

[研究ノート]

学習者コーパスによる中間言語語用論分析： UAM CorpusTool Version 3.2 を用いた 要求の発話行為のケーススタディ

三 浦 愛 香

東京農業大学

This paper illustrates an annotation scheme for extracting requestive speech acts from a learner corpus, the National Institute of Information and Communication Technology Japanese Learner English (NICT JLE) Corpus, with the use of the Universidad Autónoma de Madrid CorpusTool (UAMCT). UAMCT is a downloadable free annotation software, and can be used for many types of pragmatic analysis. The tool's stand-off markup enables the construction of a multi-layered annotation scheme for requests produced by Japanese learners of English at different proficiency levels, making manual annotations less laborious than XML annotations. This study partly confirms the results of previous research in interlanguage pragmatics: that the use of conventionally indirect strategies increases with a developing proficiency (Kasper and Ross 2002; Trosborg 1995).

キーワード： 学習者コーパス、UAM CorpusTool、アノテーション、要求の発話行為

1. はじめに

本研究ノートでは、UAM CorpusTool(以下 UAMCT) と呼ばれるテキストにアノテーション(言語情報の付与)を施す無料のツールが、どのように語用論研究に役立つかを事例を用いて述べる。これは、Universidad Autónoma de Madrid(マドリード自治大学)の Mick O'Donnell 氏が開発したツールで、手動や半手動、そして自動的にテキストに言語情報を付与する機能を持つ。オリジナルのテキストには手を加えず、独立した XML ファイルにタグ情報を持つ「Stand-off markup」と呼ばれる構造を持つため、一つの言語特徴に異なるタグを多重に付与することが可能であり、多層のアノテーション構造を確立できる(O'Donnell 2013)。特に、文脈に依存した言語分析を必要とする意味分析や語用論分析、学習者コーパスのエラー分析に多く活用され、これまでに 90 本近くに及ぶ研究論文が発表されている(O'Donnell 2016)。

本稿では、UAMCT Version 3.2 を用い、日本人英語学習者の話し言葉コーパスである

The NICT JLE Corpus¹ から、学習者が要求の発話行為で使用した言語項目を抽出するアノテーション・スキームとその結果を提示する。The NICT JLE Corpus は、Standard Speaking Test (SST) と呼ばれるインタビューテストを受験した 1, 281 名の日本人学習者の発話を書き起こしたものである。

2. 研究の背景

2.1. 語用論とコーパス研究の融合

母語話者の発話例に基づいて論じられる語用論や談話分析の分野には、大規模データから得られる語彙や統語的特徴の頻度情報に基づくコーパスの手法は適合しなかったとされる (Rühlemann 2010; Thornbury 2010)。語用論的機能を文脈によって特定し手作業でタグを付与する必要がある、時に分析者にとって文脈から推測することが容易とは限らず、部分的に自らの直観に頼って分類する傾向にあるのは否めないからである (Adolphs 2008, Romero-Trillo 2008)。一方、発話の語用論的な解釈に、コーパスがもたらす計量的な結果を融合させる有用性が指摘されている (Rühlemann 2010; Rühlemann and Aijmer 2015; Romero-Trillo 2008; Thornbury 2010)。次節において、中間言語語用論の観点からその有用性について述べる。

2.2. 中間言語語用論におけるコーパス研究の可能性

学習者の語用論的能力を検証する中間言語語用論のデータ収集の手法としては、発話内行為やポライトネスに焦点を置いた談話完成タスクである Discourse Completion Task (DCT) が代表的である。DCT では、研究対象となる発話行為を直接的に抽出することを目的とし、発話行為が生じる状況や人間関係、対象となる学習者の変数の統制を可能にする。しかし、参加者のデータは限られ、その発話行為の内観に焦点を置くため、実験によって結果が多様化する傾向にあり、結論を一般化することが難しい (Callies 2013; Green 2012)。一方、コーパスを活用した外国語習得の研究では、語彙やコンコーダンスラインに比重を置く傾向が強いのが現状であるが (Adolphs 2008)、学習者コーパスが提供する大規模な被験者データにアクセスすることによって、頻度情報を中心としたより信頼性の高い統計データを分析することが可能になった (Granger 2002)。近年、学習者コーパスを用いて談話標識や発話行為の観点等から語用論的能力を検証した研究論文も増えてきている (Aijmer and Rühlemann 2015; Romero-Trillo 2014)。

