学習者コーパスを使用したレベル別頻度比較の方法

森, 秀明

(Citation)
Learner Corpus Studies in Asia and the World,3:303-322

(Issue Date)
2018-03-12

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81010134

(URL)
https://hdl.handle.net/20.500.14094/81010134

学習者コーパスを使用したレベル別頻度比較の方法

森　秀明（東北大学・大学院生）
hideaki.mori.p1@dc.tohoku.ac.jp

Method of Frequency Comparison According to the Level in Learner Corpus

MORI　Hideaki (Tohoku University, Graduate Student)

概要

本稿では，学習者コーパスで検索した頻度を習得レベル別に比較する場合，どのような分析法を用いるのが有効かについて検討した。データは主に，タグ付き KY コーパスと I-JAS で検索した条件表現「タラ」の頻度を使用した。先行研究で推奨されている①「調整頻度」を学習者コーパスに使用するのは原理的に問題があり，学習者個々に算出した「個別調整頻度」を使用するべきである。しかし個別調整頻度を使用しても，②「平均値」は外れ値が多く，比較が難しい。③「中央値」，④「タラの使用者数割合」は原理的には妥当だが，一部の情報しか有効活用できない。学習者データはばらつきが大きく，習得状況を一つの代表値でとらえるのは困難である。学習者の実態を面的にとらえるには，作図による分析が適している。最終的にマーカーが重複しない散布図が描ける⑤「蜂群図と箱ひげ図の合成図」で観察する方法が，タラの習得状況を有効に比較できる方法だと考えられた。

キーワード

学習者コーパス，頻度比較の方法，個別調整頻度，蜂群図と箱ひげ図の合成図

1. 研究の目的と先行研究

学習者コーパスを使用する目的の一つに習得研究がある。例えば日本語の条件表現「タラ」が初級～超級のどのレベルで習得されるかを知りたいとする。これを調査するには，タラ・ダラを検索してレベル別の頻度を調べ，それを比較すれば良さそうに思える。しかし，レベル別の人数や語数が異なる場合，コーパスから得られたままの粗頻度では比較できない。このような場合，多くの先行研究では調整頻度（換算頻度，正規化頻度）の使用が推奨されている（バイバーほか，2003，pp.38-41；石川ほか，2010，pp.27-8；石川，2012，pp.114-5；マケナリーほか，2014，pp.74-6 など）。

調整頻度とは粗頻度÷コーパスの総語数×一定数で求められる値である。粗頻度÷コーパスの総語数×100 は使用率と呼ばれ，ある単語の頻度があるコーパスの総語数の中で何%を占めているかを表している。総語数が異なったコーパス間で単語の頻度を比較する場合は，この使用率を用いても比較できるが，頻度が低い単語の場合，使用率では値が小さくなりすぎて扱いにくい。そこでパーセント単位で比較する代わり，それより大きい 1 千語，1

万語，10 万語などの一定数をかけて分かりやすい値にしたものが調整頻度である。語数の異なるコーパス間で比較する場合はコーパスの総語数で割るが，学習者のレベル間で比較する場合はレベルごとの総語数で割って平準化する。

しかし調整頻度は，基本的に個体（テキスト）ごとの語数が一定の均衡コーパスの比較で用いられる調整方法である。一方，学習者コーパスは，個体（学習者）ごとの語数が異なっている。詳細は第 2 節で論じるが，均衡コーパスで使用されている調整法を，性質の異なる学習者コーパスにそのまま適用することには問題がある。調整頻度が使えないとしたら，他にはどんな方法が考えられるであろうか。レベルごとの平均値か，中央値か？本稿では，学習者のレベル別頻度比較というごく基礎的な分析方法の有効性を，改めて問い直す。

これまで日本語教育研究で最も多く使用されてきた学習者コーパスは，KY コーパスである（李, 2009, p.60）。本稿では第 2 節で KY コーパスを使用し，タラの検索データを用いた分析法の比較を行う。しかし KY コーパスは総語数約 17 万語の小規模コーパスである。このため第 2 節の結論は小規模コーパスでのみ成り立つ結論に限定される恐れがある。そこで第 3 節では現在公開されている I-JAS（International Corpus of Japanese as a Second Language：多言語母語の日本語学習者横断コーパス）の二次データ約 122 万語を使用し，第 2 節の結論が大規模コーパスでも変わらないことを示す。最後に第 4 節でまとめと今後の課題を述べる。

2. KY コーパスを使用した頻度分析法の比較

2.1 使用するデータの説明

KY コーパスは，Oral Proficiency Interview（OPI）のデータを基に，鎌田修氏と山内博之氏によって構築された学習者コーパスである（鎌田, 1999, 2006;山内, 1999）。OPI は学習者に対して最長 30 分のインタビューを行い，米国外国語教育協会（ACTFL）が定めた外国語能力基準によって学習者の能力を初級下～超級の各段階に判定するインタビューテストである。KY コーパスの構築が始まった 1996 年段階では 9 段階，その後改訂されて 10 段階にレベル分けされている。初級は決まり文句や習い覚えた語句，単語の羅列で最小限のコミュニケーションが行える程度のレベルだが，超級は具体的・抽象的な話題で議論ができるレベルである。KY コーパスはこの OPI を書き起こした発話データで，語数はおよそ 17 万語，学習者の内訳は初級下～超級までの 9 レベルに分けられた英語・韓国語・中国語母語話者各 30 名の合計 90 名となっている。各レベルの学習者数はレベルの大分類である初・中・上・超級の 4 分類では，それぞれ 15・30・30・15 のようにまとまっているが，下位分類では，初級下：3 名，初級中：5 名，初級上：6 名のように一定ではなく，同じレベルの学習者数も母語によって異なる場合がある。

この KY コーパスをもとに，2008 年には李在鎬氏等によって形態素や誤用情報などを付与した「タグ付き KY コーパス」の提供が始まり，2013 年には検索システムを備えてウェブ公開された（http://jhlee.sakura.ne.jp/kyc/）。本稿ではこのタグ付き KY コーパスを利用し，

文字列「タラ・ダラ」を検索し，目視で調査目的にそぐわない用例（いわゆるゴミ）と誤用を除き，580 の用例を得た。以後の分析にはこの 580 のデータを用いる。

学習者コーパスを使用した計量的な分析では，グレンジャー等によって提唱されている対照中間言語分析（contrastive interlanguage analysis：CIA）が主流で，学習者コーパスと母語話者コーパスを比較して学習者の過剰・過少使用を探る分析が一般的である（グレンジャー（編），2008，pp.14-6；石川，2017，pp.67-8）。しかし KY コーパスには母語話者のデータが付随していないため，本稿では名大会話コーパスを使用して母語話者のデータを取得した。名大会話コーパスは科学研究費基盤研究（B）（2）「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成 13 年度～15 年度　研究代表者大曾美恵子）の一環として作成された会話コーパスで，10 代から 90 代までの日本人母語話者（女性 161 名、男性 37 名）による雑談 129 会話，約 142 万語のデータである。現在は国立国語研究所の検索インタフェース『中納言』による検索が可能になっている。本稿ではこれを利用し，語彙素「た」，品詞「助動詞」，活用形「仮定形」で検索したタラ 4,491 のデータを使用する。

