テキストマイニングによる学習者作文における談話能力の測定と評価

大阪府／大阪大学大学院在籍　小林 雄一郎

概要　本研究では，日本人中学生，高校生，大学生の英作文を集めた「学習者コーパス」を解析し，⑴中学生，高校生，大学生と学年が上がるにつれて，談話標識の頻度や使用傾向はどのように変化するのか，⑵学習者（中学生，高校生，大学生）と母語話者の間には，談話標識の頻度や使用傾向にどのような違いがあるのか，という 2 点に光を当てて，量的分析と質的分析を行った。また，談話標識に関して，先行研究ではさまざまな定義や構成要素が提案されているが，本研究では Hyland（2005）の metadiscourse markers の定義とリストに準拠した。

その結果，接続表現（transitions, frame markers），視点（self-mentions），心的態度（hedges, boosters）といった多くの点において，習熟段階の異なる学習者の間，そして学習者と母語話者の間に頻度や用法に関する大きな差異が見られた。

1　はじめに

学習者が実際に使用した言語の特徴や誤用についての知識を持つことは，英語教師にとって重要なことである。第ニ言語習得の分野では，誤用は，1）その学習者が対象言語をどれくらい学んだかについての情報を教師に提供する，2）どのように対象言語が学習されるかについての情報を研究者に提供する，3）学習者が正用と誤用に関する情報をもとに対象言語の規則を発見できる，という 3 つの点で有意義であるとされている（Corder, 1967）。

近年，英語教育の現場では実践的コミュニケーション能力の育成を図ることが求められており，中学校や高校の学習指導要領にも同様の記述が見られる。しかし，円滑で効果的なコミュニケーションをするためには，「何を」伝えるかよりも，「いかに」伝えるかが必要不可欠となる。具体的に効果的なコミュニケーションを達成する 1 つの方法は，談話標識によって，談話のユニット間の論理関係や意味関係を表すことである（Altenberg & Tapper, 1998, p.80）。それらの表現には，談話の結束上の手がかりを与えて（Leech & Svartvik, 1994, p.177），読み手や聞き手がユニット間の一貫性や，命題内容に対する書き手の評価や態度を見つけることを手助けし，テキストの意味理解ができるようにする働きがある。しかしながら，対象言語とは異なる言語的背景や文化的背景を持つ学習者にとって，談話標識の「適切な」使用は非常に難しい。接続語の使用数が増えても結束性の質が上がるわけではなく（Crewe, 1990），対象言語の「規範」から「逸脱」したモダリティの使用がテキストの理解を妨げることもある。したがって，実践的コミュニケーション能力の育成を図る上で，学習者による談話標識の使用傾向を調査し，彼らの談話構造における特徴や誤用を究明することは極めて重要である。しかしながら，これまでの研究では，手作業による談話分析のコストが高いこともあって，限られた数の学習者データしか扱うことができず，そこから得られた結果がどこまで普遍的なものかを検証することが難しかった。さらに，大規模な調査を行う場合は，多くの分析者が必要となり，どうしても結果が個々の分析者による主観に影響されてしまうという欠点があった（e.g., Baker, 2006）。

それに対して，本研究では，日本人中学生，高校生，大学生の英作文を集めた「学習者コーパス」をテキストマイニングの手法（3.3参照）を用いて客観的に解析し，学習者作文における談話標識の使用傾向を統計的に俯瞰（ふかん）する。また，量的分析から得られた結果を手がかりに，質的分析を有機的に組み合わせ，学習者談話に特徴的なパターンを抽出する。

2 研究の背景

2.1 コーパス言語学

「コーパス」（corpus）という語は，元来「身体」（body）を表すラテン語であるが，*Oxford English Dictionary* によれば，「言語分析のための言語資料の集積」を意味する使用例は，1956年と比較的新しいものである。また，現代のコーパス言語学において，「コーパス」とは，「機械可読な」形式で集積された大規模で「真正な」言語データベースのことを指し，特定の言語変種やジャンルに対する「代表性」を持つ点で単なるアーカイブとは異なるものである（Baker, Hardie, & McEnery, 2006, pp.48-49）。そして，コーパス言語学は，Leech（1992）によれば，次のような特徴を持つとされている（p.107）。

- 言語能力よりも，言語運用に中心を置く
- 言語の普遍的特性の解明よりも，個別言語の言語記述に中心を置く
- 質的な言語モデルのみならず，量的な言語モデルにも中心を置く
- 言語研究における合理主義的な立場よりも，より経験主義的な立場に中心を置く

2.2 学習者コーパス

学習者コーパスとは，外国人学習者によって産出された言語データを機械可読な形式で集積したものである（Leech, 1998, p.xiv）。1990年代以降，さまざまな学習者コーパスが構築されてきたが，第二言語習得（SLA）の研究のために言語データを比較検討するという方法論は必ずしも新しいものではない。1950年代後半，対照分析の研究者たちは，量的データこそ使っていなかったものの，目標言語と第一言語の類似度が言語学習の困難度と関係があると予想していた（e.g., Lado, 1957）。1960年代に入ると，誤用分析の研究者たちが，学習者言語の発達過程を解明するために，学習者がどの段階でどのような誤用をするかに注目し，そのパターンの記述を試みた（e.g., Corder, 1967）。1970年代には，運用分析の研究者たちは，誤用だけでなく，正用も含めた「中間言語」（Selinker, 1972）の運用全体を分析対象とするようになった（e.g., Dulay & Burt, 1973）。学習者コーパスとは，このような SLA における歴史の流れの中で生み出されたものにほかならない。

学習者コーパスの先駆的な試みとしては，コペンハーゲン大学で1970年代に行われた PIF Corpus（Færch, Haastrup, & Phillipson, 1984），1970～80年代にドイツの移民を対象に調査した ZISA Project（Clahsen, 1980），ヨーロッパ5か国における移民を対象に調査した ESF Database（Perdue, 1993），幼児の母語習得データベースである CHILDES（MacWhinney, 1995）などがある。そして，コーパス言語学の知見を十分に取り入れた現代の学習者コーパスの代表的なものとして，さまざまな母語を背景とする英語学習者の作文を集めた ICLE（International Corpus of Learner English）（Granger, 1998a）をはじめ，LLC（Longman Learner's Corpus）や CLC（Cambridge Learner's Corpus）といった商用コーパスがある（e.g., Pravec, 2002）。

2.3 談話標識

「談話標識」（discourse markers）とは何か。その定義や構成要素をめぐって，これまで多くの議論が成されてきた。まず，その名称に関しても，discourse markers 以外に，cue phrases，discourse connectives，discourse connectors，discourse operators，discourse particles，discourse signaling devices，phatic connectives，pragmatic connectives，pragmatic expressions，pragmatic formatives，pragmatic markers，pragmatic operators，pragmatic particles，semantic conjuncts，stance connectives など，枚挙にいとまがない（e.g., Fraser, 1999）。そして，談話標識の定義についても，多くの議論が存在する。しかしながら，いずれの研究においても，限られた数の談話標識しか分析対象とされておらず，網羅的なリストが提示されていない。

「メタ談話標識」（metadiscourse markers）も，広義での談話標識の一種である。かつてメタ談話は

「談話についての談話」（Vande Kopple, 1985, p.83）と考えられていたが，最近の研究では「書き言葉，あるいは話し言葉のテキストにおける言語要素で，命題内容に何かを付け加えるものではなく，聞き手や読み手が与えられた情報を系統立て，解釈し，評価することを助けるためのもの」（Crismore, Markkanen, & Steffensen, 1993, p.40）と定義されるようになった。

メタ談話標識の研究において，最もよく使われる枠組みは，おそらく Hyland list（Hyland, 2005）であろう。このリストは，Vande Kopple（1985）や Crismore et al.（1993）による研究をベースとして，10種類のカテゴリー（表１）に分類される約400種類の談話表現を網羅的に収録したものである。また，このリストは，コーパスに基づく統計的研究を想定して作成されたものであり，これまでにアカデミック・ライティングをはじめ，教科書，学位論文，ビジネスレターなど，さまざまな言語データの分析で成果を上げている。

