンとしては「(節 $_1$) 事件が起き、(節 $_2$)」,「(節 $_1$) 声に応じて、(節 $_2$)」のような，連体修飾が関与するパタンに限定し，連体修飾節 (節 $_1$) とその外の節 (節 $_2$) に記述された事態間にある談話関係が成り立つのかを調査する．これは，談話関係に関する既存研究では，一般に文内の談話関係を考える場合に主節と従属節 (副詞節・並列節) の間の関係を考察する場合が多く，また，連体修飾節を考える場合 [9] でも単純に詳細化 (elaboration) の関係としてしか扱われていないという問題に動機づけられている．さらに，我々が先行研究で対象にした日本語文章の談話分割 [12] でも節が連体修飾節に埋め込まれているために，それ以上の分割ができないという問題が生じており，これを解決しなければならないという点も本研究の動機となっている．例えば，例 (1) では文中に「購入する」「分かる」「追及する」の 3 つの事態が含まれているが，単純に命題の内容を保持するように節分割すると「分かる」と「追及する」を主辞とした 2 つの談話単位にしか分割することができないという問題が起こる．

(1) また、山下巡査長の胸に刺さっていた包丁は、容疑者が事件直前に近くの金物店から 購入し た **こと**も 分かり、 捜査本部で動機など詳しく 追及し ている。

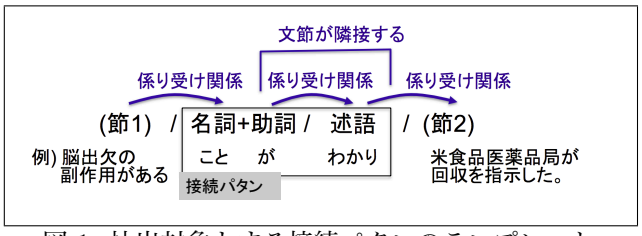

図 1: 抽出対象とする接続パタンのテンプレート

ただし，この文を例 (2) のように「ことも分かり」を「ために」と言い換えることで,「購入する」ことと「追及する」という関係が明示的に理解できる．

(2) また、山下巡査長の胸に刺さっていた包丁は、容疑者が事件直前に近くの金物店から 購入し た **ために**、 捜査本部で動機など詳しく 追及し ている。

このように，ある特定の接続パタンを言い換えることで，その表現が談話関係を明示的もしくは非明示的に表す接続表現となるか否かを判断できると考えられる．特に連体修飾節を対象にした単純な接続表現への言い換えを考えることで，文の簡単化 (sentence simplification) [11] や連体節主節化 [15] を行うための手がかりにもなる．そこで，本研究では「連体節が修飾する名詞句の主辞とその主辞を項として持つ述部の文節」という特定の接続パタンのみを対象に，そのパタンがある特定の談話関係を示唆するか否かを調査する．まず，事例を調査するため，あらかじめ荒い粒度で文内の談話関係をアノテーションした結果について 2 節で紹介する．次に，3 節で，作成したデータの一部を対象に，さまざまな談話関係のうち特に因果関係について細かく事例を分析し，何が談話関係として認識するための手がかりとなっているのかを考える．さらに，4 節で，分析した結果と既存の言語資源を使い，接続パタンの一般化を考える．

2 データの収集

まず，連体修飾をともなう接続表現の候補パタンを文章集合から収集する．可能であれば，多様な現象を収集するために日本語書き言葉均衡コーパス [5] のような均衡コーパスからの事例収集が望ましいが，レジストリによっては会話文を多く含み，また記述の揺れも大きいため，収集した結果をまとめあげるのが困難であると考えられる．そこで，本研究ではまず，新聞記事を対象に候補パタンの収集を行った．新聞記事 13 年分[1]を係り受け解析器 *CaboCha*[2] で解析した結果を利用して，図 1 に示

[1] 毎日新聞 1991 年〜2013 年
[2] https://code.google.com/p/cabocha/

表 1: 頻出接続パタン

ことになり，性があると，ところによると，ことを受け，必要があると，性もあると，ことにより，ことが分かり，ことを決め，連絡によると，ことができ，こともあって，ことになると，ことができると，ことを明らかにし，性があり，ことに加え，ことがあると，ことがあり，ことになるが，ことで合意したと，ことを知り，ことを受けて，こともあると，ことを認め，疑いがあると，疑いが強まり，狙いがあると，ことができず，恐れがあると，調査によると，ことになっており，事件があり，こともあったと，ことになれば，必要があり，ことを踏まえ，方針を固め，ことがあるが，ことにし，ことになったと，ことに関連して，ことになったが，ことがわかり，恐れがあり，よう求め，こととし，ことを挙げ，用意があると，こともあったが，ことに関連し