名大会話コーパスは一人の話者が複数の会話に登場する場合もあれば，一つの会話で短い発言しかしない場合もある。データは一つの会話が 30 分～1 時間程度になるように録音されたもので，データの取得単位は話者ではなく会話になっている。そこでデータは会話単位で集約し，1 会話当たりの平均値や中央値などを算出した。KY コーパスと名大会話コーパスを比較する意味は，母語話者が様々な雑談を行った場合に平均的に使用するタラの頻度に比べ，学習者のレベルによってどれぐらいの過少・過剰使用があるかを評価することにある。ただし KY コーパスは純粋な雑談ではなく，面接官によって管理されたインタビューであるため，厳密な比較は困難で，あくまで参考程度の比較にとどまる。

なおタグ付き KY コーパスは形態素解析エンジン ChaSen（茶筌），辞書 IPAdic で解析されている（李，2009，p.62）。一方，名大会話コーパスや第 3 節で使用する I-JAS は形態素解析エンジン MeCab，辞書 UniDic で解析されている[1]。この 2 つの辞書では言語単位の認定方法が異なり，例えば「経済学部」という語なら，IPAdic は 1 単位，UniDic では「経済／学／部」の 3 単位で認定される（小木曽，2014，p.100）。また単位名も IPAdic では形態素，UniDic では短単位という名称になる。このため，これらも厳密な比較は困難で，データ数の名称もそれぞれ形態素数，短単位数と呼ぶのが正確だが，本稿ではこれらを単純化して語数と呼ぶ。

2.2 ①調整頻度，②個別調整頻度の平均値，③中央値，④タラの使用者数割合の調査結果

初めにタグ付き KY コーパスと名大会話コーパスの基礎統計量，およびタラの調査結果に基づいて算出した①～④の指標の分析結果を表 1 に示す。レベルで「母語」とあるのが名大会話コーパスのデータである。

KY コーパスのレベル別学習者数は，先に述べたように初級下 4 名，初級中 5 名のように

一定数にはなっていない。語数合計はレベル別の学習者データの語数合計で，初級下では692 語，語数平均はこれを 4 名で割った値の 173.0 語である。173.0 語という語数は非常に少ないデータ量で，初級下は面接官の質問に対し，単語や句で回答するのがやっとのレベルであるため，面接官と学習者で 40〜50 回のやり取りがあっても，学習者の発話部分のみでは本稿の数行分にしかならない。しかしレベルが上がるに連れて語数が増え，中級上以上では 2,000 語以上と初級下の 10 倍以上のデータ量になっている。

表 1　タグ付き KY コーパス・名大会話コーパスの基礎統計量と 4 種類の分析結果の比較

				⓪	①	②	③	④	①'	②'	③'	④'
レベル	学習者数	語数合計	語数平均	タラ粗頻度	調整頻度	個別平均	中央値	使用者数	調整頻度割合	個別平均割合	中央値割合	使用者割合
初級下	4	692	173.0	0	0.0	0.0	0.0	0	0%	0%	0%	0%
初級中	5	2,542	508.4	0	0.0	0.0	0.0	0	0%	0%	0%	0%
初級上	6	3,965	660.8	0	0.0	0.0	0.0	0	0%	0%	0%	0%
中級下	9	11,338	1,259.8	13	1.1	0.8	0.0	3	33%	24%	0%	33%
中級中	14	20,854	1,489.6	42	2.0	2.0	1.6	13	57%	56%	50%	93%
中級上	7	14,004	2,000.6	76	5.4	5.2	5.0	7	154%	150%	156%	100%
上級	12	24,269	2,022.4	80	3.3	3.3	2.3	12	94%	96%	70%	100%
上級上	18	49,213	2,734.1	186	3.8	3.6	3.4	18	108%	103%	106%	100%
超級	15	42,990	2,866.0	183	4.3	4.2	3.4	15	121%	121%	106%	100%
母語	*129	1,419,729	11,005.7	4991	3.5	3.5	3.2	129	100%	100%	100%	100%

*：母語の 129 は人数ではなく名大会話コーパスの会話数。この行は 1 会話当たりの指標を算出している。

⓪粗頻度は，タラの用例からゴミと誤用を除く以外，何も調整していない頻度である。粗頻度で見ると中級上の粗頻度は 76，上級の粗頻度は 80 であるから，一見中級上より上級の方がタラを多く使用しているように見える。しかし中級上の学習者は 7 名，上級の学習者は 12 名であるから，一人当たりのタラの頻度で考えると中級上の方が多い。また語数合計で考えると中級上は 14,004 語発話した中で 76 回タラを使用しているのに対し，上級は 24,269 語発話した中で 80 回タラを使用しているのであるから，両者の語数を調整して比較すれば，やはり中級上の方が多くなる。つまり，レベルごとに人数や語数に違いがある場合，粗頻度で比較しても無意味であり，レベルごとの条件を一定にした上で比較する必要がある。第 1 節で述べたように，異なる条件でも比較が可能なように一定数で平準化した語数を調整頻度という。表 1 の①調整頻度は「粗頻度÷語数合計×1,000 語」で平準化した 1,000 語当たりの調整頻度である。これを見ると中級上では 5.4，上級では 3.3 で，上級より中級上の方が 1.63 倍ほど多く使用されている。

②個別平均（個別調整頻度の平均値）は学習者一人ひとりの調整頻度を出し，その合計を人数で割って一人当たりの平均を求めた値である。調整頻度と個別平均はよく似た値になるが，中級下では 1.1 と 0.8，中級上では 5.4 と 5.2 のように若干異なっている。この理由は次節で詳しく述べるが，調整頻度が合計数を使用して簡便に計算しているため，正確な値にならないことに起因している。

③中央値は，学習者ごとのタラの頻度を大きさの順に並べた時，真ん中に位置する値である。この学習者の頻度は粗頻度ではなく個別調整頻度を使用している。学習者が偶数の場合は真ん中の 2 名の平均である。データの中に少数の極端な値（外れ値）が存在する場合は，平均値より中央値の方が代表値としてふさわしいとされている（吉田，1998，p.45）。

④使用者数はタラを 1 回でも使用した学習者が何名いるかを数えた値である。初級下，初級中，初級上の使用者は 0 名，中級下では 9 名のうち 3 名がタラを使用している。中級上からは全員がタラを使用している。このように各レベルの学習者全員のうち，何割の学習者がタラを使用しているかを計算した値が，表 1 の右端の④’ 使用者割合である。

①’ 調整頻度割合，②’ 平均値割合，③’ 中央値割合は，各指標の母語の値を 100%とした場合の各レベルの割合を示した値である。ここでは名大会話コーパスの母語話者が行った雑談の値を基準にし，それと比較することで各レベルがどれぐらいの割合になっているかを算出している。

調整頻度，平均値，中央値，使用者割合などの指標は，どれも基礎的な統計量であり，これまでどの指標を使用するかは分析者の判断に任され，それぞれの指標の妥当性が問題にされることは少なかった。しかし表 1 を見ると，どの指標を使用するかで学習者がタラを習得した時期の判断が変わる可能性がある。例えば①’ 調整頻度割合や②’ 平均割合なら，中級上から習得に至ったと判断できそうである。中級上の調整頻度割合は 154%，平均割合は 150%で，母語よりも多くタラが使用されている。上級ではわずかに割合が下がるが，ほぼ母語話者と同じレベルである。一方③’ 中央値割合はこの二つより判断が難しい。中級上では母語の 156%，上級では 70%と割合が大きく揺れている。中級上で過剰使用されているだけなら，習得したばかりで使用に偏りがあったとも考えられるが，上級で逆に過少使用になるのであれば，データの信憑性も含めて詳しく検討する必要がある。あえて表 1 の値だけで判断するなら，中央値では上級上と超級で母語話者と同レベルの 106%になるため，上級上で習得に至ったと考えるのが妥当かもしれない。④’ 使用者割合は中級中で 93%，中級上以上は 100%となっており，中級中から習得に至ったと判断してよさそうである。中級中では 1 名＝7%がタラを使用していないが，学習者コーパスの場合，未出現＝未習得とは必ずしも考えられない（石川，2012，p.220）。未出現の学習者の場合，コーパスのデータではたまたまタラが出現していないものの，実際はタラを習得している可能性を否定できない。中級中では 1 名の学習者を除いて 93%の学習者がタラを使用していることから，このレベルではタラの習得に至っていると判断しても良さそうである。

