文化　第 81 巻 第 1・2 号 ―春・夏― 別刷
平成 29 年 9 月 25 日発行

KY コーパスを使用した計量的分析法の現状と課題

森　　秀　明

KY コーパスを使用した計量的分析法の現状と課題

森　　秀明

1．研究の目的

KY コーパスは日本語教育学で最も使用されてきた学習者コーパスである（李，2009:60）。Oral Proficiency Interview（以下 OPI）のデータを利用して鎌田修・山内博之の両氏によって構築され、1999 年に一般公開された。2008 年には李在鎬氏等によって形態素解析と誤用タグが施されたデータが公開され、2013 年に検索システムを備えた「タグ付き KY コーパス」として Web 公開されている（http://jhlee.sakura.ne.jp/kyc/）。これまで KY コーパスを使用してどのような計量的研究が行われてきたかを振り返れば、日本語教育学における学習者コーパスを使用した計量的分析法のおおよその現状を知ることができる。

学習者コーパスはデータが集めにくく、集めたデータも学習者の言い直しや誤用を多く含むため、単語の認定や品詞分解が困難である。データを集める場合も母集団を定めてそこからサンプルデータを無作為抽出できるわけではない。データは限られた学習者に対し、限られたタスクやインタビューを行って採取するのが現実的に可能な方法で、学習者のレベルや採取場所、タスクの課題による偏りも存在する。このため、学習者コーパス研究は有望な研究分野ではあっても、データ分析や解釈手法については制約や課題が残されていると指摘されている（石川，2012:218）。特に KY コーパスはレベル別の学習者の人数や発話量（語数）が異なり、計量的な分析が難しいコーパスである。

学習者コーパス研究の嚆矢とも言えるグレンジャー（編）（2008:18）では、言語学者が統計的・計量的な分析を行う際の問題に触れて、「言語学者と数値の組合わせは危険な混合火薬のようなもので，なんら意味をなさない場合でも「統計的に有意である」と言ってはばからないことが多い」と述べられている。多くの労力を費やして構築されたコーパスのデータであっても、考察を裏付ける分析方法に問題があっては十分に研究の目的を果たすことはできない。KY コーパスは計量分析が難しいコーパスでありながら、どのような方法を使

用すれば妥当な分析ができるかといった方法論上の議論がこれまで積み重ねられてこなかった。

そこで本稿では計量的な分析法に焦点を当てて KY コーパスを使用した研究を概観するとともに、代表的な研究のいくつかについて簡単な追試を行い、KY コーパスを使用した計量的な分析法の現状と課題を明らかにする。

2．KY コーパスの概要と分析上の問題点

KY コーパスの内容やコーパスの構築に OPI データを使用していることによる分析上の注意点については、鎌田（1999，2006）、山内（1999）に詳しく述べられている。本節ではこれらを参照して KY コーパスの概要と質的な問題点を確認するとともに、稿者が「タグ付き KY コーパス」を使用して取得した品詞数に基づいて算出した学習者レベルごとの語数平均を示し、量的な分析上の問題を検討する。

2．1．KY コーパスの概要と質的な問題点

KY コーパスは『第 2 言語としての日本語の習得に関する総合研究』（平成 8 年度～ 10 年度科研費基盤研究 (A)，代表研究者カッケンブッシュ寛子）の研究を遂行するために収集した日本語学習者の発話データである。それまで日本語の第 2 言語習得研究では「日本語学習者の自然で，かつ，標準化された能力判定がつき，また，横断研究を行うのに十分な量のデータ」が不足していた（鎌田，2006:43）。そこで、それらの条件を満たすデータとして作成者を含めて 25 名のテスターから OPI データの提供を受け、KY コーパスが構築された。

KY コーパスは OPI データに基づいて構築されているため、使用に当たっては OPI に関する理解が必要になる。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判定する評価法である。学習者のレベルは KY コーパス作成当時の基準で初級（下・中・上）、中級（下・中・上）、上級（上級・上級上）、超級の 9 段階になっており、これらの総合判定規定は以下の通りである。

初級（Novice）：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

中級（Intermediate）：自分なりに言葉が使え，なじみ深い話題について簡

単な質問をしたり，答えたりでき，また簡単な状況や，やり取りに対処できる

上級（Advanced）：すべての時間的枠組で叙述，描写ができ，かつ，複雑な状況が処理できる

超級（Superior）：意見の裏付け，仮説構築，具体的・抽象的話題について議論ができ，そして言語的に不慣れな状況が処理できる（鎌田，2006:47 より引用。ただし、稿者の判断で初級～超級の提示順を逆にした。）

テスターは学習者がこれらのレベルのどれに該当するのかを判定するため、学習者が発話し続けるのが困難なレベルにまで会話を誘導する「突き上げ」を行って学習者の言語的挫折を引き出す。またテスト中に学習者が中級以上の能力を持つと判断された場合には、ロールプレーが行われる。このためインタビューはできるだけ自然な会話を目指しながらも、通常の雑談とは異なった内容を含むことになる。このため分析に当たっては以下のような点などが問題になると指摘されている（鎌田，1999:234-236）。

(a) OPI は学習者能力の上限と下限を判定するため、データに言語的挫折が生じている部分や問題なく話し続けられる部分の両方を含むが、上限探し下限探しは学習者やテスターによってまちまちであるため、誤用・正用の割合を学習者間で比較することはあまり意味を有さない。

(b) 各レベルではそれに必要とされる種類のタスクが決まっているため、例えば上級では「記述・叙述・意見」を求めるタスクに必要な「～と思う」などの形式は頻出するが、伝達文の「～そうだ」などの形式は自然発生的にしか出てこない。

