談話行為タグ付きウズベク語談話コーパスの設計と予備的観察

日髙　晋介
筑波大学
hidaka.shinsuke.gt@u.tsukuba.ac.jp

概要

本稿では、ウズベク語談話コーパスの基本設計と、試行注釈に基づく予備的観察を報告する。本コーパスは自由会話 4 本 (合計 80 分 46 秒，9857 語) からなり、ELAN 上で発話を「長い発話単位」に区切ったうえで、転記テキスト、日本語訳、英訳、グロス、談話行為タグを付与する。談話行為の注釈には ISO 24617-2 (2020) を参照する。予備的観察では、肯定的反応表現 *aha* と *xo‘p* に着目し、同じく肯定を表す形式でも発話連鎖上の位置によって他者フィードバックあるいは自己フィードバックの機能が分化し得ることを示す。

1 はじめに

本稿では、中央アジアチュルク諸語談話コーパス構築の一環として進めているウズベク語[i]談話コーパスの基本設計について報告する。

本コーパスでは、各発話単位に転記テキスト、日本語訳、英訳、形態素区切りとグロス、談話行為タグを付与しており、対話における談話機能を多層的に記述・参照できるようにしている。さらに、試行注釈に基づく予備的観察として、談話に頻出する肯定的な反応表現 (例：*aha*、*xo‘p*)の機能を取り上げ、同一形式であっても発話連鎖上の位置に応じて談話行為が異なり得ることを示す。

2 コーパスの基本設計

2.1 目的

筆者は、これまで中央アジアのチュルク諸語五言語 (トルクメン、カザフ、キルギス、ウズベク、現代ウイグル) において、書き言葉コーパスと母語話者への聞き出し調査を用いて、記述研究を行ってきた。しかし、書き言葉コーパスでは会話文が相対的に少なく、話者の心的態度や聞き手への働きかけを表すモダリティ要素ならびにフィラー等の談話的要素の出現を十分に観察できないという課題があった。そこで本稿では、従来の手法では扱いにくかった、対話に高頻度で現れる要素を分析する基盤として、談話コーパスの構築を行う。本稿では、特に注釈作業が進行している、ウズベク語の談話コーパスを取り上げる。

2.2 収録方法

談話を撮影・録音するにあたり、談話に参加する二名の話者には次の 2 点のお願いをした：1. 約 20 分間、協力者 2 名に当該の言語で自由に話してもらうこと、2. 個人情報へのマスキングの負担を最小限に抑えるため、個人的な話題はなるべく話さないこと。なお、全ての協力者は、録画前に、筆者による研究の説明を受けたうえで、同意書への署名を行った。

録音場所は、筑波大学内にある筆者の研究室あるいは筑波大学東京キャンパス内の教室である (図 1 参照)。撮影時には、全体撮影用として、話者二人が画角全体に入るように話者の横に ZOOM 社製 Q8n-4K 一台を設置し、バックアップ用として、身振りが画角に入るように各話者の斜め前に ZOOM 社製 Q2n-4K 一台 (計二台) を設置した。ヘッドマイクは、SHURE 社製 WH20TQG ダイナミック型ヘッドウォーンマイクロホンを Q8n-4K に接続して使用した。いずれのカメラでも 44.1kHz, 16-bit で撮影を行った (ISO 24617-2 2020: 15-17) 参照)。談話の撮影方法については、伝・榎本 (2014) を参考にした。

[i] ウズベク語は、チュルク諸語のひとつであり、現代ウイグル語に系統的に最も近い。主にウズベキスタン共和国で話され、母語話者数は約 2800 万人である。いわゆる膠着的な言語であり、語順は「修飾語—被修飾語」「主語　目的語　述語」である (以上は日髙 2025 の要約である)。

