医学生物学談話依存構造コーパスの構築のための アノテーションガイドライン ver 3.1

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 9 月 15 日

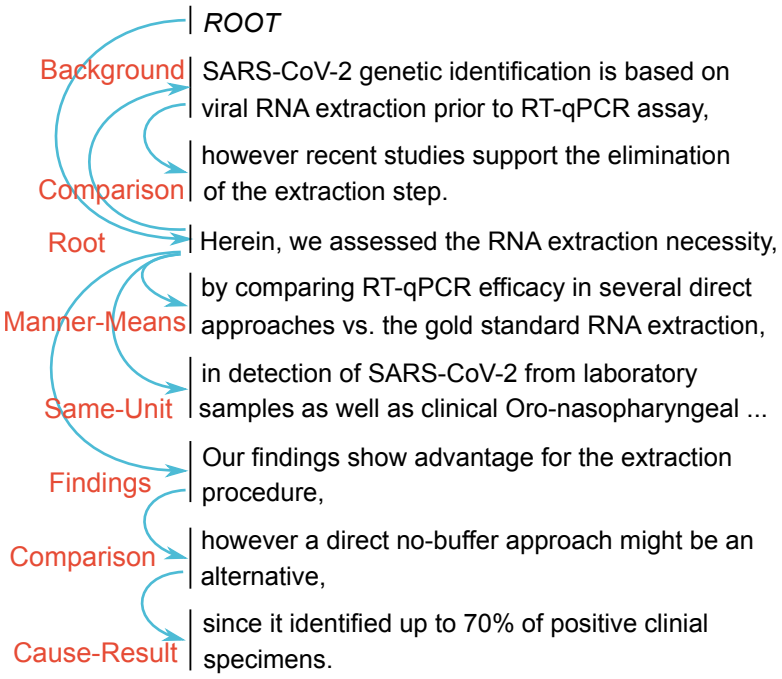

図 1: 医学生物学分野の論文要旨の談話依存構造の例 [3]。

1 コーパス構築の背景と目的

一般的に、文章には意味的、論理的な一貫性 (coherence) があり、一貫性のある文章では文や節 (clause) が作用しあいながら文章全体の論旨の展開を伝える。**談話依存構造** (Discourse Dependency Structure) は、節間の係り受け (談話関係) に基づいて文章がどのように構造化されているのか表現する [5, 4]。具体的には、文章は Elementary Discourse Unit (EDU) と呼ばれる節レベルのテキストスパン (ノード) に分解され、EDU 間の談話関係 (ラベル付き有向リンク) に基づいて一つのグラフ (木構造) として表現される。談話依存構造の例を図 1に示す。談話構造は、文書要約や極性分類、情報抽出など様々な自然言語処理タスクで有用であることが知られている。

談話依存構造解析を実用的な技術にするためには、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重要である。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コーパス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、科学論文は一般的なドメイン (ブログや SNS など) の文章に比べてより強く一貫性が要求されるため、談話依存構造の自動解析技術を研究するのに適している。また、大量にある医学生物学論文から有用な知識・知見を自動で抽出し、体

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.,

系化し、人間が容易にアクセスできるように整備する技術の開発は、特に昨今の世界情勢を踏まえると社会的に喫緊の課題であり、医学生物学論文に的を絞った談話構造解析技術の開発の意義は大きい。

科学技術論文要旨の談話依存構造を収録するコーパスとして SciDTB [6] が既に存在する。SciDTB は自然言語処理分野の論文要旨 798 件に対して人手で談話依存構造をアノテーションしている。しかし、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 を用いて訓練した談話構造解析システムの解析精度は、医学生物学論文要旨に対しては著しく低下する。

本ガイドラインは次の内容で構成されている。2章では、EDU 分割の基準と指針、注意すべき例外事項について記述する。3章では、談話依存構造の構造に関する制約と、本プロジェクトで定義する 15 種類の談話関係カテゴリーについて、例を用いながら詳しく説明する。4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。5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。

2 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。

2.1 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.]

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.]

2.2 例外

Carlson ら (2001) のマニュアル [2] 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)

(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.]

(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.]

