医学生物学談話依存構造ツリーバンクの構築のための アノテーションガイドライン Version 4.1

西田 典起
理化学研究所 革新知能統合研究センター
noriki.nishida@riken.jp

2021 年 10 月 24 日

1 更新内容

- Version 4.0 → 4.1 の更新内容
  - Elaboration と Addition を区別するための例を追加 (図 6)。
  - アノテーションツールの更新にともなう 6 章の変更。
- Version 3.0 → 4.0 の更新内容
  - Elaboration を分解 → Elaboration (詳細化、例示)、Addition (添加・累加・系列・同列)。
  - Joint の削除 (Addition に統合)。
  - Attribution の削除。
  - Same-Unit チェインとの接続ルールの変更: 文内、文間を問わず常にチェインの始点を接合点とする。

2 ツリーバンク構築の背景と目的

一般的に、文章は意味的、論理的に一貫している。そのような文章の一貫性 (coherence) は、節 (clause) や文、段落が相互作用しあうことによって実現されている。**談話依存構造** (Discourse Dependency Structure) は、Elementary Discourse Unit (EDU) と呼ばれる節レベルのテキストスパン (ノード) 間の関係性 (談話関係) に基づいて文章をグラフ構造として表現する [6, 5]。談話依存構造の例を図 1 に示す。談話構造は、文書要約や文書分類、質問応答、情報抽出など様々な自然言語処理技術で有用であることが知られている。

談話依存構造の自動解析は、人手によって構築されたツリーバンクを学習データとして必要とする。しかし、既存の談話構造ツリーバンクの規模は限られている。例えば、最もよく使われている RST-DT [2] には 385 文書しか収録されておらず、実用的な談話構造解析器を構築するのには不十分である。SciDTB [8] には自然言語処理分野の論文要旨が 798 件収録されているが、分野によって論旨の展開傾向や語彙は大きく異なるため、SciDTB で訓練した談話構造解析器の解析精度は他の分野の論文要旨に対しては著しく低下する。

談話依存構造の自動解析技術を実用的なものにするためには、高品質で大規模な談話依存構造コーパスを整備することが必須である。本プロジェクトでは、医学生物学分野の論文要旨に談話依存構造をアノテーションした大規模ツリーバンクを構築することを目的にする。医学生物学に焦点をあてることにはいくつかの利点がある。一つは、医学生物学論文の文章は、一般ドメイン (ブログや SNS、対話など) の文章に比べて論旨の展開が明確であることが期待でき、談話構造の研究対象として望ましい。また、昨今の世界情勢を踏まえると、大量にある医療文献から有用な情報を抽出して体系化する知識獲得技術の開発は喫緊の課題であり、医学生物学論文に対象を絞った談話構造解析技術の開発の意義は大きい。

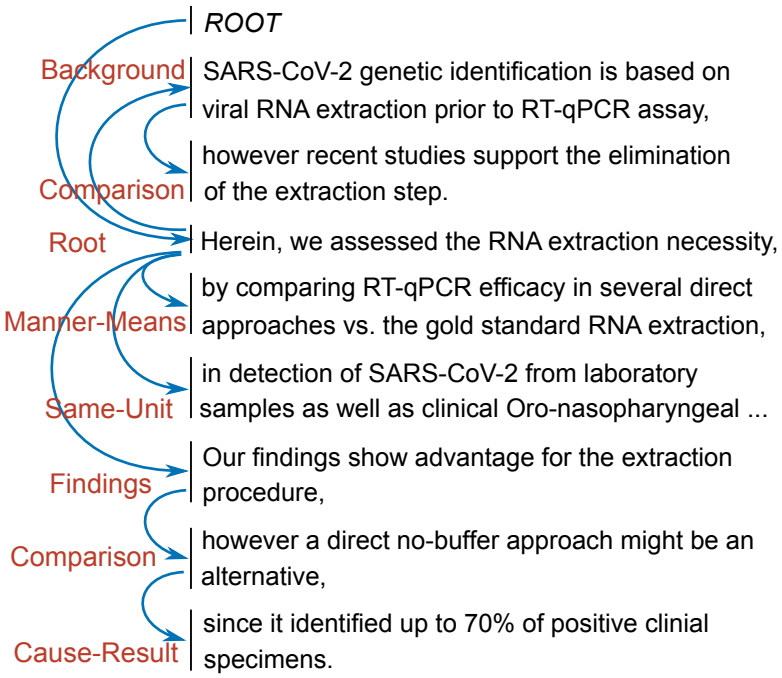


図 1: 医学生物学分野の論文要旨の談話依存構造の例 [4]。

本ガイドラインは次の内容で構成されている。3 章では、医学生物学分野における論文要旨の談話依存構造の典型例と傾向について説明する。4 章では、EDU の分割基準と指針、注意すべき例外事項について記述する。5 章では、談話依存構造の仕様について説明する。最後に、6 章では、本プロジェクトで用いるアノテーションツールについて説明する。

3　医学生物学分野の論文要旨の談話依存構造

本章では、医学生物学分野における論文要旨の談話依存構造の典型例や傾向について共有する。

論文要旨の基礎項目　医学生物学分野だけに限らないが、一般的に論文要旨では大きくわけて以下の情報が含まれている。

- 背景 (Background)
- 目的 (Objective)
- 手法、実験設計 (Method)
- 結果、考察、結論 (Results, Discussion, Conclusion)

実際、論文要旨中でこれらの項目ごとに見出し (e.g., “Background:”, “Results:”) を設けている論文も数多く存在する。

典型例 1　大多数の論文要旨では、上記の順番通りに基礎項目を説明している。それらは図 2, 図 3 のような談話依存構造になる。

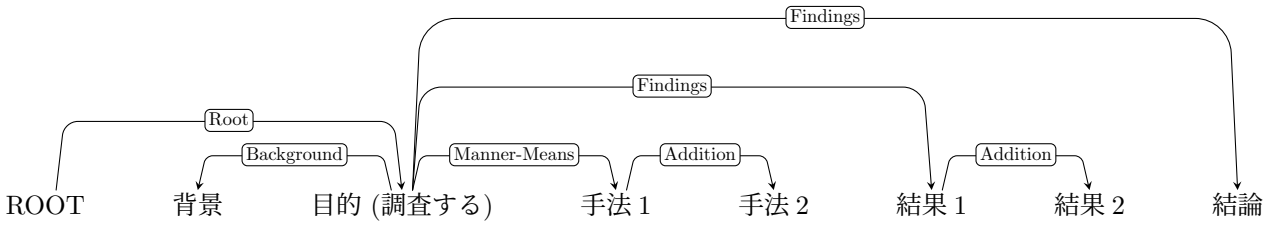


図 2: 典型例 1 (a)