図 1　対話風景

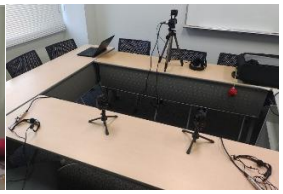

図 2　撮影機材

次に、談話に参加したウズベク語母語話者の情報を生年順に以下に並べる。冒頭のアルファベットは各話者のファーストネームの頭文字である：

1. U: ウズベキスタン・アンディジャン出身、1997 年生、男性
2. B: ウズベキスタン・ナマンガン出身、1998 年生、男性
3. S: ウズベキスタン・シルダリア出身、1999 年生、男性
4. R: ウズベキスタン・ホラズム出身、2003 年生、女性

2.3 コーパスの規模

本稿で報告するウズベク語談話コーパスは、合計 4 本の対話データから構成されている。各対話の収録時間数とのべ語数は以下のとおりであり、合計で 80 分 46 秒、9857 語である。各談話の冒頭に付した談話名は「撮影年月日_整理番号_参加者 1_参加者 2」を指している:

1. 20241223_01_B_S: 20 分 33 秒、 2349 語
2. 20241224_01_B_U: 18 分 33 秒、 2453 語
3. 20241227_01_B_R: 19 分 29 秒、 2249 語
4. 20241227_02_U_R: 22 分 11 秒、 2806 語

現段階では非常に少ないサイズのコーパスとなっているが、今後、談話行為注釈等の付加情報を整備しつつ、データの拡充を進める。

2.4 研究用付加情報

本コーパスにおいては、会話の映像と音声を収録した上で、注釈ソフト ELAN を用いて、情報を付与する。図 3 に ELAN における注釈画面を示す。ここに示されているのは、談話 20241224_01_B_U 冒頭部分である。

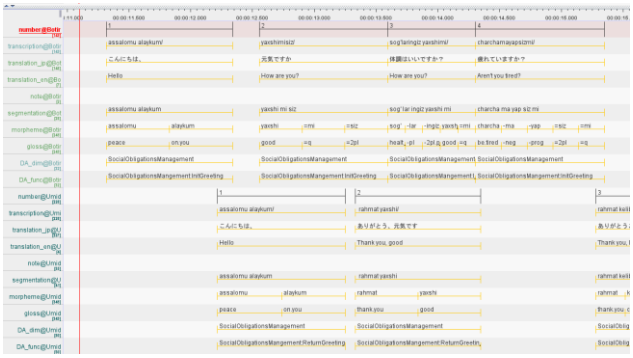

図 3　ELAN における注釈画面

本コーパスでは、各発話に、転記テキストに加え、日本語訳、英訳、グロス (形態論情報の略号)、談話行為タグを付与する。図 4 に注釈の例を挙げる。以下では、図 4 中の赤字で示された情報について詳述する。

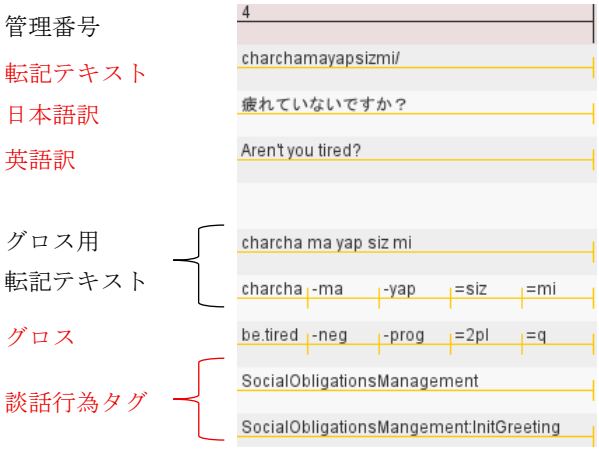

図 4　注釈例

転記テキストは、基本的にウズベク語のラテン文字正書法にしたがって書き起こす。言いさし・言い誤りについては、伝・榎本 (2014: 7)、Japanese Discourse Research Initiative. (2017: 2) による転記記号を採用する。ただし、言語音とはみなせない音声については <> で囲む。音声であればラテン文字による簡易的な音声表記で記し、笑い声は hhh、呼吸は hh で表記する。