¹ 詳細情報及び入手方法は、国立研究開発法人 情報通信研究機構が提供する以下のページを参照されたい。 https://alaginrc.nict.go.jp/nict_jle/

3. 本研究ノートの目的

本稿の目的は、前述の抽出タスクの実験を中心とする先行研究の結果が示す学習者の語用論的言語特徴や傾向を、コーパスから得られる頻度情報を使って再確認することである。具体的には、日本人英語学習者の話し言葉コーパスとしては最大の The NICT JLE Corpus の買い物のロールプレイのデータにおいて、UAMCT を用い要求の発話行為²を示す言語項目を手作業で特定し言語情報を付与した。要求の発話行為のアノテーションには、中間言語語用論の研究で多く取り入れられてきた Blum-Kulka, House and Kasper (1989) の分類スキームを参照にした。ただし、これは、母語話者に焦点をおいた対照言語学的な発話行為の分析に基づいたものであるため、学習者言語の分類に応用できない場合は改良を加えた。

なお、本稿では、表層的な言語情報を基に要求の発話行為を特定した「語用言語学的能力」に着目している。つまり、表層的な言語情報が手掛かりにならず、話し手や聞き手ではない第三者による特定が恣意的となる間接的な発話（「ヒント」等）は対象外とした。また、発話の適切さをポライトネスの観点から検証する「社会語用論的能力」の習得については触れていない。³ 当該学習者コーパスは、インタビューテストの書き起こしデータのため、発話者本人の発話の意図や実際の音声データを確認することはできないからである。

4. 要求の発話行為を抽出するアノテーション・スキーム

4.1. Blum-Kulka, House and Kasper (1989) による分類スキーム

本スキームは、要求の発話行為において、話し手が選択する文法や語彙項目に着目することにより異なるストラテジーに分類するものである。要求の発話行為は、「呼びかけ (Alert)」、「主要行為 (Head Act)」と「補足 (Supportive Move)」の3つに分けられる。⁴ また主要行為は、(1)「直接的ストラテジー (Direct Strategy)」、(2)「慣習的な間接的ストラテジー (Conventionally Indirect Strategy)」、そして (3)「非慣習的な間接的ストラテジー (Non-conventionally Indirect Strategy)」に分類される。(1)のストラテジーは、命

² 本研究で「要求の発話行為」に着目した理由は、従来の中間語用論で最も多く研究されているからである。学習者にとって日常的な場面における重要性があり使用頻度が高いことや、ポライトネス、FTA、敬意の表現や和らげ表現など語用論の理論に基づいて検証できるということが挙げられる (Schauer 2009)。

³ これらの用語や定義については Leech (2014) を参照にした。

⁴ 「呼びかけ」と「補足」は「主要行為」を任意に修飾する。

令形や動詞 want 等を使って示される。(2) は、助動詞 can や could を使った疑問文等がある。聞き手が、助動詞が元来持つ「可能」や「能力」の表層的な意味ではなく、文脈内や文脈外の情報や経験に基づき、要求の発話行為の機能があると捉える。(3) は、(1) と(2) とは異なり、要求の発話行為が文法や語彙の言語項目の選択によって明示的に示されない間接的な発話である。

主要行為は、「内的調整 (Internal Modification)」または「外的調整 (External Modification)」によって修飾することにより、発話行為が聞き手に及ぼす度合いを緩和または増長することができる。内的調整は、主要行為内に出現する語彙や統語的特徴に表れ、「Lexical downgrader (又は upgrader)」又は「Syntactic downgrader (又は upgrader)」と呼ばれる。例として、ポライトネス・マーカーの please がある。外的調整は、主要行為の言語項目外に見られる「補足 (Supportive Move)」として特定され、その機能例として「理由 (grounders)」等がある。⁵