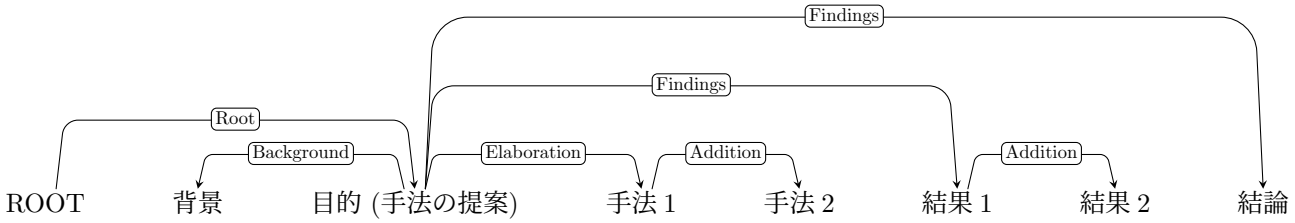


図 3: 典型例 1 (b)

典型例 2　工学系分野では少ないが、医学生物学などの分野では研究によって発見された知見や事例そのものを論文 (論文要旨) の中心的な情報として置き、論文要旨はその知見と意義について記述していることがある。そのような場合は、談話依存構造は図 4 のようになる。

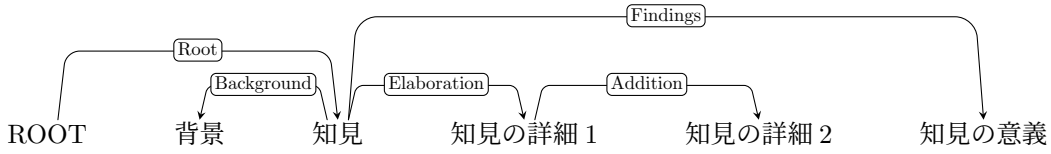


図 4: 典型例 2

4　EDU への分割

本プロジェクトのアノテーションでは、まず論文要旨を **Elementary Discourse Unit (EDU)** と呼ばれる節 (clause) レベルのテキストスパンに分割するところから始める。各 EDU は連続した領域であり、EDU 間にオーバーラップはない。以下では、EDU 分割のための基準と例外について述べる。

4.1　EDU 分割の基準

あるテキストスパンが節 (EDU) に対応するかどうかは、主に**動詞** (述語) に基づいて判断する。また、表 1 に載せているような**ディスコースマーカー**を伴う句については、独立の EDU として認める。結果的に、以下のようなケースでは EDU 分割を行う。

1. 主節、並列節

    (1)　[John cleaned the kitchen] [and Paul vacuumed the dining room.]

表 1: ディスコースマーカーの例。

in spite of	in stead of
despite	irrespective of
regardless of	in contrast to
in comparison with	because of
due to	as a result of
such as	not only ⋯ but also
for the purpose of	i.e., e.g.,

2. 接続詞で結合される従属節

(2) [Although I have a plan to go back home,] [I took a ticket to Hawaii.]

(3) [I took a ticket to Hawaii,] [because I have a plan to go back home.]

3. 分詞構文 (participle clause)

(4) [Having nothing to do,] [I went to bed early.]

(5) [A typhoon hit the city,] [causing big destruction.]

4. 「目的」「結果」の意の to 不定詞、“in order to” 節, so that 節

(6) [He hurried back home] [to get his laundry in.]

(7) [In order to get his laundry in,] [He hurried back home.]

(8) [He hurried back home] [so that he could get his laundry in.]

5. 副詞的役割の「前置詞 + 動名詞」

(9) [He figured out the location of the restaurant] [by using a map.]

(10) [While reading that book,] [he was not alone.]

6. 名詞を後置修飾する { 分詞、to 不定詞、「前置詞 + 動名詞」 }

(11) [I know the woman] [sitting at the chair.]

(12) [This is the book] [stolen by the man.]

(13) [He has a plan] [to go back home.]

(14) [Sleep deprivation increases the risk] [of committing cognitive errors.]

7. 関係節

(15) [I have a friend] [who speaks five different languages.]

(16) [I was born in Kyoto,] [which has many historical buildings.]

(17) [I visited the office] [where my father works.]

(18) [The rain continued for three days,] [which caused a landslide.]

8. 同格の that 節

(19) [I hear the news] [that she is coming to Tokyo today.]

(20) [His comment is based on the fact] [that sleep deprivation increases the risk of health problems.]

9. 動詞を含む相関従属節 (correlative subordinators)

- [... 比較級 ...] [than ...]
- [... as ...] [as ...]
- [... so/such ...] [that ...]
- [... enough ...] [to ...]
- [... too ...] [to ...]
- [The 比較級 ...] [the 比較級 ...] など

(21) [It's a lot cheaper and quicker to buy a plan] [than to build one.]

(22) [Adults under age 30 like sports cars far more] [than their elders do.]

(23) [It was as easy] [as collecting shells at Malibu.]

(24) [Marni Rice plays the maid with so much edge] [as to steal her two scenes.]

(25) [The problem is so vast] [that we need to try innovative solutions.]

(26) [A private market like this just isn't big enough] [to absorb all that business.]

(27) [There were too many phones ringing] [to expect market makers to be as efficient as robots.]

10. 括弧やダッシュによる括り

(28) [Sleep deprivation increases levels of ghrelin] [(the hunger-stimulating hormone).]

11. ディスコースマーカーを伴う句

(29) [In spite of the rain,] [they went out for a picnic.]

(30) [They couldn't go on a picnic] [due to the typhoon.]

4.2 例外

Carlson ら (2001) のマニュアル [3] に従い、本プロジェクトでも以下のようなケースでは**独立した EDU とは認めない**。前節の基準を満たす場合でも、これらの例外に該当する場合は EDU 分割しない。

1. 動詞の主語・目的語・補語や、前置詞の目的語としての節 (clausal subject, clausal object, clausal complement)[1]

(31) [Making computers smaller often means sacrificing memory.]

(32) [To deceive him will make him mad.]

(33) [He started digging.]

(34) [We need to find a solution in our project.]

