英語科

コーパスを応用した英語教育研究と英語学習
―現状と可能性の考察―

藤田高弘

【抄録】　近年の自然言語処理やコーパス言語学の発展にともない、コーパスを利用しての英語教育研究や英語学習が注目されている。本稿ではまず、コーパスの定義、英語学習者コーパスと英語学習用コーパスについての概略をおこない、次に日本の英語教育という環境のなかでどのような可能性と問題点があるのかを、1）第2言語習得・外国語学習研究への応用　2）英語学習への応用という2つの観点から現在進行中のプロジェクトの現状、また研究や実践の可能性を考察をした。

【キーワード】　コーパス、第2言語習得、英語学習者コーパス、英語学習者用コーパス、英語教育、中間言語
パーソナルコーパス

はじめに

パソコン、インターネットの発達・普及により近年急速に興起したのがコーパス言語学（Corpus Linguistics）である。言語の科学的研究を行うのが言語学であり、英語という特定言語の科学的研究を行うのが英語学であるとするならば、コーパス言語学は言語学の、英語コーパスは英語学の一分野[1]である。本稿では、英語コーパス研究の中でも英語教育とコーパスとの観点からその現状と新しい可能性、または残されている問題点をまとめ考察を試みた。

1．コーパスの定義と英語学習者コーパス／英語学習用コーパスの概略

コーパスとは、広義にはテキストの集成（McEnery & Wilson 1996）で、普通はコンピュターで処理可能な電子化されたテキストの集成である。近年になって登場した学習者コーパス（Learner Corpus）とは、英語を第2言語、および外国語として学ぶ学習者の産出した（話し言葉・書き言葉）データを目的に応じて組織的に収集し電子化し（品詞情報、エラー情報等が入ったものもある。）、コンピュター処理が可能なコーパスのことである。基本となる英語データは、発達段階にある中間言語（Interlanguage）を反映した言語能力であり、言語運用であるのが特徴である。英語教育研究における学習者コーパスの目的は、大量にデータを収集することによって学習者の誤り分析、中間言語の特徴、言語獲得の発達段階等を明確にし、一般化を目指し、第2言語習得、英語教育研究に役立つデータを提供することにある。

英語学習用に利用されるコーパス（Corpus used for learning）とは、多様な目的と多様な学習者に応じて豊かなコンテクストにある大量の英語用例を、コンピュターで柔軟な検索が可能となるように組織的に収集し電子化されたコーパスを意味する。基本的に収集し記録されるデータは、英語母語話者が書いたり、話したりした大量のデータであり、または英語教材として利用される書き言葉、話し言葉の集成となる。従って、その英語は母語話者の完成した言語能力、言語運用のものとなる。

英語学習分野における英語コーパス利用の意義は、コンピュター利用の英語学習CALL（Computer Assisted Language Learning）の一部として、学習者個人の学習レベルや目的に柔軟に対応できる英語学習支援としての利用と、さらには，日常的に授業等で使用する教材をコーパスにして簡易的な調査や検証を実施したり、また英語教師が教材やそれ以外のソースから目的に応じて組織的に収集した英語用例収集や教材開発支援への応用にある。

2．英語学習者コーパス（Learner Corpus）の第2言語習得・外国語学習研究への応用例

現時点での英語学習者コーパスの第2言語習得・外国語学習研究に応用されている、または応用が可能な方法は、大きく2つにまとめることができる。1つは、General Learner Corpusと言われるものと、Specitic Learner Corpusと言われるものである。それぞれの特徴と実際のコーパス構築の実態、研究例と応用可能な例をそれぞれ具体的に見ていきたい。

General Learner Corpus

このコーパスは、単純均一なタスクを学習者に課し、学習レベルを統一した外国語学習者から横断的にサンプリングした大量データーコーパスである。データ量の大きさ、学習者に課したタスク、学習レベルの均一性によって、目的とする研究の一般化や実証性を高めるという利点がある。英語教育研究、特に実験を通しての実証的研究や調査において、その綿密なリサーチ・デザインのもと収集した学習者コーパスを利用することにより、データ量の不足により一般化や科学性に欠けるという従来からの批判に応えうる可能性を持っている。また、特定学習者の包括的、体系的な学習文法の実態を実証的なデータに基づいて捕らえることが可能となる。さらに、このコーパスを利用して、学習者の文法構造、語彙使用、エラーの一般的特徴や発達段階別の特徴を把握するこで、語彙、文法、英作文の治療的指導や、診断テストに応用することも可能となる。