4.2. UAMCT によって構築したアノテーション・スキーム

まずは、前節の Blum-Kulka ら (1989) が提唱した要求の発話行為の分類スキームを参照にしてトップダウン的にドキュメントに手作業でアノテーションを施した。しかし、実際に各ドキュメントを読んで発話行為を特定していくとスキームに改良を加える必要があったため、ボトムアップ的にも分析し、最終的に図 1 のスキームを完成させた。これは、学習者の発話データにはエラーが含まれること、発話データは不完全な話し言葉であること、そして要求の発話行為が買い物の場面に特化していることが、Blum-Kulka et al. (1989) の分類スキームには適合しなかったからである。

4.2.1. UAMCT によるアノテーション・スキームの構築の基本手順

UAMCT では、基本的に以下のような手順に沿って任意のテキストにアノテーションを施す。

- (1) UAMCT をホームページからダウンロードする。⁶
- (2) UAMCT を立ち上げ新しいプロジェクトの名称をつけ任意の場所に保存する。
- (3) 研究対象のコーパスデータ(テキストファイル形式で保存されたドキュメント)を UAMCT にアップロードする。
- (4) テキストに言語情報を付与するためのアノテーション・スキームを構築する。ス

⁵ 「前置き (preparatory)」や「脅し (threat)」や「約束 (promise)」等もある。

⁶ 現在最新版の UAMCT Version 3.3 (2016 年 5 月 6 日に公開) を以下のサイトより無料でダウンロードできる。<http://www.wagsoft.com/CorpusTool/index.html>

キームに名称をつけたあと、以下を任意に選ぶ。

- i. タグ付与の方法：文法構造と品詞の自動タグ付与か、手作業のタグ付与か
- ii. アノテーション・スキームの構築方法：スキーム⁷を自分でデザインするか、既に組み込まれたスキーム⁸を使うか、既に自分で構築したスキームを再利用するか
- iii. アノテーションの範囲：ドキュメント単体にラベルを付与するか、ドキュメント内の任意のセグメントにアノテーションを施すか

本研究では、「i. 手作業のタグ付与」で「ii. スキームを自分でデザイン」し、「iii. ドキュメント内の任意のセグメントにアノテーションを施し」た。

⁷ UAMCT ではスキームを Layer と呼んでいる。

⁸ 「Clause Grammar」、「Appraisal」、「Rhetorical Structure Theory」及び「Error Analysis」のスキームがある。