(c) それぞれの OPI はそれぞれ独自の話題で展開していくため、話題に左右される語彙の使用頻度はそれぞれのデータによって大きく異なり、一般化が困難である。

これらの注意点に従えば、KY コーパスを使用したほとんどの分析が困難に

なると思われるが、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）のような均衡コーパスでさえデータは様々な偏りを持っているのが当然で、コーパス言語学とはそのような雑多で揺らぎを持ったデータから何らかの意味を持った情報を引き出す学問分野だと言える。上記の問題点は作成者自らが KY コーパスに寄せられる可能性がある批判を先取りして厳し目に記述していると思われるが、KY コーパスから得られた頻度をそのまま鵜呑みにすることなく、何らかの偏りがある可能性を常に考えながら分析していく必要があることだけは確かであろう。

なお、鎌田（1999, 2006）、山内（1999）では触れられていないが、KY コーパスの質的な問題点として日本語母語話者のデータが組み込まれていないことがあげられる。学習者コーパスの研究はグレンジャー等によって提唱された中間言語対照分析（contrastive inter-language 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。KY コーパスを使用して CIA を行うためには、別途他の母語話者のコーパスを使用する必要がある。

2.2. KY コーパスの量的な問題点

次に KY コーパスの量的な問題点を考える。1999 年公開の KY コーパス Ver.1.1 は、形態素解析を行っていないプレーンテキストのデータであるため、当時はデータ量がどれぐらいあるか確定できていなかった。鎌田（2006:43）では OPI のインタビューにかかる時間の大半が 20〜30 分で、総数 90 名分であるところから、中間の 25 分×90 名で概算し、「KY コーパスは総時間 2250 分ほどの音声データを文字化したものといえる。」と述べられている。

KY コーパスの形態素解析は複数の研究者によって試みられているが、2008 年に李在鎬氏等によって形態素解析された研究が最も精度が高いと考えられる（山内, 2015:50）。李在鎬氏等の研究に基づいて報告されたデータ量は、李ほか（2008）では 173,198 形態素、李（2009）では 232,605 語と記されている。稿者が「タグ付き KY コーパス」を使用して学習者の全品詞をダウンロードして集計した形態素数は 170,454 であった。李在鎬氏によればこのような違いが起きる原因はその都度修正しながら作業しているためで、現時点でダウンロードできる形態素数を総語数と考えて良いとのことである（2017.07.15 のメールによる）。このため稿者がダウンロードした語数を現時点の確定した語数であ

ると考えることにする。なお、形態素数と語数は厳密には異なるが、本稿では簡単のため語数と呼ぶ。

表1は稿者がダウンロードした品詞数のうち、記号を除いた語数を使用して算出した形態素数の統計量である。記号を除く理由は、記号が「うん、海、海、＊＊＊＊＊海、〈うん〉スイミング／泳ぐ、泳ぐ」（learner ID：KNM01, 初級，韓国語）などの発話の「＊」が主なもので、この発話で聴取不能部分を表わす「＊」が4つ記されているからと言って、これを4語と認定するのはふさわしくないと判断したためである。

表1を見るとKYコーパスではレベル別の人数や語数にかなりのばらつきがあることが分かる。レベルの大分類では初級5、中級10、上級10、超級5のようにある程度人数がそろえられているが、下位分類でははばらつきがある。また学習者の語数も大きく異なっている。BrownコーパスやBCCWJ固定長のような均衡コーパスでは、データの語数や文字数が一定にそろえられているため、統計的な分析に適している。学習者コーパスはBCCWJで言えば可変長に当たり、データの長さが様々に異なっているため統計分析が難しい。例えばAという学習者とBという学習者で何らかの頻度が2倍違っていたとしても、元々のデータ量がそもそも2倍違っているのなら、一定量当たりの頻度は同じだと考えられる。このためコーパスから得られた生の頻度（粗頻度）で比較するのは困難である。

表1　KYコーパスの母語とレベル別の人数・語数平均・標準偏差

	英語			韓国語			中国語		
レベル	人数	語数平均	標準偏差	人数	語数平均	標準偏差	人数	語数平均	標準偏差
初級下	1	108	-	2	189	93	1	207	-
初級中	2	763	615	1	513	-	2	252	104
初級上	2	656	465	2	622	181	2	705	121
中級下	4	1,122	326	2	1,183	477	3	1,495	665
中級中	4	1,509	556	6	1,477	297	4	1,490	333
中級上	2	1,908	630	2	1,890	700	3	2,136	421
上級	3	1,755	414	6	1,921	303	3	2,493	435
上級上	7	2,808	610	4	2,671	592	7	2,697	584
超級	5	2,757	808	5	3,257	812	5	2,584	370

先行研究では異なる語数のコーパス同士を比較する場合は調整頻度を使用することが推奨されている（石川・前田・山崎，2010:28；石川，2012:114-115；

マケナリー＆ハーディー，2014:74-76 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，粗頻度をコーパスの語数で割ることでコーパスの大きさの影響をなくし，値が小さくなりすぎて扱いにくくならないように一定数をかけた頻度のことである。しかし、調整頻度はサンプルの語数が一定でなければ正確には調整できない。それぞれの語数が異なる学習者コーパスで頻度を調整する場合は、各学習者ごとに調整頻度を算出する必要がある（森，2017a）。このため正確な調整頻度を算出するためには、あらかじめ正確な形態素解析を行って各学習者の語数を確定させておく必要がある。現在は「タグ付き KY コーパス」が公開されているため、時間さえかけて集計すれば誰でも各学習者の語数を得ることができるが、李（2008）のデータ公開以前の研究では自ら形態素解析を行い、膨大な解析ミスを人手で修正する必要があったため、正確な調整頻度の算出は困難であった。KY コーパスにおける量的な問題点のうち最も重要な問題は、粗頻度ではレベルごとの比較が困難で、しかもそれを標準化する調整頻度も簡単には算出しにくい点である。

