F-4

日本語コーパスの談話構造アノテーションに向けた予備的研究

野元　裕樹　　大久保　弥　　佐近　優太
（東京外国語大学）
{nomoto, okubo.wataru.m0, sakon.yuta.n0} @tufs.ac.jp

要旨

談話は単なる発話の連続ではなく、階層構造を成す。英語では、談話構造情報付きコーパスが複数ある。一方、日本語では同様のコーパスはないようである。本研究では、議論下の疑問 (Question Under Discussion; QUD) の枠組み (Roberts 2012, Büring 2003) に基づき、3 名が国立国語研究所統語・意味解析コーパスの 2 つの文章に対し、互いに独立にアノテーションした結果を報告する。QUD の枠組みでは、談話は疑問 (Q) と答え (A) により構造化される。明示的 Q がない場合、非明示的 Q を仮定する。本研究では、非明示的 Q の恣意的設定排除を追求するRiester (2019) に従ったが、アノテーション一致率は依然低かった。以下の改善策を提案する。[1] 可能な構造関係および非明示的 Q の中で使用可能な語彙に制限をかけて理論を精緻化する。[2] アノテーションを 2 段階で行う。

1　はじめに

語や文と同様に、談話も基本単位の単なる時系列上の連続ではなく、基本単位が集まって階層構造を成す (e.g. Mann & Thompson 1988, Asher & Lascarides 2003)。英語では、これらの研究を踏まえた談話構造情報付きコーパスが複数、一般公開されている (cf. Prasad et al. 2008, Wolf & Gibson 2005)。一方、日本語では研究者間で共有される同様のコーパスはないようである。本研究では、議論下の疑問 (Question Under Discussion; QUD) の枠組み (Roberts 2012, Büring 2003) に基づき、2 つの文章に実際にアノテーションを行った結果を報告する。それにより、アノテーションにおける諸問題を明らかにするとともに、改善策を提案し、今後のコーパスアノテーションにつなげる。

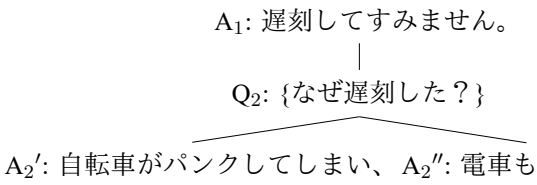
本稿の構成は次の通りである。2節では、QUD の枠組みについて、本研究に直接関係する範囲で概要を紹介する。3節では、本研究で行ったアノテーションの方法について述べる。4節では、アノテーションの結果とそれに対する考察を行う。5節では、今後の本格的アノテーションにつながる提案を行う。

2　QUD の枠組み概要

QUD の枠組みでは、談話は疑問 (Q) と答え (A) により構造化されると考える。明示的 Q が存在しない場合、非明示的 Q を仮定する。例えば、(1) の談話は (2) のような構造を持つと分析する[1]。非明示的 Q は { } に入れて示す。

(1)　遅刻してすみません。自転車がパンクしてしまい、電車も遅れて…。

(2)



[1]Q および A に関する表記法については、注4と樹形図 (15) を参照されたい。

まず、談話全体が基本談話単位 (elementary discourse unit; EDU) に分割される。(1) の 2 文は、節に対応する 3 つの EDU に分割される (A_1、A_2'、A_2'')。(1) には明示的 Q が存在しないので、次に、これらの関係を捉えるものとして、間に非明示的 Q が立てられる (Q_2)。その結果、談話の階層構造と A_2'·A_2'' が A_1 の理由であるという修辞関係が捉えられる。

本研究で QUD の枠組みを採用するのは以下の 3 つの理由による。第一に、QUD の枠組みは近年、英語のみならず様々な言語の意味研究において広く用いられている (北京語: Constant (2014); タガログ語: AnderBois (2016), Latrouite & Riester (2018); スンバワ語: Riester & Shiohara (2018); ペルシア語: Okubo (予定), 日本語: 大久保 (2020))。第二に、分節談話表示理論 (Segmented Discourse Representation Theory; SDRT) (Asher & Lascarides 2003) などに基づく修辞構造を導出可能とされ、さらに詳しい修辞関係が表現できる (Onea 2019)。第三に、QUD の枠組みに基づくコーパスアノテーションが英語やドイツ語で進みつつある (De Kuthy et al. 2018)。