表 2: 談話関係のラベル

関係	説明	置換のための表現
因果	arg_1 の事態の結果，arg_2 の事態が生じる	（〜した）結果
対比	arg_1 と arg_2 が事態レベルで逆の内容を表現している	（〜した）が，しかし
背景・理由	arg_2 が起こっている背景・前提・理由として arg_1 が起こっている	（〜した）背景（理由）により
詳細・補足・例示	arg_2 が arg_1 の詳細な説明となっている	（〜した）ことを詳細に言うと
連続事態	arg_1 の事態の後に arg_2 の事態が生じる	（〜した）後に
目的	arg_1 が arg_2 を行うための動機・目的	（〜した）目的で

arg_1 は連体節の主辞に係る節を表し，arg_2 は連体節の主辞が係る述語を主辞とする節を表す．

すテンプレート[3]に適合する接続パタンを収集した．この結果，565,643 事例（異なりで 288,404 接続パタン）を収集した．接続パタンの出現頻度は概ねジップの法則にしたがって減少するため，接続パタンによっては分析のための十分な量の事例を確保できないという問題が起こる．そこで，本研究では分析対象を頻出 2,000 パタンに限定し，各接続パタン 10 事例ずつ，合計 20,000 事例に対して人手で談話関係を表すかどうかの判断を行い，網羅性とパタン単位の事例数を確保する．頻出上位の接続パタンの例を表 1 に示す．このパタン集合を見てわかるように，多くの接続パタンは名詞「こと」を含み,「ことがあり」のような広義のモダリティ相当の表現も含まれている．このようなモダリティ相当の表現は区別して扱うべきという意見もあるが，それを弁別する明確な判断基準を設定することは難しいため，今回は収集できたパタンをすべて利用する．

上述の 20,000 事例に対して，談話関係のアノテーションを考える．まずは，粒度の大きい談話関係を把握するために，表 2 に示す 7 種類の談話関係を各事例に対してアノテーションした．アノテーション作業は作業者を 2 名雇用し，2 名に 20,000 事例をアノテーションさせた[4]．作業結果を表 3 に示す．この結果より，約 2 割の接続パタンが因果関係をともなって利用されていることがわかる．一方，他の談話関係は因果関係と比較して相対的に

[3]テンプレート中の'/' が文節のデリミタを表す．また，図の名詞に該当する箇所は文節内の名詞句の最右の形態素のみとし，また固有名は品詞に応じて〈固有名詞地域〉のように抽象化を行った上で収集した．

[4]各作業者の作業事例数はそれぞれ 8,140 事例と 11,860 事例である．

表 3: 談話関係アノテーションの結果

ラベル	関係無し	因果	対比	内容	目的	詳細
頻度	14,701	3,754	802	397	317	29
割合	0.735	0.188	0.04	0.040	0.016	0.001

少ないため，必ずしも接続パタンが談話関係の把握に影響している可能性が低い．このため，以降の分析では因果関係にのみ着目し，具体的にどのような接続パタンが関係の把握に役立っているのかを実例を詳しく調査する．

3 因果関係となるパタンの分析

以降の分析では，人手で分析を行った時点で作業が完了していた 10,032 事例を対象に調査を行った結果について報告する．一つの接続パタンに対し作業者はその接続パタンを含む 10 事例をアノテーションの対象とするが，このうち，10 事例すべてに対して因果のラベルが付与された 77 の接続パタン，770 事例を調査対象とした．ただし，アノテーション時には例 (3) の接続パタン「情報が流れ」について，前文脈「両国は〜合意している」と後文脈「市民に〜みられている」の全体に関してある談話関係が成り立つかを判断させたが，その際,「とみられている」という判断などに関する知覚動詞が主辞になっている場合，それを含めた場合と含めない場合では関係を認定するかどうかの判断が異なる場合がある．例えば,「〜かもしれない」のような推定を表す文末表現が含まれる場合，その表現が含まれることで事態の成立の判断が揺れ，その結果，例えば，前文脈と後文脈の事態対が因果関係にあるか否かの判断が異なる．このような揺れを無くすために，各文脈から「とみられる」のようなモダリティ相当の部分を削除した上で，談話関係が成り立つか否かを再度判断するとともに，因果関係が成り立つ場合はその接続パタンがどのような特徴を持っているかを考察した．例えば，例 (3) の場合は,「〜と合意する」と「〜に不満が高まる」という二つの事態対の間の因果関係を考える．