量的な問題では、初級の学習者数が少なく、語数が非常に少ない点にも注意が必要である。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」（鎌田，2006:47）というレベルであるから、発話量（語数）がごく少ない。OPI の場合、レベル判定が確定すればそこでインタビューは終わり、制限時間の 30 分間まで会話をし続けることはない。このため初級で未出現であった学習項目は、学習者が未習得であるから出現していないのか、データ量が少ないために単に出現していないのかの判定が特に難しい。グレンジャー（編）（2008:12）では最適なコーパスサイズは研究対象を何にするかで決まり、高頻度の語彙や構造の研究ならば、2 万語ぐらいのサンプルで十分だと述べられている。この基準に照らすと初級は 7,199 語しかないため、高頻度の学習項目であっても高い精度は期待しにくいと考えられる。一方、中級は 46,196 語、上級は 73,482 語、超級は 42,990 語であるため、高頻度の学習項目の研究では問題がないと思われるが、出現数がそれほど多くない項目については精度が低くなる可能性に留意する必要がある。

3．KY コーパスを使用した計量的分析の分類

KY コーパスを使用した計量的研究の現状を概観するにあたり、国立情報学研究所の「NII 論文情報ナビゲータ CiNii」（http://cini.ac.jp/）と国立国語研究

所の「日本語研究／日本語教育文献データベース」（https://bibdb.ninjal.ac.jp/bunken/data/）で「KY コーパス」をキーワードに検索したところ、それぞれ 32 件と 25 件の論文情報が得られた。これらを中心に『日本語教育』、『第二言語としての日本語の習得研究』、『日本語／日本語教育研究』等に掲載された KY コーパスを使用した論文やそれらの参考文献等を参照し、用例分析を中心とした論文やごく短い成果報告の論文を除いて、書籍所収論文を含む 43 編の論文を取得した。

KY コーパスを使用した研究ではレベル別の頻度合計や平均を比較する分析が多いが、レベル別の人数や発話量（語数）が異なるため正確な分析が難しい。このため頻度ではなく学習者の習得状況の「分布」を観察する研究やレベル別の「割合」を観察する分析も行われている。また少数ではあるが回帰分析やクラスター分析などの統計分析が使用されている研究もある。

そこで本稿では全 43 編の論文を粗頻度分析（16 編）、分布分析（10 編）、割合分析（12 編）、その他の統計分析（5 編）の 4 種類に分類し、代表的な論文を紹介するとともに、その分類で特徴的な計量分析の問題点にはどのようなものがあるかを検討する。なお、取り上げた論文は「［番号］書誌情報」のように本文に表記し、参考文献には記載しない。

4．粗頻度分析

粗頻度とはコーパスから取得した生の頻度のことで、何の調整も加えていない頻度を指す。先に述べたように KY コーパスを使用した研究ではこの粗頻度分析の割合が最も多く、論文中のおよそ 4 割となっている。

［1］渋谷勝己（2001）「学習者の母語の影響」『日本語学習者の文法習得』大修館書店，pp.83-100.

［1］は、学習者の母語転移について論じた研究で、KY コーパスを使用して学習者の母語別に、「（ラ）レル」「可能動詞」「デキル」などの可能形式の頻度が調査されている。表 2 の「渋谷 2001」はこのうち「〜スルコトガデキル」という頻度を調査した値を引用したものである。［1］は表 2 の韓国語・中級の頻度が 15 になっていることに着目し、韓国語母語話者が他の母語話者より「スルコトガデキル」を多用するのは韓国語に「する＋こと（＋が）＋できる」に対応

する組み合わせ形式が存在するからではないかと指摘している。

しかし「渋谷 2001」では学習者のレベル別頻度の絶対数が最大で 15 とかなり少ない。また使用されている頻度は粗頻度であるため、このままでは正確な比較が困難である。そのため「タグ付き KY コーパス」を使用して調整頻度を算出する追試と、大規模な学習者コーパスを使用した追試を行った。

「追試粗頻度」は稿者が「タグ付き KY コーパス」を使用して形態素「デキル」で検索し、目視で「スルコトガデキル」の用例に絞り込んだ頻度である。「追試調整頻度」はこの「追試粗頻度」を使用して学習者ごとに学習者平均語数である 2,300 語当たりの調整頻度を計算し、10 人ごとの頻度に標準化した値である。なお「渋谷 2001」と「追試粗頻度」との粗頻度が食い違う理由は不明である。

表 2 の「追試調整頻度」の英語・中級を見ると、「追試粗頻度」では 2 であった頻度が 6.5 になり、頻度を標準化すると中級の英語母語話者でも一定量使用されることが分かる。

表 2 「～スルコトガデキル」の頻度比較：渋谷（2001）とその追試

		渋谷 2001	追試 粗頻度	追試調 整頻度
中国語	初級			
	中級			
	上級	2	1	1.0
	超級	1	1	1.6
韓国語	初級			
	中級	15	12	17.8
	上級	12	10	10.7
	超級		1	1.1
英語	初級			
	中級	3	2	6.5
	上級	5	5	4.7
	超級	1	1	1.5

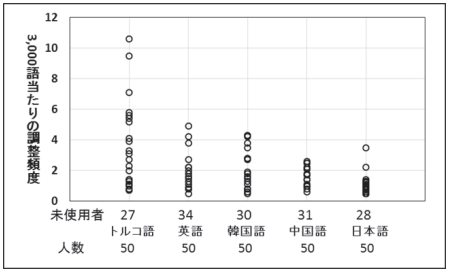

図 1 「スルコトガデキル」の散布図：I-JAS

図 1 の散布図は国立国語研究所によって構築中の大規模な学習者コーパスである 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）の二次データを使用して追試を行った結果である。I-JAS は 12 言語の母語話者 1,000 名に対し綿密に設計されたタスクを行ってデータが集積されており，精度の高さが期待されている。追試ではトルコ語、英語、韓国語、中国語、日本語の母語話者各 50 人のデータを使用した。図 1 の横軸は母語の別、縦軸は 3,000 語当たりの調整頻度である。横軸上部の「未使用者数」は 50 人のうち「スルコトガデキル」を 1 回も使用していない