- ただし attribution verb (reporting verb, cognitive verb) の目的節 (that 節、疑問代名詞で始まる節) については独立した EDU とする。
  - 注: 疑問代名詞 + to 不定詞は EDU と認めない
  - 注: 帰属先が不明の場合は EDU と認めない

(39) [The paper showed] [that sleep deprivation exacerbates health problems.]

(40) [The paper showed] [why sleep deprivation exacerbates health problems.]

(41) [The paper showed how to avoid health problems in sleep deprivation.]

(42) [It is shown that sleep deprivation exacerbates health problems.] (帰属先が不明のため)

(43) [Sleep deprivation exacerbates health problems,] [according to the paper.]

2. 分裂文・疑似分裂文 (強調構文)、外置構文など

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

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

表 2: 本コーパスで採用する談話関係カテゴリー。

	談話関係カテゴリー	意味
0.	Root	最もトップの EDU (研究の目的、主要報告内容など)
1.	Elaboration	話題の展開 (深化、発展)
2.	Comparison	逆説、対比
3.	Cause-Result	原因 (理由、根拠)、結果
4.	Condition	仮定的条件、前提
5.	Temporal	時間、状況
6.	Joint	並列
7.	Enablement	目的、可能化
8.	Manner-Means	方法、手段、道具
9.	Attribution	帰属 (主張、報告、認識)
10.	Background	研究の背景
11.	Findings	研究の結果・結論
12.	Textual-Organization	文書構造 (e.g., タイトル、タグ)
13.	Same-Unit	埋め込み EDU によって分離された EDU の結合

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

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

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

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

3 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。

まとめると、**談話依存構造のアノテーションは (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カテゴリーを同定することで行う。**

本プロジェクトでは、談話関係を 14 種類にカテゴライズする。表 2に談話関係カテゴリーの一覧を示す。これらは、SciDTB および RST Discourse Treebank [2], ISO 24617-8 [1] を参考に、特に科学技術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、各談話関係カテゴリーについて例を使って説明する。**より多くの例については、アノテーションツールの「ランダムサンプル」ボタンから確認することができる。**

3.0 Root

Root は、文書の先頭に挿入されている Root EDU (親) と、論文要旨で最も代表的な EDU との間の関係カテゴリーとする。一般に、論文要旨で最も代表的 (重要な) 箇所は研究目的部分である。談話依存構造では上述の木構造制約を仮定するため、ここでの子 EDU から Root EDU 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。Root 関係は、各文書に**必ず一度だけ**現れるとする。“In this paper, ...” や “This study shows” などの表現を伴うことが多い。

例

(50) [**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*]$_{\texttt{dep}}$ ...

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

(52) [**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.*]$_{\texttt{dep}}$

(53) [**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 ...*]$_{\texttt{dep}}$

上の例 (53) では、一見 [Here, we report] が Root EDU の子になりそうであるが、“report” は attribution verb であり、それはここで that 節を目的にとっているため、[Here we report] とその後ろの that 節の間に Attribution 関係があり、Attribution 関係では内容側 (that 節) を親と規定しているため (つまり [Here we report] は既に親を一つもっているため)、Root EDU の子もそれにしたがって that 節にシフトしている。

3.1 Elaboration

Elaboration は、親 EDU と、親 EDU の話題を展開する子 EDU との間の関係カテゴリーとする。Elaboration は話題の「チェイン」を表すと考えてよい。Elaboration は談話関係カテゴリー中で最も基礎的であり、出現頻度は最も高い。その他 14 種類のいずれのカテゴリーも該当しない場合は、Elaboration の適用を検討する。Elaboration はディスコースマーカーを伴わないことも多いが、“ and ” や “ also ” (後述の Joint との区別に注意), “ then ”, “ finally ’ , “ moreover ”, “ furthermore ”, “ for example ” などを伴うこともある

注意 1: 話題の深化と発展　親 EDU の話題の展開方法はさらに細かく、話題の**深化** (詳細化、縦展開; Vertical Elaboration) と話題の**発展** (横展開; Horizontal Elaboration) に分けることができる。ここで、話題の深化とは既存の話題 (親 EDU) をさらに詳しく述べるために新情報を追加する展開方法のことであり、一方、話題の発展とは共通の話題 (共通祖先) に関して一連の新情報を (親 EDU とともに) 追加していく展開方法のことである。しかし、これまでの経験からこのような展開方法の区別は多くの事例で判断が難しく、アノテーション時間の増大とコーパスの一貫性の低下の原因となりうる。そのため、展開方法の区別は今後の課題とし、本プロジェクトでは**これら二種類の展開方法を Elaboration カテゴリーとして統一してアノテーションする**。