(3) 両国は米国の中東和平新構想を評価する姿勢を見せていたが、エジプトやサウジがアラファト議長追放で米国と 合意し ているとの観測**情報が流れ、**市民に不満が 高まっ たためとみられている。

この結果，調査した 77 の接続パタンは表 4 に示す分類のいずれかに該当することがわかった．以降で，各パタンのタイプの特徴を具体例とともに紹介する．

〈1〉**前件にモダリティの情報を付加**：最頻出の接続パタンは「疑い」や「見方」など，前文脈中の事態に推定の情報を付与する接続パタンであった．ただし，この接続パタンはパタン中の事態と後文脈内の事態が因果関係になる場合と，前文脈の事態と後文脈の事態が因果関係になる場合の曖昧性がある．例えば，例 (4) では,「脱税していた疑いが強」まった結果,「家宅捜索に乗り出した」という因果関係が成り立つ．

(4) 石油製品加工会社「常盤興産」（佐竹正俊社長）が不正軽油を製造販売し、軽油引取税を 脱税し ていた**疑いが強まり、**県警生活環境課と境署は１４日、同県税務課と合同して、地方税法違反の疑いで同社の家宅捜索に 乗り出し た。

表 4: 因果関係を想起する接続パタンの人手分類の結果

パタンの種類	頻度	パタンの具体例
⟨1⟩ 前文脈の事態にモダリティの情報を付加	31	見方を強めており，疑いがあるとして
⟨2⟩ パタンに前文脈の内容を端的に伝える表現が含まれ，それを受けて後文脈の事態が発生	12	事故を受け，結果となり，流れを受けて
⟨3⟩ 前文脈の事態の知覚が後文脈の事態の発生につながる	6	ことが発覚し，ことを知り
⟨4⟩ 前文脈の事態が後文脈の事態が発生する前提となる	8	事態となれば，〜性があれば
⟨5⟩ 接続助詞に因果を想起させる表現が含まれる	3	ことになるので
⟨6⟩ 前文脈が要求や重要視すべき内容を表し，それに応じて後文脈の事態が発生する	10	声に応え，ことを重視し，影響を受け
⟨7⟩ 前文脈の事態が増加することで後文脈の事態が成立する	1	人が増えれば
⟨8⟩ アノテーション誤り	6	ことを聞かないので

一方，例 (5) では，前文脈の「海砂を海中に投入し」た結果，後文脈の「海砂納入をストップした」事態につながっている．

(5) 関西国際空港の２期事業で、工事業者が大阪府から使用許可の出ていない韓国産の海砂を海中に 投入し ている**疑いが強まり、**関西空港用地造成会社が、一部業者から海砂納入を ストップし たことが２８日分かった。

このため，この ⟨1⟩ の接続パタンは最頻出であるが，実際に知識獲得や談話関係解析に適用する場合には，その曖昧性の解消が必要となる．

⟨2⟩ パタンに前文脈の情報を端的に伝える表現が含まれる：例えば，例 (6) の「報復事件が続発している事態」のように「報復事件が続発」することを端的に表現するコトに関する名詞が存在する．この種の名詞が接続パタンに含まれ，さらに「受けて」などの事態の発生を表す表現がパタンに出現することで，その事態が実際に起こったことを表し，その結果，後文脈の事態が発生するという因果関係が成り立つ．例 (6) では，例えば,「ヒンズー教徒の報復事件が続発」することで「インド陸軍が現地入り」するという事態が起こっている．

(6) インド西部グジャラート州のアーメダバードで、列車放火事件を機にイスラム教徒へのヒンズー教徒の報復事件が 続発し ている**事態を受け、**インド陸軍は１日、現地入りした。

ただし，この種の接続パタンはパタン中の「事件」や「事故」などの表現が前文脈中の述語の項となる場合には，因果関係を想起させない．例えば，例 (7) では「(爆発) 事故」が前文脈中の述語「起きる」の項となっているため，パタンを除いた前文脈と後文脈で因果関係が成り立たない．

(7) 三重県企業庁のごみ固形燃料（ＲＤＦ）発電所の貯蔵槽で 起き た爆発**事故を受け、**環境省が実施した全国調査の概要が１４日 まとまっ た。

⟨3⟩ 前文脈の事態の知覚：「ことが発覚し」,「ことが分かり」のように，パタン中に前文脈中の事態の知覚を示唆する述語が含まれる．つまり，前文脈中の事態の生起を読み手に伝えることで，後文脈に書かれる事態の内容を納得する度合いを高めることに貢献している．これはまさに修辞構造理論 [6] で因果関係（volitional cause, non-volitional cause）を認定するための中心（nucleus）と周辺（satellite）の 2 つの談話単位に関する制約に対応する．例えば，例 (8) では，前文脈の「冷却水が漏れる」事態を知覚することで，後文脈の「手動停止の作業を始め」たことの把握が容易になる．この際，同時に前文脈の事態と後文脈の事態との間に因果関係が成り立つことがわかる．

