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 3.0

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 8 月 18 日

1　コーパス構築の背景と目的

文章は一般に意味的、論理的に一貫している。一貫性のある文章では、節 (clause) はお互いに関係しあい、孤立した節は存在しない。**談話依存構造** (Discourse Dependency Structure) は、節間の係り受けによって一つの文章がどのように構造化されているのか表現する方法である [4, 3]。より具体的には、談話依存構造では文章を Elementary Discourse Unit (EDU) と呼ばれる節レベルのテキストスパン (ノード) に分解し、EDU 間の談話関係 (ラベル付き有向リンク) に基づいて文章全体を一つのグラフとして表現する。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、情報抽出などの自然言語処理タスクで有用であることが知られている。

本プロジェクトでは、医学生物学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、科学技術論文の文章は他のドメイン (ブログや SNS など) の文章とくらべてより厳密な一貫性を要求されるため、談話依存構造のスキームや解析技術について研究するのに適している。また、昨今の世界情勢を踏まえると、特に医学生物学分野の論文集合から有用な知識・知見を自動で抽出し、それを体系化したいという潜在的な要望は大きく、医学生物学分野に焦点をあてた談話構造解析システムの開発の社会的な意義は大きい。

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 [5] 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。

refsec:segmentation 章で EDU 分割の基準と指針、注意すべき例外について記述する。3章では、談話依存構造の制約と各種談話関係カテゴリーについて例を用いて説明する。4章では、本プロジェクトのアノテーションで用いる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。また、文書の先頭には便宜的な Root EDU を挿入する。以下では、EDU 分割のための基準（指針）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) に着目することで判断することができる。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。以下のようなケースでは EDU 分割を行う。

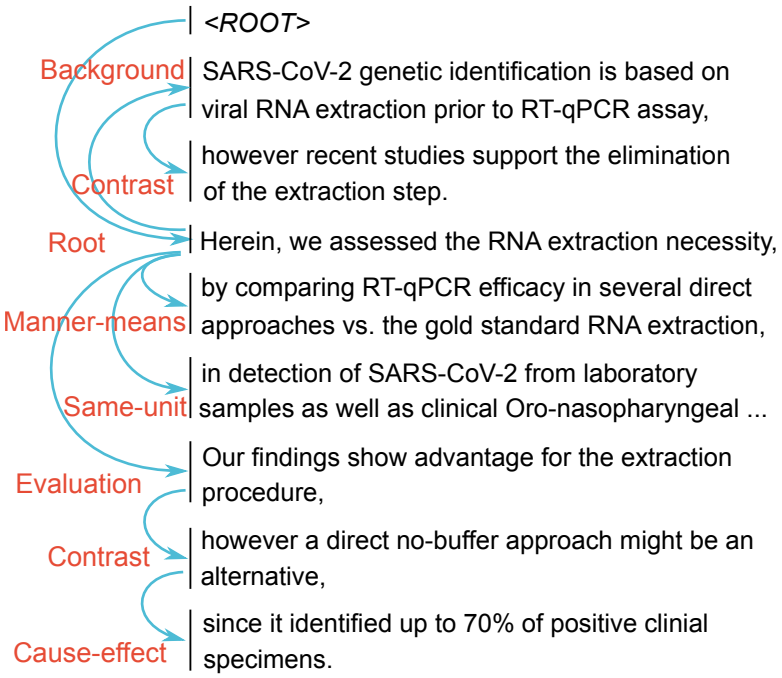

図 1: 医学生物学分野の論文要旨の談話依存構造の例 [2]。

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.,

1. 単文、重文

   (1) [May cleaned the kitchen] [and she vacuumed the dining room.]

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

2. 分詞構文

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

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

3. 「目的」の意の不定詞節 (infinitival clause) や so that 節

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

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

4. 節修飾の「前置詞 + 動名詞」

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

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

5. 関係節

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

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

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

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

6. その他の連体修飾句 (後置修飾句)

(13) [They discussed their plans] [to cut costs.]

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

7. 同格の that 節

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

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

8. 括弧やダッシュによる括り

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

9. ディスコースマーカーを伴う句

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

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

2.2 EDU 分割しないケース

Carlson ら (2001) のマニュアル [1] 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**。

1. 動詞や前置詞の主語、目的語、補語としての節 (clausal subject, clausal object, clausal complement)、

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

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

(22) [He started digging.]