以上の例から分かることは，基礎的な統計量であっても，どれを使用するかで結論が変わる可能性があるということである。分析者が無批判に何らかの指標を使用した場合，その指標が妥当な結論を導き出しているとは限らない。これまで当たり前のように使用されてきた指標ではあっても，その妥当性を検討してみる必要がある。

2.3 ①調整頻度，②個別調整頻度の平均値，③中央値，④タラの使用者数割合の妥当性

表 1 の検討により，①〜④の指標ではどれを使用して分析するかで結論が変わる可能性があることが分かった。この 4 つの指標の妥当性を評価するため，本節では表 1 の中級下の詳細データをまとめた表 2 を使用し，それぞれの指標の計算方法を検討する。表 2 の 1 行目は学習者個々の発話の語数である。最も少ない学習者は 714 語，最大は 1,991 語で，同じ中級下でも 3 倍近くの幅がある。2 行目の粗頻度は学習者がタラを使用した回数で，何も調整していない頻度である。1 回もタラを使用していない学習者が 6 名いる。3 行目の個別調整頻度は，学習者個々にタラの 1,000 語当たりの調整頻度を求めた値で，2÷1,520×1,000＝1.3 のように計算してある。

縦線より右側は，左側のデータに基づいて算出した値である。合計は 9 名の合計数，人数はいずれも 9 名，平均値は合計を人数で割った値である。右端の調整頻度はタラの粗頻度合計 13.0 ÷語数合計 11,338×1,000 で，1,000 語当たりの調整頻度である。これは粗頻度の平均値 1.4÷語数の平均値 1,259.8×1,000 と同じ値で，調整頻度とは，粗頻度平均を 1,000 語当たりの頻度に直した値と同じである。3 行目の個別調整は，左側で一人ひとりの値を個別に調整しているため，右側で調整頻度を算出する必要はなく，合計を人数で割った値が個別調整頻度の平均値（個別平均）になる。

表 2 KY コーパス中級下の詳細データ：学習者 9 名の発話語数・タラの粗頻度・個別調整頻度

	中級下の学習者の個別データ									合計	人数	平均値	調整頻度
発話語数	714	739	846	1,018	1,299	1,455	1,520	1,756	1,991	11,338	9	1259.8	
粗頻度	0	0	0	0	0	0	2	9	2	13.0	9	1.4	1.1
個別調整頻度	0	0	0	0	0	0	1.3	5.1	1.0	7.4	9	0.8	

調整頻度も個別平均もどちらも平均値であるが，値が異なる（表 2 の網掛け箇所）。それは個別平均が個々の学習者ごとに調整頻度を算出した上で平均値を出しているのに対し，調整頻度はタラの粗頻度と学習者の語数を合計して割っているためである。

表 2 の個別データの特徴は，語数の多い学習者だけがタラを使用しており，語数が少ない学習者はタラを使用していないところにある。この特徴の意味するところを，学習者の用例で具体的に観察してみよう。中級下で最も語数が少ないのは，ID が EIL01（E＝英語，IL＝中級下，01=№01）の学習者である。次の(1)はこの学習者のデータの一部で，T が面接官，S が学習者である。引用部では面接官が学習者に対し，学習者が現在住んでいる京都で面白いところはどこかを尋ねている。これに対する学習者の応答は，ほとんど単語でなされている。

(1) T：どんなところが、面白いですか、銀閣寺の面白いところ
S：どんな建物
T：建物、〈うん〉寺の形ね、のうち、〈うん〉でなにがあなたは面白いですか
S：おもしろい
T：きょうみ、興味がありますか

S:神社
T:神社で、あのー、銀閣寺、庭がいいですか、あのー
S:建物
T:建物がいいです
S:いいです （KYコーパス, EIL01より引用）

次の(2)は，中級下でタラの頻度が最も多い中国語母語話者 CIL03 のデータで，電話をして面接官をご飯に誘うロールプレイを行っている会話の一部である。

(2) S:あ、そ、んー、よかったらー、んー焼き肉いきましょうか
T:うーん、えー高いですか、あのあまりお金がないんですけど
S:大丈夫よー、んー、はん、うんー、、少し食べてー、〈ええ〉話、したいです
T:えーで、でも、おなかすいてるから、〈あ、そうですか〉えーえー
S:うーんー、うん、どうしようかなあ、先生は、なに食べたいですか
T:んいやな、なんでもいいですけど、えーえー
S:なんでもいいですでも、〈ええ〉高いのも、ものー
T:うん高いのはちょっと困るなあ
S:困るなあ、〈うんうんうん〉焼き肉高いですか
T:んーちょっと高いかも知れませんね
S:そうですか （KYコーパス, CIL03より引用。太字・下線は稿者による。以下同じ。）

(1)では，学習者の発話が単語レベルであったのに比べ，(2)では複文も使い，勧誘の目的を遂行する発話ができている。この2名の学習者はともに中級下のレベルで判定されているが，発話の量も質もかなり異なることが分かる。(1)の学習者は単語レベルの発話が多いため，語数が少なく，条件節を作るタラも出現しない。タラの出現には複文が発話できる能力が必要であり，その能力があれば逆に語数は一定以上の量になる。(2)のように，一定以上の語数があってはじめて出現するタラの調整頻度を算出するのに，(1)のようなタラが出現しない短い語数のデータを合計して計算するのは原理的に問題がある。タラの調整頻度は，タラが出現した学習者の語数で割ってこそ意味を持つ。

調整頻度は，各学習者の語数が同じなら個別平均と同じ値になる。この場合，個々の学習者ごとに調整頻度を算出する手順を省いて，レベル別の合計で計算しても問題はない。しかし，個々の学習者によって語数が異なる学習者コーパスの場合，合計を用いて計算しても正しい値にはならない。個々の学習者ごとに個別調整頻度を算出し，その平均を求めるのが適切な方法である。

ただし，個別調整頻度の使用に当たっては，注意が必要な場合がある。それは，語数が非常に少ない学習者が，たまたまある調査語を使用していた場合などのケースである。語数が100語しかない学習者の1回の頻度を1,000語当たりの個別調整頻度に直すと10回の頻度

になるが，それが実態に合っているとは限らない。語数が少ない学習者のデータの場合，単純な割り算や掛け算で語数が平準化できないことには留意する必要がある。

個別平均は個別調整頻度を使用して算出した平均値である。この平均値も妥当な代表値であるとは考えにくい。中級下では 9 名中 6 名の学習者の頻度が 0 だが，1 名の学習者が 5.1 回使用しているため平均値が 0.8 回になっている。平均値はデータの中心的傾向を代表した値だが，1 名の学習者のためにその傾向が歪んでしまっている。平均値で比較する際の適用条件は，データが正規分布に従い，比較するデータ同士の分散が一致していることである。データ量が少なく外れ値が多い学習者コーパスは，この条件に合いにくい。

中央値には調整頻度や平均値に見られるような適用条件の問題はない。しかし中央値は学習者の調整頻度を順番に並べた時の真ん中の学習者の頻度情報しか使用していないため，小規模のデータでは偏った値になる場合がある。表 1 の中央値割合を見ても中級上で 156%，上級で 70%と不安定な値を示していた。