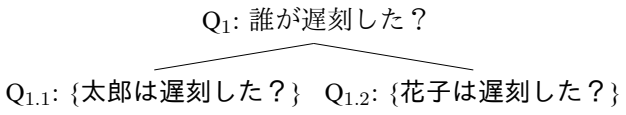
3 アノテーションの方法

QUD の枠組みの詳細は著者により異なるが、本研究では、実際のコーパスアノテーションを視野に入れたRiester (2019) に従った。

筆者 3 名が国立国語研究所統語・意味解析コーパス（NPCMJ）の「お礼の言葉」(nonfiction_rei1508\01O) と「桃太郎」(misc_momotaro) に対し、互いに独立にアノテーションした[2]。このコーパスでは、現代日本語のテクストに対して句構造、文法関係等の統語解析情報が付されている。2 つの文章のサイズは、「お礼の言葉」が 18 ツリー (432 語)、「桃太郎」が 35 ツリー (387 語) である。

Roberts (2012) の枠組みでは、Q は、(3) のように、他の Q の下位疑問としてのみ生じる。

(3)



一方、Riesterは、(2) の Q_2 のように、答えにより惹起される Q も認める[3]。Riester (2019:174) は、非明示的 Q の立て方が恣意的になるのを防ぐために、(4)–(6) の 3 つの制約を提案する。

(4) **Q-A 合致** (Q-A-Congruence)：QUD は、それが直接支配する主張 (assertion) により答えられるものでなければならない。

(5) **Q-所与性** (Q-Givenness)：非明示的 QUD は、所与 (given) の（あるいは少なくとも際立ちの高い）もののみから成り得る。

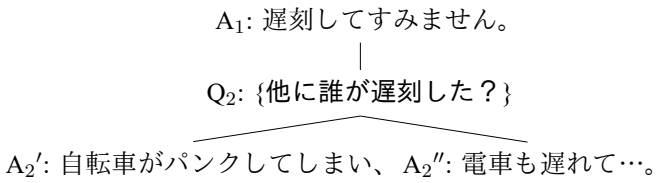
(6) **Q-照応性最大化** (Maximize-Q-Anaphoricity)：非明示的 QUD は、可能な限り多くの所与の（あるいは際立った）ものを含まなければならない。

Q-A 合致 (4) は、(7) の Q_2 のような非明示的 Q を排除する。この Q_2 は A_1 に続く発話としては十分可能であるものの、それが直接支配する主張、すなわち A_2' および A_2'' が答えられるものではない。

[2] http://npcmj.ninjal.ac.jp/（2020 年 1 月 20 日公開）

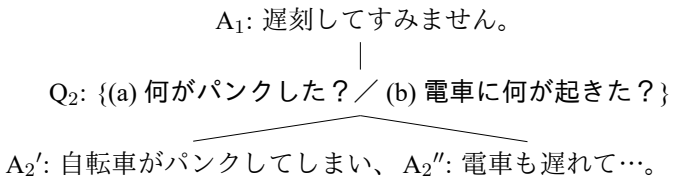
[3] van Keppevelt (1995) は、疑問の意味論を基にした談話構造の分析の中で、このように疑問を生じさせる主張を feeder と呼んでいる。

(7)



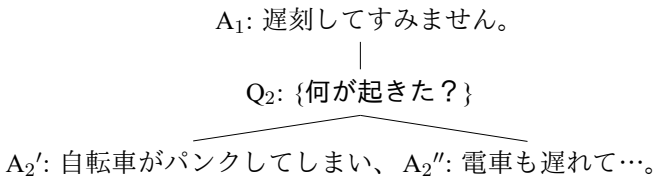
Q-所与性 (5) は、(8) の Q_2 のような非明示的 Q を排除する。A_2' や A_2'' は Q_2 に答えるものである。しかし、A_1 まで考慮に入れると談話が不自然になる。それは、情報構造上の問題、すなわち、新情報として提示されるべき $A_2' \cdot A_2''$ が Q_2 のせいで旧情報になることによる。このような問題を回避するため、非明示的 Q は疑問詞等の一部機能語を除き、新情報を含むべきではない。さらに、Riesterは、Q-所与性は意味だけでなく、非明示的 Q の中で使用される表現形式についても適用されるとする。