Hyland list は書き言葉の分析を想定したリストである。また，多くの研究で「談話標識」と見なされている接続詞や接続副詞の多くは，Transitions（TRA）や Frame markers（FRM）に含まれているが，Hedges（HED）や Boosters（BOO）のような stance markers（Biber, Johansson, Leech, Conrad, & Finegan, 1999, p.979），Self-mentions（SEM）のような人称代名詞のみならず，書き手の態度や評価を表す動詞なども含まれているため，非常に多角的な分析が可能である。以下, 本研究においては, Hyland（2005）によってリスト化されたメタ談話標識を「談話標識」と呼び，分析の対象とする。

3 分析方法

3.1 リサーチ・クエスチョン

本研究の目的は，日本人中学生，高校生，大学生の英作文を集めた学習者コーパスをテキストマイニングの技法を用いて解析し，そこから得られた結果を母語話者のコーパスと比較することである。その目的を達成するために，本研究では，以下の２つの research questions（RQ）を設定する。

■ 表１：メタ談話標識の意味カテゴリー

Category	*Function*	*Examples*
Interactive resources	***Help to guide reader through the text***	
Transitions（TRA）	Express semantic relation between main clauses	in addition / but / thus / and
Frame markers（FRM）	Refer to discourse acts, sequences, or text stages	finally / to conclude / my purpose here is to
Endophoric markers（END）	Refer to information in other parts of the text	notes above / see Fig / in section 2
Evidentials（EVI）	Refer to source of information from other texts	according to X / (Y, 1990) / Z states
Code glosses（COD）	Help readers grasp functions of ideational material	namely / e.g. / such as / in other words
Interactional resources	***Involve the reader in the argument***	
Hedges（HED）	Without writer's full commitment to proposition	might / perhaps / possible / about
Boosters（BOO）	Emphasize force or writer's certainty in proposition	in fact / definitely / it is clear that
Attitude markers（ATM）	Express writer's attitude to proposition	unfortunately / I agree / surprisingly
Engagement markers（ENG）	Explicitly refer to or build relationship with reader	consider / note that / you can see that
Self-mentions（SEM）	Explicit reference to author(s)	I / we / my / our

- 中学生，高校生，大学生と学年が上がるにつれて，談話標識の頻度や使用傾向はどのように変化するのか
- 学習者（中学生，高校生，大学生）と母語話者の間には，談話標識の頻度や使用傾向にどのような違いがあるのか

これら2つのアプローチに関して，Granger (1998a) の表現を用いるならば，前者は「中間言語の異なる段階の比較」(IL-IL comparison) であり，後者は「母語と中間言語の比較」(NL-IL comparison) である (pp.12-13)。また，本稿では，テキストマイニングを用いて談話標識の全体的な使用傾向を俯瞰したのち，日本人学習者の英作文に特徴的な談話標識に光を当てる。

3.2 分析データ

本研究では，JEFLL Corpus，ICLE-JP，LOCNESS の3種類のコーパス（総語数は661,043語）をデータとして使用する。

JEFLL Corpus は，日本の中学生と高校生による自由英作文を集めた学習者コーパスである（約60万語)。本研究では，ICLE-JP および LOCNESS とのデータの整合性を考慮し，小学館コーパスネットワーク（SCN）で無償公開されているデータのうち，論説文（argumentative essay）のみを分析対象とする。

ICLE-JP は，日本の大学生による論説文を集めた学習者コーパスである（約17万語)。本研究では，著作権者の許諾を得てプレリリース版を分析対象とする。

LOCNESS は，英米の母語話者による英作文を集めたコーパスであり，ICLE の参照コーパス (reference corpus) として設計された（約30万語)。これは，コーパス作成者にコンタクトを取ることで入手可能（有償）である。本研究では，アメリカ人大学生による論説文のみを分析対象とする。

表2は，本研究で使用するデータの概要である。なお，表中の JH，SH，UNI，NS は，それぞれ中学生，高校生，大学生，母語話者を表している。また，作文タスクの詳細については，投野 (2007) および Granger (1998a) を参照されたい。

■表 2：分析データの概要

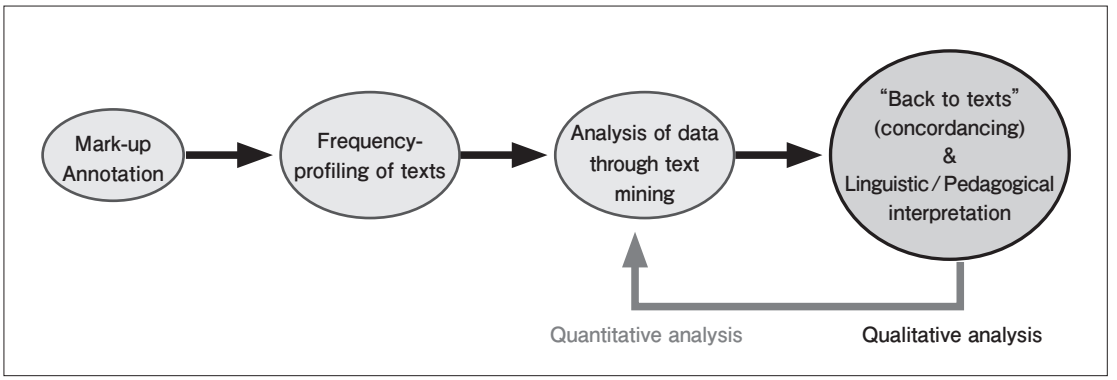
	JH	SH	UNI	NS
Corpus	JEFLL		ICLE-JP	LOCNESS
N	2921	2453	327	176
Tokens	162919	179750	168800	149574

3.3 テキストマイニング

テキストマイニングとは，テキストデータをコンピュータで計量的に解析し，有益な情報を抽出するためのさまざまな手法の総称であり，統計学，データマイニング，人工知能，自然言語処理で開発された技術を背景に持っている。テキストマイニングは，大規模なテキストデータを統一的な視点から少ない労力で客観的に分析することを可能にする (e.g., 松村・三浦, 2009)。現在のコーパス言語学では，インターネット上の膨大な言語データからコーパスを自動生成する技術が導入されつつあり，個人の研究者であっても数億語から数十億語のコーパスを構築することが可能となり，100億を超えるウェブページに含まれる言語データを蓄積している研究グループも存在する。このような大規模データを前にしたとき，手作業でデータを解析することは極めて困難となる。しかしながら，分析に利用されるデータが大きければ大きいほど，データ縮約やテキスト分類といったテキストマイニング技術の必要性は高まり，その精度も安定していく。したがって，テキストマイニングは，言語研究において，分析データの量および分析結果の質を飛躍的に高めるブレイクスルーをもたらす可能性を秘めている。

本研究は，このようなテキストマイニングの技法を用いて，英作文における談話構造を解析するものである。具体的には，コーパス中の文章から品詞情報・構文情報・談話情報などを自動抽出し，それらの情報から得られる頻度パターンに対してさまざまな量的分析を行う。しかしながら，コーパス言語学においては，量的分析と質的分析が常に相補的な関係になければならない。したがって，テキストマイニングによって全体的な傾向が把握された後は，「テキスト」そのものに戻って，コンコーダンスの精緻な読みがなされなければならない。そして，テキストの読みから得られる知見は，新たなる量的分析のための手がかりを与えてくれる。このようにして，量的分析と質的分析は有機的に循環していく（図1)。