使用者数割合も原理的な問題はなく，情報を単純化しているだけに分かりやすい。しかし使用者数割合は，タラを 1 回でも正しく使用した学習者は，タラを習得していると見なす考え方である。これは 1 回の使用も 10 回の使用も同一視するため，タラの習得を甘い基準で認定してしまう可能性がある。中央値や使用者数割合は，本来利用できる情報の一部しか使用しないため，取得した情報量を生かし切れていない。

2.4 ⑤蜂群図と箱ひげ図の合成図を使用した分析法

調整頻度や平均値は，算出の原理や適用条件に問題がある。中央値や学習者数割合は，情報の一部しか有効活用できていない。これらの統計指標が有効に機能しないのは，使用したデータの性質によるところも大きい。KY コーパスはデータ量が少なく，学習者ごとのばらつきが大きいデータである。そのようなデータを使用しているにも関わらず，データが正規分布に従っている時に最も威力を発揮する平均値を算出しても，指標が有効に機能しないのは当然であろう。

調整頻度や平均値などは，データを一つの点で代表させる代表値である。代表値はデータの特徴を端的に表すため多くの分析で使用されるが，学習者コーパスのようにばらつきの大きいデータの場合，学習者の分布を全体的に俯瞰する分析法の方が適している。これには，散布図による観察が有効である。図 1 は横軸を学習者のレベル，縦軸をタラの個別調整頻度にし，統計ソフトの R を使用して作図した散布図である。図 1 で中級下のマーカーを観察すると，用例(2)で引用した学習者の 5.1 という値が，他の学習者から大きく離れていることが確認できる。

ただし図 1 のような一般的な散布図では，1 つのマーカーに何名の学習者が重なっているのかよく分からない。このため，R で beeswarm のパッケージを使用し，マーカーが重ならないように蜂群図で描いたのが図 2 である。母語のマーカーが多いのは，データ数が 129 の名大会話コーパスを使用したためである。図 1 では頻度 0 のマーカーが重なり，学習者が何名いるのか分かりにくいが，図 2 ではマーカーが横に広がっているため，把握しやすい。

しかし図 1, 2 では，レベルが上がるにつれ，タラの頻度が直線的に上昇しているように感じられ

る。これは，無意識のうちにグラフの頂点付近のマーカーをつないで観察してしまうからである。グラフの頂点付近には外れ値も含まれるため，このような観察は好ましくない。外れ値の影響を受けないグラフを描くには，縦軸を頻度ではなく順位に直して作図する方法が考えられる。しかし順位に直すと，せっかくの頻度情報を失ってしまうため，図 3 のようにデータの分布比率を描画できる箱ひげ図を重ね書きする方法の方がより優れている。

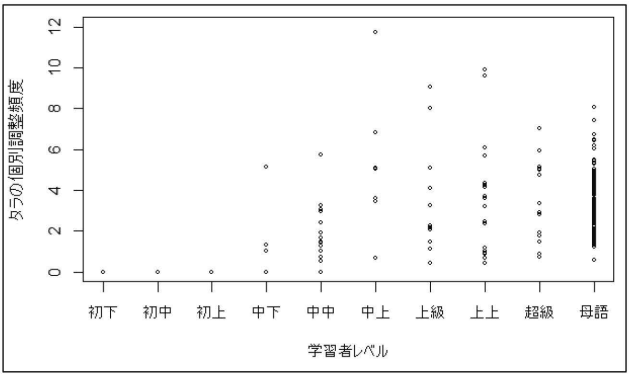

図 1　学習者レベルと個別調整頻度の散布図

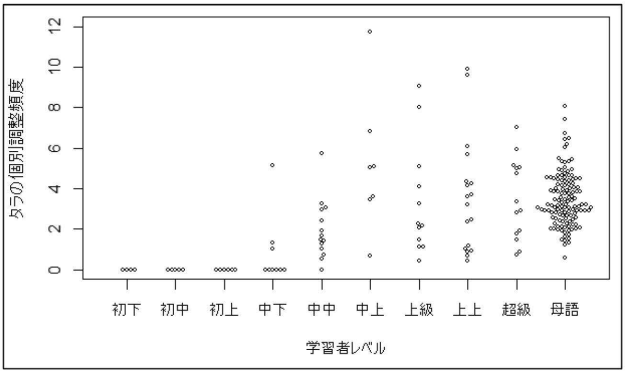

図 2 学習者レベルと個別調整頻度の蜂群図

箱ひげ図は横の太線が中央値，箱の下限が第 1 四分位点，上限が第 3 四分位点を表す。つまり中央値を挟んで，25%から 75%の学習者が箱の中に納まる。ひげの長さは中央値から箱の長さの 1.5 倍までとすることが多く，このひげの外側のデータは外れ値と見なすのが一般的である。図 2 では，頂点の推移に目が向きやすいが，図 3 なら中心的傾向の推移やデータのばらつきの具合，

外れ値の見極めなどが容易で，レベルごとの比較がしやすい。また，単なる箱ひげ図とは異なり，学習者の値がマーカーで示されるため，箱の位置や大きさが何に基づいて描かれているのか理解しやすい。図 3 では，中級下から上級上にかけて，タラの使用が徐々に増加し，やがて頭打ちになっていく様子が観察できる。

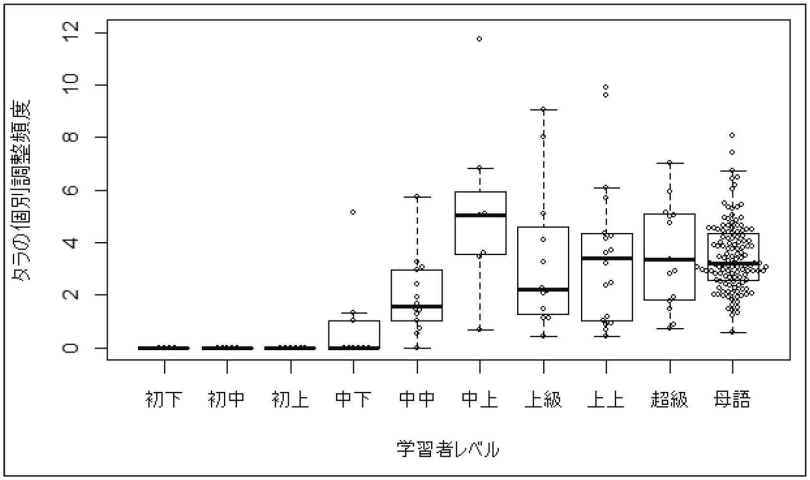

図 3 学習者レベルとタラの個別調整頻度の合成図（蜂群図＋箱ひげ図），母語：名大コーパス

ただし，中級上だけは他のレベルと大きく傾向が異なっている。これはなぜであろうか。第 2.2 節の表 1 で，中央値割合を比較した際は，中央値割合が 50%→156%→70%と大きく揺れることしか分からなかったが，図 3 では，中級上で最も頻度の低い学習者は 1 回程度，他の多くの学習者は 5 回前後，最も多い学習者は 12 回程度と，学習者によって大きなばらつきがあることが見て取れる。このような違いがなぜ生じるのか，学習者の用例で確認してみよう。

次の (3)は，タラの個別調整頻度が最も多かった中国語母語話者の発話の一部である。ここでは日本の大学生が中国の大学生に比べて授業を欠席しがちであり，先生も注意しないことに戸惑いを覚えているというＳに対し，Ｔが質問している。