学習者の人数を示した。I-JAS は日本語母語話者のデータも用意されており、図中の「日本語」はこの値である。

これを見ると、韓国語と英語母語話者の使用傾向はほぼ同じで、「スルコトガデキル」を多用する 3，4 名を除けば全体的な分布は中国語や日本語母語話者と大差ない。表 2 の「渋谷 2001」で韓国語母語話者だけが突出して多く見えたのは、小規模なコーパスで低頻度の学習項目を調査したことによる精度の低さと、粗頻度を使用したことによる英語学習者頻度の過小評価が原因だと思われる。データ量が少ない KY コーパスを使用する場合、ある程度頻度が出現する対象でなければ、結果を一般化することは難しく、粗頻度を調整頻度で標準化しないと、レベル別の比較を行うのは困難だと考えられる。

［2］山内博之（2003）「OPI データの形態素解析―判定基準の客観化・簡易化に向けて―」『実践女子大学文学部紀要』45, pp.1-10.

［2］は KY コーパスに初めて形態素解析を施した研究で、以後多くの研究に大きな影響を与えた。研究の目的は示準化石のように、ある化石（形態素）が見つかれば地層（学習者のレベル）が分かるような、学習者のレベルを特徴づける形態素を探すことである。この研究は山内（2009）で再び取り上げられ、文法の難易度を考えること、言語活動から文法を眺めること、理解のための文法と使用のための文法を区別することの実証面を支える分析として紹介されている。また、山内（2015）ではこれらを発展させ、文法シラバスの構築に利用されている。

［2］では KY コーパス全てのデータが使用されているわけではなく、レベル別の差を明確にするという理由で中級では上・中・下の中のみ、上級は（普通の）上級と上級上の 2 つのうち、上級のみが選択されている。しかし、レベルが連続したデータから任意のデータだけを抜き出しても正確な分析はできないと考えられる。そこで本稿では全データを使用して追試を行った。

表 3 の「コウ」、「ケレドモ」、「（ッ）テイウ」は［2］で超級を特徴づける文法形態素と認められた形態素である。発話の中にこれらが出現すれば、ほぼ超級レベルと考えられるという。表 3 では各文法形態素の 1 行目が［2］p.8 に掲載されている表 7 からの引用で、－は、［2］では調査の対象になっていないレベルである。2 行目の「タグ付き KY コーパス」は、稿者が調査した粗頻度の値、3

行目はそれに基づいて算出した3,000語当たりの調整頻度を、全てのレベルの人数の15人でそろえた頻度である。

表3　超級レベルの判定に寄与する形態素

		中級下	中級中	中級上	上級	上級上	超級
こう	山内（2003）	—	2	—	1	—	173
	タグ付きKYコーパス	1	0	1	0	94	168
	調整頻度	4	0	4	0	76	168
けれども	山内（2003）	—	1	—	8	—	104
	タグ付きKYコーパス	0	1	0	7	81	96
	調整頻度	0	4	0	17	59	97
（っ）ていう	山内（2003）	—	2	—	10	—	164
	タグ付きKYコーパス	0	1	13	7	135	154
	調整頻度	0	2	36	14	110	159

表3では[2]では調査されていない上級上で一定数の出現があるため、これらの文法形態素が「超級を特徴づける」とは考えにくい。上級上はほぼ超級と見なせるのかもしれないが、OPIの基準で明確に区分されている以上、上級上を超級と見なすことには問題がある。KYコーパスが多くの研究に利用されているのもOPIの判定が支持されているからであり、この特色を生かすためにも全てのデータを使用して分析する方法が望ましいと考えられる。

[3] 堀恵子（2012)「第二言語としての日本語習得過程研究における学習者コーパスの制約：KYコーパスとインタビューコーパスとの比較から」『東洋大学人間科学総合研究所紀要』14，pp.95-118.

[3]は、条件表現のタラとバを分析対象としてKYコーパスと堀氏が作成した堀コーパスにおけるレベル別出現傾向を比較した研究である。分析方法には粗頻度の平均が使用されている。分析の結果、条件表現の使用は、堀コーパスの方が初出時が早く、全体の使用頻度においても高かったと報告されている。[3]はこの分析結果に基づいてKYコーパスの分析から学習者の習得順序を一般化するには問題があると指摘している。

[3]は2つのコーパス間で相違点が生じた原因として、「（1）コーパスサイズが小さいこと、（2）KYコーパスが口頭能力試験であることから来る学習者の不安と、言語使用の回避が見られる可能性、（3）口頭能力試験であるため

の話題の偏りの影響、の3点が考えられる。」（p.95）と述べている。コーパスはどのようなコーパスであれ、データ採取方法の違いによって何らかの偏りを持っている。このため、異なったコーパス同士の分析結果を比較し、それぞれのコーパスの性質を明らかにする［3］のような研究は非常に重要である。

［3］では、堀コーパスとKYコーパスにおいて特に中級下の条件表現の初出や頻度が異なることが指摘されている。この理由として［3］では先に引用した3つの要因が上げられているが、特に「（1）コーパスサイズが小さいこと」の意味についてはさらに検討が必要だと思われる。