注意 2: Joint との区別　話題の発展 (Horizontal Elaboration) は、後述する Joint (並列関係) としばしば類似する。これも後述するが、コーパスの一貫性を高めるために、本プロジェクトでは Joint に対して二つの条件を設定し、それらの条件をすべて満たす場合のみ Joint を適用し、その他の場合では Elaboration を適用する。具体的には、以下の二条件を設定する: (1) 各 EDU の間に意味的、論理的、修辞的な順序性が**ない**こと (= 各 EDU の順番を入れ換えても文章として一貫性を損なわないこと); (2) 各 EDU の間になんらかの並列性が示されていること。

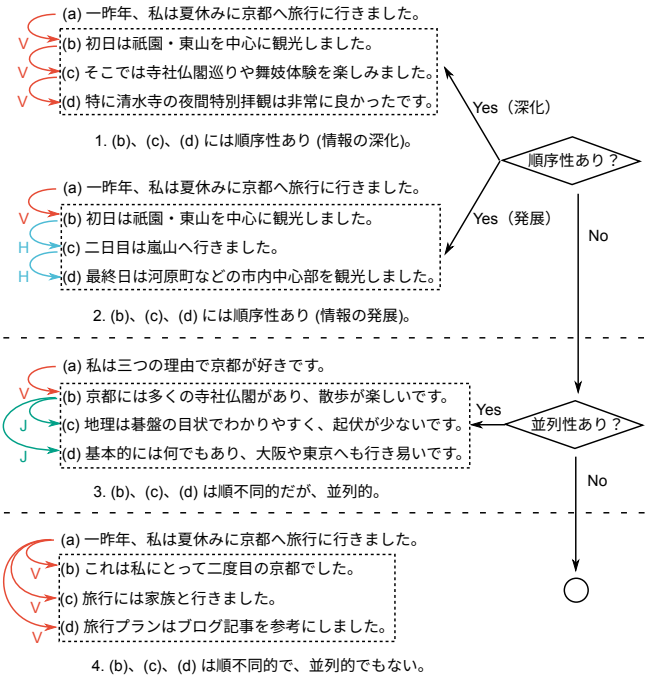

図 2: Elaboration による話題のチェインの例。V は Vertical Elaboration (話題の深化) を、H は Horizontal Elaboration (話題の発展) を、J は Joint (並列; 後述) をそれぞれ表す。本プロジェクトでは V と H の区別をしないため、重要なのは (1) 順序性があるかどうか、(2) 並列的があるかどうか、の二点になる。

注意 3: フローチャート　以上の事項を踏まえて、Elaboration および Joint のアノテーションに関するフローチャートを図 2に載せる。図 2 (1) の例では、(b)、(c)、(d) の間に意味的な順序性 (依存関係) があり、(b) は (a) の 、(c) は (b) の、(d) は (c) のそれぞれ Vertical Elaboration (話題の深化、詳細化、縦展開) である。この例では、(a) で導入された話題は (a) → (b) → (c) → (d) というチェインで (縦) 展開されていることがわかる。図 2 (2) の例でも、(b)、(c)、(d) の間に順序性があり、(c)、(d) は共通祖先 (a) に関する (b) から続く Horizontal Elaboration (話題の発展、横展開) である。先述の通り、本プロジェクトでは Vertical Elaboration と Horizontal Elaboration を区別せず Elaboration カテゴリーとして統一的に扱うため、図 2 (1) と (2) のアノテーション結果は実質的には等しくなる。図 2 (3) の例では、(b)、(c)、(d) の間に順序性はないが、並列性があるため、Joint 関係とする (詳しくは Joint の定義を参照)。図 2 (3) の例では、(b)、(c)、(d) の間に順序性はなく、かつ並列的でもないため、(b)、(c)、(d) はそれぞれ独立に (a) の Vertical Elaboration とする。すなわち、(a) で導入された話題は (a) → (b), (a) → (c), (a) → (d) という独立な三つのチェインで (縦) 展開されていることがわかる。

例

(54) 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*]$_{\texttt{dep}}$ ...