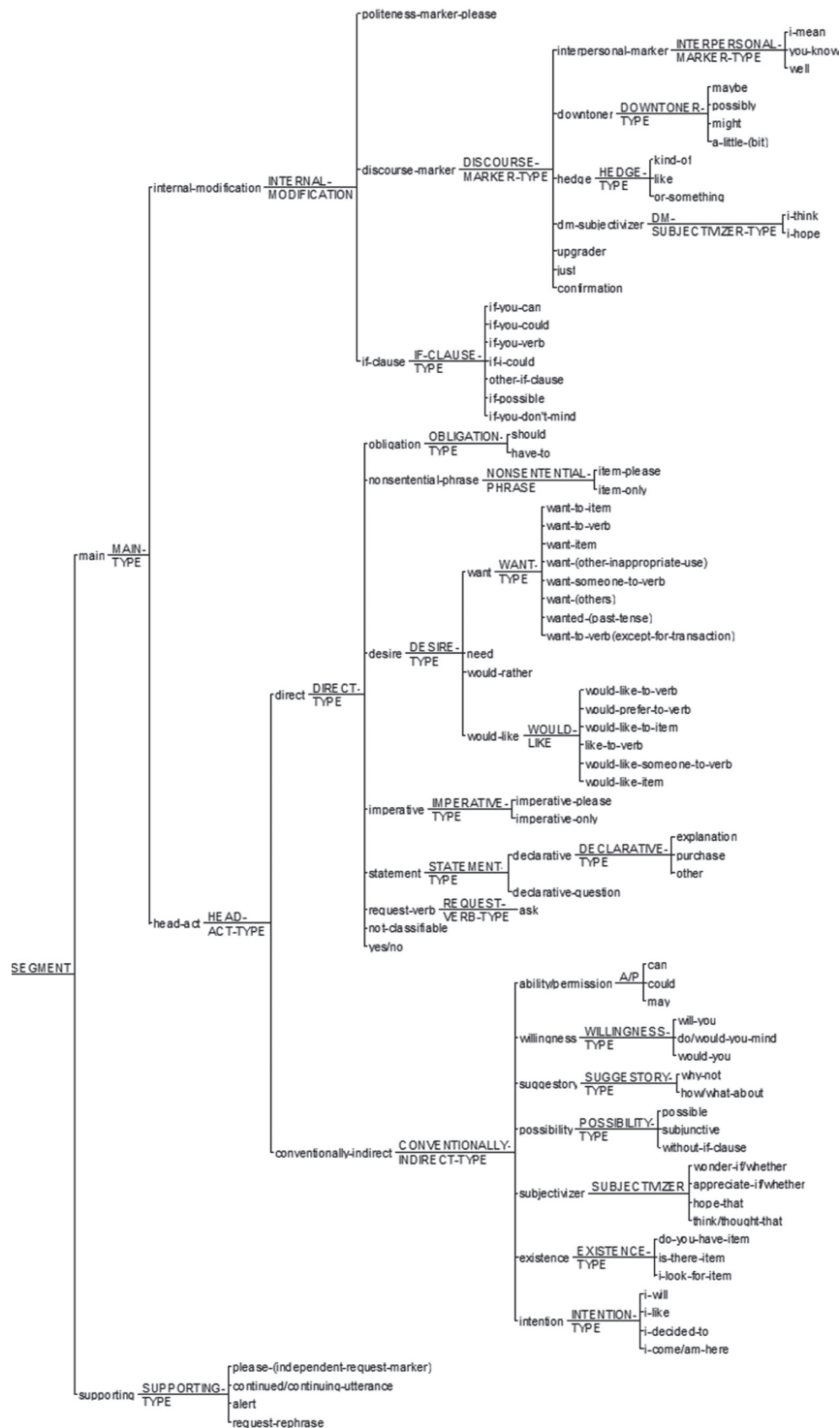


図 1. The NICT JLE Corpus における要求の発話行為を抽出するアノテーション・スキーム

4.2.2. 構築したアノテーション・スキームの詳細

当該コーパスは、試験官及び受験者の対話データから成る。⁹ SST には、ロールプレイが含まれており、そのうち「買い物 (Shopping)」は複数あるトピックのうちの一つである。試験官が店員を演じ、受験者 (学習者) は客として指定された品物を購入する。本研究では、「買い物」のロールプレイに取り組んだ学習者データのみを分析対象とした。

図1の SEGMENT がアノテーションの起点であり、UAMCT では、上層から下層に沿って分類項目が順を追って表示されるので、該当の項目を選び、要求の発話行為にアノテーションを施す。その手順を以下に示す。

- (1) SEGMENT : ドキュメントを読み進め、要求の発話行為の機能を持つ言語項目を特定する。当該項目をマウスで選択し、ハイライトする。その後、以下の Main または Supporting のどちらかを選択する。
 - i. Main : Head-act 及びそれを内的に修飾する Internal-modification に分類する。
 - ii. Supporting : Head-act に外的に付随する言語項目で、Please-(independent-request-marker) (Head-act から統語的に切り離して使用された 'please')、Continued/continuing-utterance (試験官の相槌等による中断後再開された発話の続き)、Alert (呼びかけ)、Request-rephrase (Head-act の繰り返し) に分類する。
- (2) Main
 - i. Head-act : a. Direct と b. Conventionally-indirect のいずれかに分類する。¹⁰
 - ii. Internal-modification : a. Politeness-marker-please (ポライトネス・マーカの 'please')、b. Discourse-marker (談話標識)、そして c. If-clause (if 節) のいずれかに分類する。
- (3) Head-act
 - i. Direct : a. Obligation (義務の助動詞)、b. Nonsentential-phrase (断片的フレーズ)、c. Desire (要望を示す動詞)、d. Imperative (命令文)、e. Statement (助動詞等の使用がない平叙文または疑問文)、f. Request-verb (要求の動詞)、g. Not-classifiable (主語がない等エラーを含む不完全な発話のため分類不可)、h. Yes/no (yes または no のみの発話) のいずれかに分類す

⁹ 当該コーパスには、話者情報として、各ドキュメントに試験官の発話が <A> のタグで、受験者 (学習者) の発話は のタグで既に付与されている。

¹⁰ 要求を示す表層的な言語情報を持たない「非慣習的な間接的ストラテジー (Non-conventionally Indirect Strategy)」は前述のように分析対象外である。