各発話は「長い発話単位」で区切る。「長い発話単位」とは、「話し手と聞き手が行為や情報を交換する際の基本単位に相当し、統語的・談話的・相互行為的な一まとまり」(Japanese Discourse Research Initiative. 2017: 2) を指す。

日本語訳は原則として筆者が付与する。ただし，筆者のみでは訳出が困難な箇所については母語話者の協力を得て作成する。英訳は、転記テキストをも

とに Google Translate により生成した訳を参照しつつ、筆者が原文を確認・修正した上で付与する。

グロス (形態論情報の略号) は Leipzig Glossing Rules (Max Planck Institute for Evolutionary Anthropology. 2015; LGR) に従い、LGR にないグロスは下地 (2025) を参照して、それぞれ付与する。

最後に、談話行為タグについて概観する。下記の (1) と (2) の発話は、与えられた文脈の中で、他者に情報を提供したり、他者から情報を要求したりなどといった様々な機能を果たす (伝 2015: 102)。

(1) chiba0132: 7.3876-8.7060
C: 中二の子たち持ってて: ←情報提供

(2) chiba0232: 150.3486-151.5822
B: 今日の発表はどうだったの? ←情報要求

伝 (2015: 102) によれば、談話行為タグは、談話交換構造の理論 (Sinclair and Coulthard 1975) と言語行為論 (Austin 1962, Searle 1969) を組み合わせた観点から、上記の例で示したような発話機能を分類したものであるという。国際的な談話行為タグは国際標準化規格 ISO 24617-2 (2020) としてまとめられており、本コーパスもこの規格で提案されている談話行為タグをもとに注釈付与を行う。この規格では、発話機能の多重性が 11 の次元にまとめられている (ISO 24617-2 2020: 15-17)。ISO 24617-2 (2020) では、多くの次元で一般的に利用される機能 (一般目的機能) のみならず、特定の次元でのみ利用される機能 (次元特有機能) も挙げられている。次元特有機能の詳細については、ISO 24617-2 (2020: 40-61) の Annex C を参照されたい。応答先の同定に関しても、機能的依存関係とフィードバック依存関係の 2 種類を区別している。

本コーパスでは、国際標準化規格 ISO 24617-2 (2020) を基準に、「長い発話単位」に対して 2 層の談話行為タグを試行的に付与する。1 層目では次元 (12 種類) でタグ付けを行い、2 層目では、タスクと判断した場合には一般目的機能 (7 種類)で、タスク以外の次元であると判断した場合には次元特有機能 (35 種類) でタグ付けを行う。なお、今後、応答先の同定についてもタグを付与する予定である。

3 予備的観察

現時点では談話全体に対する注釈作業は進行中であるが、本節では試行的な注釈の結果に基づき、特徴的な事例について予備的に示す。以下では、談話に頻出する肯定的応答表現である *xo‘p* と *aha* を取り上げ、同じ肯定的反応を表すように見えても談話上の機能が異なる点を示す。

xo‘p は「承認小詞」(Begmatov et al. 2008: 436) であるとされ、ウ日辞書 (中嶋 2015: 258) には「①よろしい、わかった、OK」という意味が記載されている。他方、*aha* は、筆者の調べた限り、辞書にも文法書にもまとまった記述は確認できない。

以下に、本コーパスからの用例を挙げる。それぞれ、(3) は、一方の発話の間に、相手が *aha* を発した例、(4) は、話し手が自分の発話を受けて *xo‘p* を発した例、(5) は、一方の発話内で、相手が *aha* あるいは *xo‘p* を発した例、である。

(3) [20241224_01_B_U: 01:06.076-01:12.959]
U: *qanaqa savol-lar-ingiz bor yaponiya-ga*
how question-PL-2PL.POSS exist Japan-DAT
kel-ib (.) *men siz-ga nisbatan*
come-CVB.SEQ 1SG 2PL-DAT compared.to
sal ko‘p-roq yasha-gan=man
little many-COMP live-PRF=1SG
「どんな質問がありますか、日本に来てから。私はあなたよりも少し長く暮らしています。」

