N-gram 分析を通した CEFR レベル基準特性の特定：動詞共起フレームに焦点をあてて

能登原 祥之
同志社大学　文学部
{ynotohar}@mail.doshiha.ac.jp

1 はじめに

1.1 背景

近年、言語教育界において CEFR (Common European Framework of References) という言語共通参照枠に準拠した教育の在り方が模索されている。特に CAN DO statements (e.g., B1 Spoken Production: I can connect phrases in a simple way …) の形で習熟度別の指導目標を透明にし、指導と評価の一体化を念頭に習熟度教育が見直され改善されている [1][2][3]。

しかしながら、CAN DO statements の記述内容は言語教育者の主観によって判断され規定されたものが多く、学習者の実際の言語使用状況をふまえたものになっていない。そこで、大規模な書き言葉学習者コーパスである Cambridge Learner Corpus (CLC) を使用し、参照枠の各レベルを実証的に記述する（RLD: Reference Level Descriptions）研究が進められている。また、CEFR の６つの習熟度レベル（A1, A2, B1, B2, C1, C2）をそれぞれ規定する言語特徴を基準特性 (criterial features) として特定する研究が進められている [4]。

1.2 本研究の目的

本研究では、上記の先行研究をふまえた上で、先行研究で網羅できていない A１レベルの記述を充実させるため、日本人英語学習者書き言葉コーパス (JEFLL-CEFR Corpus) を用いて、先行研究に従い動詞共起フレーム (e.g., NP-V-NP) に焦点をあて、探索的に N-gram 分析を通して CEFR レベルの基準特性を特定する研究を行っている。

そして、(1) 先行研究で示されている基準特性が日本人英語学習者書き言葉コーパスの場合でも実際に各レベルを特定できるのか、(2) 先行研究で A2 の基準特性とされているものが A1 ですでに使用されているものか、の２点を明らかにすることを目的とする。

2　RLD における動詞共起フレーム研究

RLD 研究では、動詞共起フレームの研究が特に進められている[4][5][6][7][8]。これらは、Cambridge ESOL によって構築された学習者コーパスで約 4500 万語の書き言葉学習者コーパス CLC の言語データに品詞タグを付け、The Robust Accurate Statistical Parser (RASP) を通して統語解析を行い、学習者が使える英語表現（positive features）の使用傾向を習熟度別に記述している（表１）。また、一部に付与した独自のエラータグ 70 種以上を通して、学習者が使えていない表現（negative features）の使用傾向も記述している（表２）。

表１ A2 動詞共起フレームの例 [8]

NP-V
e.g., *He went*
NP-V (reciprocal Subj)
e.g., *They met*
NP-V-PP
e.g., *They apologized* [*to him*]
NP-V-NP
e.g., *He loved her*
NP-V-Part-NP
e.g., *She looked up* [*the number*]
NP-V-NP-PP
e.g., *She added* [*the flowers*] [*to the bouquet*]
NP-V-NP-PP (P = *for*)
e.g., *She bought* [*a book*] [*for him*]
NP-V-V(+*ing*)
e.g., *His hair needs combing*
NP-V-VPinfinitival (Subj Control)
e.g., *I wanted to play*
NP-V-S
e.g., *They thought* [*that he was always late*]

表2　エラーコードの例 [8]

RN　Replace noun
e.g., *Have a good travel* (journey)
RV　Replace verb
e.g., *I existed last weekend in London* (spent)
MD　Missing determiner
e.g., *I spoke to President* (the)
AGN　Noun agreement error
e.g., *One of my friend* (friends)
AGV　Verb agreement error
e.g., *The three birds is singing* (are)

これらの先行研究の問題点として、(1) CLC を利用した一連の研究には A1 レベルの記述が少ないこと、(2) A2 の言語特徴とされていても A1 レベルで既に使えている特徴があると思われるがその記述が少ないこと、(3) CLC 内において動詞共起フレームを抽出する際に使用された基軸動詞が全体の語彙のどの程度の割合を占めているものか明記されていないこと、(4) 各レベルの動詞共起フレーム頻度はどの基軸動詞のフレームの割合が多いのか明記されていないこと、の４点が挙げられる。