▶図 1：分析の手順



4 多変量アプローチ

4.1 頻度集計

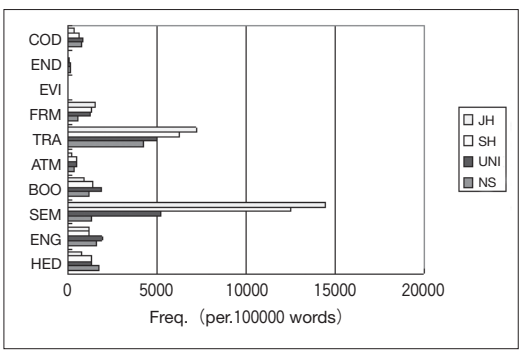
まず，一部の記号（括弧，疑問符，ハイフンなど）を除いて，Hyland list におけるすべての談話標識の頻度を調査したところ，355種類の談話標識が延べ138,988回使用されていた。表 3 は，355種類の談話標識を10種類のカテゴリー（表 1 ）に分類した生頻度を集計したものである。また，図 2 はそれを10万語当たりの相対頻度に変換した後に視覚化したものである。表中の “chi-squared” と “*p*-value” は，それぞれ χ^2 統計量と有意確率を表している。

■表 3：カテゴリー別の頻度集計表（生頻度）

	JH	SH	UNI	NS	chi-squared	*p*-value
COD	568	1133	1475	1175	1480.964	***
END	114	195	334	306	496.623	***
EVI	0	4	44	82	1685.416	***
FRM	2573	2442	2213	853	162.497	***
TRA	11788	11325	8443	6455	741.850	***
ATM	419	985	887	503	419.840	***
BOO	1550	2506	3170	1766	1735.963	***
SEM	23622	22549	8869	2018	12992.040	***
ENG	1985	2261	3242	2470	2422.122	***
HED	1343	2385	2291	2645	3037.304	***

df = 3, *** = $p < .001$

▶図 2：カテゴリー別の頻度集計グラフ（相対頻度）



4.2 相関分析

中学生，高校生，大学生，母語話者の 4 グループは，談話標識の使用パターンに関して，どの程度の類似性を持っているのであろうか。表 4 は，談話標識のカテゴリー別頻度集計表（表 3）に対して，各ケース間における Pearson の積率相関係数を求めた結果である。

■表 4：Pearson の積率相関係数

	JH	SH	UNI	NS
JH	1.000			
SH	0.998	1.000		
UNI	0.910	0.916	1.000	
NS	0.455	0.466	0.760	1.000

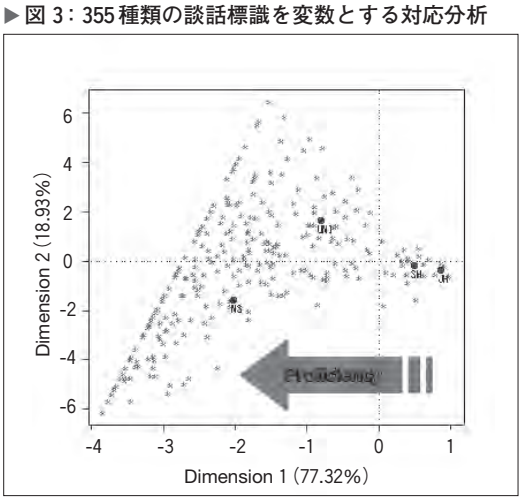
表 4 によれば，学習者（中学生，高校生，大学生）の間では相関係数は0.910～0.998と極めて高く，談話標識の使用傾向という点に限って言えば，これら 3 つのグループはかなり類似した言語特徴を持っている。それに対して，母語話者と学習者の相関係数は0.455～0.760と相対的に低い。

4.3　対応分析

次に，分析データ（中学生，高校生，大学生，母語話者）における談話標識の全体的な使用傾向を俯瞰するために，対応分析を行う。この手法は，大量のデータを分類，整理，縮約することでデータの全体像をつかみ，ケース間の関係，変数間の関係，ケースと変数の関係を多次元空間上に視覚化するものである。数学的には林知己夫の数量化Ⅲ類や西里静彦の双対尺度法と同じ計算方法で，クロス表における行と列の相関（正準相関）を最大化する解析法である。なお，この解析法は，1回の固有値計算や特異値分解によって解が得られるために計算が簡便であり，広く利用されている。さらに，計算過程のオプションがないため，因子分析や主成分分析といった解析法よりも再現性が高い。

図3は，データ中に生起するすべて（355種類）の談話標識を変数として，対応分析を実行した結果である。この図は，解析の結果として得られる得点のうち，最も寄与率の高い第1次元（77.32%）と第2次元（18.93%）を2次元散布図に布置したものである（第2次元までの累積寄与率は96.25%）。図中で近接する項目は類似した性質があることを示し，図中の項目間を隔てる距離が大きいほど異質性が高いことを示す。ただし，解析結果を解釈するにあたっては，各次元が直交している（無相関である）ことに注意しなければならない。なお，視認性を重視し，変数（談話標識）のラベルは非表示としている。

▶図3：355種類の談話標識を変数とする対応分析



この図を見ると，第1次元（横軸）の右から左へ向かって，中学生，高校生，大学生，母語話者の順でケース（サブコーパス）が布置されており，談話標識の使用傾向と英語習熟度が関連性を持っていることがわかる。また，変数の分布を見ると，第1次元の負の帯域（左側）に数が多く，学年が上がって，習熟度が上がっていくにつれて，産出できる談話標識の種類が増えていくこともわかる。

図4は，10種類のカテゴリー別の頻度行列（表3）に対して，対応分析を実行した結果である。この図は，解析の結果として得られる得点のうち，最も寄与率の高い第1次元（93.84%）と第2次元（4.60%）を2次元散布図に布置したものである（第2次元までの累積寄与率は98.44%）。

▶図4：10種類の意味カテゴリーを変数とする対応分析

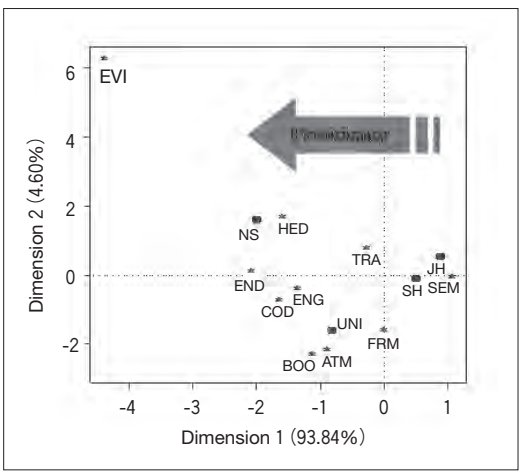


図4においても，第1次元に習熟度が反映されており，談話標識の使用傾向と習熟度が強い関連性を持っていることがわかる。そして，ケースと変数の分布を見ると，中学生や高校生の談話にSEM，FRM，TRAなどが顕著である一方，母語話者の談話にはHEDやENDが顕著であることがわかる。

5　言語学的分析

前節では，多変量アプローチを用いて，日本人学習者と母語話者による談話標識の使用傾向を俯瞰した。本節では，日本人学習者の英作文に特徴的な談話標識に光を当てて，詳細な量的・質的分析を行う。

紙面の都合もあり，低頻度のカテゴリー（COD, END, EVI）や，学習者の作文と母語話者の作文に大きな差が見られないカテゴリー（ATM, ENG）は割愛する。

5.1 Interactive resources

Hyland（2005）によれば，interactive resourcesとは，TRA, FRM, END, EVI, CODの上位カテゴリーであり，命題に関する情報を構造化し，テキストに結束性を与える役割を持っている（p.50）。本稿では，比較的高頻度なTRAとFRMを中心に，多くの先行研究で「談話標識」と見なされている接続詞や接続副詞を分析していく。