B: [***aha***
INTR
「そうですね。」

U: [*shu-ning uchun imkon bor-i=cha* men
that-GEN for chance exist-3.POSS=ADVL 1SG
javob ber-a ol-a=man deb
answer give-CVB.CNT take-NPST=1SG QT
o‘yla-y=man
think-NPST=1SG
「だから、チャンスがあれば私は答えることができると思います。」

(4) [20241227_02_R_U: 04:52.590-05:03.425]
U: *mashhur emas edi men o‘z-im*
famous NEG PAST1SG own-1SG.POSS
unaqa-dan vodiy-dan bo‘l-ganlig-im
like.that-ABL valley-ABL be-PTCP.PAST-1SG.poss
uchun (.) *unaqa ham* [<hh>
for like.that also
「有名ではありませんでした。私自身、あれだ、谷出身だから、そのように (吸気音)」

R: [*aaa* *mmm*
FIL FIL
「ああ、えーっと」

U: ***xo‘p*** *televizorlar-da* *shu* *biz* *vaqt-imiz-da*
yes television-LOC that 1PL time-1PL.POSS-LOC
naruto *degan* *uncha* *bo‘l-ma-gan*
Naruto called like.that be-NEG-PRF
「はい、テレビで、それが、我々の時には、ナルトという、そんなものはありませんでした。」

(5) [20241224_01_B_U: 01:26.000-01:35.453]

B: *shipment* *aaa* *biz-gacha* *kel-a=di* *ekan*
shipment FIL 1PL-until come-NPST=3 EVID
「荷物が、あー、わたしたちのところまで来るそうです」

U: ***aha***
FIL
「そうですね。」

B: *shipment* *aaa* *bizga* *kel-guncha*
shipment FIL 1PL-DAT come-CVB.TER
「荷物が、あー、我々のところにくるまで」

U: ***aha***
FIL
「そうですね。」

B: *transportirovka-si* *qanchadir* (D_bal) *katta*
shipping.fee-3.POSSshow=INF big
mahsulot-lar-ni *ol-sa-k*
product-PL-ACC take-COND-1PL
「送料は、どんな大きさの製品を買えば、」

U: ***xo‘p*** ***xo‘p***
yes yes
「はい、はい。」

B: *qimmat-roq* *o‘sha* *bepul* *bo‘l-ar* *ekan*
expensive-COMP that free be-PTCP.FUT EVID
[*deb* *ayt-gan=di-ngiz*
QT say-PTCP.PAST=PAST-2PL
「より高い、それが無料になるだろう、とあなたは言った。」

U: [***xo‘p***
yes
「はい。」

(3) と (5) の *aha* は、相手の長い発話単位の途中あるいは終わりに位置し、相手に発話継続を促していることから、他者フィードバック (聞き手が先行発話をどのように処理しているかを示す行為; ISO 24617-2 2020 参照) として機能していると考えられる。他方、(5) の *xo‘p xo‘p* や *xo‘p* も相手が発話している最中あるいは終わりに挿入される点では *aha* と共通し、いずれも他者フィードバックとして機能する。ただし、*aha* が主として発話継続を促すあいづち的機能を担うのに対し、*xo‘p* は同意・受領を明示する応答として用いられる点で異なる。さらに、(4) の *xo‘p* のように話者自身の発話を組み立てる過程で現れる場合には、自己フィードバック (話し手が先行発話に対する自身の処理を示す行為) としても理解できる。以上より、*aha* と *xo‘p* のように同じ発話途中に挿入されるとしても談話上の機能は一様ではないこと、*xo‘p* のように多機能な談話行為を表すものもあることが示された。したがって、談話行為タグの付与にあたっては、発話の連鎖上の位置づけを踏まえて機能を判定することが重要である。