(8) 福井県大飯郡高浜町の関西電力高浜原発１号機（加圧水型軽水炉、出力八十二万六千キロワット）で、冷却水が 漏れ ている**ことが分かり、**十八日午前九時、手動停止の作業を 始め た。

⟨4⟩ 前文脈の事態が後文脈の事態の前提となる：「〜れば」のような条件節は前文脈の事態が後文脈の事態が成立するための前提条件となる．この際，パタン中の名詞句が「こと」のような明確な事態を導入する表現の場合であっても,「可能性」などの前文脈の事態にモダリティの情報を付与する表現であっても，同様に前提となる関係を表すことになる．例えば，例 (9) では,「これ（=途上国向けの環境分野の援助）を引き上げる」ことが「事態が改善する」ことの必要条件となっている．

(9) これを向こう５年間、３０％台後半、５０００億円程度に 引き上げる **ことをすれば、**事態は大きく 改善される。

ただし，ここであげた前提の関係は修辞構造理論では条件（condition）の関係として定義されているため，因果関係の区分に含めるべきではない．アノテーションされた関係の検証はこのような人手分析を進めることで進めていきたい．

⟨5⟩ 因果関係を想起させる接続助詞が含まれる：「ので」や「ため」はそもそも因果関係を直接的に明記するための接続表現であるが，これがパタン中に含まれる場合も存在した．例えば，例 (10) の「こともあるので」のような場合がこれに該当する．この場合,「寝つけない」ので「静かに過ごす」という因果関係を表している．

(10) 興奮しやすい性格で、電話で話し込むと 寝付け ない**こともあるので**夜は読書などで静かに 過ごし ます。

⟨6⟩ 前文脈の事態が要求や重要視すべき内容を表す：「声が強まり」や「重要性を強調し」のように,「(〜する) 声」で表される要求や「重要性」という句で強調される前文脈の事態の内容などを「強まる」や「強調する」といった程度を高める（もしくは低める）表現で伝える場合，前文脈の事態と後文脈の事態の間に因果関係が成り立つ場合がある．例えば，例 (11) では,「フルオーケストラの音を聴きたい」という要求に対し,「(オーケストラの演奏が) 実現」したという因果関係を表している．

(11) これまでは１０人以下のアンサンブルばかりだったが、「一度、フルオーケストラの音を 聴きたい」という**声に応え、**大阪音楽大の協力を得て 実現する ことになった。

ただし，このパタンの場合では「聴きたい」の「たい」という要求を表す助動詞を含めた事態を前文脈の事態とする必要がある．このような事態の範囲の認定は他の分類においても問題になるため，今後どのようにその範囲を自動検出するかを検討したい．

⟨7⟩ **前文脈の事態の増加**：「(〜する) 人が増えれば」のように前文脈の事態が増加することによって，後文脈の事態が成立する．例えば，(12) では，「たばこをやめる」事態が増えることで，「税収減とな」ることを伝えているが，つまり，前文脈の事態が一度でも起これば微細ではあるが，後文脈の事態が成立することになる．この例では，「たばこをやめる人が一人いることで，税収が減る」という関係が成り立つ．これは，因果関係という意味では正しい解釈ではあるが，程度によっては因果関係が成り立たない場合があると考えられるため，さらに調査が必要であると考えられる．