このようなGeneral Learner Corpusの日本の現状は、東海大学のCorpus of Japanese Learner Corpus[2]、東京学芸大学のTGU Learner Corpus[3]の２つのプロジェクトが現在進行中である。前者は、中・高・大の広範なデータを採集予定で、インターネット上でのデータの共有化も行っている。後者は、中学２年から高校３年までの同一のトピックで自由英作文のデータを100万語まで電子化しこれもインターネット上で公開予定である。

さて、このGeneral Learner Corpusを利用してどのような研究や学習に活用できるのかを実際の研究例と可能性という観点から論じてみたい。

まずは、第２言語習得・外国語学習研究への応用研究として、形態素・文法項目習得順序の追試と検証がある。形態素の習得研究として形態素習得の自然順序を唱えた先行研究であるDulay&Burt（1978）の追試を、前述のTGU Learner Corpusを利用して投野（1997）が追検証を試みた。（図１を参照）また、同じくTGU Learner Corpusで、桃井（1998）は日本人英語学習者の基礎語彙のコロケーションを母語話者と比較しその発達段階を分析した。

海外の研究ではInternational Corpus of Learner English[4]のプロジェクトで、11の異なる母語の英語学習者の作文データを収集し、エラー分析を試みている。その目的は、母語の違いを超えて共通する普遍的エラーと、母語の干渉によるエラー分析をおこない言語習得の研究に役立てようとするものである。

さて次に、このようなGeneral Learner Corpusを利用してどのような研究や学習指導に活用できる可能性があるのかを考えてみたい。統語レベルでは、General Learner Corpusから、特定構文の過度・過小使用の比較、分析を通して、英語表出レベルの特徴やその原因と対策に活用できる。また、日英パラレルコーパス（日本語と英語の両方を電子化し収集したコーパス）を構築することによって日英の発想の違いによる統語レベルのエラー分析、学習者の母語の干渉によるエラー分析をより精密に分析し、学習者の学習困難点の予測や言語材料（ここでは語彙、構文、文法を意味する）の選定と配列について提言できる可能性がある。また、語彙に関して、量的には発表語彙と受容語彙の語彙数、質的には語彙タイプやコロケーションの特徴を客観的で一般性のあるデータとして語彙使用の実態を把握し、学習語彙の量的、質的検証が可能となる。さらに、前述の日英パラレルコーパスを利用し日英の語彙の意味の違いによる学習の実態等も客観的なデータとして把握し、学習指導に役立てることができる。

また、英語母語話者のコーパスとGeneral Learner Corpusを比較することによって英語を母語とする人の統語、語彙使用と外国語として学ぶ学習者の統語、語彙使用の違いを知ることができ、言語材料（ここでは語彙、構文、文法を意味する）の選定、配列、学習診断テストへの活用が可能となる。

Specitic Learner Corpus

このコーパスは目的に応じた多様なタスクを課し、目的に応じた多様な調査項目を、目的に応じた外国語の学習者から、目的に応じた抽出法でサンプリングしたコーパスと言える。このコーパスでは、タスクの多様性が収集の目的を明確にし、さらに調査項目、調査対象、研究方法、データの抽出法の多様な組み合わせ、比較によって一般化や実証性を高めようとする。研究方法とは、Cross-sectional（横断的）な研究手法やLongitudinal（縦断的）な研究手法を取り入れたりすることを意味している。このコーパスは、実証的研究で問題となる変数、つまり、調査項目、調査対象、研究方法、データの抽出法、学習者から見ると課せられたタスクを目的に応じて組み合わせた小規模ではあるが、明確なサンプリングの目的を持った学習者コーパスと言える。

例えば、特定の文法項目、特定の語彙をデータとして抽出できるようなタスクを特定の集団の学習者に与え、その結果をコーパスにすることによって目的に応じたデータを比較、分析できるのがこのコーパスの利点である。