4 おわりに

本稿の前半では、ウズベク語の談話コーパスの設計について概観した。本コーパスでは、二人の母語話者の会話を「長い発話単位」に区切り、各発話単位に、転記テキストに加え、日本語訳、英訳、グロス (形態論情報の略号)、談話行為タグを付与すると述べた。

後半では、談話に頻繁に見られる肯定的な形式二種類について、談話行為を分析しながら、それらの機能を同定した。その結果、類似の意味・談話内での挿入位置を持ちながらも異なる談話的な機能を持つことが明らかとなった。したがって、談話行為タグの付与は、発話の連鎖を考慮に入れながら、慎重に機能を判定する必要があると結論付けた。

謝辞

本研究は JSPS 科研費 JP24K22461 の助成を受けたものである。

参考文献

Austin, J. L. (1962) *How to Do Things with Words.* Oxford: Oxford University Press.

Begmatov, E., and A. Madavaliyev, N. Mahkamov, T. Mirayev, N. To‘xliyev, E. Umarov, D. Xutoyberganova, A. Hojev. (2008) *O‘zbek tilining izohli lug‘at. To‘rtinch jild*.［ウズベク語用例付き辞典 第 4 巻］Toshkent: “O‘zbekiston milliy entsiklopediyasi” Davlat ilmiy nashriyoti.

伝康晴 (2015)「第 5 章　対話への情報付与」『講座日本語コーパス　3. 話し言葉コーパス―設計と構築―』東京: 朝倉書店. 101-130.

伝康晴・榎本美香 (2014)『『千葉大学 3 人会話コーパス』使用説明書』(https://research.nii.ac.jp/src/files/Chiba3Party_manual.pdf [最終閲覧日: 2026/1/7])

日髙晋介 (2025)「《総説》ウズベク語について」『アジア・マップ：アジア・日本研究 Web マガジン』(https://www.ritsumei.ac.jp/research/aji/asia_map_vol03/uzbekistan/country01/ [最終閲覧日: 2026/1/7])

ISO 24617-2 (2020) *Language resource management — Semantic annotation framework (SemAF) —Part 2: Dialogue acts*, *2nd ed*.

Japanese Discourse Research Initiative. (2017)「発話単位ラベリングマニュアル version 2.1.」(https://www.jdri.org/resources/manuals/uu-doc-2.1.pdf [最終閲覧日: 2026/1/7])

Max Planck Institute for Evolutionary Anthropology, Department of Linguistics. (2015) Leipzig Glossing Rules. (https://www.eva.mpg.de/lingua/resources/glossing-rules.php [accessed: 2026/1/7])

中嶋善輝 (2015)『簡明ウズベク語辞典』大阪: 大阪大学出版会.

Searle, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

下地理則 (2025) SearchGloss ver 1.0. (DOI: 10.5281/zenodo.16419404 [最終閲覧日: 2026/1/7])

Sinclair, J. M. and Coulthard, R. M. (1975) *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford: Oxford University Press.

転記記号

(.)　0.1 秒未満の発話単位内休止
(D_bal)　言いさし (意図された語が不明)
[　重複始まり
<hh>　呼吸音

略号一覧

1	first person	一人称
2	second person	二人称
3	third person	三人称
ABL	ablative	奪格
ACC	accusative	対格
ADVL	adverbalizer	副詞化
CNT	continuative	継続
COMP	comparative	比較級
COND	conditional	条件
CVB	converb	副動詞
DAT	dative	与格
EVID	evidential	証拠性
FIL	filler	フィラー
FUT	future	未来
GEN	genitive	属格
INF	infinitive	不定
LOC	locative	処格
NEG	negative	否定
NPST	non-past	非過去
PAST	past	過去
PL	plural	複数
POSS	possessive	所有
PRF	perfect	パーフェクト
PTCP	participle	形動詞
QT	quotative	引用
SEQ	sequential	継起
SG	singular	単数
TER	terminative	終結