

International Corpus of Crosslinguistic Interlanguage (ICCI) : 英語初学者データの 国際比較

投野 由紀夫

(東京外国語大学大学院総合国際学研究院)

要 旨

東京外国語大学グローバルCOEプログラムの傘下で進められている双方向英語学習者コーパス・プロジェクト International Corpus of Crosslinguistic Interlanguage (ICCI) の中間報告である。ICCI は英語初学者データの欠けを満たすために世界 7 地域 (国) の学習者コーパス研究者と共同で構築が進められている。すでに公開済みの日本人データ JEFLL Corpus との比較が可能なように英作文課題などが設定され、2010 年度末で約 6700 件、53 万語に及ぶデータが収集できた。本論ではその基本的なコーパス設計とデータのフォーマットやタグ付与に関する情報、および開発中の検索インタフェースについて紹介する。

1. はじめに

International Corpus of Crosslinguistic Interlanguage (以下 ICCI) は、東京外国語大学グローバル COE プログラム「コーパスに基づく言語学教育研究拠点」の傘下で計画された、英語初学者の英作文データを収集する国際学習者コーパス・プロジェクトである。過去 20 年間学習者コーパス研究 (learner corpus research) は飛躍的に進展してきた。1990 年代初頭、第 2 言語学習者コーパスがほぼ皆無だった状態から、Louvain 大学の Sylviane Granger の提唱により、International Corpus of Learner English (ICLE) というプロジェクトが発足し、世界 10 数カ国の英語専攻大学生による英作文コーパスが構築された (Granger et al., 2009)。同時に、ロングマンや Cambridge University Press などの出版社主導で Cambridge などの英語技能試験の自由英作文データを中心として Longman Learner's Corpus, Cambridge Learner Corpus など 1000 万語以上の大規模作文コーパスも構築が進んだ。本プロジェクトの責任者である投野も 1990 年代より英作文データのコーパス化に取り組み、1 万人の中高生の英作文データを収集した JEFLL Corpus (投野 2007)、1200 人以上の英語スピーチ・データをコーパス化した NICT JLE Corpus (和泉他 2005) などの構築に携わってきた。

世界的にも学習者コーパスは利用できるデータも充実してきてはいるが、データの特徴

を見ると決定的に欠けている点が、初学者データの分量である。これにはいくつか理由がある。第一に、大学生のデータの方が一人の学習者がまとめた分量を書けるので少数で一定量のコーパスを採取しやすい。第二に、大学生レベルでは入力がほとんどワープロまたはオンラインでの入力であるので、書き起こしの手間が不要。第三に、ICLE に代表される初期のプロジェクトが大学生中心のデータだったため、その後の比較データも大学生のものを採取した物が主流だった、という点がある。

今回の ICCI はこの初学者データを中心に収集することで、従来の学習者コーパスのデータ収集上の欠けを補うということが 1 つの大きな目的である。幸い、投野が構築した中高生 1 万件の英作文データ JEFLC Corpus はすでに公開して利用可能であり、そのデータと比較するという意味でも大きな意義がある。本論では、ほぼ完了したデータ収集の全容とコーパス・データの仕様、検索ツールの改良点などを中心にその成果を報告する。

2. コーパス・データの概要

2.1. サブコーパス構成

2011 年 3 月現在で収集した ICCI データの内訳を表 1 に示す：

表 1 ICCI データの内訳

国 \ 学年	3	4	5	6	7	8	9	10	11	12	合計
オーストリア	0	0	60	148	167	139	118	69	68	0	769
中国	0	0	0	81	494	139	137	102	0	0	953
香港	108	28	23	306	161	56	29	32	36	0	779
イスラエル	0	0	99	556	200	290	282	259	201	148	2035
ポーランド	0	18	76	84	87	101	101	99	100	85	751
スペイン	0	0	23	30	134	114	164	107	88	77	737
台湾	0	0	0	0	100	117	89	72	193	104	675
学年合計	111	50	286	1211	1350	964	929	750	697	426	6699

データは学年不明の台湾データ 12 件を含めて、6,711 ファイル、総語数は 534,856 語である¹。イスラエルのデータだけは現地で電子化したものを送ってもらっているため総ファイル数も多いが、その他はすべて手書きの作文データを日本に送付してもらい、東京外国語大学において学生アルバイトなどを用いて人手で書き起こしを行った。ポーランドも 2000 件を超える規模のデータを送付してもらっているが、2010 年度の書き起こしの規模的に他とだいたいそろえて 750 件程度としてある。内訳を見ていただければわかるとおり、必ずしも学年毎に均一ではないが、各地域の英語教育の現状にある程度即してデータ採取を行っ