5.1.1 Transitions (TRA)

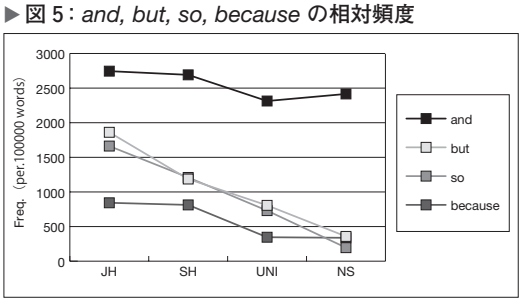
学習者による産出言語の中核を担うのは内容語であるが，機能語も文構成上で重要な働きを持っている。とりわけ，接続詞は，初期段階から英語学習教材などで多く提示され，学習者になじみのある語彙である。同時に，接続詞は，学習者が英作文などで多用する表現でもあり，さまざまな学習者コーパスにおいて高い頻度で現れている（e.g., 小林, 2009a, 2009b, 2009c）。だが，「たくさん使っていること」は必ずしも「正しく使っていること」を意味してはいない。学習者の作文は，それが上級の学習者のものであったとしても，ときに母語話者の目には奇妙に映る。それは，多くの場合，学習者が「適切な種類の接続詞を使用できていない」（McCarthy, 1991, p.50）からであり，接続詞の過剰使用によって「人工的で機械的な文章」（Zamel, 1983, p.27）となっているからである。接続詞や接続副詞のような接続語を適切に使用することは，語彙に関する知識のみならず，構文や意味，そして談話に関する多様な知識が要求されるため（Tankó, 2004, p.159），必ずしも容易ではない。

学習者による接続語の使用を分析した先行研究は多い（e.g., Altenberg & Tapper, 1998; Granger & Tyson, 1996; Tankó, 2004）。これらの先行研究の多くは，母語話者と学習者のデータを統計的に比較し，学習者が有意に過剰使用／過少使用している表現を報告したものである。そこで用いられている方法論は，異なる母語を持つ学習者のデータを比較しているために，第二言語習得における母語の影響などを探るには有効である。その一方，同一の母語を持つ異なる学習段階の学習者データを分析しているものは極めて少なく，あったとしても，扱っているデータ規模は数十人規模と極めて小さい。その限りで，中学生，高校生，大学生による大規模コーパスを縦断的に解析する本研究の意義は小さくない。

5.1.1.1 高頻度接続詞の頻度

図5は，中学生，高校生，大学生，母語話者の4グループにおけるTRAの上位4語（*and, but, so, because*）の相対頻度（10万語当たり）をまとめたものである。これら4語の頻度を合計すると，分析データにおけるTRAの頻度全体の90.03%を占めている。なお，本来ならばヒストグラムで描かれるべきデータ形式であるが，各グループ間の推移をとらえやすいように折れ線グラフで描いている。

▶ 図5：*and, but, so, because* の相対頻度

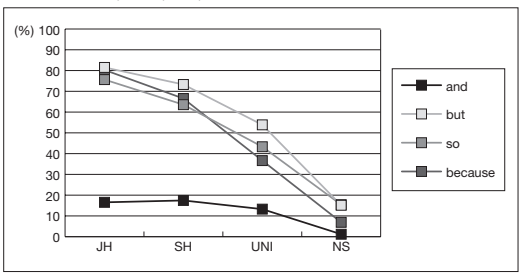


この図を見ると，4語とも習熟度（学年）が上がっていくにつれて，有意に頻度が減少していく（*and*: $\chi^2 = 88.407$, df = 3, $p < .001$; *but*: $\chi^2 = 1818.366$, df = 3, $p < .001$; *so*: $\chi^2 = 1958.101$, df = 3, $p < .001$; *because*: $\chi^2 = 660.092$, df = 3, $p < .001$）。ICLEのフランス人学習者データとLOCNESSを統計的に比較したGranger and Rayson（1998）によっても報告されているように，接続詞の過剰使用は，学習者言語の顕著な特徴の1つである（p.127）。そして，母語話者は，学習者が *and* や *but* ばかりを産出するのに対して，*also* や *however* などの接続副詞を巧みに使い分けることで，より効果的な文章を構築している。

5.1.1.2 高頻度接続詞が文頭で生起する割合

次に，図6は，4グループにおけるTRAの頻度上位4語が文頭で生起する割合をまとめたものである。

▶図 6：*and, but, so, because* が文頭で生起する割合



この図を見ると，4 語とも，習熟度が上がるにつれて文頭で生起する割合が小さくなっていく。それぞれの語に関して，文頭での頻度と文中での頻度の比率には有意差が見られる（*and*: $\chi^2 = 588.916$, df = 3, $p < .001$; *but*: $\chi^2 = 1104.453$, df = 3, $p < .001$; *so*: $\chi^2 = 669.402$, df = 3, $p < .001$; *because*: $\chi^2 = 979.620$, df = 3, $p < .001$）。日本語では「また」,「しかし」,「なぜなら」のように，接続語で文章を始めることが多いため，学習者が文頭の接続詞を好むのは，母語の干渉である。

また，Biber et al.（1999）が示しているように，文頭に生起する *and* や *but* は，話し言葉を特徴づけるものであり（p.84），一般的にフォーマルな書き言葉で使ってはならないとされている。それにもかかわらず，母語話者が *but* の約15.14％を文頭で使っている点は注目に値する。これは，おそらく，語用論的用法，つまり語と語をつなぐ統語論的機能ではなく状態と状態をつなぐ語用論的機能を担う用法（van Dijk, 1977）であるが，この点については今後の詳細な分析が必要である。

そして，*so* と *because* に関しては，単に文頭の頻度を見るだけではなく，構文レベルや談話レベルの分析が求められる。まず，*so* という語には，一般的に，副詞としての用法（e.g., We behaved **so** stupidly.）と従属接続詞としての用法（e.g., It was quite windy, **so** we had to button our coats up.）があるとされている。さらに，Fraser（1993）は，第 3 の用法として，引用(1)のように文頭に現れる *so* を談話標識の *so* と定義し，引用(2)のような従属接続詞の *so* などと区別している（p.6）。

(1) John was sick. **So** don't expect him.
(2) John was sick, **so** he went to bed.

しかしながら，Fraser（1993）が談話標識の *so* と定義した用法は，リーチ・池上・上田・長尾・山田（2006）などにも明記されているように，通常は話し言葉で用いられる用法である（p.1594）。それにもかかわらず，学習者が文頭で *so* を多用することは，彼らの言語使用が話し言葉の影響を受けていることの証左となる。

次に，*because* は，多くの文法書や辞書にも書かれているように，「主節の後ろにくるのが普通」（江川, 1991, p.386）であり，文頭で使うのは「先に理由や原因を述べる場合に限られる」（リーチ他, 2006, p.128）。しかし，もっと問題にされるべきは，学習者が文頭で *because* を用いた場合の多くが，引用(3)〜(5)のような主節を持たない断片文であることである。

(3) * **Because** it is my dream.（JH）
(4) * **Because** I am Japanese!（SH）
(5) * **Because** English is used all over the world.（UNI）

本研究の分析データでは，中学生，高校生，大学生が文頭で *because* を使った場合，それぞれ97.55％，94.75％，84.58％の割合で断片文となっている。これも「なぜなら, それは私の夢だからです」のような文章が許容される日本語の影響であると考えられるが，主節を持たない *because* の文も，話し言葉における *why* で始まる疑問文の答えとしては許容される。したがって，学習者が断片文を多く産出する原因として，彼らが書き言葉と話し言葉というモードの違いを理解していないことが挙げられる。さらに，小林（2009c）で明らかにされているように，中学校検定教科書で提示されている *because* の例文の数が極端に少なく，その半数以上がダイアログ部分で提示される主節を持たない文に現れていることも見逃せない。しばしば，「日本人学習者は，あたかも話しているように，英語を書く」（Asao, 2006; 竹蓋, 1982）と指摘されてきたが，学習者による接続詞の使用頻度や用法などにも，その特徴は顕著に見られる。

5.1.1.3　接続詞の質的分析

ここまで，あらゆる学習者コーパスは「さまざまな言語形式に関する独特の頻度行列」（Krzeszowski, 1990, p.212）を持っているという仮説に基づいて，

学習者言語の計量分析を行ってきた。しかしながら，計量的なアプローチのみで解析できるのは，学習者言語の一面にすぎない。それゆえに，個々の使用例を質的に精査することによって，質的な特徴抽出や誤用分析も併用しなければならない。以下，いくつか典型的な特徴や誤用を示す。