そこで、(1) A1-A2 レベルの多いデータを用いること、(2)どの CEFR レベルにも見られる共通の高頻度動詞に焦点を絞り、その割合や各動詞の各共起フレーム頻度への影響を確認しながら基準特性の特定作業を進めること、(3) 統計的にも多角的に確認した上で基準特性を抽出すること、(4) Parser で拾えないところをボトムアップ的に拾っていくため N-gram 分析を利用すること、の４点を改善点とした。

3　JEFLL-CEFR Corpus の調査

3.1 調査目的

JEFLL-CEFR Corpus に見られる 30 の高頻度動詞から基軸動詞を選定し、その 4-gram の言語的特徴の中から基準特性の候補になり得るものを特定する。そして、(1) 先行研究で示されている基準特性が日本人英語学習者書き言葉コーパスの場合でも実際に各レベルを特定できるのか、 (2) 先行研究で A2 の基準特性とされているものが A1 ですでに使用されているものか、の２点を明らかにすることを目的とする。

3.2 調査

3.2.1 使用したコーパス

本調査では、JEFLL Corpus を CEFR レベルに再編した JEFLL-CEFR Corpus を利用した。WordSmith ver.6 (2012) [9]を通して語彙の基本統計値を算出した（表 3）。

表3　JEFLL-CEFR Corpus の基本統計値

	A1 (*N*=3507)	A2 (*N*=4956)	B1 (*N*=1529)
Tokens	137,880	315,814	218,134
Types	5,572	9,692	8,130
STTR	23.35	26.56	29.05
MWL	3.55	3.71	3.75

STTR (1000 語で正規化した異なり語比率)
MWL（平均語長数）

3.2.2 調査手順

調査手順として、(1) WordSmith Ver.6 (2012) と Someya (1998) の Lemma list [10]を用いて Wordlist と品詞タグ情報を参照し、CEFR３レベル（A1-B1）全てで出現する高頻度動詞 30 語を選定する。(2) 高頻度基軸動詞を IBM SPSS Statistics Ver.22 (2013)を使い、クラスター分析（Ward 法）を通して動詞を分類する。(3) 分類したクラスターからそれぞれ代表的な動詞を基軸動詞として選定する。(4) 選定された基軸動詞を基に WordSmith を用いて 4-gram 分析を行う。(5) CEFR３レベル（A1-B1）全てで出現する基軸動詞に関する高頻度 4-gram を基準特性の候補とし、IBM SPSS Statistics を通して、クラスター分析、コレスポンデンス分析、判別分析の３種類の分析を行い、その結果をふまえ基準特性を特定する。

3.3. 結果

3.3.1 基軸動詞の選定

高頻度動詞 30 語の CEFR レベル別の調整頻度（100 万語あたり）を基にクラスター分析を行った。その結果、クラスター１（BE）、クラスター２（DO、HAVE)、クラスター３（LIKE）、クラスター4 (EAT, THINK, GO, WANT)、クラスター5 (BUY, BRING)、クラスター6 (PLAY, LIVE BECOME, USE, SEE, COME, SAY)、クラスター７（GET, TAKE, MAKE）、クラスター８（DIE,

DRINK, NEED, FEEL, RUN, KNOW, LOOK, HELP, ENJOY, GIVE）の８種類に分類できることが明らかとなった。その結果をふまえ、各クラスターの中で高頻度かつ動詞共起フレームとしても特徴的なものを持つ動詞に注目し、BE, HAVE, LIKE, THINK, BRING, BECOME, TAKE, KNOW の８動詞を基軸動詞として選定した。