具体的には、発表語彙の調査で特定語彙の抽出が可能なタスクを課しコーパス化し、英語発表語彙知識の量的、質的な違いをcross-sectional（横断的）に調

査し、ある一時期の多くの中学生、高校生の発表語彙知識の発達段階別のリサーチが可能となる。

さらに、同じ発表語彙知識の量的、質的変化をlongitudinal（縦断的）に追跡し、同一学習者の語彙知識の習得（学習）の発達の変化を半年、1年といった定期的な観察によって調査することによって、量的、質的変化をより詳細に新しい観点からとらえるリサーチが可能となる。また、受容語彙と発表語彙の量的、質的な違いや変化を比較することも可能となる。

全く同じように特定文法項目を横断的にまたは縦断的に調査し、特定文法項目の発達段階別、発達の変化を客観的で一般性のある形で把握できる。

これらの客観的で一般性のあるデータをもとにして語彙、構文、文法の学習内容、配列の検証や、導入と定着の実態、さらには長期的な観点からの診断テストの開発への応用と様々な可能性が考えられる。

次に、言語活動の違いによる学習者のコミュニケーションパターンの違いの客観的な分析なども可能である。具体的には、教室でのアクションリサーチに多様な形で応用できる可能性がある。例えば、教室内での英語のteacher talkをコーパス化することによって、有効なインプットとなるteacher talkの分析への応用や、ディスカッションの授業での事前学習で異なったタスクを与えることによる学習者のoral performanceへの影響を（図2を参照）、学習者のディスカッションをコーパス化して調査、分析することに応用できる。教室内の学習で学習者、教師と学習者、またはペーアワーク、グループワークの活動で学習者同士の間でどのようなコミュニケーションが交わされているのか、そのコミュニケーションパターンを客観的に理解することによってよりよい言語活動のあり方を探ることができる。

もっと簡易な利用法としては、2種類の違ったディクテーションを課したあとの、学習の目的とする文章の表出テストをコーパス化して比較し最も効果的なディクテーションのあり方を考察することも考えられる。様々なアクションリサーチで得られたデータを蓄積しコーパス化していくことで、さらにより説得力のあるアクションリサーチのデータを提供できる可能性がある。

3．英語学習分野におけるコーパスの応用

学習者の英語学習支援としてのコーパス利用

英語学習支援ツールとしてのコーパス利用は、大きくわけて2つある。1つは文法や語法を帰納的、発見的に教える教材作成支援としてのコーパス利用である。まず、対象学習者の英語レベル、対象となる学習項目にあった簡易コーパスを作成する。例えば、OUPのGraded Readerシリーズや中・高の教科書（すでにテキストデータになっている）をコーパスにし、コンコーダンスソフトを利用して目的とする語句をキーワードとしてKWIC（Kew Words In Context）形式で出力し学習者に提示し（図3を参照）、機能的学習が効果的であると考えられる文法項目、語法（定冠詞、不定冠詞、機能語の意味、用法等）の帰納的、発見的な学習が可能となる。実際に、筆者は中学生に不定詞の用法を整理させる為に400語レベルの英語教材をコーパスにしてコンコーダンスソフト[5]を利用した教材を作成し、帰納的学習を試みた。

2つ目に、ライティング支援のツールとしてのコーパス利用である。学習者（教師も含めて）が英文を書くさいに自分の書いている用法を確かめたり、テーマ別英作文での必要な語句や機能別（依頼、賛成、反対等）の表現形式をまとめて提示したりすることができる。いわゆる辞書とは違い、豊かなコンテクストにある大量の英文を対象となる学習者とその目的に応じて編纂し検索したり、教材にできるのがその特徴である。よって、まずは目的に応じた電子化されたテキストを大量に集めることが必要となるが、インターネットの普及によりこのような電子化されたテキストは容易に入手できる環境[6]になってきている。インターネット上で電子テキストを入手したり教科書の電子テキストを集め目的に合ったパーソナルコーパスを構築することができる。そのさい、収集の目的と規模を明確にして集めることが重要となる。

しかしながら、学習者自身が自由に検索ソフトを利用して文法や語法を調べたり、英語を書くという環境にはパソコンの整備、情報教育の充実を待たなければならない。現段階では、教師の教材作成支援ツールとして充分に利用できる環境が整ってきたといえる。

学習教材のコーパス化利用

日常的に授業等で使用する教材をコーパスにして、教材の言語材料（ここでは語彙、構文、文法を意味する）を調査、検証したり、さらに、日常的に使う教材をコーパスにして簡易教材の作成の具体例を考えてみたい。例えば、コースブック（教科書）、英語Ⅰ，Ⅱや中・高の教科書、英語Ⅰとオーラルの教科書のテキストデータそのままをコーパス化して、その教材の語彙再出指標（density index）の調査することができる。教材の素材となる語彙がコースごとに