¹ 数字は研究メンバーでデータ処理担当の Huaqing Hong 氏のレポートによる。

たため、一定の制約があったことは致し方ない。それでも6年生～11年生くらいまでは比較的均等にデータが集まっている。

2.2. タスク別内訳

英作文課題に関してはできるだけ課題の統一を心がけた。JEFLC Corpus の英作文データとの相互比較の可能性を考えて、作文データには共通のトピックおよび各国の教育事情に合わせた課題の両方を盛り込むようにした。

各国のトピック別語数は表2の通りである：

表2 ICCI トピック & 国別テキスト語数内訳

国	トピック	学年	3	4	5	6	7	8	9	10	11	12	学年不明	総語数
台湾	food					1665	4253	6683	1803	8176	10610			33190
	money					9302	4572	398	4767	11235	278			30552
	総語数					10987	8825	7081	6570	19411	10888			63742
香港	food	1904		361	10748	8407		1662		1073			717	24872
	money	1950	2584	318	12367	7150	5933	1738	3294	3525			515	39374
	総語数	3854	2584	679	23115	15557	5933	3400	3294	4598			1232	64248
中国	birthday				1188	7808	4010	127	5152					18263
	food				2733	5298	3594	3510	2444					17579
	film				170		1075	3438	355					5038
	funny story				458	4296	6270	1679	1418					14121
	money					4100	2164	7691	3554					17509
	school					7641								7641
	総語数				4527	29143	17113	16445	12923					80151
ポーランド	film			887	922	2228	3463	4560	3318	543	10275			26194
	food		998	1743	1929	1728	2270	1548	2605	5625	1719			20165
	money			46	1594	68	230	603	2949	3937	209			9638
	総語数		998	2678	4445	4022	5963	6711	8872	10105	12203			55995
スペイン	film			267	590	2851	10668	3949	6776	6617	8265			39983
	food			153	330		111	575	1296					2465
	money			130	161	2723	3455	4358	979	2254	1638			15694
	総語数			550	1081	5574	14234	8880	9051	8871	9901			58142
オーストラリア	after school				625	2208	1326		571					4730
	food				1645	2960	8185	1369	79	152				14390
	important event									312				312
	money				5372	2233	2357	2631	5684	11300				28577
	strange thing						3389		971	192				4552
	describing yourself			2141	3172									5313
	postcard			2589	1780	1014								5393
	exciting dream					6534	4313							10847
	good or bad day					10963	3525	12463	3519					30470
	eating out						888	2769	1280					4947
	総語数			4730	12604	25912	23993	19232	12104	11956				110531
イスラエル	money			837	6408	2923	6653	5719	8832	4532	4029	1233		39933
	funny story				4515	586	4761	1759	4761	1483		833		17865
	breakfast picture				3434	592	3078	1134	3725	502		414		12465
	schoolyard picture				25									25
	event or funny story			1136	1404	3740	4458	5505	194	3820	5017			25072
	food			1589	3838	3864	4769	5735	3429	4744	5647			33615
	eating out			30										30
	after school			37	37									74
	birthday				106									106
	総語数			3735	19661	11705	23717	19852	20941	14881	14893	2480		129185

学年によりトピックに多少の変動があるが、全体として JEFLC Corpus と比較可能な題材として、好きな食べ物または普段の食事に関するテーマ（表中の food）およびお小遣いをまとめた額もらったら何に使うか、といったお金に関するテーマ（表中の money）の2つ

は最低トピックとして統一するようにした。この2つはエッセイのタイプの的にも前者の food の方はどちらかというとなりて・記述 (narrative-descriptive) なもの中心になるが、後者の money の方は欲しい物とその理由という論説的 (argumentative) な傾向の強いものとなると期待された。

タスクは全体的には JEFLL Corpus と比較可能なように統一された。すなわち、英語授業内での20分間の辞書を使わないで書く自由英作文、という課題である。準備の時間はなく、終了後すぐに回収された。JEFLL Corpus の場合には中学1年生など英語を習い始めたばかりの学習者もある程度書けるように、模範英文をタスク指示文中に用意したが、ICCI では模範英文の影響が強く出てしまうことを懸念して、タスクには例は載せなかった。各地域ともほぼ同様のタスクの与え方で実施をしたが、次節で述べるように、学習者のメタデータの収集に関しては学年や学校の事情によっても異なり、できる範囲で詳しい情報を集める、というのに留まった。

3. コーパスへの情報付与

本節では収集した学習者の英作文データに付与した情報とその方法を紹介する。コーパス・データではテキストの外的情報 (メタ情報) をマークアップ (mark-up)、テキスト内の言語情報をアノテーション (annotation) と区別することが一般的である。ICCI では表3に挙げるようなメタ情報を学習者変数として操作できるように付与している：