前述のように，日本人学習者は，母語話者と比べて，*and*，*but*，*so*，*because* のような接続詞を有意に過剰使用し，とりわけ文頭でそれらの語を使用する傾向がある。これは，単文を連続させ，それらを文頭の接続詞で連結しようという意識の現れであると思われる。さらに，学習者が文頭で接続詞を用いている例の中には，引用(6)〜(9)のような，不必要なコンマを挿入する誤用が多く見られる。紙面の都合で中学生の用例のみを示すが，高校生や大学生のデータにも同様の例が現れている。なお，本節では，接続語以外の誤用については言及しない。

(6) * And, I like books and games.（JH）
(7) * But, it's very interesting.（JH）
(8) * So, I want it very much.（JH）
(9) * Because, we need money.（JH）

この種の誤用は，とりわけ *and* と *but* に多い。接続詞ではなく，*also* や *however* のような接続副詞の場合はコンマが必要となる（e.g., However, it is very interesting.）。不必要なコンマの挿入は，接続詞と接続副詞の違いに関する知識が学習者に不足していることに起因するものである。

また，個々の語の使用例を精査していくと，さまざまな誤用が散見される。まず，引用(10)〜(12)を参照されたい。

(10) * I don't have tennis shoes and a hat! (JH)
(11) * I don't have rice and miso-soup in the morning. (SH)
(12) * Though, we can't still speak and write English fluently. (UNI)

このような場合，否定語の後では *and* ではなく，*or*（あるいは *nor*）を用いるとされている（e.g., I don't have rice or miso-soup in the morning.）。この用例については多くの辞書や文法書にも記述があるが，学習者の理解度はそれほど高くない。

そして，学習者作文には，引用(13)〜(14)のように，順接の意味が希薄な *and* や逆接の意味が希薄な *but* の使用例が見られる。ちなみに，JEFLL では，学習者の流暢さを確保するために，英語が浮かばない箇所を日本語で書くことを許可している。

(13) I was really fun and tired.（SH）
(14) I usually have bread.
But I like rice.
I eat rice sometimes.
But it's めったにない.
And I don't like milk.
But I like 味噌汁.
But milk and 味噌汁 drink めったにない.（JH）

引用(13)における *fun* と *tired* は対比されるべき表現であることから，この文章は，逆接の接続詞である *but* などを使って，"I [really had] fun but [I was] tired." や "[It] was [real] fun but [I was] tired." などと書かれるべきものである。そして，引用(14)では，7文のうち，4つもの文頭で *but* を使用しており，2文続けて用いている箇所もある。しかし，（I like rice, too. のようなニュアンスにも読める）2行目の "But I like rice." のように，必ずしも逆接とは言えない表現がある。この点について，Scollon and Scollon（1995）は，日本人，韓国人，中国人などの学習者が用いる *and* や *but* がしばしば結束性を持っていないと指摘している。日本語の「が」は，逆接だけでなく順接の場合にも用いられるため（綿貫・宮川・須貝・高松, 2000, p.596），逆接の意味が希薄な *but* も母語の干渉と見ることができる。

前述した断片文にも，いろいろな種類がある。もちろん，引用(3)〜(5)のような単純に主節が脱落したものが数としては多いが，引用(15)〜(20)のような断片文も散見される。

(15) * Because I have a camera, but it is not so good.（JH）
(16) * Because it is very useful and it has my memories.（SH）
(17) * Because they suddenly barked and chased me.（UNI）
(18) * Because, if we don't have a breakfast.（JH）
(19) * Because I have no time every morning, so I

can eat bread for breakfast very quickly.（SH）

⑳ * Because when I go to foreign country, I will can speak to them (foreigners) and attempt to communication.（UNI）

引用⑮〜⑰は，*because* で始まる従属節が等位接続詞（*and*, *but*）によって拡張されているタイプの断片文である。これらは，文と節の違い，あるいは主節と従属節の違いなどを理解していないために起こる誤りである。そして，引用⑱〜⑳は，*because* で始まる従属節の中に，他の従属接続詞（e.g., *if*, *when*）によって導かれる従属節が入れ子のように埋め込まれているタイプの断片文である。また，引用⑲のような *because* と *so* の二重使用も多い。本来ならば，これは，*because* か *so* のどちらか一方だけで因果関係を表せる文である。この種の誤用は，学習者が *because* や *so* を等位接続詞と同じように用いていることを反映している。ちなみに，若干タイプは異なるが，接続詞の二重試用という点では，引用㉑のような *but* と *although* の二重使用も見られる。

㉑ * Although he died soon, but his story, and his name as a story teller, lasted forever.（SH）

さらに，従属接続詞には主節と従属節の両方が必要であるという認識がない学習者は，引用㉒のように，しばしば非常に複雑な埋め込みがなされている文を産出する。

㉒ * If I take all then I will die because there are very heavy, so I can't escape.（JH）

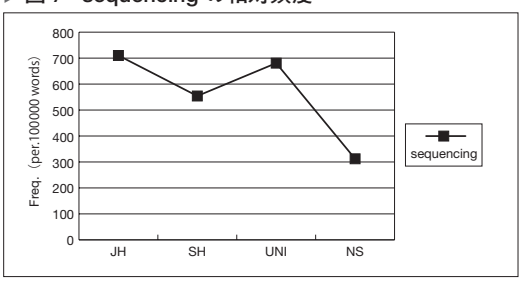
Turton and Heaton（1996）によれば，最初の節が *if*，*since*，*as*，あるいは *because* で始まる場合は，その節に続く節を *so* で始めてはならない（p.304）。しかし，引用㉒には，*if*，*because*，*so* という 3 つの従属接続詞が用いられており，それに加えて，接続副詞の *then* も用いられている。これまで学習者による接続語の過剰使用や，主節と従属節の違いに関する知識の不足について見てきたが，引用㉒はその極端な例である。

5.1.2 Frame markers (FRM)

Hyland list において，接続語のほとんどは TRA に含まれているが，談話の順序を表す *firstly*，*lastly*，*then* といった表現は，FRM の sequencing という下位カテゴリーに含まれている。

図 7 は，中学生，高校生，大学生，母語話者の 4 グループにおける sequencing の相対頻度（10万語当たり）をまとめたものである。

▶ 図 7：sequencing の相対頻度



この図を見ると，学習者が母語話者の約 2 倍も sequencing を用いていることがわかる。引用㉓は，中学生による作文の全文引用である。なお，学習者による綴りの誤りはそのままにしてある。

㉓ I'll bring our よきんつうちょう first.
I'll bring our アルバム second.
I'll bring my 金 thrid.
I'll bring my game ソフト seventh.
I'll bring my 服 a bittle fourth.
I'll bring our ちょっとの food sexth.
These are in some of my fag.（JH）

引用㉓では，なぜ 4 番目の文に *seventh* とあるのかは不明であるが，最後を除くすべての文に順序を表す談話標識がついている（*thrid* と *sexth* は，それぞれ *third* と *sixth* の誤り）。おそらくこれは，物事を順序立てて書かなければならないという意識が空回りしたものであろう。事実，上級の学習者になると，談話標識を明示的に使わなくても，論理的な文章を書けるようになる（Intaraprawat & Steffensen, 1995, p.271）。しかしながら，日本人学習者の場合は，彼らがそれまでに受けてきた作文指導や教材の影響もあってか，接続語を過剰使用する傾向がある。