(3) T：あなたが大学の先生だっ**たら**どうしますか
S：だっ**たら**、〈うん〉そうですねー、日本へ、来る前、〈ええ〉ちゅ、たぶん厳しい先生、〈ええ〉自分が先生だっ**たら**、学生いなかっ**たら**、ちょっと気持ち悪い、〈んー〉でも日本へ来て、{笑}今、そこで、日本の教育、制度、ちょっとだけ、わかるようになっ、なったので、〈はい〉んー、帰っ**たら**、先生た、になっ**たら**、どうにかなる、まだわかりませんでした、〈はー〉わかっ

ておりません、たぶん中国の先生より、厳しくない、〈うん〉日本の先生より厳しい、〈あー〉その間かもしれません　(KY コーパス, CIH01 より引用)

次の(4)は, タラの個別調整頻度が中央値に近い, 5.1 の学習者が, バスケットボールのルールの説明を求められている会話である。

(4) T:あー、バスケットボールは、見たことはあるんですけど、わたしはルールがよくわからないんですけども、などういうふうになったら、勝つんですか
S:やー、バスケットボールは、それはやっぱり、あのー、点数が高くなって、〈ええ〉勝ちます、あの、自分の、あの、自分、いや、あれは、なんという、わからないです、あの、ボールは、自分の、方に、なげて、〈ええ〉あのー、あたったら、あの、点数が、とります
T:あたったら
S:はい
T:どこにあたるんですか
S:あの、バスケットボールは、あの丸いの、〈ええ、丸い〉日本語、何という分からないです
T:んーどんなものですか
S:たかいのものは、〈はい〉あのー、ボールをそのなかに投げて、〈ええ〉もし当たる、入れます、〈はい〉あの、ボールは、いれ、たら、点数が、あげます　(KY コーパス, CIH02 より引用)

(3)や(4)では, 面接官のタラを含んだ質問に答えるために, 学習者にもタラが多く使用されている。これに対し次の(5) はタラの粗頻度が 1 回と, 中級上で最も少なかった英語母語話者の会話の一部である。ここでも学習者はトライアスロンの内容や筋肉トレーニングなどのスポーツの説明を求められているが, 面接官にタラの使用がなく, 話題も仮定の内容を含まないため, 引用部分にはタラの使用が見られない[2]。

(5) T:トライアスロンって何ですか
S:トライアスロンは、バイクとか、〈はい〉水泳とか、〈はい〉running、とか、〈ええ〉ずっと、race、全部で、〈はい、ふーん〉race、ですね、長い、とても、大変、〈ふーん〉ですね
T:それを、まあやってらっしゃるんですね
(中略)
T:うん、私、実は、そのウェイトトレーニングっていうのを、よく知らないんですが、〈うん〉あのー、具体的にはどんなことをするんですか
S:あのねー、いろいろ、あの筋肉は、いついつ全部違いますね、〈ええ〉だから special、練習あります、〈あーん〉この1つずつ、〈えーえー〉だから、あの、この、この、あの、筋肉は、こんな練習、だから、こんな筋肉は、こんな練習、〈うん〉全部違いますから、〈あーそうですか〉うん

T: でも、いつも、あのー、決まったことをやっているんですね（KY コーパス, EIH03 より引用）

中級上 7 名のスクリプトを観察すると，面接官のタラの使用に合わせて学習者がタラを使用しているケースが多く見られる。学習者のタラの使用には一定の負荷がかかっている可能性があり，この頻度分布は例外と見なすのが妥当だと思われる。

中級上のデータを例外と見なすと，タラは中級下から一部の学習者で使用が始まり，中級中から徐々に使用が多くなり，上級上で頭打ちとなる。超級の分布は母語話者の分布とよく似ており，超級では母語話者と同じレベルでタラが使いこなせるようになっていると考えられる。表 1 のような代表値ではこの変化の詳細が分からないが，すべての学習者を一つのグラフ上にプロットできる蜂群図と，その統計的な分布の比率を図示できる箱ひげ図との合成図を使用すれば，学習者は大きなばらつきを持ちながらも，一定の傾向性を持ってタラを習得し，母語話者と同レベルになっていく状況が的確に比較できる。

3. I-JAS を使用した頻度分析法の比較

前節では KY コーパスを使用して，タラの検索データを用いた分析法の比較を行った。しかし KY コーパスは総語数約 17 万語の小規模コーパスである。このため前節の結論は小規模コーパスでのみ成り立つ結論に限定される恐れがある。そこで本節では，現在公開されている I-JAS の二次データを使用して同様の比較を行い，前節で得られた結論が大規模なコーパスでも変わらないことを示す。

3.1 使用するデータの説明

I-JAS は，国立国語研究所によって構築が進められている大規模な学習者コーパスで，12 言語の母語話者 1,000 名に対し，インタビューやタスクなどを行ってデータが集積されている（迫田（編），2016；迫田ほか，2016）。2016 年 5 月の一次公開に続き，2017 年 5 月には学習者 400 名，日本語母語話者 50 名の二次データが公開された。I-JAS には発話データのほか，若干の作文データが含まれている。本稿では OPI を基にした KY コーパスと比較するため，発話データのみの約 122 万語を使用する。I-JAS も名大会話コーパスと同様に，国立国語研究所が開発した検索インタフェース『中納言』による検索が可能になっている。本稿ではこれを利用し，語彙素「た」，品詞「助動詞」，活用形「仮定形」で検索したタラ 3,725 語のデータを使用する。

I-JAS では 17 名の学習者を除き，各学習者に J - CAT（Japanese Computerized Adaptive Test）の聴解，語彙，文法，読解の各 100 点満点中の得点と合計点，および SPOT （Simple Performance - Oriented Test）の 90 点満点中の得点が付与されている。したがって，OPI のレベルと J-CAT や SPOT の得点の対応付けができれば，基本的にはレベルを合わせた比較が可能になる。ただしこれらは異なった枠組で作成されたテストであるため，機械的に対応付けられるわけではない（李ほか，2015，pp.53-4）。本稿では J-CAT の「スコア互換表」

（http://www.j-cat.org/page/interpret, 2017.1.7 閲覧）を利用し，J-CAT の合計点を KY コーパスのレベル区分に読み替えて分析を行う。区分は 100 点以下：初級，101～150：中級下，151～200：中級中，201～250：中級上，251～300：上級，301～350：上級上，351 以上：超級の 7 種類とする。この区分はあくまでも大まかな目安であり，この区分と OPI のレベルにどれぐらい密接な対応があるのかは，今後多くの学習項目に基づく比較を通して検証していく必要がある。

3.2 ①調整頻度，②個別調整頻度の平均値，③中央値，④タラの使用者数割合の調査結果

表 3 は表 1 と同様の方法で集計した分析結果である。この結果には J-CAT 未受験者の 17 名は含まれていない。学習者数の欄を見ると，I-JAS では初級と上級上，超級の人数が少なく，中級中と中級上に学習者が集中しているのが分かる。

表 3 で①調整頻度と②個別平均を比較すると，KY コーパスと同様に少しずつ値が異なっている。中級中や中級上といった学習者数が 120 名を超えるレベルでも，両者の値は一致しないことが確認される。②'個別平均割合と③'中央値割合は，中級下の 70%と 37%，中級中の 74%と 55%のように大きく異なっている。この理由は，外れ値を含んだまま平均を算出したため，平均値が高く出てしまったことによる。表 3 を見ると，学習者数が多くなったからと言って，各指標が必ずしも安定するわけではないことが分かる。