(8)



Q-照応性最大化 (6) は、談話の結束性を保証するもので、この制約により (9) ではなく、(2) の方が望ましくなる。(9) では、Q_2 の答えとして A_1 と関連しない主張も可能である。非明示的 Q は、その前後の主張がどのような主題の下で関連するのかを最大限に反映するものでなければならない。

(9)



アノテーションでは、上記 3 つの制約に従い、文章の流れに沿って、各アノテーターが適切と考える非明示的 Q を追加した。A については、焦点 (F)、主題 (T)、対比的主題 (CT)、non-at-issue 要素 (NAI) をマークする。NAI とは、通常の主張とは意味的に独立して存在し、否定などの演算子の影響を受けない要素である。挿入句、非制限的関係節、待遇表現、証拠性標識などがこれにあたる (cf. Potts 2005)。F は Q の直接の答えとなるが、T および CT はならない。NAI は F が含まれる at-issue 要素とは独立しているため、答える疑問も F とは異なる (cf. Simons et al 2010)。(1) の 2 番目の文がそうであるように、文全体が A になるとは限らない。そのため、同時に原文を基本談話単位 (EDU) に分割する必要もある。

4 結果と考察

4.1 Q/A の設定

アノテーションの結果、Q/A の数と 3 人のアノテーター間の EDU 分割一致率は表1のようになった（括弧内は標準偏差)。いずれの文章にも明示的 Q は含まれなかったため、A の数がそのまま EDU の数となる。EDU 分割一致率は、Wolf & Gibson (2005) に従い、「共通の分割の数／（共通の分割の数＋異なる分割の数)」という式により計算した。

「お礼の言葉」は「桃太郎」に比べ一文が長く複雑であるため、原文分割数とA/EDUの数の差が大きくなる。例えば、文(10)は、二人が(11)のような4分割をした。もう一人は(11a)と(11b)を一つのAとする3分割をした。

表1: Q/Aの数とEDU分割一致率

	お礼の言葉		桃太郎	
原文分割数	18		35	
Aの数平均	28.7	(2.1)	40.0	(3.6)
Qの数平均	31.0	(2.0)	43.0	(16.1)
EDU分割一致率平均	.67	(.06)	.62	(.23)

(10) 思い起こしてみると、私が大学院生であった頃は、自分のやりたいことをどう進めていいのか自信を持てず、また周囲にも理解してもらえず、苦しい時が続きました。

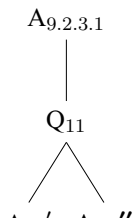
(11) a. [思い起こしてみると、]NAI [私が大学院生であった頃は、]T
b. [自分のやりたいことをどう進めていいのか自信を持てず、]F
c. また[周囲にも理解してもらえず、]F d. [苦しい時が続きました。]F

一方、原文分割数が多い「桃太郎」では、特に非明示的Qの数に、アノテーター間で大きな差が生じた。

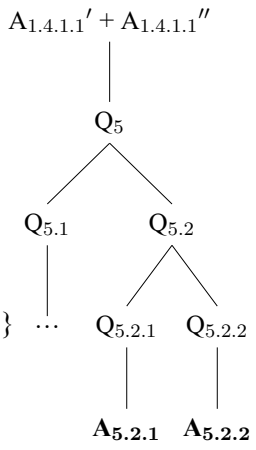
一致率が低い主な原因は、複数の文から成るセリフと従属節の扱い、T/NAIの認定の違いであった。例えば、文(12)は、一人はα = NAI、β = Fとして2つのAと、別の一人は両方Fであるとして2つのAと分析した。それぞれ、(13)、(14)のような異なるQおよびそれを含む談話構造を設定している。