る。なお、b、e、g 及び h は学習者特有のエラーを含む発話として著者がオリジナルのスキームに追加した。

- ii. Conventionally-indirect : a. Ability/permission (可能・許可の助動詞)、b. Willingness (相手の意思を確認する文)、c. Suggestory (相手に提案する文)、d. Possibility (可能性を尋ねる文)、e. Subjectivizer (I hope/wonder/appreciate 等を使った依頼文)、f. Existence (購入したい物の存在を尋ねる文)、g. Intention (自分の意思を明示する文) のいずれかに分類する。なお、f 及び g に関しては買い物に特徴的な要求としてオリジナルのスキームに追加した。¹¹

図2はアノテーションの一例である。“May I help you, sir?”という試験官の発話に対する学習者の応答である“<F>Um</F> I’m looking for a book about architecture.”¹²を、要求の発話行為として特定し、本セグメントを Head-act とアノテーションする。次に、Conventionally-indirect を選び、提示される7つの分類から Existence を選ぶ。さらに、具体的な言語特徴が数種類提示されるので、I-look-for-item を選ぶ。本事例のように、スキームの最下位項目の多くは、具体的な言語特徴例を示す。

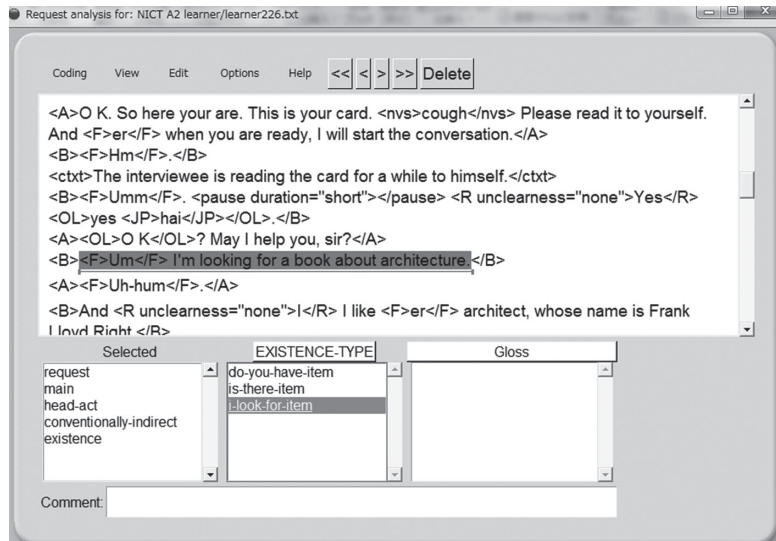


図2. 要求の発話行為のアノテーションの手順例

¹¹ 「Blum-Kulka, House and Kasper (1989) の非慣習的な間接的ストラテジーには、“Got a pen?” のように実際は慣習的に使用されることが多い項目が含まれる」という Leech (2014, pp. 142–143) の指摘を参考にして、f と g の2項目を「Conventionally-indirect」の分類に加えた。