[1] バージョン 3.1 以前では、Attribution verb の目的節については、目的節が that 節または疑問代名詞で始まる節 (疑問視代名詞 + to 不定詞を除く) で、かつ目的節の帰属先が明らかである場合のみ、目的節を独立した EDU として認め、attribution verb を含む EDU と目的節との間に Attribution 関係をラベル付けした。しかし、バージョン 4 以降では Attribution 関係を削除するため、Attribution verb の目的節についても他の動詞の場合と同様に独立した EDU としては認めない。

表 2: 本コーパスで採用する談話関係とその意味、代表的なディスコースマーカー。

	談話関係	意味	代表的なディスコースマーカー
0.	Root	研究の目的、研究の主要内容	
1.	Elaboration	詳細化、例示	
2.	Addition	添加、累加、系列、同列	also, as well as, moreover, furthermore, besides, in addition, next, then
3.	Comparison	逆説、譲歩、対比、比較	but, however, although, yet, despite, whereas, instead of, alternatively, on the other hand
4.	Cause-Result	原因、理由、結果	because, so, therefore, thus, due to, consequently, as a result, leading to
5.	Condition	条件、仮定	if, as long as, unless, when
6.	Temporal	時間、状況	when, before, after, while
7.	Enablement	目的、可能化	in order to, for ...ing, so as to, so that, which enables to, which allows to
8.	Manner-Means	方法、手段、手法セクション	by, using
9.	Background	背景セクション	
10.	Findings	実験結果セクション、結論セクション	
11.	Textual-Organization	文書構造 (e.g., 見出し、タグ)	
12.	Same-Unit	分離した疑似 EDU の結合	

(35) [He tried to get the work done as quickly as possible.]

(36) [He made me what I am.]

(37) [He is interested in climbing Everest.]

(38) [He was cautious about making a fatal mistake.]

2. 分裂文・疑似分裂文 (強調構文)、外置構文など

(39) [It is sleep deprivation that exacerbates health problems.]

(40) [What exacerbates health problems is sleep deprivation.]

(41) [It is obvious that we cannot read the book.]

(42) [It is difficult to read the book.]

(43) [This book is difficult to read.]

(44) [I found it difficult to read the book.]

5 談話依存構造のアノテーション

談話依存構造のアノテーションは、EDU 間の**談話関係 (Discourse Relation)** を同定することによって行う。談話関係は**親 EDU (中心部)** に対する**子 EDU (周辺部)** の働き・役割を表し、親から子へのラベル付き有向リンクとして表される。一つの論文要旨に対して一つの談話依存構造をアノテーションするために、統語的依存構造

と同様の木構造制約を設ける。すなわち、**Root EDU を除くすべての EDU は必ず親を一つだけもち、かつすべての EDU はグラフ上で連結されているとする (任意の二つの EDU 間にパスが存在する)**。Root EDU だけは親を持たず、論文要旨全体で最も重要な EDU をその唯一の子としてもつ。

したがって、談話依存構造のアノテーション作業は、(Root EDU を除く) 各 EDU についてその親を (木構造制約を満たしながら) 選択し、親と子の間の談話関係を同定することによって行う。

本プロジェクトでは、談話関係を 13 種類にカテゴライズする。表 2 に談話関係の一覧を示す。これらは、SciDTB および RST Discourse Treebank [3], Penn Discourse Treebank [7], ISO 24617-8 [1] を参考に、特に医学生物学論文からの情報抽出のために応用されることを念頭に設計した。本章の以降では、各談話関係について例を使って説明する。

5.0 Root

Root は、文書の先頭に挿入されている "ROOT" (親) と、論文要旨中で最も重要な EDU との間の従属的 (subordination) な談話関係とする。一般的に、論文要旨中で最も重要な EDU は研究目的、あるいは研究の主要内容について記述する EDU である。Root 関係は、各文書に**必ず一度だけ**現れるとする。

談話依存構造では木構造制約を仮定するため、Root 関係における子 EDU から ("ROOT" を除く) 他のすべての EDU へはリンクを矢印の向きに辿ることで到達することができる。

例

(45) [**ROOT**]$_{\texttt{head}}$ Mucosal vaccination is an effective strategy for ... [*In this study, Lactobacillus plantarum strains NC8 and WCFS1 were used as oral delivery vehicles*]$_{\texttt{dep}}$ ...

(46) [**ROOT**]$_{\texttt{head}}$ A Resequencing Pathogen Microarray (RPM) is a single, highly multiplexed assay ... [*In this study, a new RPM (RPM-IVDC1) was developed*]$_{\texttt{dep}}$ ...

(47) [**ROOT**]$_{\texttt{head}}$ The Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3a locus encodes a 274 a.a. novel protein ... [*We established a transgenic fly model for the SARS-CoV 3a gene.*]$_{\texttt{dep}}$

(48) [**ROOT**]$_{\texttt{head}}$ RNA dependent DNA-polymerases, reverse transcriptases, are key enzymes for retroviruses and retroelements. ... [*Here, we report that certain RNA template structures and G-rich sequences can be strong stimulators for ...*]$_{\texttt{dep}}$

5.1 Elaboration

Elaboration は、親 EDU と、その話題 (トピック) を詳細化、あるいは例示によって掘り下げる子 EDU との間の従属的な談話関係とする。Elaboration は従属的な談話関係の基本形である。分詞や関係節などによる名詞句の後置修飾節と、修飾先の名詞句を含む EDU との間の関係も、他の談話関係があてはまらない場合は Elaboration とする。[2]

例

(49) [**Here we show that human coronavirus (HCoV) NL63 and severe acute respiratory syndrome (SARS) CoV papain-like proteases (PLP) antagonize innate immune signaling**]$_{\texttt{head}}$ mediated by STING. [*STING resides in the endoplasmic reticulum*]$_{\texttt{dep}}$ ...

[2] バージョン 3.1 以前では、Elaboration は本バージョンにおける Elaboration (詳細化、例示; 話題の深化) と Addition (添加、累加、系列、同列; 話題の発展) を一つの談話関係として統一してラベル付けしていたが、本バージョンではこれらを区別する。

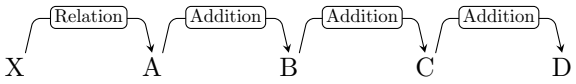
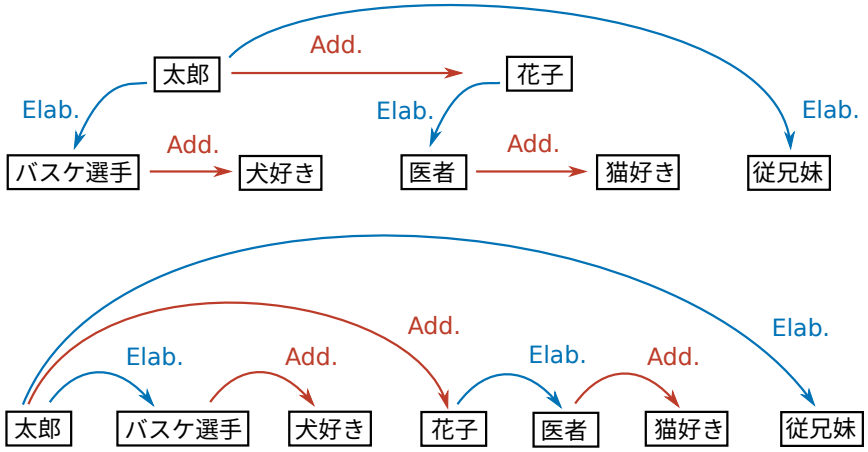


図 5: Addition 関係のアノテーション。EDU A, B, C, D は共通の話題 (トピック) X に対して、同じ重みで情報を添加・累加するとする。A, B, C, D は系列 (順序あり)、または集合 (順序なし) の要素と見なすことができる。

私には太郎という幼馴染がいます。
彼はバスケの選手です。
彼は犬が好きです。
花子も私の幼馴染です。
花子は医者です。
花子は猫が好きです。
彼らは従兄妹です。

(a) 例文章



(b) 談話依存構造。上と下は等価。

図 6: Elaboration と Addition による談話依存構造の例。

(50) [**In this study, a new RPM (RPM-IVDC1)**]$_{\texttt{head}}$ [*that consisted of 224-bp detector tiles*]$_{\texttt{dep}}$ was developed ...