表 3　I-JAS の基礎統計量とタラ頻度に基づく 4 種類の分析結果の比較

レベル	学習者数	語数合計	語数平均	⓪ タラ粗頻度	① 調整頻度	② 個別平均	③ 中央値	④ 使用者数	①' 調整頻度割合	②' 個別平均割合	③' 中央値割合	④' 使用者割合
初級	21	37,318	1,777	39	1.0	1.0	0.6	12	40%	37%	23%	57%
中級下	39	98,491	2,525	202	2.1	1.8	0.9	26	79%	70%	37%	67%
中級中	121	320,681	2,650	668	2.1	1.9	1.3	104	80%	74%	55%	86%
中級上	134	416,589	3,109	1106	2.7	2.6	2.2	128	102%	100%	91%	96%
上級	53	182,043	3,435	628	3.4	3.3	2.9	50	132%	126%	120%	94%
上級上	13	49,099	3,777	185	3.8	3.8	3.2	13	144%	146%	132%	100%
超級	2	7,224	3,612	38	5.3	5.2	5.2	2	202%	198%	212%	100%
母語	50	240,819	4,816	628	2.6	2.6	2.4	49	100%	100%	100%	98%

3.3 ①調整頻度の妥当性

本節では中級中のデータを使用し，調整頻度と個別平均の値が食い違う理由を考察する。この理由は図 4 を見ると理解しやすい。図 4 は中級中の語数とタラの個別調整頻度の散布図である。この図で，タラの頻度が 0 の学習者に着目すると，タラの未使用者は 3,200 語以下の語数が短い学習者に限られることが分かる。特に 1,500 語以下の学習者に，タラを使用する学習者はいない。この現象は，第 2.3 節で観察した KY コーパスと同一であり，タラの頻度を平準化するのに，短い発話しかできないタラ未使用者の語数を使用しているため，食い違いが生じると考えられる。

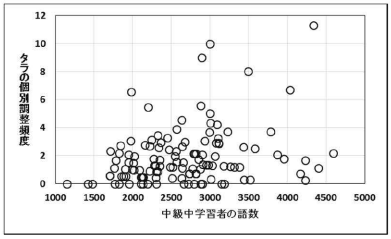

図 4　中級中学習者の語数とタラ頻度の散布図

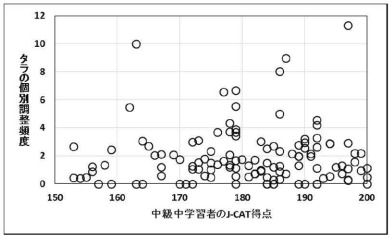

図 5 J-CAT 得点とタラ頻度の散布図

図 5 は J-CAT 得点とタラ頻度の散布図である。これを見ると，J-CAT 得点とタラの頻度にはほとんど相関がない。得点が低くてもタラを多用する学習者もいれば，得点が高くてもタラを使用しない学習者もいる。中級は初級や上級に比べ多様な学習者が存在し，そのばらつきも大きいことが確認できる。このように発話の語数もタラの頻度も大きくばらついているデータを使用して，合計数で調整頻度を求めても，正確な値にはならない。頻度を調整するには学習者ごとに個別調整頻度を算出するのが適切な方法である。

3.4　3 指標の妥当性と合成図の有効性

本節では，合成図の観察を通し，平均値，中央値，使用者割合の妥当性と合成図の有効性を検討する。 図 6 は，図 3 と同じ方法で描いた合成図である。I-JAS には日本語母語話者のデータが含まれているため，「母語」は学習者と同条件で採取した母語話者のデータを使用している。これに参考値として，図 3 で使用した名大会話コーパスのデータを加えた。

図 6 は，それぞれの箱の大きさがコンパクトで，中央値も箱の真ん中に位置するものが多い。この理由は，データ数が多く，学習者の傾向性が安定するまでデータが出そろっているからだと思われる。中央値や使用者割合はもともと外れ値に強い指標であるが，人数が増えたことによって，より指標の信頼性が高まったと考えられる。ただし，当然ながらこの 2 つの指標が限定した情報しか使用していないことには変わりはない。

一方の平均値は，データが増加したことによって信頼性が増したとは考えにくい。その理由は図 6 を見れば明らかなように，多くの外れ値が存在するからである。平均値がデータの代表値としてふさわしいのは，母語のデータのように中央値を境に上下がほぼ対称になる場合である。母語でも外れ値がひとつだけ見られるが，他のデータに強い正規性があるためほとんど影響はない。これに対して，学習者のデータはひげの長さも非対称で，ひげの上に多くの外れ値が続いている。この場合，平均値を使用しても中心的な傾向を示すことができない。中級に外れ値が多いのは，先に述べたようにこのレベルに多様な学習者が存在するためだと考えられる。したがって，いくらデータを増加させても，中級におけるばらつきは解消しない可能性が高い。このため大規模な学習者コーパスを使用しても，平均値や平均値の一種であるレベル別の調整頻度を使用することは妥当ではないと考えられる。

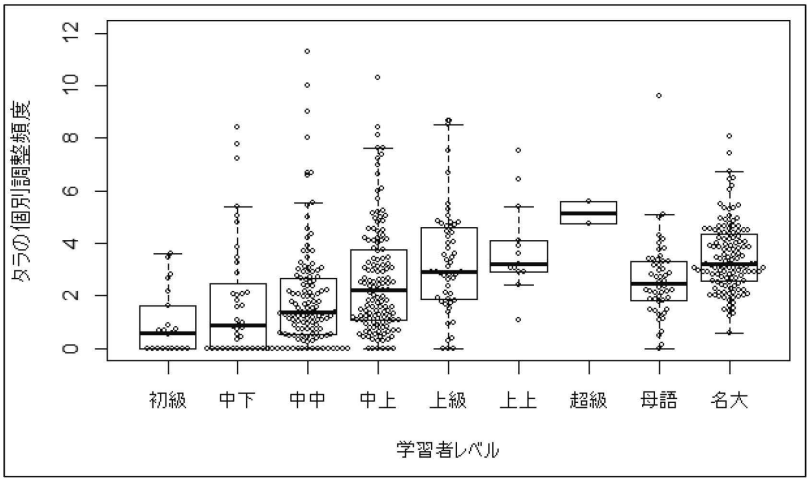

図 6 学習者レベルとタラの個別調整頻度の合成図(蜂群図＋箱ひげ図),
母語:I-JAS, 参考:名大会話コーパス

次に合成図の有効性を考察する。合成図を観察する利点の一つは, 図 6 を観察することによって中央値の正確さや平均値の不正確さの理由が分かったように, グラフを観察することでデータの正確性が評価できる点にある。例えば超級では, タラの頻度が飛びぬけて高いのは, 超級にわずか 2 名の学習者しかおらず, たまたま頻度の高い学習者がそろったためだと思われる。その根拠を明確に示すのは難しいが, 初級から上級上までの頻度増加の傾向から考えると, ここで超級だけ極端に頻度が増加するとは考えにくい。表 3 の指標を使用した場合, 算出された値が正しいという前提に立って分析するしかないが, 合成図ではデータがどれほど正確であるかの評価も行いながら全体の分布を考察することができる。

このようなデータの正確性の評価を行いながら分析できる利点は, 異なったコーパスの比較を行う際にも有効に働く。一般的に言って, コーパスはどのようなコーパスであれ, それぞれの設計方針に基づく何らかの偏りを持っている。このため単独のコーパスで,「日本語学習者」といった巨大な母集団の実態を推定することは困難である。しかも, それぞれのコーパスがどのような偏りを持っているかは, 異なるコーパスを比較してみないと分からない場合が多い。図 7, 8 は図 3 の KY コーパスと図 6 の I-JAS のタラ頻度を比較したグラフである。