(12) [言語学の教育と研究を続けていくには厳しい状況も出て来てはいますが、]$_\alpha$ [次の世代に手渡せるようなしっかりしたものを作るために、まだもう少し与えられた時間を使いたいと思っています。]$_\beta$

(13) $A_{9.2.3.1}$: 年々新しく入ってくる学生の真剣な取り組みを見るにつけ、[心身にいつの間にか張り付いたサビが消えるのを感じます。]F
Q_{11}: {年々新しく入ってくる…サビが消えるのを感じ、その結果、どうなった？}
A_{11}': そういう時、[教師というのは得な職業だなと思います。]F
$\mathbf{A_{11}''}$:[言語学の…いますが、]NAI [次の世代に…と思っています。]F



(14) $A_{1.4.1.1}'$: [今年8月に私が60才になることから、8月1日に異文化間教育論講座の学生や卒業生、スタッフの皆さんが集まって、還暦記念パーティーを開いていただきました。]F
$A_{1.4.1.1}''$: また、[当日出席できない方々を中心として、素敵なプレゼントやメッセージをいただきました。]F
Q_5: {[これまでは]T どのような状況だったか？}
$Q_{5.1}$: {[東北大学赴任以前は]T どのような状況だったか？}
$Q_{5.2}$: {[今の状況は]T どうような状況か？}
$Q_{5.2.1}$: {[今は]T 何が起きている？}
$\mathbf{A_{5.2.1}}$:[言語学の…いますが、]F

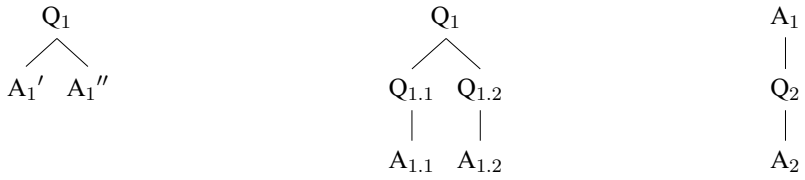


$Q_{5.2.2}$: {[今は]T どのように思っている？}
$A_{5.2.2}$:[次の世代…と思っています。]F

4.2 直前の A との構造関係

EDU への分割が同じであっても、非明示的 Q の立て方次第で前後の EDU との談話構造上の関係は異なり得る。そこで次に、分割が一致した A について、直前の A との構造関係を調べた。各関係は (15) に示すような配置を取る[4]。

(15) a. 姉妹（Q を共有）　b. いとこ（上位疑問を共有）　c. 支配（Q を惹起）



各関係の生起数と 3 名の間の一致率の平均は表2のようになった（括弧内は標準偏差）。

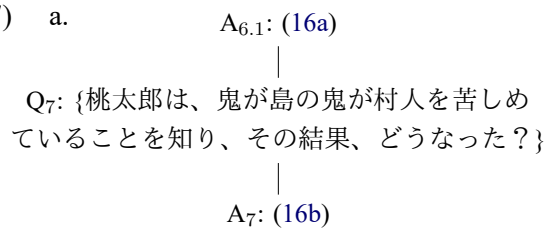
一致率低下の主な要因は、A が答える Q の由来が直前の A か上位の Q かの違いであった。例えば、(16) の 2 つの A は分割も付与されている T/F も同じである。しかし、(17a) と (17b) では異なる構造関係にある。

(16) a. やがて、[大きく育った桃太郎は、]T
[鬼が島の鬼が村人を苦しめていることを知り、]F
b. [鬼退治に行くことにしました。]F

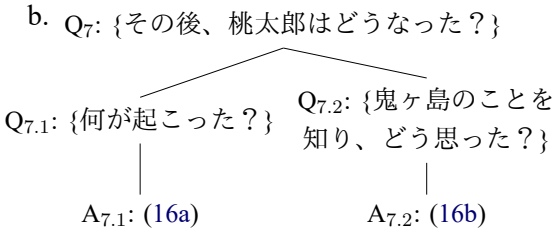
表 2: 直前の A との構造関係

	お礼	桃太郎
姉妹	3.0 (4.2)	10.5 (2.1)
いとこ	6.5 (0.7)	4.0 (4.2)
支配	9.5 (2.1)	15.0 (7.1)
その他	9.5 (3.5)	11.5 (6.4)
一致率	.69 (.12)	.60 (.14)

(17) a.



b.