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

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

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

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

3.2 Comparison

Comparison は、逆説や対比など、EDU 間の差異や類似点に焦点をあてる関係カテゴリーとする。“however”, “although”, “while”, “instead”, “in spite of”, “comparing with” などのディスコースマーカーを伴うことも多いが、ディスコースマーカーを伴わなくても意味的に逆説・対比の関係にあるならば Comparison を適用する。

例

(59) [**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.*]$_{\texttt{dep}}$

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

(61) [**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.*]$_{\texttt{dep}}$

3.3 Cause-Result

Cause-Result は、原因 (理由、根拠) と結果の関係にある EDU 間の関係カテゴリーとする。原因 (理由、根拠) 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。“because” や “since”, “as”, “due to”, “because of”, “as a result” などのディスコースマーカーを伴うことが多い。

例

(62) [**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.

(63) 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.

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

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

(66) [**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.*]$_{\texttt{dep}}$

3.4 Condition

Condition は、親 EDU と、その仮定的条件、前提を表す子 EDU との間の関係カテゴリーとする。“if” や “when”, “as far as”, “, given that”, “depending on” などのディスコースマーカーを伴うことが多い。

例

(67) [*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,**]$_{\texttt{head}}$ ...

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

(69) [We envisage] [**that our hypothesis,**]$_{\texttt{head}}$ [*if used clinically as an adjuvant,*]$_{\texttt{dep}}$ [may significantly improve the therapeutic outcomes of the current treatment regimen] ...

3.5 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 との間の関係カテゴリーとする。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 Conditional 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 関係を適用する。もし when 節の内容が仮定的であるならば、Conditional 関係を適用する。

注意 2: Elaboration 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、Elaboration を適用する。

例

(70) [**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).*]$_{\texttt{dep}}$

(71) [**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.*]$_{\texttt{dep}}$

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

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

3.6 Joint

Joint は並列関係にある二つ以上の EDU をまとめるための関係カテゴリーとする。**次の二つの条件をすべて満たす場合のみ、**Joint 関係を適用する。[1]

1. 各 EDU の間に意味的、論理的、修辞的な順序性が**ない**こと (各 EDU の順番を入れ換えても文章として一貫性を損なわないこと)。
2. 各 EDU の間になんらかの並列性が示されていること。

[1]前のバージョンでは独立性、順不同性、並列性の三つの条件としていたが、順不同であるときは常に独立であり、また独立性で分けられる Vertical Elaboration と Horizontal Elaboration は今回は区別しないため、上記の二つの条件にまとめた。

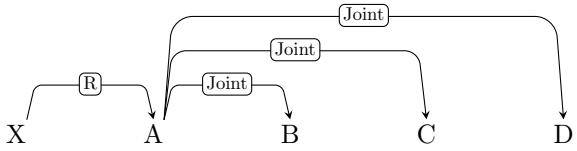

図 3: Joint 関係のアノテーション。EDU A, B, C, D は並列関係にあり (三つの条件を満たし)、最も早く出現する A を親として、B, C, D をその子とする。

一つめの条件では、EDU x, y の内容がお互いに依存せず、各 EDU の順番をシャッフルしたとしても、文書としての整合性が破綻しないことを要求する。例えば、EDU y に含まれる語句が x の中で定義されている場合は、x と y の間には順序性がある。二つめの条件では、x と y が並列であることが何かしらの方法で示されていることを要求する。“(1)”, “(2)”, “(3)” のような局所的なマーカーでなくても、各 EDU の統語構造の対応などで示されることも稀にある。

Joint の親の選択について　本プロジェクトでは Universal Dependency にならい、三つ以上の EDU が並列関係にある場合、並列関係にある EDU のうち**最も前に出現した EDU を共通の親とし**、並列関係にあるその他の EDU はすべてこの共通の親の子として接続する。例を図 3に示す。

例

(74) [**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.*]$_{\texttt{dep}}$

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

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

3.7 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 との間の関係カテゴリーとする。“in order to” や “so as to”, “so that”, “which enables ...” などのディスコースマーカーや、「目的」を表す to 不定詞、for + 現在分詞によって明示的に示されることが多い。