3.3.2 基軸 4-gram の選定

８種類の基軸動詞を基に 4-gram を確認し、それらの CEFR 3 レベル別の使用頻度を記述した。その結果、CEFR 3 レベル全てで出現し動詞共起フレームとしても特徴的なものを持つ 4-gram を確認したところ、12 種類 *I do n't have, I do n't like, I will bring a, so I will bring, I will take out, I m go to, I do n't know, will take out my, I think it be, I do n't think, became a old man, do n't know what* を抽出した。

3.3.3 基軸 4-gram の分類

調整頻度（100 万語あたり）を基に 12 種類の 4-gram のクラスター分析を行った結果、クラスター１（*I do n't have*）、クラスター２（*I do n't like, I will bring a*）、クラスター３（*so I will bring*）、クラスター4（*I will take out, I m go to, I do n't know*）、クラスター5（*will take out my, I think it be, I do n't think, became a old man, do n't know what*）の 5 種類に分類できた。

3.3.4 基軸 4-gram と CEFR レベルの対応関係

まず、12 種類の基軸 4-gram と CEFR の３レベルとの関係をさぐるため、コレスポンデンス分析を行った。$\chi 2$ 検定（両側）の結果、4-gram と CEFR レベルとの関係に有意差が確認された（$\chi 2$ (22)= 2071.144, p=.000, Cramer's V=.238, p=.000, ２次元までの累積イナーシャの寄与率 100%）。その関係性を図示すると図１となる。

3.3.5 基軸 4-gram による CEFR レベルの判別

次に、５種類の基軸 4-gram が CEFR 3 レベルのどれを区別する基準特性と言えるのかを３レベルで確認するため、正準判別分析を行った。その結果、３つの次元が有意であることが分かった（第１次元 p=.994, 第２次元 p=.985, 第３次元 p=.832）。特に、次元１は A1 の判別（.854*）、次元 2 が A2 の判別（.839*）、次元３が B1 の判別（.982*）にそれぞれに関連があり、３次元による正判別率は 83.30%であった。4-gram の判別関数得点の分布を視覚化すると図２のようになる。

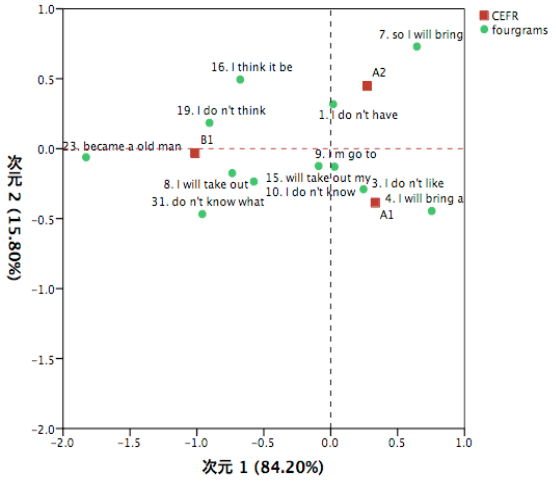

図１ 基軸 4-gram と CEFR レベルの対応関係

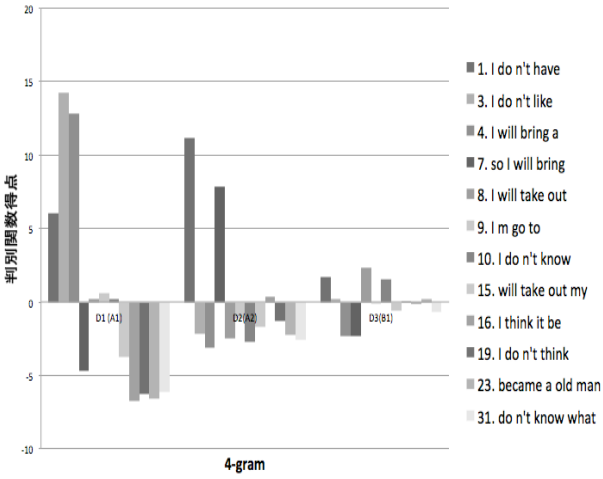

図２ 基軸 4-gram の判別関数得点の分布

4 考察

JEFLL-CEFR Corpus を使い、4-gram を軸に先行研究で示されている基準特性を検証したところ、A1 の基準特性として、*I don't like* (NP-V-NP) と *I will bring a* (NP-Aux-V-NP) が候補となる。これらは、先行研究で A2 の基準特性として示されているが、A1 の基準特性として指摘できる。また、A2 では *I don't have* (NP-V-NP), *so I will bring a* (Conj-NP-Aux-V) が基準特性の候補となる。前者の方は、学習者コーパスの用例を見る限り、ま