表3 ICCI のメタ情報

情報の種類	コード	値 (例)	必須項目
国	country	Austria	X
学校種	school	secondary school	X
学校名	schoolname	GRG 13 Wenzgasse	
学校レベル	schoollevel	high	
海外経験	overseas	1 (=1 year)	X
週英語授業	classhour	4 (class hour)	X
課外学習	extralesson	2 (hours)	X
教科書	textbook	教科書名を書く	
学年	year	5	X
クラス	class	1F	
学生番号	studentid		
氏名	name		
性別	sex	male	
母語	mothertongue	German	X
英語試験	otherscore		
媒体	medium	written	X
ジャンル	genre	descriptive	X
トピック	topic	describing yourself	X
辞書使用	dicuse	YES/NO	X

準備	preparation	YES/NO	X
時間	time	20min	X
フィードバック	feedback	YES/NO	
日付	date		
書き起こし者	transcriber	氏名	X

これらのメタ情報は最も初期の書き起こし時点では、書き起こし作業者が Web の書き起こしツールを利用して本文を入力し、入力結果を送信すると、メタデータが付与された形式で保存される。それが図 1 のような形式である：

```

<file id="20825">
<head>
<cat id="20825">
  country='Austria'
  school='Secondary school'
  schoolname='Rahlgasse'
  schoollevel=""
  overseas='0'
  classhour=""
  extralesson=""
  region=""
  textbook=""
  year='8'
  class='4'
  studentid="" name=""
  sex='female'
  mothertongue=' German'
  otherscore=""
  medium='written'
  genre='descriptive'
  topic='good or bad day'
  dicuse=""
  preparation=""
  time='20min'
  feedback=""
  date=""
  transcriber='Antoniya Mangova'
</cat>
</head>
<body>
<s>The day was last year on the 2nd November.</s>
<s>One week before, my mother drove to "Graz" because she wanted to sell her pictures.</s>
<s>She stayed at a little farm from an old "bekannter".</s>
<s>I had a cat and the cat was sad because she was all time alone.</s>
<s>So we thought we should buy another.</s>
<s>So on this farm, there was a little cat.</s>
<s>The cat's mother was dead and a dog looked up at her.</s>
<s>When my mother saw this tiny cat, she had the idea to take it with her.</s>
<s>She phoned me, and felt very happy.</s>

```

```
<s>And on the 2nd November I saw her the first time.</s>
<s>Every day I love her more.</s>
<s>She's the best little cat.</s>
</body>
</file>
```

図1 書き起こし作業ファイルの初期状態

この初期ファイルをもとに、検索システム別に複数のファイル・セットを作成した。

- (a) 市販のデスクトップ使用のコンコーダンサー用テキスト・ファイル
- (b) Xaira での利用を考えた XML フォーマットとバイナリー・データ
- (c) MySQL + web 検索システム形式の XML フォーマット
- (d) Corpus Encoding Standard for XML (XCES) 形式フォーマット
- (e) Wmatrix で処理した品詞+意味タグ付与フォーマット

(a) は通常のコンコーダンサーによる使用を考えたもので、ヘッダ部分はエクセル形式による別ファイルで用意し、必要なファイルを選択してコンコーダンサーに読み込んで使用する。AntConc, MonoConc Pro, WordSmith などでの汎用的な使用を考慮に入れている。

(b) の Xaira は XML フォーマットの言語コーパス向けの汎用コンコーダンサーで、British National Corpus XML Edition 用に開発された Windows 検索ソフトの 1 つである。データ本体をサーバーに置いて検索できるため比較的パソコンそのものへの負荷が低くて済む。XML の検索は非常に複雑な条件検索が可能である一方、XML 構造を十分理解している人が使わないと上手に検索するのは難しい。Xaira のフォーマットでは個々の単語に付与された品詞タグ、見出し語情報、意味タグなどを複合的にからめた検索が可能。

(c) の MySQL + php または java などを用いた web 検索システムは web 上でデータベースを扱う汎用的な形式で、最近比較的大規模コーパスの web インタフェースでよく使用されている組み合わせである。一般コーパスで有名なものでは Mark Davies の Corpus of Contemporary American English (COCA)², Bill Fletcher の Phrases In English (PIE)³などがこの形式でデータ管理を行っている。

(d) は Corpus Encoding Standard for XML (XCES) という XML 標準のコーパス格納方式でフォーマットされたもの。EAGLES という言語データの配布形式の標準化プロジェクトでの成果を受けて Vassar 大学の Nancy Ide が中心で設計したもの。American National Corpus の中間リリースのものはこのフォーマットになっている。ただし大規模データでなければ構造が複雑すぎて利用が困難である。今回は一般配布用にこのフォーマットのものも準備した (図2 参照)。

(e) はランカスター大学 UCREL で開発している Wmatrix によるアノテーション・データ、XML フォーマットで各単語に品詞、意味タグを付与した形式のもの (図3 参照)。この形