体系的に取り扱われているのか、またそうでない場合はその語彙を教材とし与える等の措置をとることができる。また、英語Ⅰとオーラルの教科書で扱う語彙の量や特徴が客観的に把握でき、学習支援に応用できる。同様に、構文、文法も前述のように調査、検証ができる。

次に、教材となるテキストデータのコーパスから得られた語彙と、大量の母語話者コーパスを構築して作られた学習辞書の語彙との比較から、教材として扱う語彙の質や有効度、学習必要度などを検証することもできる。

例えば、Collins COBUILD English Language Dictionary（以下COBUILD）は、The Bank of Englishという約３億語からなるコーパスを構築し辞書を編纂している。同辞書の一つの特色として、各見出し語に頻度上位700語、1,200語、1,500語、3,200語、8,100語といった頻度情報がついている。同辞書の頻度上位700語、1,200語という頻度の語彙と、日本で使われる教科書の語彙頻度とを比較することがコンコーダンスソフトのマッチング機能を利用することによって容易に比較ができる。その結果をもとに学習語彙項目の違いや、語彙の頻度数の違いを検証したり、教科書の語彙の有効度、学習必要度なども考察できる。また、教科書で使われる基本語彙のコロケーション（例えばmake）の特色をコンコーダンスソフトで調査し、COBUILDのコロケーション辞典（CD）と比較をして、コロケーションの特徴、妥当性、学習必要度なども調査できる。

次に、Longman Dictionary of Contemporary English（以下LDOCE）では、話し言葉、書き言葉別に、1,000ごとにそれぞれS１，S２，S３、W１，W２，W３と頻度の違いを表示している。また、棒グラフを使いコロケーションの典型例がひと目でわかるようになっている。このような情報と中・高の教科書のテキストデータをコーパス化し比較することによって、話し言葉、書き言葉別の教科書語彙の頻度、コロケーションの特徴を分析し、学習教材としての語彙選定への考察が可能となる。

教材となるテキストデータのコーパスを構築し、中学・高校で目標とする語彙数、語彙項目の選定を考えるさいの、客観的根拠となりうる基準作りを考察する調査研究に活用できる。

おわりに

本稿は、コーパスの第２言語習得・外国語学習研究への応用、英語学習への応用という観点から現在進行中のプロジェクトと応用研究の可能性を考察した。これらのプロジェクトや応用研究はその緒についたばかりである。従って研究の可能性の拡大、充実の為には、コーパス・デザインを明確に定義し、確立していく必要がある。具体的には、研究方法、データ採取対象の定義、データ抽出法、データ処理法、技術的課題などにおいて解決すべき課題がまだ多く残されている。例えば、第２言語習得・外国語学習研究で利用される学習者コーパスの構築において、サンプルを抽出する対象となる母集団の厳密な定義[7]が必要となる。現状では、特に学年で整理する程度であるが、実際に同じ学年でも英語力に違いがあるので、客観的な英語力や英語学習環境を比較的容易に示せる共通指標の整備が必要である。コーパス構築の為の学習者データープロファイルの作成が望まれる。

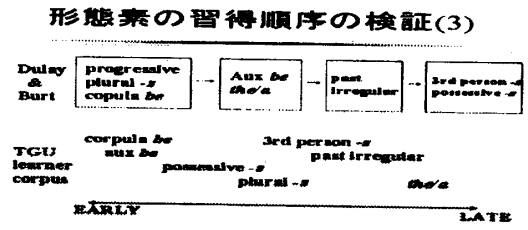
次に、目的に応じてコーパス化する学習者の英語データの抽出法の整備も急務である。現状では、話題別コーパス、課題別コーパスといった学習者の英作文を通してのデータ抽出法でデータが集められている程度である。