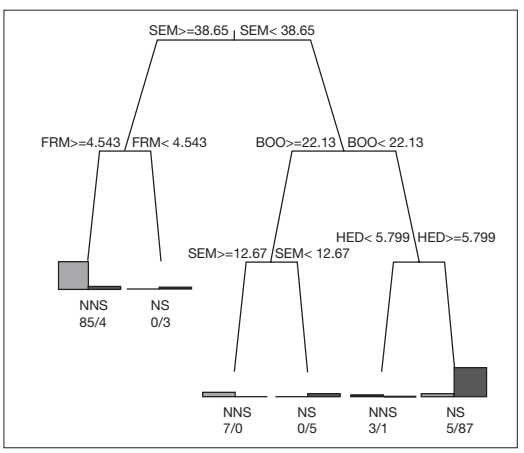
5.2 Interactional resources

Hyland（2005）によれば，interactional resources とは，HED，BOO，ATM，ENG，SEM の上位カテゴリーであり，命題内容に対する書き手の視点を示し，読み手を議論に巻き込んでいく働きを担っている（p.52）。本稿では，最も高頻度なカテゴリーである SEM に加えて，書き手の心的態度を表す HED と BOO を中心に分析していく。

5.2.1 Self-mentions (SEM)

SEM とは，書き手に対する明示的言及を指し，主に 1 人称の代名詞（e.g., *I*, *we*）のことである（Hyland, 2005, p.53）。また，小林（2010a）が回帰木（CART）を用いた分類モデルで明らかにしているように，SEM は，学習者の談話と母語話者の談話を最もよく判別する変数である（図 8）。

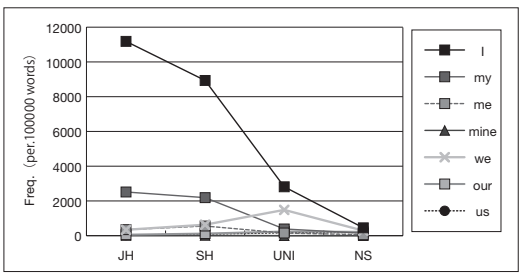
▶図 8：回帰木による NS/NNS 分類モデル（小林，2010a, p.319）



この図を見ると，個々の英作文に対して，まず SEM の値（相対頻度）に基づいた分類が行われている。そして，SEM の値が一定以上であったデータに対しては，次に FRM の値に基づいた分類が行われる。また，SEM の値が一定未満であったデータに対しては，BOO の値に基づいた分類が行われる。さらに，BOO の値が一定以上であったデータでは再び SEM の値を参照し，逆に BOO の値が一定未満であったデータでは HED の値を参照している（この分類モデルの精度は95%）。

図 9 は，4 グループにおける 1 人称代名詞の相対頻度（10万語当たり）を視覚化したものである。

▶図 9：1 人称代名詞の相対頻度



この図を見ると，習熟度が低いほど，1 人称単数主格の *I* を有意に過剰使用する傾向にある（χ^2 = 21680.57, df = 3, p < .001）。また，1 人称単数所有格の my にも，有意差が見られる（χ^2 = 5448.33, df = 3, p < .001）。

引用(24)は，中学生の英作文（全文引用）である。

(24) I often eat rice in the morning.
I drink milk everyday.
I like milk very much.
I sometimes eat bread.
I like bread a little.
I eat breakfast everyday.（JH）

論説文においては，書き手の存在はテキストの背後にあり，それゆえ，書き手がテキストの前面に出てきたときに修辞的効果が生まれるのである（e.g., Hyland, 2001）。言い換えれば，基本的に「客観的」なトーンを持ったテキストに突如「主観的」なトーンが現れるからこそ，そこが強調されるのである。だが，学習者の作文では，基本的なトーンが「主観的」であるため，SEM の修辞的効果は生まれていない。むしろ，書き手の存在が常に前面に出ているために，論説文に必要とされる客観性がほとんど見られない。

このような 1 人称代名詞の過剰使用は，先行研究（e.g., Petch-Tyson, 1998）が報告しているように，学習者による英作文の特徴である。1 人称代名詞は，話し言葉に顕著な言語項目であり（Biber et al., 1999, p.333），このこともまた，学習者による書き言葉が話し言葉に近い特徴を持っていることの証左となる。

5.2.2 Hedges (HED)

学習者にとって，対象言語で懐疑や確信を適切に

表現することは非常に難しい。しかしながら，それらを適切に表現することは，外国語習得やアカデミック・ライティングに不可欠なものである。特に，HED は，自らの主張（claim）を弱めることで，逆に議論（argument）そのものを強くする働き（Meyer, 1997）を持っており，それを巧みに使いこなせるかどうかが母語話者と学習者の間の“rhetorical gap”（Hyland, 1995, p.39）である。

学習者は，さまざまな言語的背景や文化的背景を持っているため，対象言語が用いられている談話共同体における「規範」に従うことが容易ではなく，しばしば大きくそこから「逸脱」してしまう。例えば，モダリティの使用に関して，中国語を母語とする英語学習者は，母語話者よりも直接的で強い表現を好み，それほど確信のある事柄でなくとも断言してしまう傾向があると報告されている（e.g., Allison, 1995）。したがって，日本語を母語とする学習者がどのような表現を好み，逆にどのような表現を苦手とするかを調査することは有意義なことである。

表 5 は，4 つのサブコーパスにおける HED の頻度上位20タイプをまとめたものである（同じ頻度の語が複数ある場合は，それらも含めている）。表中の頻度は，10万語当たりの相対頻度である。なお，中学生，高校生，大学生，母語話者による20タイプの使用頻度の合計は，HED 全体の使用頻度に対して，それぞれ100％，98.16％，92.06％，87.71％を占めている。

中学生が最も多く使う HED は *sometimes* であり，この語の順位は習熟度が上がっていくにつれて下がっていき，母語話者の使用頻度は中学生の使用頻度の1/15程度である。また，中学生は，*maybe* を好む一方で，*may* はその1/3 程度しか用いていない。*maybe* と同様に書き手の推量を表す表現として，*perhaps* がある。これらはほとんど同じ意味を持っている語であるが，学習者は前者を好み，母語話者は後者を好んで用いている。例えば，表 5 を見ると，中学生は，*maybe* を母語話者の約 5 倍多く

■ 表 5：HED の頻度上位 20 タイプ

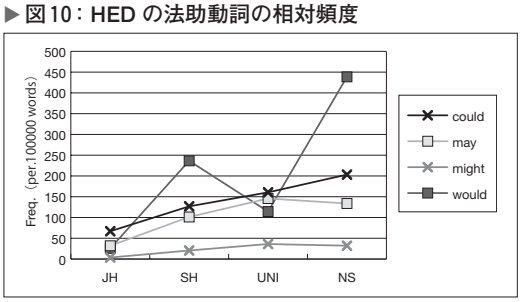
	JH		SH		UNI		NS	
1	sometimes	319.18	would	227.54	about	313.39	would	430.56
2	about	171.86	feel	177.47	may	145.73	about	216.62
3	maybe	108.64	sometimes	174.69	could	120.26	could	193.88
4	often	57.08	about	161.89	would	111.97	may	133.71
5	could	40.51	may	101.25	feel	65.17	feel	102.29
6	may	31.92	could	95.69	often	59.83	claim	58.17
7	couldn’t	26.39	often	81.22	maybe	50.95	often	51.48
8	would	23.32	maybe	72.88	almost	49.76	around	42.79
9	almost	14.73	almost	40.06	sometimes	47.99	argue	36.77
10	guess	6.14	couldn’t	31.15	couldn’t	40.28	possible	32.76
11	around	5.52	possible	24.48	possible	39.69	might	32.09
12	felt	4.30	around	21.14	might	36.14	seems	30.09
13	might	3.68	might	20.58	around	33.77	probably	28.08
14	possible	3.07	probably	17.80	felt	26.07	perhaps	26.07
15	wouldn’t	2.46	perhaps	15.58	probably	24.29	sometimes	21.39
16	perhaps	1.84	felt	13.35	tend to	22.51	likely	20.73
17	generally	1.23	wouldn’t	8.90	seems	19.55	maybe	20.73
18	frequently	0.61	quite	5.01	perhaps	16.59	felt	20.06
19	possibly	0.61	likely	3.89	quite	8.89	almost	19.39
20	probably	0.61	mostly	3.89	generally	8.29	claims	17.38
	quite	0.61	seems	3.89	in my opinion	8.29		
					suppose	8.29		