² <http://www.americancorpus.org/>

³ <http://pie.usna.edu/>

式を先の (c) など加工することができる。

```
<?xml version="1.0" encoding="utf-8"?>
<Doc>
<Header>
<TrsName>icci0043</TrsName>
<Country>Austria</Country>
<Grade>5</Grade>
<Gender>Male</Gender>
<Mother_Tongue>German</Mother_Tongue>
<Genre>Descriptive</Genre>
<Topic>Postcard</Topic>
<Transcript_Word_Count>43</Transcript_Word_Count>
<FileName>icci0043</FileName>
</Header>
<Text>
<T id="t_1" turn_number="1" speaker_name="1" gender="unknown" start_time="0.0"
end_time="0.0" ><w pos="NP1">Peter</w> <w pos="CC">and</w> <w pos="APPG">his</w>
<w pos="NN1">family</w> <w pos="VBR">are</w> <w pos="II">in</w> <w
pos="NP1">London</w> <w pos="UKN">.</w> </T>
<T id="t_2" turn_number="2" speaker_name="1" gender="unknown" start_time="0.0"
end_time="0.0" ><w pos="PPHS2">They</w> <w pos="VV0">go</w> <w pos="II">to</w> <w
pos="AT">the</w> <w pos="NN1">sea</w> <w pos="CC">and</w> <w pos="VV0">hire</w>
<w pos="AT1">a</w> <w pos="NN1">leaved</w> </T>
<T id="t_3" turn_number="3" speaker_name="1" gender="unknown" start_time="0.0"
end_time="0.0" ><w pos="PPHS2">They</w> <w pos="VV0">go</w> <w pos="RP">on</w>
<w pos="II">to</w> <w pos="AT">the</w> <w pos="NN1">island</w> <w pos="CC">and</w>
<w pos="VV0">make</w> <w pos="AT1">a</w> <w pos="NN1">fire</w> <w
pos="UKN">.</w> </T>
<T id="t_4" turn_number="4" speaker_name="1" gender="unknown" start_time="0.0"
end_time="0.0" ><w pos="RT">Then</w> <w pos="PPHS2">they</w> <w
pos="VVZ">walks</w> <w pos="II">around</w> <w pos="AT">the</w> <w
pos="NN1">Irohle</w> <w pos="NN1">island</w> <w pos="UKN">.</w> </T>
<T id="t_5" turn_number="5" speaker_name="1" gender="unknown" start_time="0.0"
end_time="0.0" ><w pos="RT">Then</w> <w pos="PPHS2">they</w> <w pos="VV0">go</w>
<w pos="II">to</w> <w pos="AT">the</w> <w pos="NN1">bout</w> <w pos="CC">and</w>
<w pos="VV0">tahr</w> <w pos="RP">back</w> <w pos="RL">home</w> <w
pos="UKN">.</w> </T>
</Text>
</Doc>
```

図 2 XCES 形式のファイル

```
<w id="1381.1" pos="FO" sem="Z99">Trn0001</w>
<w id="1381.2" pos="FO" sem="Z99">spk1</w>
<w id="1381.3" pos="APPGE" sem="Z8">My</w>
<w id="1381.4" pos="NN1" sem="E2++">favourite</w>
<w id="1381.5" pos="VBZ" sem="A3+">is</w>
<w id="1381.6" pos="NN2" sem="Z99">jiaozi</w>
<w id="1381.7" pos="." sem="PUNC">.</w>
<w id="1382.1" pos="FO" sem="Z99">Trn0002</w>
<w id="1382.2" pos="FO" sem="Z99">spk1</w>
<w id="1382.3" pos="CS" sem="Z5/A2.2">Because</w>
<w id="1382.4" pos="PPH1" sem="Z8">it</w>
<w id="1382.5" pos="VBZ" sem="A3+">'s</w>
<w id="1382.6" pos="MC1" sem="T3">one</w>
<w id="1382.7" pos="IO" sem="Z5">of</w>
<w id="1382.8" pos="AT" sem="Z5">the</w>
<w id="1382.9" pos="RGT" sem="A13.2">most</w>
<w id="1382.10" pos="JJ" sem="E2+">popular</w>
<w id="1382.11" pos="NN1" sem="F1">food</w>
<w id="1382.12" pos="II" sem="Z5">in</w>
<w id="1382.13" pos="NP1" sem="Z2">China</w>
<w id="1382.14" pos="." sem="PUNC">.</w>
<w id="1383.1" pos="FO" sem="Z99">Trn0003</w>
<w id="1383.2" pos="FO" sem="Z99">spk1</w>
<w id="1383.3" pos="II" sem="Z5">On</w>
<w id="1383.4" pos="AT" sem="Z5">the</w>
<w id="1383.5" pos="NNT1" sem="T1.3-/N4">eve</w>
<w id="1383.6" pos="IO" sem="Z5">of</w>
```

図3 Wmatrix 形式のファイル (一部)