例

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

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

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

3.8 Manner-Means

Manner-Means は、親 EDU と、そのための方法、手段、道具を表す子 EDU との間の関係カテゴリーとする。“using ...” や「手段」の “by” 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 7のようにそこに Manner-Means 関係を、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 8のように Elaboration 関係を、それぞれアノテーションする。また、“Method:” などの見出しによって研究目的のための方法であることが明示化されている場合は、Manner-Means 関係をアノテーションする。

例

(80) [**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.*]$_{\texttt{dep}}$

(81) [**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*]$_{\texttt{dep}}$ ...

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

例 (82) は少し複雑である。一見、[Our work offers ...] と [by demonstrating] の間に Manner-Means 関係を適用したくなるが、“demonstrate” は attribution verb であり、それがここでは that 節を取るため、[by demonstrating] を子、[that small-worldness ...] を親とした Attribution 関係が既に存在する ([by demonstrating] は既に親を持つ)。したがって、この例では “demonstrate” の内容である that 節に対して Manner-Means 関係をアノテーションする。

3.9 Attribution

Attribution は、主張や報告、認識の内容を表す親 EDU と、その帰属先を表す子 EDU との間の関係カテゴリーとする。内容側を親、帰属先を子とする。*attribution verb* (reporting verb, cognitive verb) と呼ばれる動詞 (know, say, show, demonstrate, indicate, argue, find, notice, investigate など) とそれが目的にとる that 節、 疑問代名詞で始まる節によって明示的に示される。ただし、**帰属先が明記されない場合**は Attribution 関係を認めない (そもそも EDU 分割しない)。また、**疑問代名詞 + to 不定詞を目的にとる場合**も Attribution 関係を認めない (そもそも EDU 分割しない)。しかし、“according to” によって帰属先が明記される場合は、“according to” で始まる句 (子) との間に Attribution 関係を認める。

例

(83) [*Our method also suggests*]$_{\texttt{dep}}$ [**that polyprotein 1ab, polyprotein 1a, S, M and N are proteins of viral origin.**]$_{\texttt{head}}$

(84) [*Our results indicate*]$_{\texttt{dep}}$ [**that AVNV is a variant of OsHV-1.**]$_{\texttt{head}}$

(85) [*This paper shows*]$_{\texttt{dep}}$ [**how the self-training can generalize learning models.**]$_{\texttt{head}}$

(86) [*We envisage*]$_{\texttt{dep}}$ [**that our hypothesis,**]$_{\texttt{head}}$ [if used clinically as an adjuvant,] [may significantly improve the therapeutic outcomes of the current treatment regimen] ...

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

3.10 Background

Background は、親 EDU (一般的には研究目的) と、研究背景を表す子 EDU との間の関係カテゴリーとする。Background は科学論文要旨のためのメタ的なカテゴリーであり、論文要旨中における研究背景部分を指定するために用いられる。ほとんどのケースでは、親は Root EDU の唯一の子である研究目的部分とする。一般的に研究背景は論文要旨の先頭から始まるため、論文要旨の先頭部分に子 EDU が現れることが多い。研究背景に関する記述がない場合は、論文要旨中で Background がなくてもよい。

例

(88) [*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 ...

(89) [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.**]$_{\texttt{head}}$ ...

3.11 Findings

Findings は、親 EDU (一般的には Root EDU の子か、提案技術) と、研究の実験・結論を表す子 EDU との間の関係カテゴリーとする。Findings は科学論文要旨のためのメタ的なカテゴリーであり、論文要旨中における結果・結論部分を指定するために用いられる。

研究の結果と結論の両方が識別可能な程度に分けられ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 4や例 (95) 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結果部分と結論部分を分けて記述していたり、さらには “Discussions:” や “Limitations:” のような見出しで結果と結論に加えて考察部分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 4 のように各部分について独立に Findings 関係をアノテーションする。

また、論文全体に対して結果・結論をまとめて記述する代わりに、図 5 のように、各技術、各実験について述べたあとに逐次その結果を記述するようなスタイルもある。例えば図 5 では、技術 A の記述 → 技術 A の結果・結論 → 技術 B の記述 → 技術 B の結果・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

(90) [**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.*]$_{\texttt{dep}}$

(91) [**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.*]$_{\texttt{dep}}$

(92) [**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*]$_{\texttt{dep}}$ ...