例えば、学習者に課せられたタスクの違いによる統語、語彙の発達的変化をLongitudinal（縦断的）な研究方法で調査する場合、まず最初に問題となるのが、目的とする調査項目の抽出が可能となる信頼性、妥当性のある実験データ抽出法である。つまり、調査対象の学習者の特定の文法項目の学習の発達段階、データ抽出に課せられたタスクの違いによって、どのような習得の実態と習得の変化があるのかを実証的に調査する為の信頼性、妥当性のあるデータ抽出法を確立する必要があるということである。第２言語習得・外国語学習研究の普遍性、実証性を高めるには、コーパス化する以前の基礎データの信頼性、妥当性をともなった抽出法をまず確立する必要がある。

次に、目的に応じて集められたデータ処理法においても、どのような観点から品詞情報やエラー情報をコーパスに入れることが、学習者の中間言語の実態を正確に把握しその結果を英語教育の実践に還元できるのかを考慮に入れて有効なデータ処理法（図４を参照）を構築しなければならない。エラー情報の自動入力などの技術的課題の解決も残されている。

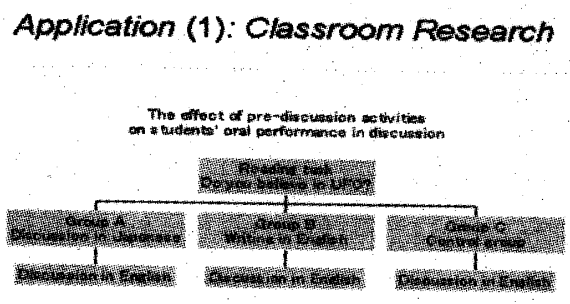
しかしながら、これらの課題を考慮し、過去の実証研究を踏まえ柔軟な発想によってコーパスを応用したリーサーチデザインを構築することで、新しい可能性を持った実証研究が、また目的に応じ学習者に合ったコーパスを構築することで、新しい可能性を持った教育実践が可能となる。

図1　形態素の習得順序の検証



投野由紀夫（東京学芸大学）のホームページより引用
（第23回全国英語教育学会　福井大会での発表資料）

図2　アクションリーサーチへの応用例



投野由紀夫（東京学芸大学）のホームページより引用（第二言語学習者のデータベース（コーパス）構築と研究方法　97／07／10　豊橋科学技術大学での発表資料）

図3　KWIC（Kew Words In Context）形式での出力

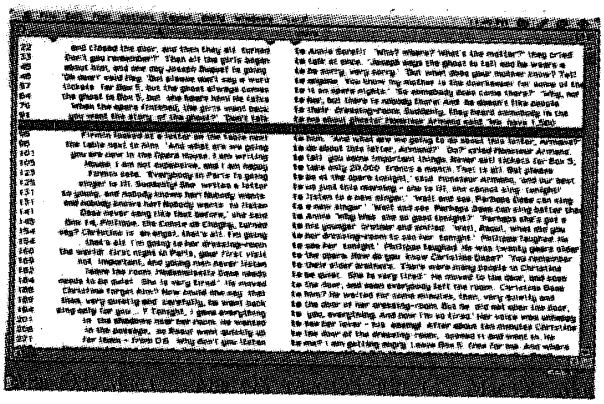


図4　タグ付きデータ

タグ付きデータの例

I love the money. So I saved it. I don't release it. Time <COP> is</COP> the money, but the money <COP> is </COP> not time. It <COP> is</COP> real? I don't think so. Money <COP>is </COP> <ART>the </ART> best thing in <ART>the </ART> world. <ART>A </ART> rich man doesn't have to work, but <ART>a </ART> poor man has to work. So, <ART>a </ART> rich man who has <ART>a </ART> lot of money has <ART>a </ART> lot of time. I love the money. I' <AUXBE>m </AUXBE> going to <ER_COP> <I = be> </ER_COP> <ER_ART> the <I = a> </ER_ART> rich man. And I play all of <ART> the </ART> life.<p>

投野由紀夫（東京学芸大学）のホームページより引用
（第23回全国英語教育学会　福井大会での発表資料）

1　『英語コーパス言語学』P.3を参照

2　東海大学のCorpus of Japanese Learner CorpusのURL: http://www.lb.u-tokai.ac.jp/lcorpus/
JACET '96で呼びかけがあった日本人英語学習者コーパス作成のプロジェクト。中・高・大と連携したデータの広範囲な採取、WWW上での共有化、エラータグ開発、音声コーパスに取り組む計画。代表は東海大学の朝尾幸次郎氏