［3］が指摘しているのはKYコーパスが390,907文字の小規模コーパスであるという全般的な指摘だが、本質はもっと別のところにあるのではないだろうか。それは堀コーパスがどのレベルの学習者についても30分のインタビューを行っているのに対し、KYコーパスのもとになったOPIでは、インタビューは「最長」30分であり、それ以前にレベル判定が終わればそこでインタビューは終わるという点である。第2.2節の表1では、KYコーパスの中級下の語数平均は上級者の5,6割に過ぎなかった。鎌田（2006）でもインタビューの時間は大半のものが20～30分とされていた。このためKYコーパスの中級下であれば、20分程度でインタビューが終わってしまった可能性がある。発話時間が長くなればそれだけ異なった単語を発話する機会が増えると考えられるため、KYコーパスの中級下では出現しなかった条件表現が、発話時間の長い堀コーパスで出現することは十分に考えられる。堀コーパスで単語の初出レベルが早くなるのはこのためだと思われる。また堀コーパスの方が初級や中級の頻度も当然多くなる。二つのコーパスの差は、互いの頻度を調整頻度に直さないかぎり正確には比較できない。堀コーパスは公開されていないため、本稿で追試を行うことはできないが、両者の頻度の比較に粗頻度に基づく平均を使用したため、見かけ上差があるように見えている可能性に留意する必要がある。

［4］鷲見幸美（2014）「中国語を母語とする日本語学習者の和語動詞の使用―KYコーパスの分析―」『言語文化論集』36-1，名古屋大学大学院国際言語文化研究科，pp. 65-79.

ここまでは粗頻度ではなく調整頻度を使用する重要性について述べてきた。しかし［4］は、調査項目の全てが調整頻度で標準化できるわけではない可能性

について考えさせられる研究である。[4]ではKYコーパスのレベルによって「ある、なる、する」などの和語動詞がどれぐらい使用されているかの異なり語数が調査され、レベルごとの平均が計算されている（表4）。この結果、「注目したいのは、中級で使用動詞のバリエーションが急激に増加している点」（p.69）だと述べられている。

しかし、表4で調整頻度に基づいた平均を出すと、どのレベルの平均もほぼ同じ値になる。図2は表4の2種類の平均を図示したものである。これを見れば調整頻度を使用した場合、和語動詞の異なり語数はほぼ一定になり、粗頻度を使用した平均とは全く異なる結論となる。

表4　使用和語動詞の異なり語数
総数と粗頻度平均は[5]より引用

和語動詞	初級	中級	上級	超級
総数	35.0	167.0	253.0	179.0
粗頻度平均	7.0	16.7	25.3	35.8
調整頻度平均	29.2	21.7	20.7	25.0

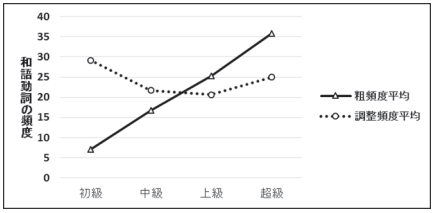

図2　粗頻度平均と調整頻度平均の比較

しかし、よく考えてみると初級では習得している語彙に限りがあるため、発話量がどれほど増えても異なり語数はそれほど増えないことが予想される。調整頻度は調査対象の頻度が発話量に比例して出現することを前提としている。延べ語数で調査した場合、この考え方でほぼ問題はないと思われるが、異なり語数の場合、必ずしも発話量に正比例して頻度が増えるとは限らない。先に述べたようにKYコーパスより初・中級者のデータ量が多い堀コーパスの方が同じ単語でも、初出の学習者レベルは低くなる。このためレベルが低目の学習者でも、発話量が多くなると異なり語数がある程度は増加すると考えられる。だからと言って図2の調整頻度のように単純に正比例で標準化してよいかどうかは難しい。この解明は、今後の重要な課題の一つである。

どれぐらいの発話量があれば学習者の言語能力を反映したデータとなるのか、何人ぐらいの学習者をそろえれば一般化できるデータが得られるのかは、まだほとんど分かっていない。この解明を直ちに行うことは難しいが、[3]のように異なるコーパス間で比較をすること、それも規模の大きなコーパスとの比較を行うことで、少しずつでも解明を進めていくことが重要だと思われる。

５．分布分析

前節で確認したように、粗頻度を使用してレベルごとの合計頻度や平均を比較する分析法には問題がある。これを回避する方法として学習者の習得状況の分布をそのまま観察する方法がある。

［5］田中真理（1999）「OPI における日本語ヴォイスの習得状況：英語・韓国語・中国語の場合」『第 2 言語としての日本語の習得に関する総合研究』（科学研究費補助金研究成果報告書 課題番号 08308019），pp.335-350.

［6］許夏珮（2000）「自然発話における日本語学習者による「テイル」の習得研究―OPI データの分析結果から―」『日本語教育』104，pp.20-29.

［5］は可能、受益、直接受身（有生物主語、無生物主語）、間接受身、使役、させていただく etc.、使役受身、られる（自発）、られる（尊敬）についての正用数と誤用および非用数をすべての学習者について一覧表にまとめ、出現の多いものから連続するようにソートして示した研究である。この表で学習者の使用の分布を確認することにより、どのようなレベルからどんな機能を持ったヴォイスが使用されるようになるのかが明瞭に観察できる。

［6］は、「テイル」の機能を運動の持続（＋長期，－長期）、性状（＋可変性，－可変性）、繰返し、結果の状態、状態の変化、経歴・経験に区分し、［5］と同様の手法で分析した研究である。［5］と異なる点は［5］が頻度を記載しているのに対し、［6］は頻度を使用せずに正用が現れた学習者のみにマーカーをつけて図示している点である。

［5］［6］の論文中には分析法が明示されていないが、これらはいわゆる Implicational Scaling の手法に習った分析法だと思われる。Implicational Scaling とは分析対象の頻度に関わらず、対象が 1 回でも正しく発話された時点を学習者の習得と見なし、学習者の一覧を習得順にソートすることでどの段階で習得が起こったかを階段状に示す手法である。論文中に Implicational Scaling を使用したと明示している研究には次の［7］［8］がある。

［7］スニーラット・ニャンジャローンスック（2001）「OPI データにおける「条件表現」の習得研究―中国語，韓国語，英語母語話者の自然発話から―」『日本語教育』111，pp.26-35.