(93) [**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*]$_{\texttt{dep}}$ ...

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

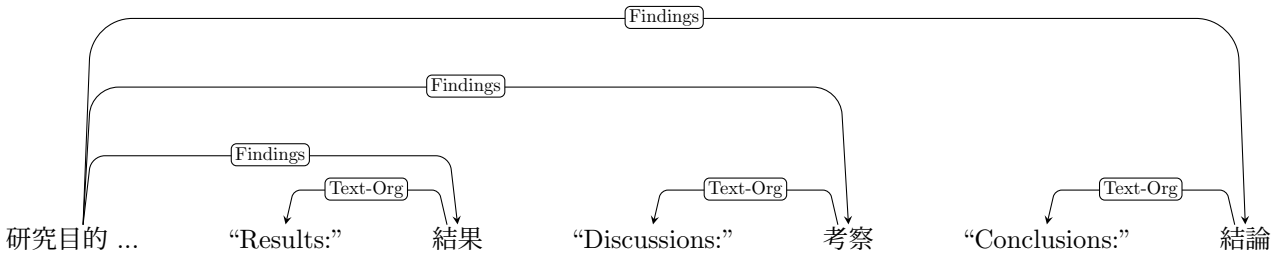

図 4: 複数の Findings 関係をアノテーションする例 1。Text-Org は Textual-Organization 関係カテゴリーを表す。

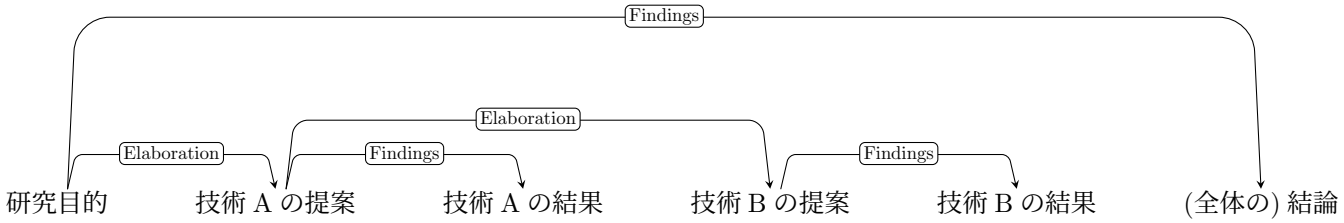

図 5: 複数の Findings 関係をアノテーションする例 2。

(95) [**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.*]$_{\texttt{dep}}$

3.12 Textual-Organization

Textual-Organization は、文書中のテキスト (親 EDU) と、文書中のタイトルやタグ (子 EDU) との間の関係カテゴリーとする。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使われ、各パートの役割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、各パートのトップの親 EDU (テキスト) と、対応する見出しとの間に Textual-organization を適用する。

例

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

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

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

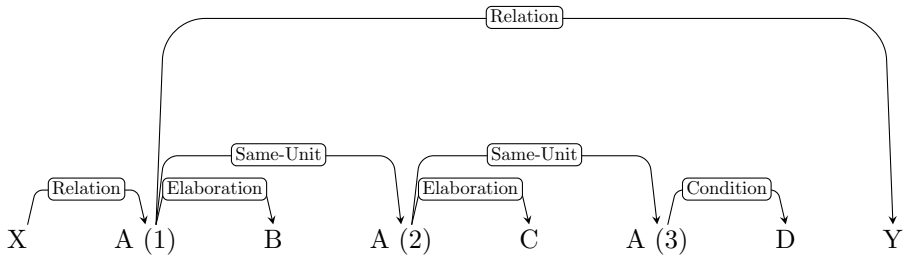

図 6: 同一の EDU が三つ以上の疑似 EDU に分離してしまう場合の結合例。“A (1)” から “D” までは一つの文に閉じているとする。

3.13 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、本当は単体の 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,” という埋め込み EDU によって、“Charles Lutwidge Dodgson was an English writer of children’s fiction.” という一つの EDU が二つの疑似 EDU に分離してしまっている。