用いている一方で，*perhaps* を母語話者の約1/14程度しか用いていない。Biber et al.（1999）によれば，*maybe* は話し言葉に顕著な表現であり，*perhaps* は書き言葉に顕著な表現である（p.869）。この事実からも，学習者の書き言葉が話し言葉によく似た特徴を持っていることがわかる。

そして，母語話者が最も多く使う HED は *would* であり，その頻度は中学生の使用頻度の約18倍である。表5を見ると，母語話者や大学生のデータでは，頻度上位5語のうちの3語は法助動詞（*would*, *could*, *may*）であり，これらの語が HED の中心的位置を占めていることがうかがい知れる。また，法助動詞は，英語教育の比較的早い段階で導入される語であるが，1つの語が文脈によってさまざまな意味で用いられるため，学習者にとっては習得の難しいものである。

図10は，Hyland list で HED とされている法助動詞（*could*, *may*, *might*, *would*）の相対頻度（10万語当たり）の分布を視覚化したものである。なお，Hyland list では，*would* と *wouldn't* や *could* と *couldn't* は区別されているが，この図では1つにまとめてある。

▶図10：HED の法助動詞の相対頻度



この図を見ると，いずれの法助動詞の場合も習熟度が上がるにつれて頻度が上昇していく傾向があり，とりわけ中学生はこれらの法助動詞をあまり使うことができない。母語話者と比べて，学習者が最も過少使用している語は，*would* である。学習者の中で最も多く *would* を用いている高校生の相対頻度は227.54回であり，母語話者の相対頻度（430.56回）の約半分程度である。

母語話者よりも使用頻度が低いものの，学習者も仮定標識としての *would* や *could* を用いている。しかしながら，学習者が用いる *would* や *could* の主語は，引用㉕〜㉖のように，1人称代名詞の *I* である場合が多い。

㉕ If a big earthquake came, I **would** take out some foods.（JH）
㉖ If I had a mobile phone, I **could** call.（SH）

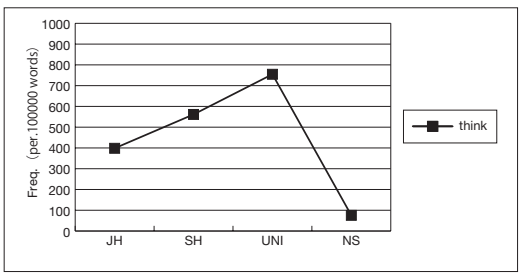
母語話者による *I would* というコロケーションの頻度は，*would* 全体の頻度の4.97%にすぎない。だが，中学生，高校生，大学生のデータでは，それぞれ78.72%，76.60%，64.00%と非常に高い。同様に，母語話者による *I could* というコロケーションの頻度は，*could* 全体の頻度の4.14%にすぎない。その一方で，中学生，高校生，大学生のデータでは，それぞれ91.42%，75.00%，4.13%であり，中学生と高校生の比率が非常に高い（大学生は，母語話者とほとんど同程度の割合である）。約言すれば，母語話者は，*would* や *could* の主語として無生物主語を好む一方，学習者は，1人称単数代名詞を好む傾向が顕著に見られる。そして，*may* と *might* の場合，それらの語の主語が *I* である割合は，中学生と高校生で20.93〜41.07%，大学生と母語話者で0.03〜0.10%と，*would* や *could* の場合と比べて低い。なお，4つの法助動詞が *I* を主語にとるか否かという比率には，4つのサブコーパス間で有意な差が見られる（*could*: $\chi^2 = 518.471$, df = 3, $p < .001$; *may*: $\chi^2 = 151.035$, df = 3, $p < .001$; *might*: $\chi^2 = 21.419$, df = 3, $p < .001$; *would*: $\chi^2 = 644.809$, df = 3, $p < .001$）。

Hyland list における法助動詞以外の HED（形容詞，副詞，動詞）に関する詳しい分析は，小林（2010b）を参照されたい。

5.2.3 Boosters (BOO)

BOO とは，HED とは対照的に，自らの意見と対立する意見を遮断し，命題に対する確信度を強調する修辞法である（Hyland, 2005, p.52）。BOO は，書き手の確信を表す動詞（e.g., *believe*, *prove*）や副詞（e.g., *obviously*, *undoubtedly*）を多く含んでいるが，最も高頻度な表現は *think* である（この語の頻度だけで，BOO 全体の頻度の34.92%を占める）。図11は，4グループにおける *think* の相対頻度（10万語当たり）を視覚化したものである。

▶図11：*think* の相対頻度



この図を見ると，学習者が母語話者の約5～10倍も *think* を用いていることがわかる。引用(27)～(29)は，学習者による *think* の使用例である。

(27) I **think** I will go there in February.（JH）
(28) I **think** it is very important to have breakfast.（SH）
(29) I **think** there is another reason.（UNI）

学習者が *think* という語，とりわけ *I think* というコロケーションを過剰使用することはこれまでも報告されてきた（e.g., Granger, 1998b; McCrostie, 2008; Ishikawa, 2009）。また，日本人大学生が母語である日本語でも「思う」という口語体を多用すると指摘されているが（黒田, 2005, pp.143-146），英語における *think* の過剰使用も，日本語の「思う」の影響である。ただ，和英辞典には，「思う」に対応する英語として，*think* 以外に，*consider*，*believe*，*expect*，*feel*，*wish*，*wonder*，*suspect*，*imagine*，*suppose*，*guess* などを含む多くの表現が載っている（e.g., 小林, 2010a）。学習者は，母語話者 のようにこれらの語を文脈に応じて使い分けることができていないことが *think* の過剰使用の原因となっている。さらに，外山（1986, 1992）が指摘しているように，日本人学習者が産出する *think* は，母語話者が BOO の意味で用いているのとは異なり，*it seems to me* のような HED に近い意味で用いている。

HED と同様，学習者は，*must* や *of course* などの数少ない例外を除いて（e.g., 小林, 2010a），頻度の面からも用法の面からも BOO を適切に使用することができていない。

6 まとめと今後の課題

本研究では，テキストマイニングの技法と Hyland の（メタ）談話標識のリストを用いて，談話分析の観点から日本人英語学習者の英作文を量的・質的に分析してきた。その結果，接続表現（TRA，FRM），視点（SEM），心的態度（HED，BOO）といった多くの点において，習熟段階の異なる学習者の間，そして学習者と母語話者の間に頻度や用法に関する大きな差異が見られた。今回扱った談話表現の多くは，文法というよりは文体にかかわるものが多く，構造的なルールというよりは確率論的なパターンにかかわるものである。その限りで，コーパスから得られる客観的データ（頻度や統計量など）が極めて有用な情報となることは間違いない。

しかし，本研究はまだ始まったばかりのものであり，今後の課題も多い。本稿では，学習者のさまざまな談話的特徴を俯瞰することを目的としたため，高頻度な表現のみに焦点が当てられ，中頻度や低頻度の表現に関してはあまり言及できなかった。また，今後は，単なる統計的な記述にとどまらず，本研究から得られた結果を実際の教室指導や教材作成に生かし，どうすれば「英語らしい」文章が書けるようになるのかという点を考えていかなければならない。そして，小林（2010a）などで試みられているように，さまざまなテキスト分類の技法を応用し，談話能力の観点から英作文を自動評価するシステムの開発も模索していきたい。

謝　辞

この研究を発表する貴重な機会を与えてくださった（財）日本英語検定協会と選考委員の先生方に厚く御礼申し上げます。また，本研究を進める過程でさまざまなコメントをくださった皆様にも，心より感謝を申し上げます。

参考文献（＊は引用文献）

＊Allison, D.（1995）. Assertions and alternatives: Helping ESL undergraduates extend their choices in academic writing. *Journal of Second Language Writing*, *4*, 1-15.