だ *have* の軽動詞的用法 (e.g., *have a look at*) が十分使いこなせない状況と解釈する必要がある。後者の方は、先行研究では言及されていないが、用例を見る限り bring 以外の様々な動詞と共にこの動詞共起フレームが使われる傾向にあるので候補として指摘する必要はある。また、B1 では、*I will take out* (NP-Aux-V-Part), *I don't know* (NP-V-NP) が基準特性の候補となる。これらは、先行研究では A2 の例として指摘されている。本研究では、先行研究で指摘されている B1 レベルの複雑な項構造は抽出できていないが、A2 レベルよりのチャンク表現が抽出されたと解釈すべきだろう。

5 結論

JEFLL-CEFR Corpus を通して調査したところ、NP-V-NP や NP-Aux-V-NP といった先行研究で A2 レベルとされる基本的な動詞共起フレームに関しては A1 の基準特性と指摘できる。また、A2 の NP-V-NP や Conj-NP-Aux-V といった動詞共起フレームは、構造的に A1 のものをやや拡張したものと想定される。また、学習者の用例から判断すると、軽動詞的用法や多義性を使いこなせているとは言いがたい。B1 になると、句動詞がチャンク的に使える状況を確認できたが、本調査では、先行研究で指摘された複雑な項構造を基準特性として確認できなかった。より厳密に探るためには、動詞の右側の項構造に絞った 4-gram 分析を行う必要があるだろう。

6 今後の課題

今後の課題として、以下３点が挙げられる。(1) 基準特性を探る際の 30 種類の高頻度基軸動詞を再吟味すること、(2) 基準特性を探る際の 12 種類の基軸 4-gram を再吟味すること、(3) 判別分析を通して選定した３レベルの基準特性の候補が妥当かを再度検証すること。

謝辞

本研究は、科学研究費 基盤研究 (A) 課題番号 24242017「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」（平成 24～27 年度 研究者代表 投野由紀夫）によるものである。

参考文献

[1] Council of Europe, (2001). *The common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

[2] North, B., Ortega, A. & Sheehan, S. (2010). *A core inventory for general English.* London, UK: British Council/EAQUALS.

[3] North, B. (2014). *English profile studies 4　The CEFR in practice*. Cambridge, UK: Cambridge University Press.

[4] Hawkins, J.A. & Filipović,　L. (2012). *English profile studies 1　Criterial features in L2 English.* Cambridge, UK: Cambridge University Press.

[5] Wiiliams, C. (2008). Challenges to parsing English text: the language of non-native speakers. *Cambridge Research Notes, 33*, 10-15

[6] Salamoura, A.& Saville, N. (2009). Criterial features of English across the CEFR levels: evidence from the English profile programme. *Cambridge Research Notes, 37*, 34-40.

[7] Saville, N. & Salamoura, A. (2010). Exemplifying the CEFR: Criterial features of written learner English from the English profile programme. *EUROSLA MONOGRAPH SERIES 1 Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 101-132.

[8] Hawkins, J.A.& Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal, 1*, 1-23, doi: 10.1017/S20415 36210000103

[9] Scott, M. (2012). *WordSmith Tools version 6.* ［Computer Software］ Liverpool, UK: Lexical Analysis Software.

[10] Someya, Y. (1998). e_lemma.txt (Ver.2 for WordSmith 4) ［Lemma List］ http://www. lexically. net/wordsmith/support/extras.html