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

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関係カテゴリーである。

リンクは常に前方向とする。同一 EDU が別の二つ以上の埋め込み EDU によって三つ以上の疑似 EDU に分離してしまった場合は、それぞれ直前の疑似 EDU と結合させ、同一 EDU 全体として一本のチェインを作るように結合させる。

注意: Same-Unit チェインとの結合　その性質上、Same-Unit は文内の談話依存構造でのみ起こる。Same-Unit でチェインされている疑似 EDU が「同一文内」の別の EDU と Same-Unit 関係以外で関係する場合は、その EDU と最も距離の近く、統語的な結びつきの強い疑似 EDU と結合させる。「同一文外」の EDU と関係する場合は、チェインのトップ (始点) の疑似 EDU と結合させる。図 6の例では、EDU D は EDU A チェインと Condition 関係にあり、EDU D と EDU A チェインは同一文内にあるため、EDU D はチェイン中で最も距離が近い EDU A (3) と結合する。一方、EDU X と EDU Y はそれぞれ EDU A チェインとは同一文には属さないため、それらはチェインのトップである EDU A (1) と結合させる。

例

(100) [**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.*]$_{\texttt{dep}}$

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

4 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。

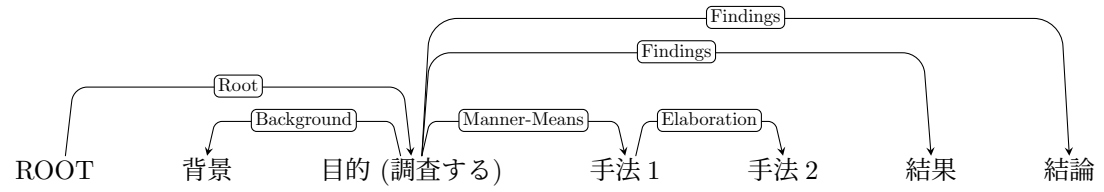

図 7: 典型例 1 (a)

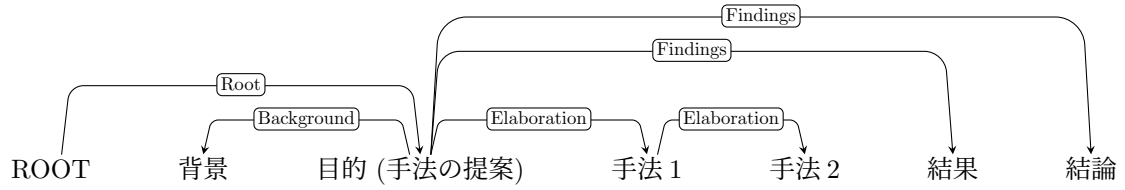

図 8: 典型例 1 (b)

論文要旨の項目　医学生物学分野だけによらず科学技術分野の論文要旨では、大きく以下の情報が含まれるべきだと考えられている。

- 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各項目を記述しており、図 7, 8 のような談話依存構造になる。

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) の中心的な情報に置き、論文要旨はその知見と意義について記述していることも多くある。そのような場合は、談話依存構造は図 9のようになる。

5　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。

5.1　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。[2]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と仮定してよい。

[2]Yang ら [6] 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。

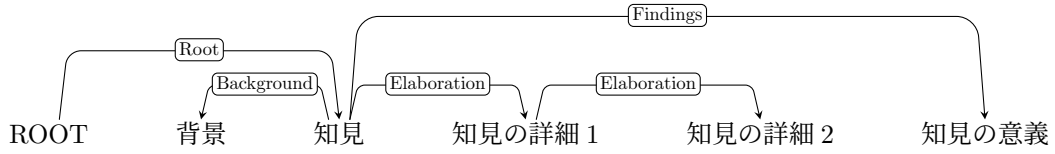

図 9: 典型例 2

5.2 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、2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 2章における EDU 分割ルールよりも優先する。例えば、ある文が一つの句からなっており、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 のうち後ろのほうの EDU の先頭単語をクリックすると、前の方の EDU と統合できる。したがって、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。

5.3 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、3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 EDU (“ROOT”) を除くすべての EDU の親が決定されるまで繰り返す