［8］森山新（2006）「JSL（第二言語としての日本語）における格助詞デの習得過程に関する認知言語学的考察」『日本認知言語学会論文集』6, pp.183-186.

［7］は条件表現を１.「仮説」２.「予定」３.「確定」４.「一般」５.「反復・習慣（－過去）」６.「反復・習慣（＋過去）」７.「反事実（－過去）」８.「反事実（＋過去）」の８項目に分け、条件表現の適切な使用の度数１以上を「1」として分析されている。表５は［7］の表１（p.29）を引用したものである。

表５　中国語母語話者における条件表現の習得順序：［7］p.29 表１より引用

	8	6	7	4	2	5	3	1
超級	0	1	1	1	0	1	1	1
上級 H	1	1	1	1	1	1	1	1
上級	0	0	1	1	1	1	1	1
中級 H	0	0	0	1	1	1	1	1
中級 M	0	0	0	0	1	1	1	1
中級 L	0	0	0	0	0	1	1	1
初級 H	0	0	0	0	0	0	0	0
初級 M	0	0	0	0	0	0	0	0
初級 L	0	0	0	0	0	0	0	0

表５の上段に示されている横軸右側の１は「仮説」で、中級Ｌ（下）の学習者から現れている。初級はすべてが０でこれは初級では条件表現が出現しないことを意味している。中級Ｌでは１.「仮説」のほか、３.「確定」、５.「反復・習慣（－過去）」が出現している。ここからレベルが上がるにつれ階段状に使用される条件表現の項目が増えていき、全ての項目が使用されているのは上級Ｈ以上である。超級は２.「予定」、８.「反事実（＋過去）」が出現していないが、もっと下のレベルで出現するため、これらはデータ不足による未出現であると考えられる。

［8］は格助詞のデの意味用法を Implicational Scaling を使用して分析し、その習得順序が「場所→道具→様態→原因→時間」の順になることを主張した論文である。これらの分析では、そのままでは比較しにくい頻度の大小に着目するのではなく、初めて正用が現れた学習者に着目して習得レベルの全体像を示すため、数値要約による誤解を招くこともなく分かりやすい。この分析法はKY コーパスが使用され始めた初期に多く、近年はそれほど使用されなくなっ

ているが、もっと着目されてよい分析法だと思われる。

ただし、1回でも正用が現れた時点を習得と見なしてよいかどうかは難しい問題である。例えば一人の学習者に正用が1回、誤用が5回ある場合、これを習得と見なすかどうかは判断が分かれると思われる。このため誤用が多く出現するような学習項目の場合、正用率や誤用率を指標にする研究も多い。

また習得という観点とは少し異なるが、ある学習項目を1回しか使用していない学習者より10回使用している学習者の方が、その項目の使い方に習熟していると考えることもできる。頻度情報を使用すれば習得した項目を使いこなせるようになっていく過程も観察することができるため、頻度情報をうまく生かしながら学習者全体の習得状況が把握できるような分布を示す工夫が望まれる。

6．割合分析

割合分析は分布分析と同じく、レベル別の学習者の人数や発話量の違いに影響されない分析法である。割合を使用した分析には次のような研究がある。

［9］山内博之（1999）「初級及び中級レベルにおける「文」の習得について」『第2言語としての日本語の習得に関する総合研究』（科学研究費補助金研究成果報告書 課題番号 08308019），pp.389-397.
［10］大関浩美（2008）『第一・第二言語における日本語名詞修飾節の習得過程』くろしお出版.
［11］東会娟（2006）「会話コーパスに見る中国人日本語学習者の縮約形の使用状況」『言葉と文化』 7，pp.51-66.

［9］は、初級上の学習者7名と中級下の学習者9名の発話を観察し、何が初級と中級を分ける要因となったのかを明らかにした研究である。表6はテスターの質問に対し学習者が応答した文がどのような質で発話されたかを、×：言語的挫折、△：単語レベルの応答、○：必要最小限の文、◎：必要最小限にとどまらない文の4段階で判定し、その割合を求めたものである。このように学習者のレベル別に頻度を合計し、その割合を比較する方法は割合分析で最も典型的に行われている分析法である。表6の実数は頻度、（ ）内の数値は％である。頻度では初級上と中級下の応答数が異なるため比較できないが、（ ）の割合で見れば比較可能である。表6では中級下で×や△が減って◎が増加してい

る様子が明らかで、第2.1節で紹介したOPIの基準に沿った結果となっている。

表6　被験者別に見た発話のタイプ：[9]p.393表2より引用

	×	△	○	◎
初級－上	26（11）	55（23）	106（45）	49（21）
中級－下	9（3）	61（19）	130（40）	124（38）

しかし、[9]のような方法で割合を算出してよいかどうかは十分に検討する必要がある。例えば[10]では[9]とは異なる方法が使用されている。[10]は学習者の名詞修飾節がどのように発達するかを調査した研究で、KYコーパスを使用した箇所では、修飾部と被修飾名詞の文法関係別の使用割合や名詞修飾節の文頭および中央埋め込みでの使用割合などが分析されている。この割合の分析には、[9]のようにレベル別で合計した頻度の割合ではなく、「平均使用比率」が使用されている。平均使用比率とは「各学習者の使用比率（%）を合計し学習者の人数で割るという方法」（p.191）である。この二つにどのような違いがあるのかを表7、8のダミーデータで確認してみよう。