＊Altenberg, B., & Tapper, M.（1998）. The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger（Ed.）, *Learner English on computer*（pp.80-93）. London: Longman.

＊Asao, K.（2006）. Spoken and written discourse of EFL learners: Findings from the sound corpus of Japanese learners of English. A paper given at the ICLE / LINDSEI Japanese Sub-Corpus Symposium. Tokyo: Showa Women's University.

＊Baker, P.（2006）. *Using corpora in discourse analysis*. London: Continuum.

＊Baker, P., Hardie, A., & McEnery, T.（2006）. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

＊Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E.（1999）. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

＊Clahsen, H.（1980）. Psycholinguistic aspects of L2 acquisition. In S. Felix（Ed.）, *Second language development: Trends and issues*（pp.57-79）. Tübingen: Gunter Narr.

＊Corder, S P.（1967）. The significance of learners' errors. *International Review of Applied Linguistics*, *5*, 161-169.

＊Crewe, W.J.（1990）. The illogic of logical connectors. *ELT Journal*, *44*, 316-325.

＊Crismore, A., Markkanen, R., & Steffensen, M.（1993）. Metadiscourse in persuasive writing: A study of texts written by American and Finnish students. *Written Communication*, *10*, 37-71.

＊van Dijk, T.A.（1977）. Connectives in text grammar and text logic. In T.A. van Dijk & J.S. Petöfi（Eds.）, *Grammars and descriptions*（pp.11-63）. Berlin: Walter de Gruyter.

＊Dulay, H., & Burt, M.（1973）. Should we teach children syntax? *Language Learning*, *23*, 37-53.

＊江川泰一郎.（1991）.『英文法解説』改訂三版. 東京：金子書房.

＊Færch, C., Haastrup, K., & Phillipson, K.（1984）. *Learner language and language learning*. Clevedon: Multilingual Matters.

＊Fraser, B.（1993）. Discourse markers across language. *Pragmatics and Language Learning, 4*, 1-16.

＊Fraser, B.（1999）. What are discourse markers? *Journal of Pragmatics*, *31*, 931-952.

＊Granger, S.（1998a）. The computer learner corpus: A versatile new source of data for SLA research. In S. Granger（Ed.）, *Learner English on computer*（pp.3-18）. London: Longman.

＊Granger, S.（1998b）. Prefabricated patterns in advanced EFL writing: Collocation and formulae. In A.P. Cowie（Ed.）. *Phraseology: Theory, analysis, and applications*（pp.145-160）. Oxford: Oxford University Press.

＊Granger, S., & Rayson, P.（1998）. Automatic profiling of learner texts. In S. Granger（Ed.）, *Learner English on computer*（pp.119-131）. London: Longman.

＊Granger, S., & Tyson, S.（1996）. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*, 17-27.

＊Hyland, K.（1995）. The author in the text: Hedging scientific writing. *Hong Kong Papers in Linguistics and Language Teaching*, *18*, 33-42.

＊Hyland, K.（2001）. Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics*, *34*, 1091-1112.

＊Hyland, K.（2005）. *Metadiscourse: Exploring interaction in writing*. New York: Continuum.

＊Intaraprawat, P., & Steffensen, M.S.（1995）. The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing*, *4*, 253-272.

＊Ishikawa, S.（2009）. Phraseology overused and underused by Japanese learners of English: A contrastive interlanguage analysis. In K. Yagi & T. Kanzaki（Eds.）, *Phraseology, corpu linguistics and lexicography: Papers from Phraseology 2009 in Japan*（pp.87-98）. Hyogo: Kwansei Gakuin University Press.

＊小林雄一郎.（2009a）.「日本人学習者の英作文における等位接続詞の使用について―“and” と “but” を例に」.『専修大学外国語教育論集』37, 21-36.

＊小林雄一郎.（2009b）.「日本人学習者の英作文における “so” の統計的分析」.『コーパス言語研究における量的データ処理のための統計手法の概観』（統計数理研究所共同研究リポート232）, 107-118.

＊小林雄一郎.（2009c）.「日本人英語学習者の英作文における because の誤用分析」.『関東甲信越教育学会研究紀要』23, 11-21.

＊小林雄一郎.（2010a）.「回帰木を用いた NS/NNS テキスト分類」.『言語処理学会第16回年次大会発表論文集』, 318-321.

＊小林雄一郎.（2010b）.「多変量アプローチによる英語学習者のレトリック分析」.『統計学的アプローチによるテクスト分析』（統計数理研究所共同研究リポート245）, 1-22.

＊Krzeszowski, T.（1990）. *Contrasting language: The scope of contrastive linguistics*. Berlin: Mouton de Gruyter.

＊黒田龍之助.（2005）.「最後はやっぱり日本語」.『その他の外国語―役に立たない語学のはなし』（pp.143-146）. 東京：現代書館.

＊Lado, R.（1957）. *Linguistics across cultures*. Ann Arbor: University of Michigan Press.

＊Leech, G.(1992). Corpora and theories of linguistic performance. In J. Startvik(Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp.105-122). Berlin: Mouton de Gruyter.
＊Leech, G.(1998). Preface. In S. Granger(Ed.), *Learner English on computer*(pp.xiv-xx). London: Longman.
＊リーチ, G.・池上嘉彦・上田明子・長尾真・山田進(監修).(2006).『ロングマン英和辞典』. 東京：桐原書店.
＊Leech, G., & Svartvik, J.(1994). *A communicative grammar of English*. 2nd ed. London: Longman.
＊MacWhinney, B.(1995). *The CHILDES project: Tools for analyzing talk*. 2nd ed. Hillsdale: Lawrence Erlbaum Associates.
＊松村真宏・三浦麻子.(2009).『人文・社会科学のためのテキストマイニング』. 東京：誠信書房.
＊McCarthy, M.(1991). *Discourse analysis for language teachers*. Cambridge: Cambridge University Press.
＊McCrostie, J.(2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal*, *32*, 97-114.
＊Meyer, P.G.(1997). Hedging strategies in written academic discourse: Strengthening the argument by weakening the claim. In R. Markkanen & H. Shroder(Eds.), *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*(pp.21-41). Berlin: Walter de Gruyter.
＊Perdue, C.(1993). *Adult language acquisition:Cross-linguistic perspectives*. 2 vols. Cambridge: Cambridge University Press.
＊Petch-Tyson, S.(1998). Writer/reader visibility in EFL written discourse. In S. Granger(Ed.), *Learner English on computer*(pp.107-118). London: Longman.
＊Pravec, N.A.(2002). Survey of learner corpora. *ICAME Journal*, *26*, 81-114.
＊Scollon, R., & Scollon, S.(1995). *Intercultural communication*. Oxford: Blackwell.
＊Selinker, L.(1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 209-231.
＊竹蓋幸生.(1982).『日本人英語の科学』. 東京：研究社出版.
＊Tankó, G.(2004). The use of adverbial connectors in Hungarian university students' argumentative essays. In J.M. Sinclair(Ed.), *How to use corpora in language teaching*(pp.157-181). Amsterdam: John Benjamins.
＊投野由紀夫(編).(2007).『日本人中高生一万人の英語コーパス "JEFLL Corpus"：中高生が書く英文の実態とその調査』. 東京：小学館.
＊外山滋比古.(1986).「『思われる』と『考える』」.『思考の整理学』(pp.218-223). 東京：ちくま文庫.
＊外山滋比古.(1992).「『私』の問題」.『英語の発想・日本語の発想』(pp.70-73). 東京：日本放送出版協会.
＊Turton, N.D., & Heaton, J.B.(1996). *Longman dictionary of common errors*. New ed. London: Longman.
＊Vande Kopple, W.(1985). Some explanatory discourse on metadiscourse. *College Composition and Communication*, *36*, 82-93.
＊綿貫陽・宮川幸久・須貝猛敏・高松尚弘.(2000).『ロイヤル英文法』. 東京：旺文社.
＊Zamel, V.(1983). Teaching those missing links writing. *ELT Journal*, *37*, 22-29.