5 よりよい談話構造アノテーションに向けて

最後に、アノテーション作業を通じて浮かび上がった問題点を整理し、改善策を提案する。

QUD の枠組みにおける非明示的 Q の立て方は、その恣意性が批判対象になってきた。実際の談話では、談話参与者の間における発話意図の解釈に差が生じることは頻繁に起こることであり、ある発話について想定可能な非明示的 Q が複数存在することは問題ではない。問題

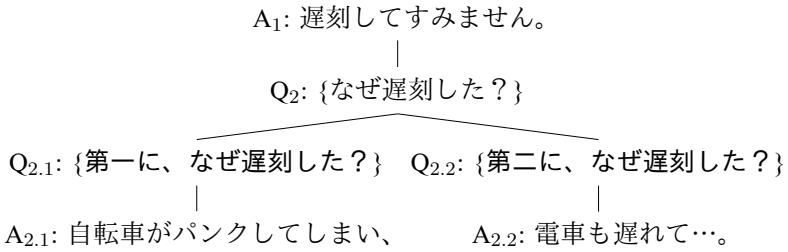
[4] 【表記法】
- Q_n に対する A は、Q と同じ数字 n を添え字とし、A_n と表記する。A が複数ある場合には、数字の後に「′」を付ける。
- Q_n の下位疑問は、Q と同じ数字の後にピリオドで通し番号を付け、$Q_{n.1}$、$Q_{n.2}$、…のように表記する。
- A に惹起される Q に添える数字は、その時点で談話において最大の数字に 1 を加えた数字を用いる。例えば、談話中に Q_7 まで存在するならば、新たに惹起される疑問は Q_{7+1}、すなわち Q_8 となる。

は可能な非明示的 Q の範囲について十分な制約が客観的に規定されないことである。本研究は、恣意性排除を追求するRiester (2019) に従ったが、3 名の分析の一致率は依然低く、批判が現状では妥当であることを裏付ける。

QUD の枠組みによるコーパスアノテーションの信頼性向上のために、以下の改善策を提案をする。

理論の精緻化　[1] 可能な非明示的 Q にさらに制限をかける。典型的な並列的答えを表す姉妹関係 (15a) は、下位疑問となる非明示的 Q を立てることで、いとこ関係 (15b) として捉えられる。例えば、(2) の 2 つの A は、これは (18) のように、リスト関係を規定する下位疑問を立てることができる。

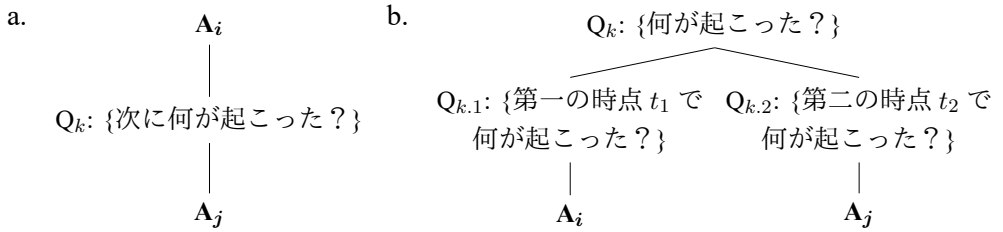
(18)



これにより、姉妹関係をなくせるだけでなく、構造関係に関するアノテーター間の一致率も向上する。

[2] 非明示的 Q に使える所与でない語彙をごく少数に限定する。Q-所与性 (5) を違反してでも用いることができる所与でない語彙としては、疑問詞がある。しかし、その他に使用可能な具体的な語彙については明確な言及がないのが現状である。本研究では事前にアノテーター間でそれについて話し合うことはしなかったために、様々な非明示的 Q が設定され、それが全体の談話構造の不一致にもつながった。例えば、物語によく見られる時間的継起の並列を考えよう。ある主張 A_i から「次に何が起こった？」という Q_k を立てると、その答え A_j は A_i と支配関係になる (19a)。一方、「次に」が使用可能な語彙に含まれなければ、(18) で用いたリストの時間軸による項目順序付けを行うような下位範疇を用い、(19b) のような構造を考えることになる。この場合、A_i と A_j はいとこ関係になる