4. 検索インタフェース

ICCI プロジェクトとして、検索インタフェースとしては2種類を開発している。1つは、COE プログラム全体でのデータ提供用に開発した簡易コンコーダンサー。もう1つは内部の研究チーム専用のインタフェースでより複雑な検索が可能なもの。この2つをそれぞれ以下に解説する。

4.1. 暫定公開用検索プログラム

これはグローバル COE プログラムのサイトから ICCI プロジェクトを選択すると実際に検索サイトが一般公開されている (図4 参照) :

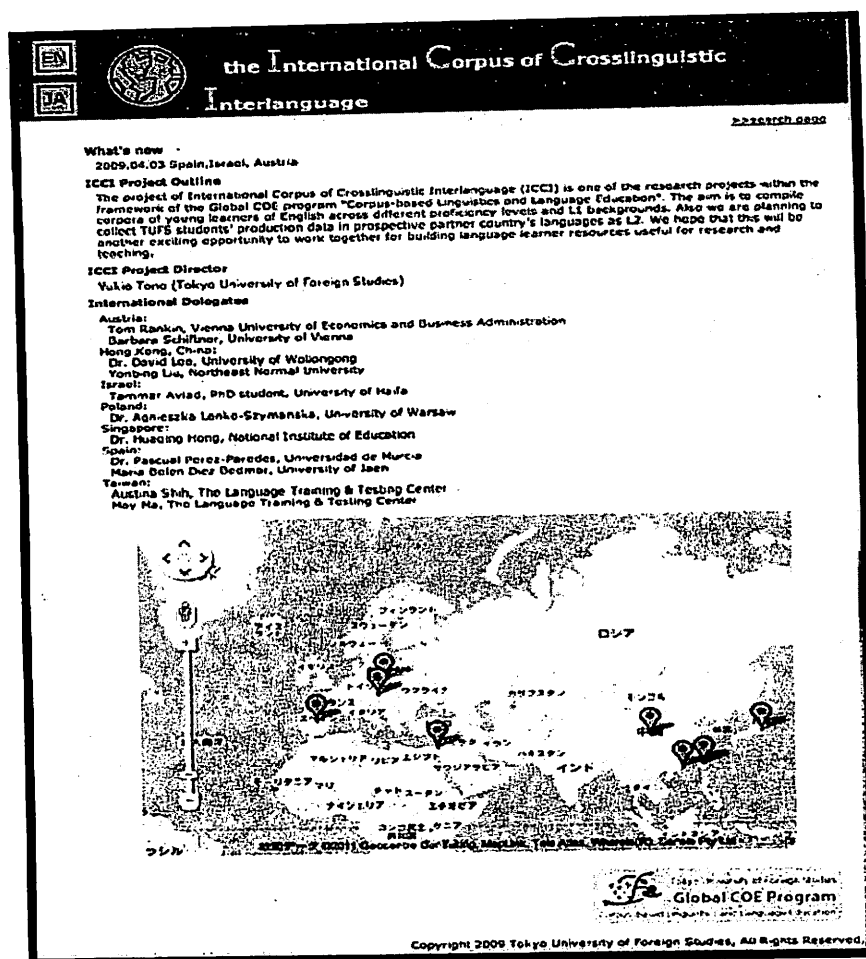


図 4 ICCI 公式 Web サイト (<http://cblle.tufs.ac.jp/llc/icci/>)

ここから右上の search page をクリックすると以下のような画面が開く (図 5) :

the International Corpus of Crosslinguistic Interlanguage

search word* country grade mothertongue topic

Search

* This is required field.

Update History

- 2010/03/31 The composition data added. (The number of composition data:523)
- 2009/04/03 Learner's Language Corpus of English released. (The number of composition data:59)

図 5 公式サイト検索画面 1

ここで検索語および国, 学年, 母語, 作文テーマなどの絞り込み検索をかけることができる。単純に buy という単語を検索した結果を図 6 に示す :

the International Corpus of Crosslinguistic Interlanguage

search word* country grade mothertongue topic

buy
* This is required field.

☐ : show author information

Or I buy clothes with the money, then I can spend my old clothes

We eat there and buy a lot of T-Shirts, jeans and bags.

I love to buy shoes.

If I'd become 100E I'd buy lots of books.

I will buy some clothes and I will go to cinema.

the other 50 euros I would go shopping with a friend and buy new cloths and in the night I would go out and have film.

because it would be the first time that people have not to buy for my drinks.

Now I would have money and I could buy for my own.

I would buy a computer game for 30 euros.

I would buy a new mobile phone or a ccam.

In my school I often buy sweets.

I often buy for me a Kebap.

Then we buy our food, then we eat it.

I also save some money because I would like to buy me a new mountainbike.

My old mountainbike is so small and so I will buy me a new mountainbike.