(12) しかし、たばこを やめる **人が増えれば**税収減と なります。

4 既存の言語資源を利用した接続パタンの解釈

3 節で導入した 7 種類の接続パタンの分類のうち，接続パタンによって因果関係が適切に同定できる ⟨2⟩，⟨3⟩，⟨6⟩ の 3 種類に対し，既存の言語資源との照合を行うことで，パタンの汎化を考える．つまり，インスタンスレベルで獲得した個々の接続パタンがそれぞれ個別にしか適用できないものなのか，より汎化して因果関係とは何によって引き起こされるのかを分析する手がかりを得られるのかを考える．この予備調査として，接続パタン内の名詞は日本語語彙大系 [13] の名詞意味体系のカテゴリを，接続パタン中の動詞は動詞項構造シソーラス [14] に付与されている意味構造の情報を参照し，各分類のパタンに共通な意味情報を獲得する．例えば，接続パタン「事故が起き」の場合，その名詞意味カテゴリとその上位概念すべて「2059_事件，2056_災難，2055_出来事，2054_事象，1235_事，1000_抽象，0001_名詞」と動詞意味構造記述「状態変化あり，生成・消滅，生成 (物理)，生成」から分類 ⟨2⟩ の特徴をできるだけ網羅的に表現できる組み合わせを人手で選択した．この結果，分類 ⟨2⟩ に関しては名詞意味カテゴリが「1235_事」，動詞の意味構造記述が「生成」に分類されている接続パタンが他と比べて多く出現していることがわかった[5]．このような分析を進めることで，人手の分析対象とはしていない「事故が発生し」といった接続パタンを因果関係を想起させる接続パタンの候補として生成することが可能になる．さらに，その結果得られた接続パタンの妥当性を調査することで，因果関係に影響する接続パタンの特性を明らかにでき，その結果，修辞構造理論で導入されている談話関係の定義をより明確にできる可能性がある．この分析の方法論はまだ予備的にしか調査していないため，今後この方針で調査を進めるべきなのかの是非を含め検討を続ける予定である．

5 おわりに

本稿では，文内の談話関係，特に連体修飾節に関係する談話構造の人手分析を行った結果について報告した．連体修飾節の内と外の事態に関して，まず荒く 6 つの談話関係をアノテーションし，その結果を用いて，因果関係を想起させる接続パタンを得た．その接続パタンを事例とともに分析することで，因果関係を想起させるパタンを 7 種に分類した．さらに，そのうちの 3 つの分類について，既存の言語資源を活用し，構成的にパタンが表現できるか否かを調査した．この結果，日本語語彙大系の意味体系と動詞項構造シソーラスを利用することで，一部の接続パタンについては汎化できる可能性を示した．

今後は，本研究で調査した結果を生かし，いくつかの方向性でさらに調査が可能である．例えば，本研究で得た 71 の因果関係に関係する接続パタンをシードとし，ブートストラップ法で事態対とパタンの獲得が考えられる．ただし，この際，事態として同定する範囲をどのように決定するかは問題となる．また，今後は本稿で導入した分析の方法論で別の接続パタンや因果関係以外の談話関係についても分析を行い，談話関係の性質を調査することを計画している．

[5] 分類 ⟨3⟩ では，動詞意味構造記述として「社会的位置付け (物) の変化」や「取得 (情報)」が関係しており，また分類 ⟨6⟩ では名詞意味カテゴリとして「1235_事」，動詞意味構造記述として「程度の変化」が関係していると考えられる．詳細な調査は今後行いたい．

謝辞

本研究の一部は，情報・システム研究機構データ中心科学リサーチコモンズ事業の助成を受けた．記して謝意を表する．

参考文献

[1] O. Biran and K. McKeown. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 69–73, 2013.

[2] L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, 2001.

[3] A. Johannsen and A. Søgaard. Disambiguating explicit discourse connectives without oracles. In *Proceedings of the 6th IJCNLP*, pp. 997–1001, 2013.

[4] Z. Lin, H. T. Ng, and M.-Y. Kan. A PDTB-styled end-to-end discourse parser. Technical report, School of Computing, National University of Singapore, 2010.

[5] K. Maekawa, M. Yamazaki, T. Maruyama, M. Yamaguchi, H. Ogura, W. Kashino, T. Ogiso, H. Koiso, and Y. Den. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the 8th LREC*, pp. 1483–1486, 2010.

[6] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.

[7] E. Pitler, A. Louis, and A. Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th ACL and the 4th IJCNLP*, pp. 683–691, 2009.

[8] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In *Proceedings of the 6th LREC*, 2008.

[9] D. Scott and C. S de Souza. Getting the message across in RST-based text generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pp. 47–73. London: Academic Press, 1990.

[10] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics, Volume 31, Number 2, June 2005*, Vol. 31, No. 2, pp. 249–287, 2005.

[11] S. Wubben, A. van den Bosch, and E. Krahmer. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th ACL*, pp. 1015–1024, 2012.

[12] 宮原聡, 飯田龍, 徳永健伸. 日本語書き言葉を対象とした談話単位分割基準の提案と自動分割の評価. 情報処理学会自然言語処理研究会 SIGNL-211-02, pp. 1–7, 2013.

[13] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1997.

[14] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤. 意味の包含関係に基づく動詞項構造の細分類. 第 14 回言語処理学会年次大会発表論文集, pp. 1037–1040, 2008.

[15] 野上優, 藤田篤, 乾健太郎. 文分割による連体修飾節の言い換え. 言語処理学会第 6 回年次大会発表論文集, pp. 125–128, 2000.