表7の学習者AとBには誤用はない。しかしCは誤用が3ある。これを合計して割合を求めると、正用率：誤用率は50%：50%になる。表7では3人のうち2人が正用率100%なのに、一人がたくさん間違えたために、50%：50%になるのはおかしい。Cがたとえ10回間違おうが100回間違おうがCの誤用の頻度はAやBには無関係である。

表7　頻度合計の割合

学習者	正用	誤用
A	1	0
B	1	0
C	1	3
合計	3	3
割合	50%	50%

表8　正用・誤用率の平均

学習者	正用率	誤用率
A	100%	0%
B	100%	0%
C	25%	75%
合計	225%	75%
平均	75%	25%

このような調査で知りたいことは、ある学習項目を学んだ時に、それをきちんと習得している人がどれぐらいいるかや、誤用を犯すとしたらどれぐらいの割合で犯すのかといった割合である。つまり観測単位は学習者なのであって、学習項目の頻度ではない。そのため表8のように学習者一人一人について割合を求め、それの平均を求める必要がある。誤用を犯す学習者がどれぐらいいる

かだけの問題であれば、誤用率ではなく学習者数で割合を出せばよい。これは前節で議論した習得を0か1かで見る方法である。しかし係助詞「ハ」の誤用など，正用率100%の学習者はごく少数しかいない場合、そのような方法をとっても有益な分析はできない。この場合，誤用率の平均を用いる方が有意義であろう。

なお、割合分析の場合は表7のように頻度が少ない場合でも100%のような大きな値になるため頻度を併記することが望ましい。[10]では分析結果の表に割合しか表示されていないため、表の解釈に誤解が生じる恐れがある。例えば[10]p.223の表6-23では韓国語や中国語母語話者の初級の名詞修飾では文頭に使用される率が100%、中央埋め込みで使用される率が0%という値が示されている。しかし、これらの用例は2例ずつしかなく、この頻度を示さないと誤解を生みやすい。

その点、[11]は「テイル→テル」、「ナケレバ→ナキャ」のような縮約形の使用状況の調査に割合を使用している研究であるが、原形＋縮約形の合計頻度が20に満たない場合は割合を表示しないという独自のルールを設けている。頻度が添えてあれば割合を表記することに問題はないと思われるが、おそらく頻度が20に満たないようなデータの精度の信頼性は低いため、割合表示を意識的に行わないことにより、読み手に注意を促す意図があると思われる。分析の目的によっては少ない頻度しか得られない場合も多いが、[11]のような分析姿勢は見習いたい。

7．その他の分析

最後にその他の分析を概観する。第5節の分布分析では、学習者90名の一覧表を掲載する以外に、学習者の分布を提示する方法を工夫していく必要性があることを述べた。[12]で使用されている散布図はそのような工夫の一つである。計量的な分析において度数分布表や散布図で全体の分布を示すことはごく基本だが、今回稿者が収集できた論文で散布図が描かれているのは[12]だけであった。

[12] 中俣尚己（2015）「生産性から見た文法シラバス」庵功雄・山内博之（編）『現場に役立つ日本語教育研究1　データに基づく文法シラバス』くろしお出版，pp.109-128.

［12］は、「テイル」や「テアル」のような文法項目が、どれだけ異なった単語に接続して使用されるかという「生産性」に着目した研究である。その中に生産性の違いによって中級学習者が使用する 103 の文項目の頻度に違いが出るどうかを散布図で示した分析がある。この散布図を見ると生産性が高い項目と低い項目が使われやすく、中程度の項目が使われにくいという傾向が視覚的に把握できる。［12］の散布図には 1 万語当たりの調整頻度が使用されており、頻度の処理も参考になる。

［13］松田真希子・宮永愛子・庵功雄（2013）「超級日本語話者の談話特性：テキストマイニングを用いた分析」『国立国語研究所論集』5，pp.43–63.

［13］はいわば［2］の談話版である。［2］は超級を特徴づける示準化石のような形態素を探すことを目的としていた。その結果「コウ」、「ケレドモ」、「（ッ）テイウ」という形態素が発見された。［13］は超級話者の談話の特徴に談話の結束性が関わっていると考え、その結束性を特徴づける分析対象の一つにコソアの指示表現を取り上げている。［2］で既に「コウ」が発見されているため、超級話者ではコ系の指示表現が多用されているという結論に目新しさはないが、［2］が形態素の頻度を観察していたのに対し、［13］は文の頻度を観察しているところが新しい。つまり［13］は指示表現の形態素がいくつ出現するかではなく、各学習者ごとに指示表現を使用した文が何文あるかを観察し、学習者の発話した総文数の中で、コ系、ソ系、ア系の指示表現が使用されている文が何％あるかを算出している。このような観察を可能にするため、テキストマイニングツールの KH-Coder が使用されており、新しい可能性を感じさせる研究である。

ただし、文数を数えた場合、上級の文数に比べて超級の文数がその 5，6 割になっていることには注意が必要である。これは超級では上級が 2 文で話す内容を、1 文程度でつなげて話していることを示唆しており、同程度に指示表現を使用していた場合、文数で計算すれば超級の使用割合の方が当然高くなると思われる。なお、［13］では文が観察単位（ケース）だとしているが、学習者ごとに文数を数え、学習者同士を比較しているのであるから、観察単位（ケース）は学習者であり、文は変量だと考えるのが妥当だと思われる（森，2017b）。

8．まとめと今後の課題

本研究ではKYコーパスを使用した研究を概観し、学習者コーパスを使用して計量的な分析を行う上での課題を検討した。計量分析の方法は粗頻度をそのまま使用する粗頻度分析、分析対象の全体的な分布を俯瞰する分布分析、分析対象の割合を観察する割合分析、その他の分析の4種類に分類できた。