図6 公式サイト検索画面2 (検索結果の表示)

各例文の左にある緑の矢印ボタンを押すと各作文のファイル情報が簡潔に表示される：

the International Corpus of Crosslinguistic Interlanguage

search word* country grade mothertongue topic

buy
* This is required field.

☐ : show author information

filename	country	grade	mothertongue	topic
astr-00001.xml	Austria	7	German	money

Or I buy clothes with the money, then I can spend my old clothes

We eat there and buy a lot of T-Shirts, jeans and bags.

I love to buy shoes.

If I'd become 100E I'd buy lots of books.

I will buy some clothes and I will go to cinema.

the other 50 euros I would go shopping with a friend and buy new cloths and in the night I would go out and have film.

because it would be the first time that people have not to buy for my drinks.

Now I would have money and I could buy for my own.

I would buy a computer game for 30 euros.

I would buy a new mobile phone or a ccam.

In my school I often buy sweets.

I often buy for me a Kebap.

Then we buy our food, then we eat it.

図7 公式サイト検索結果画面3 (ファイル情報の表示)

現状ではサンプリングされた 523 ファイルがデモ版として利用可能。今後、公式サイトでの検索機能を向上させて、プロジェクト終了時点では全データを公開する予定。

4.2. 研究メンバー用検索プログラム

研究メンバー用にはシンガポール国立教育研究所研究員の Huaqing Hong 氏の協力を得て、専用の web 検索ツールの開発を行っている。検索プログラムのトップ画面を図 8 に示す。インタフェースは直感的で、一番上のボックスがそれぞれ国、学年、性別、母語、作文ジャンル、作文トピックの 6 種の学習者プロフィール変数に対応しており、これらを柔軟に組み合わせてサブコーパスを絞り込み、Query Item で検索式を指定して検索を行う。

検索の際の変数内のリストの選択は複数項目の選択が可能なので、中国と香港の小学校レベル（3 年～6 年）のみを調べる、といった柔軟なサブコーパス選択ができる。トピックに関しては前述のように、JEFLC Corpus との比較のために food, money という 2 つの課題がメインであるが、それ以外は各国の研究者の調査目的に応じて比較的タスクに忠実なタイトルのまま作文を選択できるようにしてある。

Country	Grade	Gender	Mother Tongue	Genre	Topic
Austria	3	Male	Albanian	Descriptive	After School
China	4	Female	Amharic	Argument	Birthday
Hong Kong	5	Unknown	Arabic		Breakfast Picture
Israel	6		Armenian		Describing Yourself
Poland	7		Bengali		Eating Out

Query Item a lot of =
 Insert POS: _____
 [Reset All] [Search] [Home]

How to use this application

Step1
 Select the categories you want in the top right panel, and input the word/phrase you want to search, and then click on the search button.

Step2
 The phrase results will be shown on the left panel.

Step3
 Click on the phrase you are interested in, the concordance results will be shown on the right bottom panel.

図 8 研究メンバー用検索プログラム 1（メイン画面）

検索フィールドは現状ではきわめて簡単な検索式のみ限定されている。検索したい単語と前後の 1 語を * で指定することによりそのフレーズの頻度表を左側に出力してくれる。たとえば, going to の後にどのような単語が来るかを知りたい場合, “going to *” と入力してやると図 9 のような出力結果になる:

The International Corpus of Crosslinguistic Interlanguage

You are searching "going to *" in 6711 files (534856 words in total), and the results are as follows:

Country	Grade	Gender	Mother_Tongue	Genre	Topic
Austria	3	Male	Abanian	Descriptive	After School
China	4	Female	Arabic	Argument	Birthday
Hong Kong	5	Unknown	Arabic		Breakfast Picture
Israel	6		Armenian		Describing Yourself
Poland	7		Bengali		Eating Out

Query Item: going to *
Insert POS: _____
Reset All Search Home

Sort: context 0 order Ascending View Statistics Page 1 of 2 pages Go to Page 1 Go

No	Left Context	Query Item	Right Context	Transcripts
1	finally mom and dad said that we were	going to buy	this dog! This was the best day	cc0324
2	, my father told me that he is	going to buy	a new bike for me. After we	cc0363
3	gave me. With the money I am	going to buy	a playstation 2 and lots of games.	cc0363
4	buy with the 2000. - I am	going to buy	a new very cool bike, which I	cc0366
5	other 1,000, I think, I'm	going to buy	lots of little things. Something like new	cc0361
6	to sure about it. Maybe I'm	going to buy	a new gaming console and a few great	cc0361
7	. Summer has already started so I'm	going to buy	me some new clothes and a ticket to	cc0361
8	a lot of money, so I'm	going to buy	thinks, which I've for a long	cc0361
9	. So I tell you what I'm	going to buy	. You know that I'm very interested	cc0362
10	recording equipment. Because of this I'm	going to buy	a western guitar with twelve strings. When	cc0362
11	guitar with twelve strings. When I'm	going to buy	a good one than I have 1400 for	cc0362
12	for recording my songs. I am also	going to buy	a good microphone. When I have all	cc0362
13	: Well, first of all I'm	going to buy	lots of sweets! You know I love	cc0362
14	these unhealthy things... Then I'm probably	going to buy	this new and awesome computer game! You	cc0362
15	. A second thing, which I'm	going to buy	for the money, is a new soccer	cc0366
16	really want to buy. Tomorrow I am	going to buy	the green motorcycle, I showed you a	cc0365
17	save so much money. I think I am	going to buy	a new computer and maybe PS3 and some	cc0365

