日英の理工系口頭発表コーパスの構築と検索サイト JECPRESE

林　洋子（大阪大学国際教育交流センター）
国吉 ニルソン（早稲田大学理工学術院）
野口 ジュディ（武庫川女子大学）
東條　加寿子 （大阪女学院大学）

Building a Japanese-English Corpus of Presentations in Science and Engineering and JECPRESE

Hiroko Hayashi (Center for International Education and Exchange, Osaka University)
Nílson Kunioshi (School of Creative Science and Engineering, Waseda University)
Judy Noguchi (School of Pharmacy and Pharmaceutical Sciences, Mukogawa Women's University)
Kazuko Tojo (Department of International & English Interdisciplinary Studies, Osaka Jogakuin College)

１．はじめに

現代日本語のコーパスは、書き言葉のみならず話し言葉についても整備されつつある。しかし、留学生を含む大学院の学生が最も必要とする修士論文口頭発表のデータは未だ明らかにされていないのが現状である。これは、知的財産権の問題があること、および学位取得のための口頭発表が専門分野に属するものとされ言語教育の対象とは思われてこなかったことなどによると考えられる。

そこで、我々は許可を得て修士論文口頭発表のデータを収集し、英語の発表と比較検討する研究に取り組んでいる。本稿ではデータ収集の歩み、およびデータを載せた検索サイト JECPRESE の開発、およびデータの解析結果の一部について概観する。

２．データ収集の歩み

我々は日本語、英語、化学の研究者の混成チームであり、本稿に述べるように 400 近くの発表データを含む理工系のプレゼンテーション・コーパスを構築することができた。どのようにそれが可能であったかと、理系の研究者との共同研究を願っている日本語の研究者から問われる機会が多い。そこで、本稿では、データ収集の歩みについて詳細に述べることにする。

1991 年当時、大阪大学には留学生のための日本語教室はなかった。そこで、工学研究科に日本語クラスが設置されることになり、偶然にも林が担当することになった。その時の留学生相談室の担当教員は「ぜひ理系という特色を生かした日本語教育を」と熱意に燃えていたが、日本語が話せない留学生と漢字能力が高い中韓の留学生が混在するクラスで、実験等で休みがちの留学生に、研究できるまでの日本語力がつくよう指導することは難しかった。試行錯誤を続けたが、特にアジア圏の留学生から「第二次世界大戦で焼け野原になったのに、欧米に飲み込まれずにトップクラスの科学技術レベルを持つにいたったのはなぜか、を知りたいため日本に留学した、それを可能にした日本の文化についても学びたい」という声が聞かれた。そのため、林が担当した日本語クラスでは、「豊かさとは何か、コミュニケーション、フェミニズム」など社会科学的トピックについてのディスカッションを通じて語彙・表現を増やすという手法をとった。（使用したテキスト「日本語で考える」は工学研究科によって発行されている。）ディスカッションには、理工系で必要な語彙・表現を使用するとしたが、「理系の日本語語彙」についての参考文献はなかったため、林は工学研究科の教員を対象として語彙・表現についてのアンケート調査を行った。得ら

れた調査結果はアンケートで恣意的である可能性もあるため、網羅性のあるデータ収集方法を模索し、各専攻公聴会の聴講を重ねた。当時の工学研究科の教官の反応は「日本語の先生には専門の語彙・表現は難しいでしょう」というものであった。個人的に論文や口頭発表の入手も試みたが、大規模なコーパス構築は難しかった。その時、偶然に、ある大学の農学部卒業論文発表の録音を入手・文字化し、序論部を解析した論文を出すことができた。それを旧知の教官に見せ「頂いたデータはこのような形で公表するため専攻に迷惑をかけることはない」とデータ収集をお願いした。その教官は専攻の許可をとってくださったが「データは序論部に限る」という限定付きであった。林は「論文すべてのデータが必要」と主張し、許可を得るため専攻の教授会に出席し必要性を訴えることになった。幸いにも許可が得られ、データの解析結果を論文で発表した。その論文を他専攻の教官にお見せしその専攻でのデータ収集の許可をお願いした。

このようなやり取りの中で、林は個人ではなくチームでコーパス構築し・解析する研究の必要性を感じ、日本語の関係者に共同研究を申し入れたが賛同はえられなかった。当時、工学教育ではますます英語教育が重視されるようになっており、「理工系では英語で指導するため専門日本語教育は不必要」と考えている日本語教員も多かった。

林は工学英語（野口）を聴講させていただき、ESP (English for Specific Purposes)の考え方に触れ、コーパスの重要性を理解していた野口と、英語による化学教育を始めようとしていた国吉とチームを組むことになった。その後、国吉が早稲田大学に転出することになり、さらに東條（英語教育）がチームに加わった。