KYコーパスはレベル別の学習者数や発話量（語数）にばらつきがあるため、粗頻度や粗頻度平均ではレベル別の比較が難しく、調整頻度に標準化して比較する必要がある。調整頻度を正確に計算するためには、学習者ごとの語数を調べ、その語数を使用して学習者個々の頻度を調整する必要がある。現在では「タグ付きKYコーパス」が公開されているため、全ての品詞をダウンロードして学習者ごとに集約することで、誰でも学習者ごとの語数を知ることができるようになった。それにも関わらずごく最近でも粗頻度をそのまま使用した研究が見られるのが現状である。

ただし、異なり語数で分析する場合などは「学習者の頻度÷学習者の語数×一定数」で、本当に標準化できるかどうか分からない。研究の目的ごとに「どうすれば条件が異なる学習者同士を比較できるか」という原点に立ち返り、そのつど分析方法を工夫することが大切だと思われる。また、KYコーパスはデータ量が少ないため、データの一部だけを使用したり、頻度の小さな分析対象を研究する場合は注意が必要である。分析結果を一般化するためには、条件の異なる複数のコーパスで検証することが重要であろう。

分布分析は、現在分析の主流にはなっていないが、学習者の全体像を度数分布表や散布図の形で示すことは、今後多くの研究で取り入れられることが望ましい。散布図の示し方などにもさらなる工夫の余地があるだろう。Implicational Scalingのように頻度を完全に捨象して分析することの是非は考えなければならないが、分析対象が出現した学習者数に着目した分析を併用する研究は有意義だと考えられる。

割合分析では、合計頻度の割合を求めた研究が多く、それでは有意義な分析になりにくい。この場合も調整頻度と同じように学習者別に計算し、その平均を出す方法が適している。結局、観察単位は学習項目ではなく、学習者だということを強く意識することが重要である。割合分析では頻度が少ない場合、データの精度も低く、割合だけの表示では誤解を与える場合が多いため、必ず頻度を併記するようにしたい。

その他の分析では、ロジスティック回帰分析やクラスター分析を使用した研究など、高度な統計分析の手法を取り入れた研究もあったが、本稿では基本的な分析の紹介と検討にとどめた。

本稿で述べた内容は、統計学のごく基礎的な水準の分析法ばかりだが、学習者コーパスを使用した計量的分析では、その水準を満たしていない研究が多いのが現状である。論文中に分析の方法を詳しく説明している研究はごくわずかで、中には何をやっているのかよく理解できない研究もあった。計量的な分析を行うときは、「平均」のように執筆者にとって分かり切った分析法だとしても、どのような目的のためにどのような方法で分析するかについて、もっと詳細な説明を記述するべきである。そのような説明を明示することで、その分析法が本当にふさわしいかどうかを内省することにもつながる。

多くの研究で基本的な方法論に問題があるのは、先行研究の方法を無批判に踏襲して分析していることにも問題があると考えられる。計量分析における方法論的な問題意識を広く活性化していくためには、率直に各自の分析法を検討し合うような場も必要であろう。学習者コーパスを使用した計量的分析においては、先進的な分析法を取り入れる以前に、基礎的な分析法について広く議論したり、論文中に分析方法の説明を詳しく述べるスタイルを定着させるなどして、方法論上の妥当性を検討する意識を高めていくことが課題だと思われる。

本稿の執筆にあたって論文の収集が不十分であるために重要な論文を見逃している可能性がある。また本稿で取り上げた重要論文に対する稿者の考察や指摘も的外れであるかもしれない。これらの不備や失礼は，ここに記してお詫びしたい。

使用データ

本研究は、『タグ付き KY コーパス』（http://jhlee.sakura.ne.jp/kyc/）、ならびに国立国語研究所のプロジェクトによる成果『多言語母語の日本語学習者の横断コーパス：I-JAS』（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。

参考文献

石川慎一郎（2012）『ベーシックコーパス言語学』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』くろしお出版.
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―』研究社.
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，pp.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，pp.42-51.
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，pp.60-72.
李在鎬・淺尾仁彦・濱野寛子・佐野香織ほか（2008）「タグ付き日本語学習者コーパスの開発」『言語処理学会第 14 回年次大会発表論文集』pp.658-661.
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』ひつじ書房.
森秀明（2017a）「学習者コーパスを使用したレベル別頻度比較の方法」『Learner Corpus Studies in Asia and the World Vol.3 Position Papers from LCSAW2017』pp.99-102.
森秀明（2017b）「コーパス間における単語使用率の比較―観察単位（ケース）は単語か文書か―」『計量国語学』31 巻 3 号（掲載予定）.
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，pp.238-245.
山内博之（2009）『プロフィシェンシーから見た日本語教育文法』ひつじ書房.
山内博之（2015）「話し言葉コーパスから見た文法シラバス」庵功雄・山内博之（編）『現場に役立つ日本語教育研究 1　データに基づく文法シラバス』くろしお出版，pp.47-66.

Current Status and Problems with the Quantitative Analysis Method Using KY Corpus

Hideaki MORI

KY Corpus is the learner corpus that has been most often used in Japanese language pedagogy research. This paper outlines the researches using KY corpus and has examined their problems with quantitative analysis. As a result, it has been found that row frequency has been used in many studies when normalized frequency should have been used. Moreover, many studies have been calculated the misusage ratio after summing up all learners misusage frequencies. However, it is more appropriate to calculate the ratio of individual learners first and then obtain the average. In quantitative analysis, there are many studies which have methodological problems in the basic stage of analysis, so the task is to raise the awareness of considering the validity of analysis.