図9 研究メンバー用検索プログラム2 (“going to *” 検索結果画面)

going to * でアスタリスクに当たる部分の単語のヴァリエーションを頻度表にする, という機能である。そして左側の頻度表からクリックしてやると図9のように KWIC 画面を見ることができる。

検索フィールドのすぐ下にある [Insert POS] というフィールドは検索条件に品詞タグを利用したい場合に用いる。これを用いるとたとえば, 以下のような検索式も簡単に書ける:

動詞の-ing 形 + to + 動詞の原形
[VVG] to [VV*]

結果は図10のようになる:

The International Corpus of Crosslinguistic Interlanguage

You are searching "[VVG] to [VV*]" in 6711 files (534856 words in total), and the results are as follows:

Country	Grade	Gender	Mother_Tongue	Genre	Topic
Austria	3	Male	Abanian	Descriptive	After School
China	4	Female	Arabic	Argument	Birthday
Hong Kong	5	Unknown	Arabic		Breakfast Picture
Israel	6		Armenian		Describing Yourself
Poland	7		Bengali		Eating Out

Query Item: [VVG] to [VV*]
Insert POS: ing participle of I
Reset All Search Home

Sort: context 0 order Ascending View Statistics Page 1 of 1 pages Go to Page 1 Go

No	Left Context	Query Item	Right Context	Transcripts
1	've for a long time. I'm	planning to buy	a new drum-kit or a computer. I	cc0362
2	the money you have cont. I'm	planning to buy	a lot of items including sport equipment and	cc0366
3	my training at home. Additionally I am	planning to buy	a new mobile phone and a brand new	cc0366
4	the money from my family and I am	planning to buy	with some cool stuff for my room.	cc0367
5	done anything with it yet though I'm	planning to buy	a new case for my PC with it	cc0367
6	buy everything what I need. I'm	planning to buy	new mobile phone and give some money from	cc0367

図10 研究メンバー用検索プログラム3 (品詞を検索式に使用した例)

この場合は be going to の going は連結動詞 (linking verb) の役割をするので異なる品詞タグ (VVGK) になるからリストには現れてこない。連結動詞と一般動詞の両方を取ってきたい場合には [VVG/VVGK] または [VVG*] と書くことで OR 検索が可能である。

検索結果画面はさらにサブコーパスによる詳細な頻度分布を一覧することができる。たとえば先ほどの going to V というパターンを検索すると 393 件あり、その内訳を View Statistics というボタンをクリックすると図 11 のような画面が出てくる。画面はサブコーパスの選択により分類が異なってくるが、大別すると、①国別、②学年別、③性別、④母語別、⑤ジャンル別、⑥トピック別の頻度分布が表になっている。

表中のサブコーパス頻度画面は以下のような統計値からなっている：

(1) Hits	該当項目のサブコーパス全体での出現頻度 (粗頻度)
(2) Hits in Files	当該項目の出現したファイル数
(3) HF_Tokens	(2) のファイル合計での総語数
(4) Files of this Type	サブコーパス全体のファイル数
(5) FT_Tokens	サブコーパス全体の総語数

この数値を用いれば、一般的な正規化頻度 $((1) \div (5))$ のような計算だけでなく、使用した個別の学習者グループ単位の分析 $((1) \div (3))$ のような比較も可能になる。

現状では、1つの検索語 (式) に関してサブコーパスでの統計を簡単に得られる、という点を重視して設計をしている。この方法ならば、言語特徴 A と B の違いの基本頻度データを容易に得ることができ、その後の統計処理に注力すればよい。また検索語や検索パターンは一通りであっても、サブコーパスを初期画面で上手に選択しておけば、その後の統計分布でいちいちサブコーパス集計しなくても簡単に内訳をまとめることが可能になる。検索式の作り方とその後のサブコーパス分析の観点をあらかじめ考えておけば、有用なデータを比較的短時間に抽出することが可能になると言う点がこのような学習者属性が複雑なデータに向けたインタフェースであると言えよう。