3　東京学芸大学のTGU Learner Corpus のURL: http://www.u-gakugei.ac.jp/~tefldpt/tonolab/index-j.html
形態素の追試の結果や世界および日本の学習者コーパスの現状や構築法についての情報を入手できる。

4　International Corpus of Learner English （ICLE）（http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/abs.html）

ICE corpus のプロジェクト傘下で進められている英語学者コーパスのプロジェクト。Louvan 大学（ベルギー）の Sylvianne Granger が中心。英語を第２言語または外国語として用いる11の異なる母語の学習者データ（英語のエッセイ）を20万語ずつ集め、品詞タグ、エラータグを施す作業を進めている。
投野由紀夫（元東京学芸大学講師、現在ランカスター大学博士課程）のホームページより引用

5　* MonoConc.
Rice UniversityのMichael Barlow氏のソフト。Mac/Windows 用がある。ソートと簡単な統計値が出る。入手先は：
（http://www.ruf.rice.edu/~barlow/mono.html）
* Conc.
Mac用のフリーウェアコンコーダンサー。入手先は：
（http://www.sil.org/computing/conc/conc.html）
* LEXA.
DOS用のコーパス分析統合ツール.
tagging, lemmatization, word frequency, counts, etc. の処理が可能．入手先は：
（gopher://nora.hd.uib.no: 70 / 11 /Programs/pc/lexa）

* WordSmith
Windows 用コーパス分析ソフト。パラレルテクストの処理やタグの処理にも結構対応している。OUP が販売している。試用版から正規版を入手可能。
(http://www 1 .oup.co.uk/oup/elt/software/wsmith)

6　パーソナルコーパス構築の為の電子テキストの入手先の 1 例を次に紹介する。

Links to Electronic Book and Text Sites
(http://www.awa.com.library/omnimedia/links.html)
電子テキストがダウンロードできるサイトが一覧になっていてリンクが張られている。ここからさまざまなサイトに行ける。

Alcuin
(http://library.ncsu.edu/drabin/alcuin)
電子テキストのサーチエンジン

CNN Transcripts
(http://www.cnn.com.TRANSCRIPT/index.html)
Talk, Business, Special Interest, Politics, CNN internationalの 6 つの分野がある。Talkで話し詞のデータ、Special Interestでは環境、科学技術、旅行、健康、料理といった多様な話題のデータの入手が可能。

新聞屋さん
(http://www.threeweb.ad.jp/"harahara/index.htm)
世界の主要な新聞の記事か概要をの電子化されたテキストを入手できる。

Project Gutenberg
(http://www.cdrom.com/pub/gutenberg/)
古典的な文学作品が電子化されている。

Drew's Scripts-O-Rama
(http://home.cdsnet.net/"nikko 1 1 /scripts.htm)
電子化された映画のスクリプトが入手可能。

7　ICLE（International Corpus of Learner English）は以下のような変数とそれに対する定義をしている：* 言語に関して
○ 媒介（medium）：作文（written）
○ ジャンル（genre）：エッセイ（essay）
○ 内容(content)：賛否を書かせるようなもの；専門的でないもの
* 学習者に関して
○ レベル（level）：上級（大学 3，4 年生）
○ 母語（mother tongue）
○ 学習環境：教室内での外国語としての英語学習を前提（Classroom）

参考文献

小池生夫．1994．『第 2 言語習得研究に基づく最新の英語教育』東京：大修館書店．

斎藤俊雄，中村純作，赤野一朗．1998．『英語コーパス言語学』東京：研究社．

投野由紀夫．1997．『英語語彙習得論』東京：河源社．

投野由紀夫．1998．「学習者コーパスと学習指導」『現代英語教育』35巻，4 月号，pp. 26-29．

投野由紀夫．1998．「学習者コーパスと学習指導」『現代英語教育』36巻，5 月号，pp. 38-40．

桃井秀知．1997．『学習者コーパスによる日本人英語学習者の発達的分析-have-を例として』
(JACET英語辞書研究会口頭発表)

Dulay, H.and Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning*, 24 , pp. 37-53.

Granger, Sylviane (1993) *International Corpus of Leaner English*. Netherlands : In Papers from the thirteenth International Conference on
English Language. Research on Computerized Corpora, Nijmegen, pp. 57-71.

Granger, Sylviane (1996) *Learner English around the World* In Comparing English Worldwide
-The International Corpus of English-. New York: Oxford University Press, pp. 13-24.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.