(19)



使用可能語彙を事前に定義しておくことは、SDRT や工学的応用で普及している修辞構造理論 (Rhetorical Structure Theory; RST; Mann & Thompson 1988) で、Elaboration、Result、Narration、Result といった EDU 間の関係の集合を事前に定義していることに対応する。違いは、QUD の枠組みでは関係自体は定義されるのではなく、疑問の性質により決定されることである。例えば、(19b) の 2 つの A は、$Q_{k.1}$ と $Q_{k.2}$ により Narration に相当する関係となる。QUD の枠組みでは、修辞関係の分析が事前に準備された関係への分類ではないので、SDRT や RST よりも細かな修辞関係の記述・分析が可能となる。

アノテーションの具体的な進め方　アノテーションは 2 段階で行う。まず、マニュアルを整備して、主題 (T)、対比的主題 (CT)、non-at-issue 要素 (NAI) の認定と EDU への分割を共通させる。特に、NAI はアノテーター間で分析が分かれることが多く、日本語の意味記述の中でも手薄な現象なので、多くの具体的な現象をマニュアルに列記する必要がある。次に、F の認定とそれを問う非明示的 Q の設定および EDU の確定を行う。EDU への分割のアノテーター間一致率が上がれば、談話構造の一致率も上がるはずである。第 1 段階では、日本語文法の先行研究や NPCMJ が提供する非明示的要素や統語構造の情報が活用できる。いずれの段階においても、複数のアノテーターによる合議が必要である。

参考文献

AnderBois, S. 2016. A QUD-based account of the discourse particle *naman* in Tagalog. In H. Nomoto et al. (eds.), *AFLA 23: The proceedings of the 23rd meeting of the Austronesian Formal Linguistics Association*, 20–34.

Asher, N. & A. Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Büring, D. 2003. On d-trees, beans, and B-accents. *Linguistics and Philosophy* 26(5). 511–545.

Constant, N. 2014. *Contrastive topic: Meanings and realizations*. マサチューセッツ大学博士論文.

van Keppevelt, J. 1995. Discourse structure, topicality and questioning. *Journal of Linguistics* 31. 109–147.

de Kuthy, K., N. Reiter & A. Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 1932–1938.

Latrouite, A. & A. Riester. 2018. The role of information structure for morphosyntactic choices in Tagalog. In S. Riesberg et al. (eds.), *Perspectives on information structure in Austronesian languages*, 247–284. Berlin: Language Science Press.

Mann, W. C. & S. A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3), 243–281.

大久保 弥. 2020. 追随的疑問における伴立:「それも」の談話構造的分析.『日本言語学会第 160 回大会予稿集』. 327–333.

Okubo, W. 予定. Contrastive Topic and directionality: A comparative analysis of CT markers in Persian. *Proceedings from the Annual Meeting of the Chicago Linguistic Society* 55. Chicago, IL. Chicago Linguistics Society.

Onea, E. 2019. Underneath rhetorical relations: The case of Result. In M. Zimmermann et al. (eds.), *Questions in discourse*, vol. 2, 194–250. Leiden: Brill.

Potts, C. 2005. *The logic of conventional implicatures*. Oxford: Oxford University Press.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi & B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2961–2968.

Riester, A. 2019. Constructing QUD trees. In M. Zimmermann et al. (eds.), *Questions in discourse*, vol. 2, 164–193. Leiden: Brill.

Riester, A. & A. Shiohara. 2018. Information structure in Sumbawa: A QUD analysis. In S. Riesberg et al. (eds.), *Perspectives on information structure in Austronesian languages*, 285–311. Berlin: Language Science Press.

Roberts, C. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics* 5. 1–69.

Simons, M., J. Tonhauser, D. Beaver & C. Roberts. 2010. What projects and why. *Proceedings of SALT* 20. 309–327.

Wolf, F. & E. Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics* 31(2). 249–287.