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

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

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

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

- ただし attribution 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

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

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

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

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

(31) [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

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

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

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

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

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

3 談話依存構造の解析

談話依存構造の解析は、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、統語的依存構造解析と同様の**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、**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、談話関係を以下の 15 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 [1], ISO 24617-8 [**?**] 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。以降で、各談話関係カテゴリーについて例を使って簡単に説明する。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

説明 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。

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 (メタ)
1.	Elaboration	話題の拡張 (横展開)、話題の深化 (縦展開)
2.	Comparison	逆説、対比
3.	Cause-Result	原因 (理由、根拠)、結果
4.	Condition	条件、仮定、前提
5.	Temporal	時間関係 (同期、非同期)
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 (メタ)
11.	Evaluation	研究の実験結果 (メタ)
12.	Conclusion	研究の結論 (メタ)
13.	Textual-Organization	文書構造 (e.g., タイトル、タグ)
14.	Same-Unit	埋め込み EDU によって分離された EDU の結合

例

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

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

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

(40) [**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 (40) 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

説明 Elaboration は、親 EDU と、親 EDU の話題を拡張（横展開）したり、または深化（縦展開）するような子 EDU との間の関係カテゴリーとする。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。例えば、研究手法を二つ提案していて、それらを “also” や “moreover” などを使って記述していくケースは横展開の Elaboration とし、名詞や節を関係節や例示などで詳細化するようなケースは縦展開の Elaboration とする。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。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　上述のように、Elaboration は話題の拡張を表す *Horizontal* Elaboration と、話題の深化を表す *Vertical* Elaboration に分けることができると考えられる。しかし、これまでの経験からそのような区別は多くの事例では曖昧性があるため判断が難しく、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、本プロジェクトではこれら二種類の話題の展開方法をまとめて Elaboration 関係とする。

注意 2　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 に対して三つの条件を設定し、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 (Horizontal) Elaboration を適用する。具体的には、以下の三条件を設定する: (1) 各 EDU は意味的にお互い独立していること (独立性); (2) 各 EDU の順番を入れ換えても問題はないこと (順不同性); (3) 並列性が明示的に示されていること。

例

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

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

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

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

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

3.2 Comparison

説明　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

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

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

(48) [**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

説明 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

(49) [**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.

(50) 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.

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

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

(53) [**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

説明 Condition は、親 EDU と、その条件や仮定、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 ...” などのディスコースマーカーを伴うことが多い。

例

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

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

(56) [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

説明 Temporal は、時間的な関係にある EDU 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”) と非同期的な関係 (“before”, “after”) 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。適用はあくまでも時間に基づくケースに限定し、プロセスの手順などの厳密には時間ではなく順序にフォーカスしたケースでは Elaboration を適用する。また、“when” などは厳密には時間的な同時性ではなく “if” と同様の条件を表すために用いられることがあり、そのようなケースでは Condition を適用する。したがって、科学論文では比較的に出現頻度は低い。

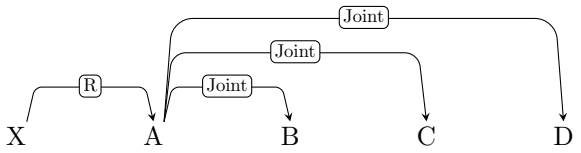

図 2: Joint 関係のアノテーション。EDU A, B, C, D は並列関係にあり (三つの条件を満たし)、最も早く出現する A を親として、B, C, D をその子とする。

例

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

3.6 Joint

説明 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。ここで並列関係にあるとは、次の三つの条件をすべて満たす場合とする。

1. 各 EDU は意味的にお互い独立していること (独立性)。
2. 各 EDU の順番を入れ換えても問題はないこと (順不同性)。
3. 並列性が明示的に示されていること。

一つめの条件では、EDU x, y の内容が他方の EDU に依存していないことを要求する。例えば、EDU x に含まれる語句が y の中で定義されている場合、x と y の間には独立性はない。二つめの条件では、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。EDU x と y の間に独立性があっても、それらの順番を入れ換えることによって不自然さが顕著になるならば、x と y の間には順不同性はない。三つめの条件では、x と y が並列であることが何かしらの方法で明示的に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 2に示す。

例

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

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

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

3.7 Enablement

説明 Enablement は、親 EDU と、その目的を表す子 EDU との間の関係カテゴリーとする。“in order to” や “so as to”, “so that” 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

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

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

3.8 Manner-Means

説明 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。論文要旨中で “Method:” などのサブタイトルによって明示的に研究目的のための方法に関する記述であることが明示化されている場合は、ディスコースマーカーがなくても Manner-Means の適用を検討する。

例

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

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

(65) [**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 (65) 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

説明 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* 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。ただし疑問代名詞 + to 不定詞を目的にとる場合は、attribution verb であってもそこに Attribution 関係を認めない (そもそも EDU 分割しない)。また、“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

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

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

(68) [*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] ...

(69) 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

説明 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

(70) [*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 ...

(71) [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 Evaluation

説明 Evaluation は、親 EDU (一般的には研究目的) と、研究結果・実験結果を表す子 EDU との間の関係カテゴリーとする。Evaluation 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実験結果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。一般的に実験結果は論文要旨の後半に記述されるため、論文要旨の後半に子 EDU が現れることが多い。

例

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

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

(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.**]$_{\texttt{head}}$ ... [Results:] [*Out of 1629 children enrolled, a total of 204 respiratory viruses were present in 171*]$_{\texttt{dep}}$ ...