二つのコーパスを比較してまず分かることは, 大きな傾向性が非常によく似ているということである。第 3.1 節では, OPI と J-CAT という異なる基準で区分されたデータを比較する難しさを述べた。OPI では単語や句のレベルでしか発話できないレベルを初級と定めているため, 複文で使用されるタラは中級以降にしか使われない。一方, J-CAT では聴解, 語彙, 文法, 読解の総合点でレ

ベル分けしているため，初級からタラを使用する学習者が観察される。このような違いはあるが，中級中から大半の習者がタラを使用するようになり，上級以上では使用数がほぼ母語話者と同レベルになるという傾向は同じである。タラの使用頻度に限れば，OPI と J-CAT の区分を同列に使用しても，それほど大きな問題はなさそうに思われる。

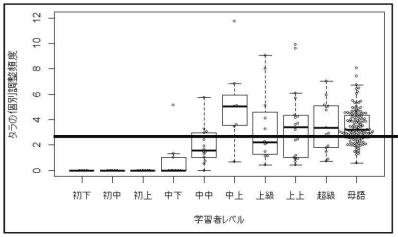

図 7　KY コーパスのタラ頻度

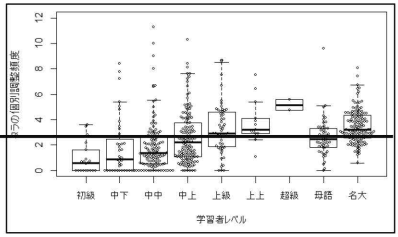

図 8 I-JAS のタラ頻度

問題は KY コーパスの中級上や I-JAS の超級のデータである。KY コーパスの中級上については，第 2.4 節でスクリプトを観察しながら，これがインタビューによって引き起こされた偏りである可能性を指摘した。しかしそれを本当に偏りと言ってよいかは判断が難しかった。ところが図 7，8 で I-JAS の中級上と比較すれば，134 名のデータに基づいて KY コーパスの中級上のデータは偏っていることが裏付けられる。I-JAS の超級データも，これが偏っているという判断は明確な根拠を示すのが難しかった。それが KY コーパスと比較することによって，15 名のデータに基づいて偏りを持っていることが裏付けられる。このような判断が可能になるのは，合成図に全学習者の情報がプロットされているため，データの前後のつながりや集積されている人数，そのばらつきの程度などを評価して判定できるからである。

さらに合成図では，この図にプロットされている学習者の情報を調べることで，分析をさらに進展させることが可能である。例えば I-JAS では，20 か所の海外地域で 850 名の調査が行われているほか，国内の教室環境学習者 100 名と国内の自然環境学習者 50 名の調査も行われている。このため，学習者を JSL と JFL に区分して分析することが可能である。

図 9 は図 6 の中からタラの頻度が 6 以上の使用者を抽出し，学習者の調査 ID を併記したグラフである。これには J-CAT の未受験者 17 名のうち，タラ頻度が 6 以上であった 5 名の学習者も加えている。図で濃い網掛けをしているのが国内の自然環境学習者（JJN），薄い網掛けが国内の教室環境学習者（JJC と JJE）である。二次データ 400 名の中で JSL 学習者は，その 20%に当たる 80 名であるが，図 9 ではタラ頻度 6 以上の学習者 30 名のうち，その 56.6%に当たる 17 名が JSL 学習者である。明らかにタラを多用している使用者には，JSL 学習者が多いことが分かる。

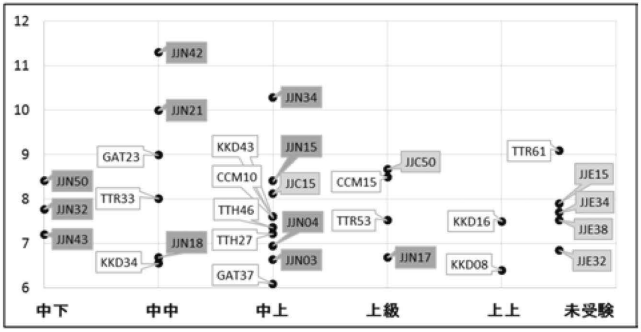

図 9 タラ頻度 6 以上の学習者とその ID

平均値，中央値，使用者割合を使用した場合，タラの頻度がレベルによってどのように異なるかは比較できるが，そこから分析を発展させていくことが難しい。合成図の場合，マーカーの一つひとつが個々の学習者に対応しているため，図 9 のような追加分析が可能になる。特に I-JAS では学習者ごとにフェイスシートが用意されているため，これを利用して，タラを多用する JSL と JFL に共通する要因をさらに追及していくことも可能である。

3.5 学習者コーパスにおけるデータ数と分布のばらつきの関係

前節前半では，学習者のデータが増加しても，平均値が妥当な指標としては使用しにくいことを確認した。本節では母語話者データを観察することにより，この理由をさらに検討する。

I-JAS の母語話者データは，学習者と同じインタビューやタスクを行って収集されている。もう一方の母語話者データである名大会話コーパスは，雑談を収集したデータであり，集約した単位も発話者ではなく会話である。これらは全く異なったデータであるが，全体的な分布のばらつきはよく似ている。図 10 はこれらを相対度数折れ線で描いた図で，両者ともほぼ正規分布に近似していることが確認できる。両者で頻度が異なる理由は，データが質的に異なることのほか，名大コーパスの調査地が名古屋を中心としていたことによる方言の影響などが考えられる[3]。しかし，ここで注目したいのは両者が非常に狭い範囲で近似している点である。

図 11 は，I-JAS の 50 名の母語話者データから，ランダムに再サンプリングした 10 名，20 名，30 名の度数折れ線を描いた図である。図 11 はたまたま選んだデータを描いたに過ぎないため，一般化はできないものの，均質な母集団からランダムサンプリングしたデータは，基本的に元の母集団に類似した分布を示すことが観察できる。このうち 10 名のグラフの形が他の 2 つと異なるのは，データ数が少ないためだと思われるが，しかしそのグラフですら，正規分布に近い形をしている。

図 10, 11 から分かることは，日本語母語話者という均質な母集団から採取したデータは，10 名や 20 名といった少ない人数でも，正規分布に近似するということである。その一方で，図 6 で観察

したように，学習者のデータはばらつきが大きく，正規分布には従わない。その理由は学習者のタラ使用が均質ではなく，様々な多様性を持っているためだと考えられる。つまり，学習者コーパスでばらつきが大きいのは，データ数の問題ではなく，学習者の多様性の問題である。このようなデータを分析する場合は，一つの値でデータを代表させるのではなく，ばらつきの状態を含めた全体の分布を観察し，どのような理由でデータがばらついているのかをさらに追求できる形で比較することが，有効な分析法だと考えられる。