(51) [**Patients with severe symptoms of COVID-19 may also present with acute neurological emergencies**]$_{\texttt{head}}$ [*such as ischemic stroke.*]$_{\texttt{dep}}$

(52) [**In this study, we identified fexaramine as a potent inhibitor of FCV**]$_{\texttt{head}}$ [*including vs-FCV strains in cell culture*]$_{\texttt{dep}}$ ...

(53) [**There is limited evidence**]$_{\texttt{head}}$ [*as to how COVID-19 infection fatality rates (IFR) may vary by ethnicity.*]$_{\texttt{dep}}$

5.2 Addition

Addition は、親 EDU と共通の話題 (トピック) に対して、親 EDU と同じ重みで情報を添加・累加する子 EDU との間の等位的 (coordination) な談話関係である。Addition は系列や集合の要素を接続するためにも使う。Addition は等位的な談話関係の基本形である。三つ以上の EDU が添加・累加・系列・同列の関係にある場合は、図 5 のように一つ前の EDU から直後の EDU に順番に接続していくことによってチェインを構成する。[3]

Elaboration と Addition の区別　Elaboration と Addition の区別は、EDU 間の**重要性の差の有無**に基づいて行う。EDU 間に重要性の差があるとき、Elaboration を適用する。重要性の差がないときは、Addition を適用する。EDU 間の依存関係や順序性の有無はそれらの区別には使えない。図 6 に Elaboration と Addition を用いた例を示す。

[3]Addition は、バージョン 3.1 以前における Horizontal Elaboration と Joint を統合した談話関係である。したがって、順序性や並列性の有無には依存せず、重要性の差の有無に基づいて Addition は判断される。

例

(54) [**Typical enteropathogenic E. coli (tEPEC) carries the highest hazard of death in children with diarrhea**]$_{\texttt{head}}$ [*and atypical EPEC (aEPEC) was recently identified as significantly associated with diarrheal mortality in kittens.*]$_{\texttt{dep}}$

(55) [**A virus must infect,**]$_{\texttt{head}}$ [*replicate*]$_{\texttt{dep}}$ [and spread] [for it to survive;]

(56) [**The results of these assays are available in 1 min**]$_{\texttt{head}}$ [*and do not require any special instrumentation.*]$_{\texttt{dep}}$

5.3 Comparison

Comparison は、逆説や譲歩、対比、比較など、EDU 間の差異に焦点をあてた従属的、または等位的な談話関係とする。ディスコースマーカーを伴わなくても、対比的な接続であれば Comparison を適用する。

例

(57) [**In both children and kittens there is a significant association between aEPEC burden and diarrheal disease,**]$_{\texttt{head}}$ [*however the infection can be found in individuals with and without diarrhea.*]$_{\texttt{dep}}$

(58) [*Although FCV vaccines are commercially available,*]$_{\texttt{dep}}$ [**their efficacy is limited**]$_{\texttt{head}}$ ...

(59) [**A virus must infect,**]$_{\texttt{head}}$ [replicate] [and spread] [for it to survive;] [*the host attempting to thwart it at every step of the way.*]$_{\texttt{dep}}$

5.4 Cause-Result

Cause-Result は、原因・理由と結果の関係にある EDU 間の従属的な談話関係とする。原因・理由側、結果側のどちらを親、子とするかは統語構造と文脈に依存する。

例

(60) [**The current coronavirus disease pandemic**]$_{\texttt{head}}$ [*caused by severe acute respiratory syndrome coronavirus 2*]$_{\texttt{dep}}$ has led to immense strain on healthcare systems and workers.

(61) In order to contain contagions [**is of supreme importance to identify asymptomatic patients**]$_{\texttt{head}}$ [*because this subpopulation is one of the main factors*]$_{\texttt{dep}}$ contributing to the spread of this disease.

(62) [**Their efficacy is limited**]$_{\texttt{head}}$ [*due to antigenic diversity of FCV strains and short duration of immunity.*]$_{\texttt{dep}}$

(63) [**The COVID-19 epidemic has spread rapidly**]$_{\texttt{head}}$ [*to become a world-wide pandemic.*]$_{\texttt{dep}}$

(64) [**Patients in the Neonatal Surgery Department have rapidly progressing diseases and immature immunity,**]$_{\texttt{head}}$ [*which makes them vulnerable to pulmonary infection and a relatively higher mortality.*]$_{\texttt{dep}}$

5.5 Condition

Condition は、親 EDU と、その仮定・条件を表す子 EDU との間の従属的な談話関係とする。

例

(65) [*If underlying health conditions are more important than age per se,*]$_{\texttt{dep}}$ [**then estimated IFR for Māori is more than 2.5 times that of New Zealand European,**]$_{\texttt{head}}$ ...

5.6 Temporal

Temporal は、親 EDU と、それに対して時間的な関係にある、あるいはその状況を表す子 EDU との間の従属的、または等位的な談話関係とする。時間関係はさらに同期的な関係 ("when", etc.) と非同期的な関係 ("before", "after", etc.) に分けることができるが、本プロジェクトではこれらを同一の談話関係カテゴリーとして扱う。

注意 1: 条件的な "when" "when" が条件的に使われている場合、Temporal と Condition のうちどちらを適用するかは、when 節の内容の仮定性・確実性に基づいて決める。もし when 節の内容が (ほぼ) 確実に起こること、あるいは既に起きたことであるならば、Temporal を適用する。もし when 節の内容が仮定的であるならば、Condition を適用する。

注意 2: Addition との区別 プロセスの手順の説明 ("First, ...", "Then, ...", "Finally, ...") などは、厳密には時間ではなく順序にフォーカスしたケースであるため、Addition を適用する。

例

(66) [**Physicians must now account for prognosis of severe COVID-19, resource utilization, and risk of infection to healthcare workers**]$_{\texttt{head}}$ [*when determining eligibility for mechanical thrombectomy (MT).*]$_{\texttt{dep}}$