図 10: EDU 分割ツールの画面。処理前と処理後。

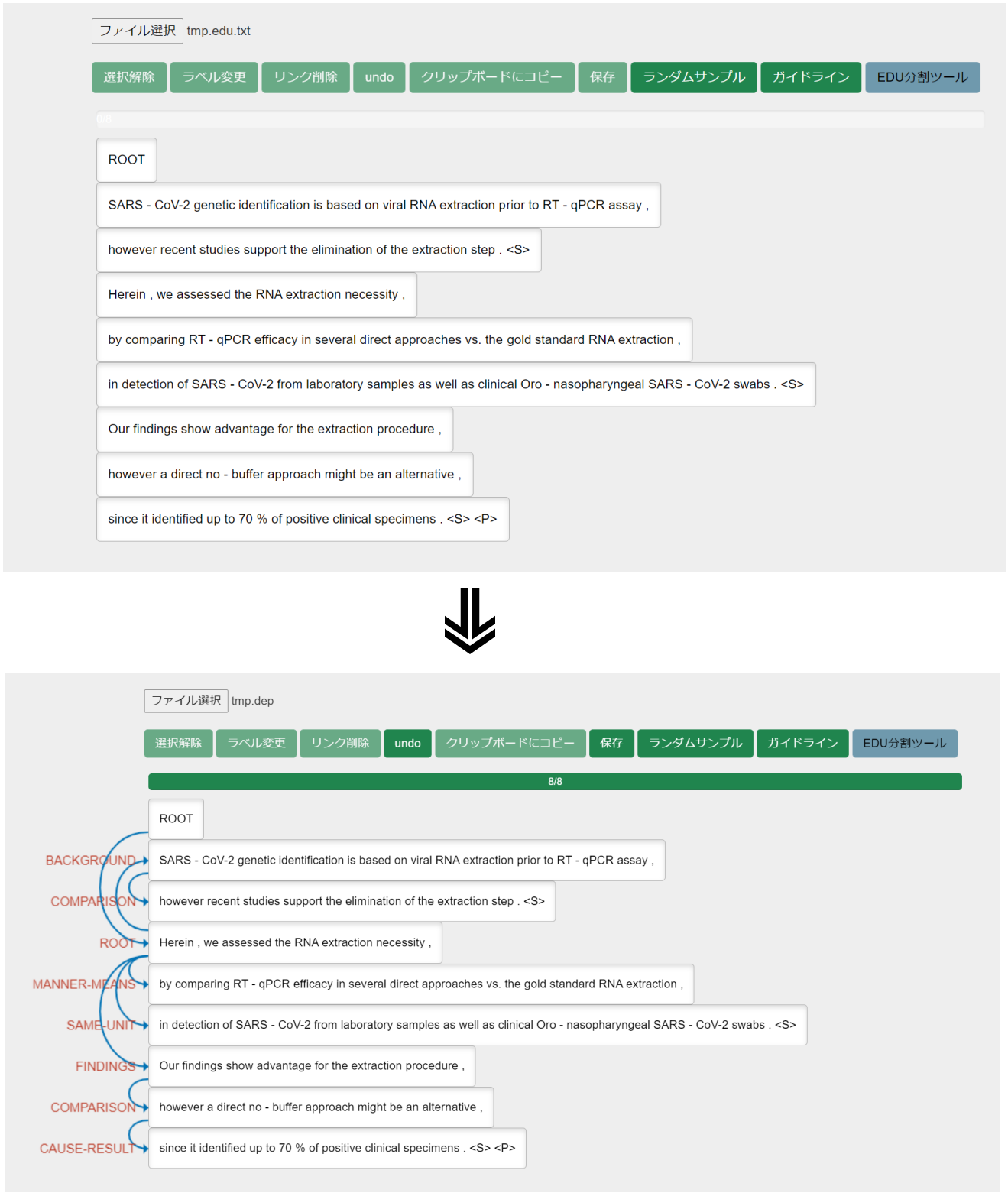

図 11: 談話依存構造アノテーションツールの画面。処理前と処理後。

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。
- 「クリップボードにコピー」: テキストをコピーしたい EDU を選択してからこのボタンを押すことで、対象 EDU のテキストをクリップボードにコピーすることができる（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。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。

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)*, 2014.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.