¹² <F></F> はフィラーを示すタグで、当該コーパスに既に付与されている。

5. 分析結果とその考察

アノテーションで抽出した要求の発話行為の形態・統語論的特徴を手掛かりにその頻度情報を提示し、異なる習得段階にある学習者が選択した要求のストラテジーの分布を表示する。三浦 (2016) では、当該コーパスで買い物のロールプレイに取り組んだ CEFR (欧州共通言語参照枠) の A1 レベル (初級) 相当の学習者 68 名と A2 レベル (初中級) 相当の学習者 114 名、総計 182 名の発話データにアノテーションを施した。¹³ UAMCT は、習得段階の情報によって各ドキュメントを分別し、付与タグの分布や頻度率を瞬時に計測する。最終的に、A1 レベルでは総計 559 タグ、A2 レベルでは 1127 タグを付与した。また、Main の割合は、前者が 93.56% (523 タグ)、後者は 91.47% (1029 タグ) であった。Main に属する Head-act と Internal-modification の出現比率は、両レベルにおいて有意差はなかった ($\chi^2=0.38$, $df=1$, $p=0.54$, Cramer's $V=0.016$)。しかし、表 1 のように、A2 になると A1 より Direct の使用頻度率が下がり、Conventionally-indirect が上がる傾向が観察され、両者はカイ二乗検定によって有意な差があると確認された ($\chi^2=15.84$, $df=1$, $p=0.00007$, Cramer's $V=0.107$)。Politeness-marker-please と Discourse-marker の使用¹⁴についても、表 2 にあるように、レベル間の有意差が示され、A1 レベルの方が A2 より Discourse-marker を使う傾向が強いと示された ($\chi^2=9.78$, $df=1$, $p=0.0018$, Cramer's $V=0.235$)。しかし、図 2 の UAMCT の KWIC (Keyword In Context) 画面を確認すると、A1 レベルでは、実際には学習者 1 名 (A1 (Intermediate)/learner1096.txt) による発話が半数を占めていることが判明した。一方、図には示していないが、A2 レベルの 9 件は異なる 9 名の学習者が発話していた。頻度数の低い言語項目については、発話者の偏りも考慮する必要がある。

表 1. A1 及び A2 レベルの学習者による Head-act のストラテジーの分布

	A1 粗頻度数	%	A2 粗頻度数	%
Direct	312	68.0	519	56.9
Conventionally-indirect	147	32.0	394	43.2
Total (Head-act)	459	100.0	914	100.0

¹³ Tschirner and Bärenfänger (2012) 等を参照し、本コーパスの SST レベルを CEFR レベルに紐付けた。買い物に取り組んだ A レベル学習者を全て分析対象とした。

¹⁴ If-clause は粗頻度数が 5 未満であったため、カイ二乗検定から除いている。

表 2. A1 及び A2 レベルの学習者による Internal-modification の分布

	A1 粗頻度数	%	A2 粗頻度数	%
Politeness-marker-please	48	75.0	104	90.4
Discourse-marker	16	25.0	9	7.8
If-clause	0	0	2	1.7
Total (Internal-modification)	64	100.0	115	100.0

File	Pretext	<Annotation feature="discourse-marker"/>	PostText
A1(Beginner)/file00404.txt	/R> twenty percent?	Yeah?	</RS> <A><
A1(Beginner)/file00757.txt	"imp"><DR str="inf">	Just	a moment, <LD m
A1(Beginner)/file00919.txt	<LD mkr="inter"><F>	well	</F></LD> <F>um<
A1(Intermediate)/learner1053.txt	please price down,	just	little. <A
A1(Intermediate)/learner1096.txt	I help you? 	<F>Well</F>	I want a
A1(Intermediate)/learner1096.txt	>Mm-hm</F> 	<F>Well</F>	do you h
A1(Intermediate)/learner1096.txt	jackets, <F>oh</F>	<F>well</F>	<pause d
A1(Intermediate)/learner1096.txt	<F>Oh</F> yeah.	<F>Well</F>	do you h
A1(Intermediate)/learner1096.txt	 <F>Oh</F>	<F>well</F>	<F>mhm</
A1(Intermediate)/learner1096.txt	on="short"></pause>	<F>well</F>	I prefer
A1(Intermediate)/learner1096.txt	 <F>Oh</F>	<F>well</F>	<pause d
A1(Intermediate)/learner1096.txt	A> <F>Ur</F> <F>	well	</F> cash card,
A1(Intermediate)/learner1168.txt	T-shirts <F>um</F>	<R unclarity="none"></R> I hope so.	
A1(Intermediate)/learner1197.txt	<F>Uhu</F> 	<laughter>I hope</laughter>	
A1(Intermediate)/learner960.txt	ion="long"></pause>	I mean	<pause durati
A1(Intermediate)/learner973.txt	unclearness="none">	Just	</R> just <

図 2. A1 レベルの学習者による Discourse-marker の発話例を示す
UAMCT の KWIC 画面

Direct か Conventionally-indirect かの戦略の選択において統計的にレベル間に有意差があったという本結果は、一部の先行研究で既に結論づけられている「習得段階が上がるほど慣習的な間接的戦略の使用が高まる」傾向と合致したと言える。これらの先行研究には、少数の第二言語学習者を対象とした縦断的研究 (Kasper and Ross 2002) や、DCT の手法を用いて異なる習熟度別にデンマーク人英語学習者グループと英語母語話者グループを比較した横断的研究 (Trosborg 1995) がある。