(67) [**There is a demand for state-of-the-art models capable of precisely segmenting chest x-rays**]$_{\texttt{head}}$ [*before obtaining mask annotations about this sort of dataset.*]$_{\texttt{dep}}$

(68) [*As SARS spreads throughout the world,*]$_{\texttt{dep}}$ [**it may become an increasingly significant problem for transplant patients and programs.**]$_{\texttt{head}}$

(69) [**Pandemics and other crisis situations result in unsettled times, or ontologically insecure moments**]$_{\texttt{head}}$ [*when social and political institutions are in flux.*]$_{\texttt{dep}}$

5.7 Enablement

Enablement は、親 EDU と、その目的、または親によって可能になることを表す子 EDU との間の従属的な談話関係とする。

例

(70) [**Chest radiography and chest CT are frequently used**]$_{\texttt{head}}$ [*to support the diagnosis of COVID-19 infection.*]$_{\texttt{dep}}$

(71) [**The present study was carried out**]$_{\texttt{head}}$ [*to apply the RNAi technology*]$_{\texttt{dep}}$ ...

(72) [**Here , we demonstrate a method**]$_{\texttt{head}}$ [*that enables such prediction*]$_{\texttt{dep}}$ ...

5.8 Manner-Means

Manner-Means は、親 EDU と、そのための方法・手段を表す子 EDU との間の従属的な談話関係とする。「手段」の “by” や “using ...” などのディスコースマーカーを伴うことが多い。

注意: 研究手法としての Manner-Means 多くの論文要旨では、研究目的の記述のあと、それについての詳細に入るのが一般的である。研究目的が “We investigate ...” のように調査等をすることとして記述されており、その後に続く EDU がその方法として解釈できるならば、図 2 のようにそこに Manner-Means 関係を認める。一方、研究目的が “We propose a new technology for ...” のように手法についてであり、その後に続く EDU がその詳細として解釈できるならば、図 3 のように Elaboration 関係をアノテーションする。また、“Method:” などの見出しによって手法に関するセクションであることが明らかである場合は、Manner-Means 関係をアノテーションする。

例

(73) [**We tested 16 different respiratory virus infections in post-surgery mild symptomatic PSP group and asymptomatic PSP group**]$_{\texttt{head}}$ [*using a quantitative real-time reverse transcriptase polymerase chain reaction (qRT-PCR) assay panel.*]$_{\texttt{dep}}$

(74) [**The aim of this study is to investigate the incidence of infection among mild symptomatic PSP group and asymptomatic PSP group after surgical procedure.**]$_{\texttt{head}}$ ... [Methods:] [*We collected the induced sputum from enrolled 1629 children*]$_{\texttt{dep}}$ ...

(75) [**Our work offers new perspectives**]$_{\texttt{head}}$ [*by demonstrating that small-worldness and non-Markovianity can stabilize a classical discrete time crystal,*]$_{\texttt{dep}}$ ...

5.9 Background

Background は、親 EDU (一般的には研究目的・研究内容セクションの代表点; “ROOT” の子) と、研究背景セクションの代表点である子 EDU との間の従属的な談話関係とする。Background は論文要旨の基礎項目である研究背景セクションを示すためのメタ的な談話関係である。研究背景セクションは論文要旨の先頭から始まることが多いため、論文要旨の先頭部分に子 EDU が現れることが多い。研究背景セクションが論文要旨中にない場合は、Background が現れなくてもよい。

例

(76) [*Mucosal vaccination is an effective strategy*]$_{\texttt{dep}}$ for ... [**In this study, Lactobacillus plantarum strains NC8 and WCFS1 were used as oral delivery vehicles**]$_{\texttt{head}}$ containing ...

(77) [Background:] [*Viral respiratory infection (VRI) is a common contraindication to elective surgery.*]$_{\texttt{dep}}$ ... [**The aim of this study is to investigate the incidence of infection among mild symptomatic PSP group and asymptomatic PSP group after surgical procedure.**]$_{\texttt{head}}$ ...

5.10 Findings

Findings は、親 EDU (一般的には研究目的・研究内容セクションの代表点; “ROOT” の子) と、実験結果セクションまたは結論セクションの代表点である子 EDU との間の従属的な談話関係とする。Findings は論文要旨の基礎項目である実験結果セクション、結論セクションを示すためのメタ的な談話関係である。

実験結果セクションと結論セクションが分けて記述されている場合は、それぞれに独立に Findings 関係を適用する。例えば、図 7 や例 (83) のように、論文によっては “Results:” や “Conclusions:” のような見出しによって実

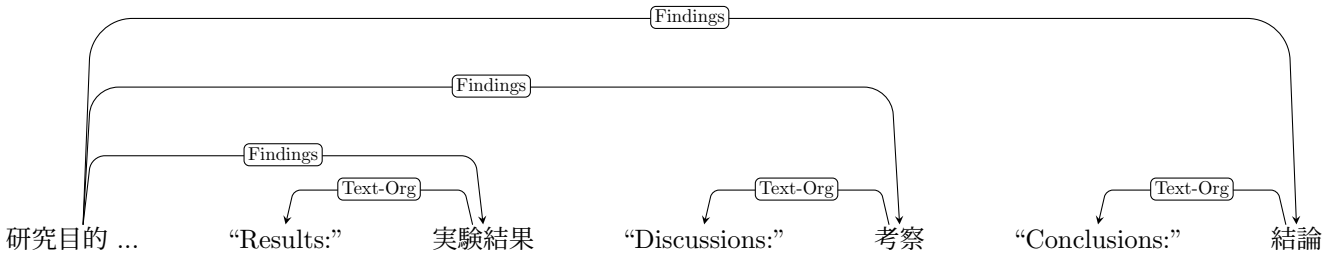


図 7: 複数の Findings 関係をアノテーションする例 1。Text-Org は Textual-Organization 関係を表す。

験結果セクションと結論セクションを明示的に分けて記述していたり、さらには “Discussions:” や “Limitations:” のような見出しでそれらに加えて考察セクションも明示的に分けているケースがある。そのようなケースでは、図 7 のように各セクションについて独立に Findings 関係をアノテーションする。

また、論文全体に対して実験結果を記述する代わりに、図 8 のように、各技術、各実験について述べたあとに逐次その実験結果を記述するようなスタイルもある。例えば図 8 では、技術 A の記述 → 技術 A の実験結果 → 技術 B の記述 → 技術 B の実験結果 → 全体の結論、という順番で論旨を展開している。このようなケースでも、独立に Findings 関係をアノテーションする。

例

(78) [**The purpose of this study was to determine the impact of aEPEC on intestinal function and diarrhea in kittens**]$_{\texttt{head}}$ ... [*Results of this study identify aEPEC as a potential pathogen in kittens.*]$_{\texttt{dep}}$

(79) [**We have developed twin assays**]$_{\texttt{head}}$ ... [*We found these assays to be useful for routine applications in kennels with large numbers of puppies at risk.*]$_{\texttt{dep}}$

(80) [**The aim of this study is to investigate the incidence of infection among mild symptomatic PSP group and asymptomatic PSP group after surgical procedure.**]$_{\texttt{head}}$ ... [Results:] [*Out of 1629 children enrolled, a total of 204 respiratory viruses were present in 171*]$_{\texttt{dep}}$ ...