現状では意味タグやレマ検索などができない (ただ品詞タグの OR 検索で代用が可能) ので、今後はそのへんの改良を加えることと、検索式で頻度リストを左側に出すのはよいのだが、検索結果が個別に出過ぎてしまい、頻度分布を集計する際にもう少し言語特徴をまとめたい、というニーズがありうる。こういった点に関する対応は今後チームでツールを使いこなしながらフィードバックをして改善するということになるだろう。

なお、現在はこのインタフェースはパスワード付でシンガポールの Huaqing Hong 氏の運営するサーバーで稼働しているが、最終的には検索系を改良したバージョンを東京外国語大学の投野研究室サーバー上で動作させ、時期を見てこのインタフェースも一般公開する予定である。

The International Corpus of Crosslinguistic Interlanguage

You are searching "going to" in 6711 files (534856 words in total), and the results are as follows:

Items
going to
going to

Hits
393
393

Country	Grade	Gender	Mother_Tongue	Genre	Topic
Austria	3	Male	Albanian	Descriptive	After School
China	4	Female	Amharic	Argument	Birthday
Hong Kong	5	Unknown	Arabic		Breakfast Picture
Israel	6		Armenian		Describing Yourself
Poland	7		Bengali		Eating Out

Query Item going to
Insert POS:
Reset All Search Home

☐ View Concordances

Country	Hits	Hits in Files (HF)	HF_Tokens	Files_of_this_Type (FT)	FT_Tokens
Austria	97	74	10254	769	34691
China	49	29	3124	954	22094
Hong Kong	25	19	2524	791	20490
Israel	127	103	6590	2035	21103
Poland	25	21	2616	751	19865
Spain	63	40	3752	737	16716
Taiwan	7	7	619	674	15328

Grade	Hits	Hits in Files (HF)	HF_Tokens	Files_of_this_Type (FT)	FT_Tokens
10	70	58	5283	740	20957
11	40	35	4044	686	21970
12	22	18	2666	414	23161
3	2	2	136	108	2053
4	2	2	224	46	3310
5	5	5	214	281	5083
6	38	35	2028	1205	12977
7	41	28	2906	1343	21727
8	94	59	5761	956	23582
9	79	51	6748	920	26983

Gender	Hits	Hits in Files (HF)	HF_Tokens	Files_of_this_Type (FT)	FT_Tokens
Female	121	96	11429	1733	33460
Male	112	89	5917	1714	23587
Unknown	160	108	10504	3264	31633

Mother_Tongue	Hits	Hits in Files (HF)	HF_Tokens	Files_of_this_Type (FT)	FT_Tokens
Bosnian	1	1	162	4	450
Chinese	82	56	6188	2425	27013
Croatian	2	1	89	9	1112
German	80	59	8596	609	32464
Greek	1	1	109	2	172
Hebrew	117	94	5783	1874	19943
Polish	30	26	3033	773	19865
Portuguese	2	1	142	3	334
Russian	9	8	802	141	8052
Slovakian	1	1	129	3	402
Spanish	63	40	3752	742	17017
Swedish	1	1	186	1	186
Turkish	3	3	483	8	1038
Unknown	1	1	54	2	101

Genre	Hits	Hits in Files (HF)	HF_Tokens	Files_of_this_Type (FT)	FT_Tokens
Argumentative	206	147	11557	2494	28583
Descriptive	187	146	13934	4217	40045

Topic	Hits	Hits in Files (HF)	HF_Tokens	Files_of_this_Type (FT)	FT_Tokens
Birthday	13	10	1209	200	11271
Breakfast Picture	33	23	1302	272	5977
Describing Yourself	1	1	60	69	3768
Eating Out	3	3	398	43	4262

図 11 研究メンバー用検索プログラム 4 (サブコーパス統計の出力画面)

5. 今後の見通し

ICCIプロジェクトは発足4年目で6000名以上の初学者データのデータ収集をほぼ完了した。そういった意味では、英語初級学習者の貴重な資源として国際的に有効なコーパス・データとなるであろう。3月に実施するはずであった G-COE 学習者コーパス国際会議は東北・関東大震災の影響で中止になった。今後は早期に研究チームによる各国の研究成果の集約と ICCI のデータならでは、というような分析成果と発表を行いたい。G-COE のシリーズの一環で John Benjamins から論文集の刊行を予定通り行うことにしている。可能であれば、それらを踏まえて夏から秋にかけて ICCI プロジェクトのメンバーだけでも再度招集し、研究成果を共有したいと願っている。

引用文献

- 和泉絵美, 井佐原仁, 内元清貴 (編著) (2005) 『日本人 1200 人の英語スピーキングコーパス』 東京: アルク.
- 投野由紀夫 (編著) (2007) 『日本人中高生一万人の英語コーパス - 中高生が書く英文の実態とその分析』 東京: 小学館.
- Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (eds.) (2009) International Corpus of Learner English v2. Louvain-la-Neuve: Presses universitaires de Louvain.