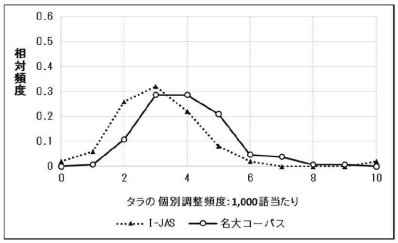

図 10 タラ頻度の相対度数折れ線（全データ）

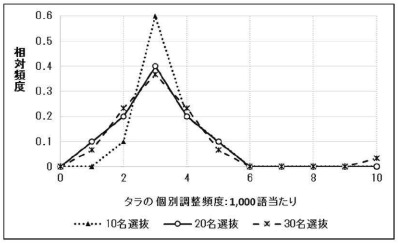

図 11 タラ頻度の相対度数折れ線（選抜）

4. まとめと今後の課題

本稿では，学習者コーパスの頻度を比較する 5 種類の方法について，その妥当性を検討した。均衡コーパスの分析で使用されている①「調整頻度」は，学習者ごとの語数が異なる学習者コーパスで使用すると不正確な値になる。レベル別の語数合計で頻度を平準化するのではなく，個別の学習者ごとに調整頻度を算出して分析するべきである。ただし，個別調整頻度を使用しても，②「平均値」は外れ値が多く不正確になりやすい。③「中央値」やタラの④「使用者数割合」は原理的には妥当だが，一部の情報しか有効活用できない。

学習者のデータは個人によるばらつきが大きく，レベル別の習得状況を一つの代表値で点的にとらえるのは難しい。学習者の実態を面的にとらえるには作図による分析が適している。散布図であれば，学習者の全体の姿を一枚のグラフで表現できる。ただし，通常の散布図はマーカーの重なりが分かりにくいため，本稿ではマーカーが重複しない蜂群図を使用し，さらにデータの四分位が分かる箱ひげ図を重ね書きする手法を採用した。蜂群図と箱ひげ図の合成図で観察すると，データの正確性を評価しながら全体の傾向性を分析できるだけでなく，個別のデータに着目した追加分析も可能になる。

数値を使用した分析を行う前に，データをグラフ化して観察するべきだという主張は，統計分析の基本であり，本稿の分析もこれを支持する結果となった。本稿の主張は，代表値を使用した分析法を否定することではなく，全体の分布観察に基づいて，使用するデータにふさわしい分析法を選択するべきだという点にある。

本稿によって新たに明らかになったことは，「学習者コーパスで調整頻度を使用するのは問題がある」という点と，「学習者コーパスは，母語話者の均衡コーパスに比べ，作図を行

って観察する必要性がはるかに高い」という点である。本稿で使用したグラフを描くには，個別調整頻度を算出する必要がある。そのためには，まず，学習者ごとの語数を知ることが必要である。しかし，『現代日本語書き言葉均衡コーパス』のような母語話者の均衡コーパスではテキストごとの語数情報が提供されているが，KY コーパスや I-JAS では，学習者ごとの語数情報は提供されていない。グラフを描こうとしても，簡単には描けないのが現状である。このような分析の基礎となる基本データから，整備をしていくことが望まれる[4]。

学習者コーパスは，学習者を無作為抽出することが難しく，データもインタビューの内容や，設計されたタスクの影響を受けやすい。母語話者のデータを統計的な設計に基づいて集積した均衡コーパスに比べると，はるかに分析が難しいコーパスである。しかし，学習者コーパスの分析をどのように行っていけば有効な分析ができるかと言った方法論は，これまでほとんど議論されてこなかった。データ量が少なく，様々な偏りを持ち，学習者が母語や教育機関などによって入れ子状の階層性を持っているデータをどのように分析していけば良いのか，さらに検討を続けていくことが今後の課題である。

注

[1] 国立国語研究所の「中納言オンラインマニュアル」に，「中納言に格納された短単位データの作成は自動形態素解析によって行われています。 形態素解析処理は形態素解析器に「MeCab」、解析用辞書に「UniDic」を使用しています。」とある。
（http://pj.ninjal.ac.jp/corpus_center/chu-00.html，2017.11.19 閲覧）

[2] タグ付き KY コーパスで検索するとこの後の会話で「筋肉だめ、〈あー〉だったら、」とタラが 1 回出現するが，山内博之氏から提供を受けたＫＹコーパス version1.2 のスクリプトでは「筋肉だめ、〈あー〉だから、」となっており，タラは使用されていない。両者で表現が異なる理由は不明である。本稿の分析はタグ付き KY コーパスの検索結果を使用しているため，EIH03 はタラを 1 回使用していると見なした。

[3] 名大コーパスで使用されている言葉は共通語が大半を占めるが，問題は共通語を使用しているつもりでいながら，タラを多用している「気づかない方言」の可能性である。

[4] 森（2017）では，論文の最後に KY コーパスと I-JAS 発話データの学習者別語数の資料を示したが，これは一研究者による語数の集計例に過ぎない。多くの研究者が共通して使用できるように，データ提供者による公認の語数情報の提供が望まれる。

使用データ

本研究は『タグ付き KY コーパス』（http: //jhlee.sakura.ne.jp/kyc/），『KY コーパス version1.2』，『多言語母語の日本語学習者横断コーパス：I-JAS』，『名大コーパス』および検索システム・コーパス検索アプリケーション「中納言」バージョン 2.2.3.1：（https://chunagon.ninjal.ac.jp/ijas/search）を利用して行われたものである。

引用文献

石川慎一郎（2012）『ベーシックコーパス言語学』東京：ひつじ書房.

石川慎一郎（2017）『ベーシック応用言語学　L2 の習得・処理・学習・教授・評価』東京：ひつじ書房.

石川慎一郎・前田忠彦・山崎誠（編）（2010）『言語研究のための統計入門』東京：くろしお出版.

小木曽智信（2014）「第 5 章　形態素解析」山崎誠（編）『講座日本語コーパス　2.書き言葉コーパス―設計と構築―』（pp. 89-115）東京：朝倉書店.

鎌田修（1999）「KY コーパスと第二言語としての日本語の習得研究」『第 2 言語としての日本語の習得に関する総合研究』科研研究報告書 08308019 代表者カッケンブッシュ寛子, 227-237.

鎌田修（2006）「KY コーパスと日本語教育研究」『日本語教育』*130*, 42-51.

グレンジャー＝シルビアン（編）（2008）船城道雄・望月通子（監訳）『英語学習者コーパス入門 ―SLA とコーパス言語学の出会い―』東京：研究社.

迫田久美子（編）（2016）『海外連携による日本語学習者コーパスの構築―研究と構築の有機的な繋がりに基づいて― I-JAS 構築に関する最終報告書(International Corpus of Japanese As a Second Language)』科研研究報告書 24251010 代表者迫田久美子, 東京：国立国語研究所.

迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016）「多言語母語の日本語学習者横断コーパス　International Corpus of Japanese as a Second Language」『国語研プロジェクトレビュー』*6*(3), 93-110.

マケナリー＝トニー・ハーディー＝アンドリュー（2014）石川慎一郎（訳）『概説コーパス言語学－手法・理論・実践』東京：ひつじ書房.

バイバー＝ダグラス・コンラッド＝スーザン・レッペン＝ランディ（2003）齊藤俊雄・朝尾幸次郎・山崎俊次・新井洋一ほか（共訳）『コーパス言語学―言語構造と用法の研究―』東京：南雲堂.

森秀明（2017）「I-JAS と KY コーパスにおける量的な性質の比較」『2017 年度日本語教育学会支部集会予稿集【東北支部】』21-26.

山内博之（1999）「OPI 及び KY コーパスについて」『第 2 言語としての日本語の習得に関する総合研究』科研研究報告書 08308019 代表者カッケンブッシュ寛子, 238-245.

吉田寿夫（1998）『本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本』京都：北大路書房.

李在鎬（2009）「タグ付き日本語学習者コーパスの開発」『計量国語学』*27*(2), 60-72.

李在鎬・小林典子・今井新悟・酒井たか子・迫田久美子（2015）「テスト分析に基づく「SPOT」と「J-CAT」の比較」『第二言語としての日本語の習得研究』*18*, 53-69.