(81) [**We have developed twin assays**]$_{\texttt{head}}$ ... [*SAT-SIT technology will find applications in rapid screening of samples for other hemagglutinating emerging viruses of animals and humans*]$_{\texttt{dep}}$ ...

(82) [**The present study was carried out**]$_{\texttt{head}}$ ... [*These results provide useful information for the development of RNAi-based gene therapy strategy*]$_{\texttt{dep}}$ ...

(83) [**Influenza A, B and coronavirus antibody titers were measured in 257 subjects with recurrent unipolar and bipolar disorder and healthy controls, by SCID.**]$_{\texttt{head}}$ ... [Results:] [*Seropositivity for influenza A ...*]$_{\texttt{dep}}$ ... [Limitations:] [*The design was cross-sectional.*]$_{\texttt{dep}}$ ... [Conclusions:] [*The association of seropositivity for influenza and coronaviruses with a history of mood disorders, and influenza B with suicidal behavior require replication in larger longitudinal samples.*]$_{\texttt{dep}}$

5.11 Textual-Organization

Textual-Organization は、文書中のテキスト (親 EDU) と、文書中の見出しやタグ (子 EDU) との間の等位的な談話関係とする。特に、医学生物学分野の論文要旨では “Background:” や “Objective:”, “Method:”, “Results:” などのような見出しが頻繁に使われ、基本セクションが明示化されているケースがある。そのようなケースでは、各セ

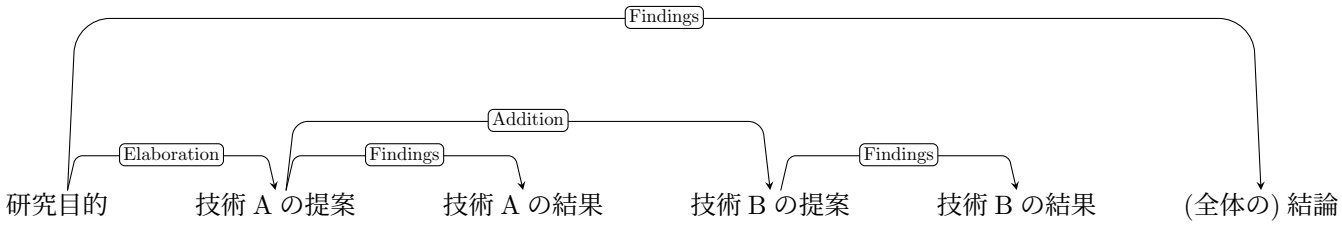


図 8: 複数の Findings 関係をアノテーションする例 2。

クションの代表点を親 EDU (テキスト) とし、対応する見出しを子 EDU としてそれらの間に Textual-Organization を適用する。

例

(84) [*Background:*]$_{\texttt{dep}}$ [**In this study we evaluated the RespoCheck Mycoplasma triplex real-time PCR for ...**]$_{\texttt{head}}$

(85) [*OBJECTIVE:*]$_{\texttt{dep}}$ [**We quantitatively examined the relationship between PERC toxicokinetics and toxicodynamics at the population level**]$_{\texttt{head}}$ ...

(86) [*Methods:*]$_{\texttt{dep}}$ [**Influenza A, B and coronavirus antibody titers were measured in 257 subjects with ...**]$_{\texttt{head}}$

5.12 Same-Unit

EDU は連続したテキストスパンであることが想定されているため、名詞の後置修飾などによって、単体の EDU が二つのスパンに分離してしまうことがしばしば起こる。ここでは、これらの分離してしまったスパンそれぞれを疑似 EDU と呼ぶことにする。例えば、次の文は “better known ... Lewis Carroll,” という EDU の埋め込みによって、“Charles Lutwidge Dodgson was an English writer of children’s fiction.” という単体の EDU が二つの疑似 EDU に分離してしまっている。

(87) [Charles Lutwidge Dodgson,]$_{\text{A}}$ [better known by his pen name Lewis Carroll,]$_{\text{B}}$ [was an English writer of children’s fiction.]$_{\text{C}}$

Same-Unit は、このように分離してしまった疑似 EDU がもともとは同一の EDU であることを指定するための便宜的な談話関係である。リンクは常に**前方向**とする。単体の EDU が三つ以上の疑似 EDU に分離してしまっている場合は、それぞれ**直前の疑似 EDU** と結合させ、一本の Same-Unit チェインを構成する。**Same-Unit チェインと他の EDU との結合は、文内、文間問わず、常に Same-Unit チェインの最初の EDU を接合点とする。**[4] 図 9 の例では、A, B, C, D は同じ文内に属し、X と Y は外部の文にそれぞれ属するとする。もともとは単体の EDU である AC は、B の埋め込みによって分離し、A→C という Same-Unit チェインが構成される。AC に接続する X, B, D, Y は、それぞれチェインの始点である A に接続する。

例

(88) [**The current coronavirus disease pandemic**]$_{\texttt{head}}$ [caused by severe acute respiratory syndrome coronavirus 2] [*has led to immense strain on healthcare systems and workers.*]$_{\texttt{dep}}$

[4]バージョン 3.1 以前では、Same-Unit チェインの文内での結合については、対象の EDU と最も距離が近く、統語的な結びつきが疑似 EDU を接合点としていたが、バージョン 4 以降では文間の結合と合わせて、常にチェインの始点を接合点とする。

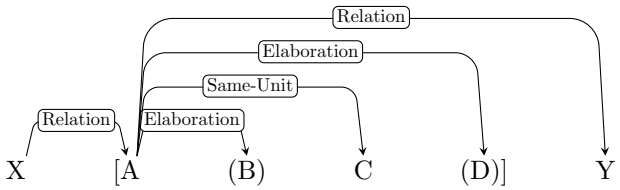


図 9: Same-Unit チェインとの接続例。

(89) [**We find that,**]$_{\texttt{head}}$ [if age is the dominant factor determining IFR,] [*estimated IFR for Māori is around 50 % higher than non-Māori.*]$_{\texttt{dep}}$

6 アノテーションツールの使い方

本章では、アノテーションツールの使い方について説明する。

6.1 準備

- アノテーションツール: `https://norikinishida.github.io/tools/discdep/`
- 論文要旨ごとのテキストファイル (*.sent.txt) (配布予定ファイル)