6. まとめ

本稿では、UAMCT を活用することにより、語用論分析とコーパス分析の融合が容易になることで、分析の再現性や結果の信頼性の向上に貢献する可能性を示唆した。著者は、本稿の要求の発話行為の抽出スキームに加え、要求の言語機能（購入の意思を示す、割引交渉をする、試着の許可を得る等）を特定するスキームや、発話が対話上適切または自然であるかを判断するスキームを同時に開発し、異なる習得段階ごとにその分布を比較している（三浦 2016; Miura 2014, 2015）。多層構造のアノテーションによって、発話行為を多面的に観察した頻度情報を提供することが可能になったと言える。今後は、当該データに限らず汎用性の高い発話の抽出スキームを開発していきたい。

参照文献

- Adolphs, S. 2008. *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins.
- Aijmer, K. and Rühlemann, C. 2015. *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.
- Blum-Kulka, S., House, J. and Kasper, G. 1989. *Cross-cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.
- Callies, M. 2013. "Advancing the Research Agenda of Interlanguage Pragmatics: The Role of Learner Corpora." In J. Romero-Trillo (ed.) *Year Book of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, 9-36. Dordrecht: Springer.
- Granger, S. 2002. "A Bird's-Eye View of Learner Corpus Research." *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3-18. London: Longman.
- Green, A. 2012. *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition across the Ability Range*. Cambridge: Cambridge University Press.
- Kasper, G and Ross, S. 2007. "Multiple Questions in Oral Proficiency Interviews." *Journal of Pragmatics* 39, 2045-2070.
- Leech, G. *The Pragmatics of Politeness*. Oxford: Oxford University Press.
- Miura, A. 2014. "An Annotation Scheme for Criterial Features of Pragmatic Competence of Japanese Learners of English." *47th Annual Meeting of the Societas Linguistica Europaea: Book of Abstracts*, 243.
- Miura, A. 2015. "Extracting Criterial Features from Requests Produced by Japanese Low-proficiency Learners of English." *LCR2015: Third Learner Corpus Research Conference*, 123-124.
- 三浦愛香. 2016. 「CEFR A1 及び A2 レベルの日本人英語学習者による語用論的能力を弁別する基準特性」、投野由紀夫（編）『平成 24 年度～平成 27 年度科学研究費補助金（基盤研究（A））研究課題番号 24242017 研究成果報告書：学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究』、292-304.
- O'Donnell, M. 2013. *UAM CorpusTool Version 3.0: Tutorial Introduction*. Retrieved from <http://www.wagsoft.com/CorpusTool/Documentation/UAMCorpusToolTutorial3.0.pdf>
- O'Donnell, M. 2016. *UAM CorpusTool: Publications*. Retrieved from <http://www.corpustool.com/research.html>
- Romero-Trillo, J. 2008. *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Berlin: Mouton de Gruyter.
- Romero-Trillo, J. 2014. *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*. Dordrecht: Springer.
- Rühlemann, C. 2010. "What can a Corpus Tell us about Pragmatics?" In A. O'Keefe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*, 288-301. Abingdon:

- Routledge.
- Rühlemann, C. and Aijmer, K. 2015. “Corpus Pragmatics: Laying the Foundations.” In K. Aijmer and C. Rühlemann (eds.) *Corpus Pragmatics: A Handbook*, 1–26. Cambridge: Cambridge University Press.
- Schauer, G. A. 2009. *Interlanguage Pragmatic Development: The Study Abroad Context*. York: Continuum.
- Thornbury, S. 2010. “What can a Corpus Tell us about Discourse?” In A. O’Keefe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*, 270–287. Abingdon: Routledge.
- Trosborg, A. 1995. *Interlanguage Pragmatics: Requests, Complaints and Apologies*. Mouton de Gruyter, Berlin.
- Tschirner, E., and Bärenfänger, O. 2012. “Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by Computer (OPIC).” *Technical Report 2012-US-PUB-1*. Leipzig: Institute for Test Research and Development.