3.12 Conclusion

説明 Conclusion は、親 EDU (一般的には研究目的) と、研究の結論を表す子 EDU との間の関係カテゴリーとする。Conclusion 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論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。一般的に研究の結論は論文要旨の末尾に記述されるため、論文要旨の終盤に子 EDU が現れることが多い。

例

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

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

3.13 Textual-Organization

説明 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:” などのようなサブタイトルが使われ、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、対応するサブタイトルとの間に Textual-organization を適用する。

例

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

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

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

3.14 Same-Unit

説明 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。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 が二つのスパンに分離してしまっている。

(80) [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スパン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。この例では EDU A (親) と EDU B (子) の間を Same-Unit で結ぶ。常に前方向のリンクとする。その性質上、Same-Unit は文内の談話依存構造でのみ起こる。

注意 Same-Unit は EDU 分割時に既に同定されているため、EDU 分割時にマーカー “$\langle \text{SU} - X \rangle$” を EDU の先頭に挿入する。$X$ は 2 以上の整数であり、マーカーが挿入されている EDU を子、そこから X 個前の EDU が親であることを表す。上の例では $X = 2$ である。

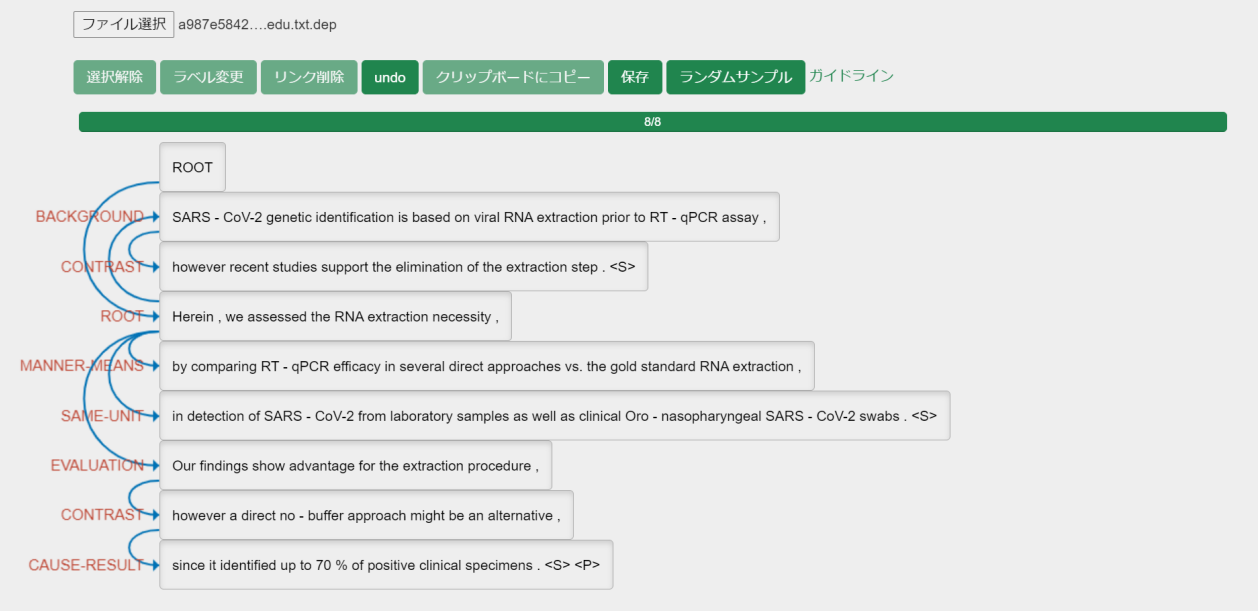

図 3: アノテーションツールの画面。

例

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

(82) [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 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。

4.1 準備

- テキストエディタか Word
- 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 (*.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。[1]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。

4.2 EDU 分割

- 処理前: テキストファイル (*.txt)

[1]Yang ら [5] 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。

- 処理後: EDU 分割済みファイル (*.edu.txt)

EDU 分割では、テキストエディタか Word を使用し、処理対象であるテキストファイルの各行 (=各文) を 2章のルールに従って EDU に分割する。EDU は各文に閉じているとし、複数の文に部分的にでもまたがることがないようにする。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注意する。これは 2章における EDU 分割ルールよりも優先する。すなわち、ある文 (テキストファイルの一行) 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。結果の EDU 分割済みファイルは、EDU が一行ごとに並んでいるとものとする。

4.3 談話依存構造の解析

- 処理前: EDU 分割済みファイル (*.edu.txt)
- 処理後: 談話依存構造ファイル (JSON ファイル) (*.edu.txt.dep)

談話依存構造の解析では、処理対象である EDU 分割済みファイルをアノテーションツールにアップロードし、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。
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
6. 談話依存構造の解析が完了すれば、「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 分割済みファイル、または談話依存構造ファイル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. でクリックした親 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 (*.edu.txt.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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.