アノテーションツールは上記 URL にブラウザでアクセスすることで使用できる。[5]アノテーションツールの開発は主に Google Chrome を用いて行っている。特別なインストール等は不要のはずである。各テキストファイル (*.sent.txt) は一つの論文要旨に対応し、単語分割と文分割は配布前に済んでおり、単語分割済みの文が一行ごとに並んでいる。

6.2 EDU 分割

- ツール: `https://norikinishida.github.io/tools/discdep/segmentation.html`
- 処理前: 文分割済みテキストファイル (*.sent.txt)
- 処理後: EDU 分割済みテキストファイル (*.edu.txt)

EDU 分割では、文分割 (+単語分割) 済みのテキストファイル (*.sent.txt) に対して、4 章のルールに従って EDU 分割を行う。結果はツールが出力する EDU 分割済みのテキストファイル (*.edu.txt) とする。

EDU は各文に閉じているため、原則として各文 (=各行) を**独立に**分割する。これは 4 章における EDU 分割ルールよりも優先する。例えば、ある文がただ一つの名詞句からなっており、その名詞句が EDU の分割ルール上では一つの EDU として認められなくても、それを一つの EDU とする。また、一つの EDU が複数の文に部分的にでもまたがることがないように注意し、例えば、k 番目の文の末尾数単語と、k+1 番目の文の先頭数単語を結合して一つの EDU を構成するということがないようにする。

EDU 分割ツールの画面を図 10 に載せる。アノテーションは、アノテーション対象のファイル (*.sent.txt、修正したい場合は*.edu.txt) を「ファイル選択」ボタンからアップロードし、EDU として分割したいスパンの「開始位置」の単語をクリックしていくことで行う。やり直したい場合 (EDU を統合したい場合) は、統合したい二つの EDU のうち後ろ側の EDU の先頭単語をクリックすると、前側の EDU と統合できる。一番目の文 (EDU) の開始位置の単語はクリック不可能になっている。

[5]Yang ら [8] によるツールの実装をベースに、本プロジェクトのために拡張している。

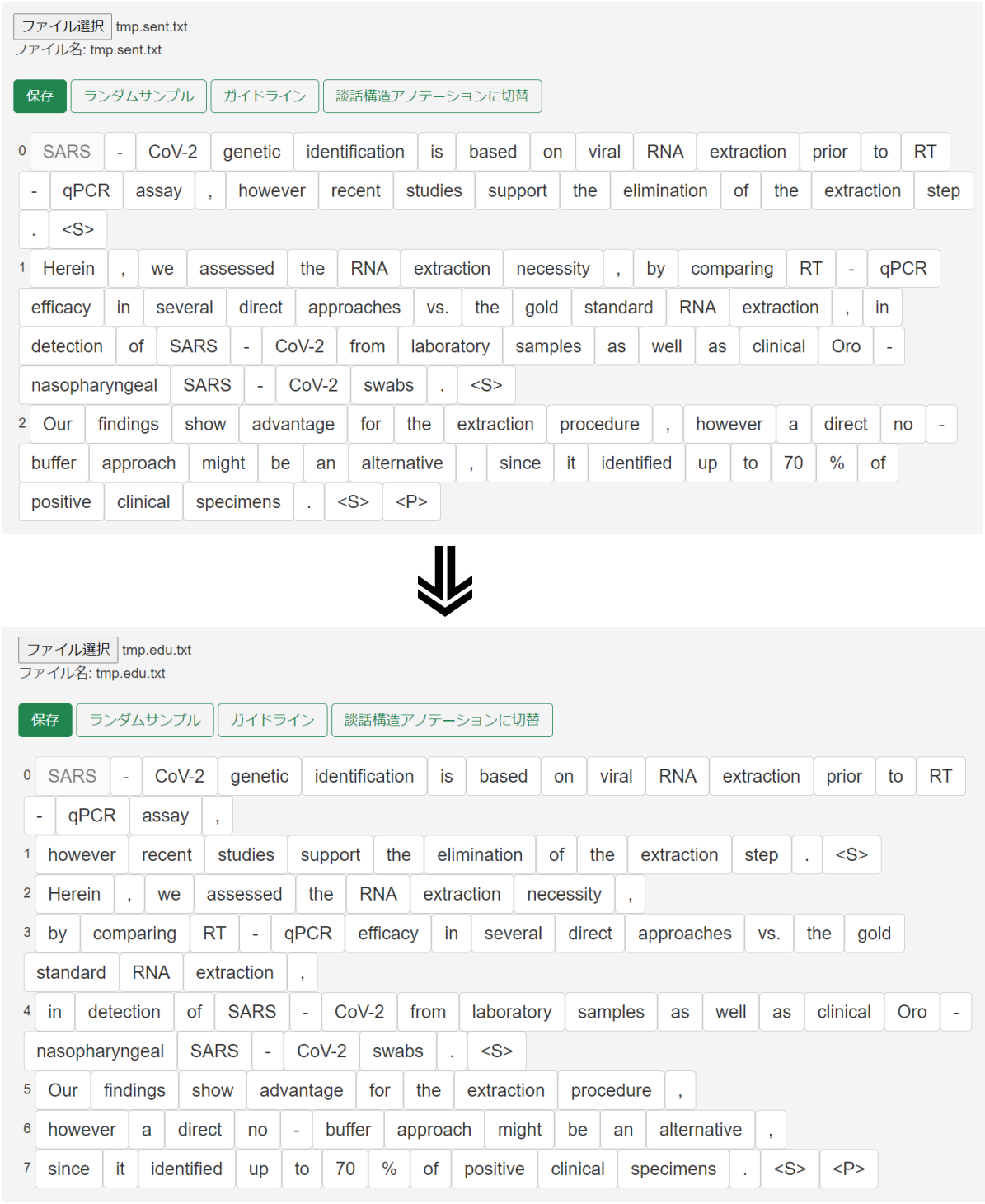


図 10: EDU 分割ツールの画面。処理前と処理後。

6.3 談話依存構造のアノテーション

- ツール: `https://norikinishida.github.io/tools/discdep/index.html`
- 処理前: EDU 分割済みテキストファイル (*.edu.txt)
- 処理後: 談話依存構造ファイル (JSON ファイル) (*.dep)

談話依存構造のアノテーションでは、EDU 分割済みテキストファイル (*.edu.txt) に対して、5 章のルールと談話関係カテゴリーの定義にしたがって談話依存構造を付与する。結果はツールが出力する談話依存構造ファイル (*.dep) とする。

談話依存構造アノテーションツールの画面を図 11 に載せる。アノテーションの手順は基本的に次のようになる。

1. ファイルをアップロード (左上の「ファイル選択」ボタンを押す)
2. 親 EDU を選択 (左クリック)
3. 子 EDU を選択 (左クリック)
4. 談話関係カテゴリーを選択 (ダイアログが現れるので、そこで選択)
5. 上記の 2. から 4. を、“ROOT” を除くすべての EDU の親が決定されるまで繰り返す
6. 談話依存構造のアノテーションが完了すれば、「保存」ボタンを押して保存。(完了するまでは「保存」ボタンは押せないようになっている)

各種ボタンの機能を説明する。

- 「ファイル選択」: EDU 分割済みテキストファイル (*.edu.txt)、または談話依存構造ファイル (*.dep) をアップロードすることができる。談話依存構造ファイルをアップロードすることで、アノテーション済みファイルの確認や修正が可能になる。
- 「選択解除」: ステップ 2 でクリックした EDU の選択状態を解除するときに使う。なにも選択していない状態に戻る。
- 「ラベル変更」: アノテーション済みの談話関係のみを修正したいときに使う。修正したい対象の子 EDU を選択してからこのボタンを押すことで、再び談話関係選択ダイアログが現れ、談話関係の再選択が可能になる。
- 「リンク削除」: アノテーション済みの談話関係（リンク含む）を削除したいときに使う。削除したい対象の子 EDU を選択してからこのボタンを押すことで、対象の談話関係リンクを削除することができる。
- 「undo」: 直近の処理をやり直す。
- 「全文をコピー」: 文書全体をクリップボードにコピーする（Ctrl+v などで他のアプリにペーストできるようになる）。
- 「保存」: アノテーションが完了した場合のみ押せるようになる。アノテーション結果は自動で JSON ファイル (*.dep) に変換されるため、基本的には単純に保存先のフォルダを指定するだけでよい。
- 「ランダムサンプル」: アノテーション済みの例をランダムでサンプリングして提示する。談話関係カテゴリーを選択できる。押すたびに異なる例を示す。
- 「ガイドライン」: 本ガイドライン (PDF) へのリンク。

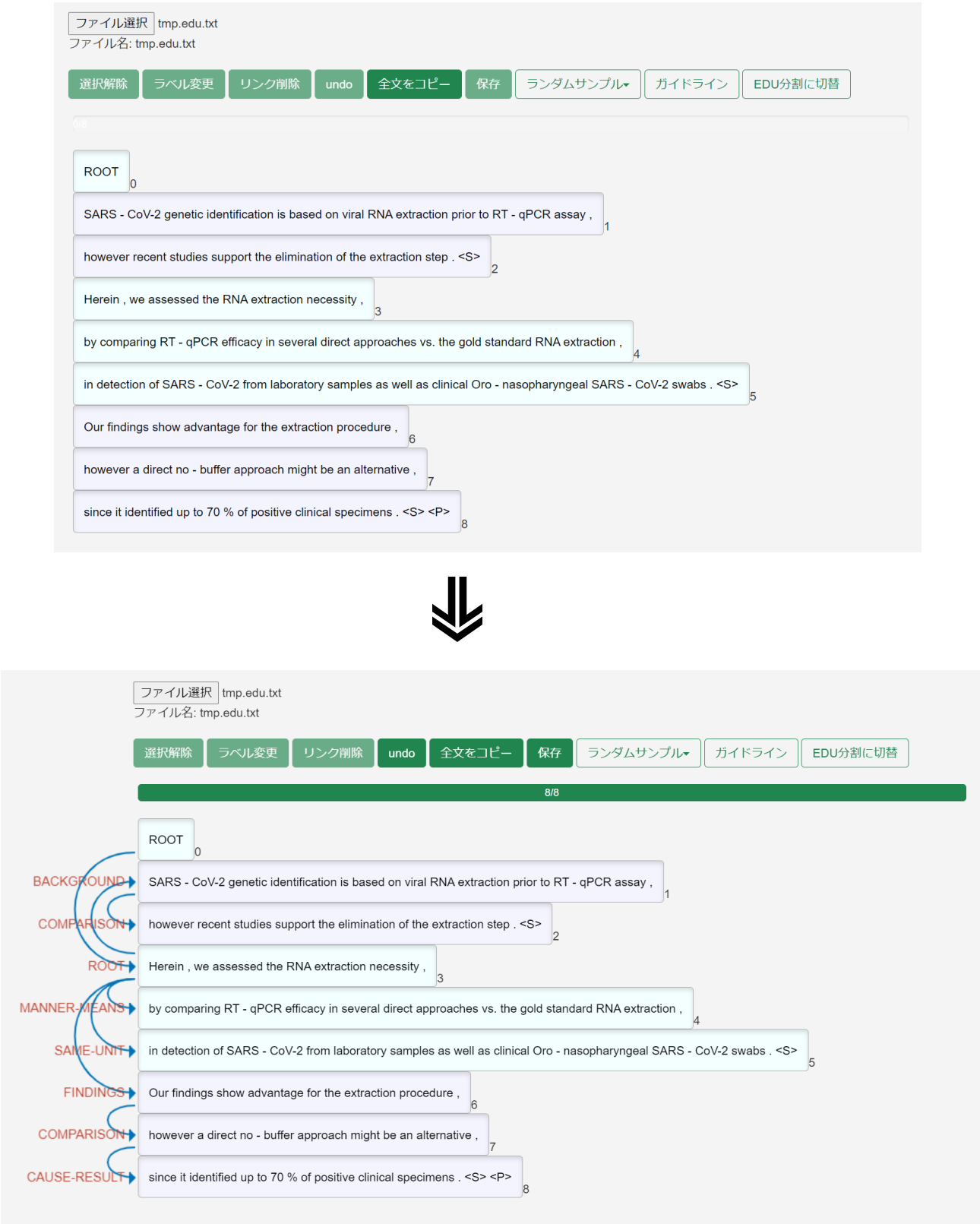


図 11: 談話依存構造アノテーションツールの画面。処理前と処理後。

参考文献

[1] Harry Bunt and Rashmi Prasad. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

[2] Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. In *Technical Report ISI-TR-545*. University California Information Sciences Institute, 2001.

[3] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 2001.

[4] Ofir Israeli, Adi Beth-Din, Nir Paran, Dana Stein, Shirley Lazar, Shay Weiss, Elad Milrot, Yafit Atiya-Nasagi, Shmuel Yitzhaki, Orly Laskar, and Ofir Schuster. Evaluating the efficacy of RT-qPCR SARS-CoV-2 direct approaches in comparison to RNA extraction. *bioRxiv preprint 2020.06.10.144196v1*, 2020.

[5] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, Vol. 8, No. 3, pp. 243–281, 1988.

[6] Mathieu Morey, Philippe Muller, and Nicholas Asher. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, Vol. 44, No. 2, pp. 197–235, 2018.

[7] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, 2008.

[8] An Yang and Sujian Li. SciDTB: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 444–449, 2018.