英語のデータ収集はチームの初めからの目標で、野口を中心にかなり働きかけをした。しかし、欧米では許可を取ることは極めて難しいことがわかった。その後アメリカの大学生の発表のデータが得られたが、日本の修士論文口頭発表とはかなり異なっていた。国際学会のデータは国吉の働きかけで得られた。日英語の比較については後に発表する予定である。

データを収集するにあたっては各専攻・発表者と厳しい契約を取り交わした。その内容はおおむね以下のようである。

1. 序論のみならず内容までに関する公聴会の聴講は許可する。
2. 公聴会での記録はテープレコーダーのみ許可する。
3. 学生発表原稿の内容は発表前日のものとする。学生によっては、手書きの学生、ワープロで作成する学生、原稿を必要しない学生、等々であるが、なるべく学生にはワープロ作成するように指導する。しかし、学生全員の原稿の収集は不可能である。
4. 公聴会において専攻内で配布する「発表内容概要（集）」は提供する。
5. 上記の研究内容に関する記録（2)、原稿（3)、概要集（4）等を第 3 者に閲覧させること及びそれらのコピーを第 3 者に配布することを厳禁とする。
6. この調査で得られたデータは加工して語彙・表現の集計として公表する。
7. 各研究内容について公表することなく，著作権・特許権等に触れることはしない
8. 調査結果を公表するに際しては事前に専攻の許可を得る。

得られたデータは録音・原稿・発表のパワーポイントなど様々である。契約でチームの 4 人以外はデータにアクセスできないため、未だ文字化できていない録音も多いが、この契約により、データは学生個人というより専攻の承認を得たものとなり、信頼性が高まったと考えている。日本においても欧米と同様、研究発表の現場データの入手は、今後さらに困難になると予想される。その意味でも本コーパスは極めて貴重なデータになろう。表 1 に本コーパスの収録データ数を掲げる。

表 1 収録されたコーパス

	専攻等	英語	日本語
2003 年	知能機能創成工学	0	30
2005 年	応用生物工学	0	39
2006 年	物質化学	0	34
	分子化学	0	31
	知能機能創成工学	0	13
	電気工学	0	13
	化学系 COE	0	5
2007 年	船舶海洋工学	7	0
	機械工学	0	69
	化学系国際学会	37	0
	環境・エネルギー工学	1	0
	バイオテクノロジー英語特別コース	4	0
2008 年	環境・エネルギー工学	0	77
	アメリカ学部学生	16	0
		65	311

３．オンライン検索サイト JECPRESE のインタフェース設計

国吉は、理工系研究者が直感的に利用できるように，JECPRESE のインタフェースを設計した．検索対象の単語・表現，言語，move(表現意図)，分野，講演者の対象言語における経験，表示結果の詳細など入力・変更する方法を容易にした．すなわち，move および研究分野の検索をクリック一つでできるようにした．また，検索方法や move 記号の詳細を「Help」ページにて，書き起こしファイルの名称の意味を「Filenames」にて，説明した．さらに，単語・表現の検索結果を，前後の文脈を基準にして簡単にソートできるようにし，検索した単語・表現がどのような文脈でよく現れるかが素早くわかるようにした．現在はさらに使いやすいインターフェースにするための変更を目指しており、今年度中には完成する予定である。

４．本コーパスの解析結果

４.１　理工系専門語彙の特定

林(2004)では「工学系の基本語彙はさらに細分化していると思われる。従って、代表的な学術雑誌に掲載された論文を抽出し分析するこれまでのような語彙調査においても、より詳細に研究分野（ロボット、マテリアルなど）を検討した上で、その分野における頻度数の高い語彙を抽出する必要があると考えられる。また、各研究分野の高頻度語彙を階層的に積み上げていくことによって機械・材料・電気・情報などの語彙のグループ、さらに工学系・医学系などの語彙のグループと、語彙の階層性が明らかになり、同時にいずれの分野にも共通の専門記述語が判明すると思われる。」とした。なお、解析には奈良先端科学技術大学院大学の自然言語処理学講座によって開発された形態素解析システム「茶筌」を用いた。林ら(2010)ではさらにコーパスを広げて解析し「頻度数上位 10 語だけで各品詞の延べ数のほぼ半数以上がカバーできる．また，それらの語彙は化学系と機械系でほぼ共通しており，さらに，全語彙中の各品詞は化学系，機械系，知能機能系，いずれの分野におい

てもほぼ同じ割合を示している．基本的な文を基本的な語彙を用いて定型化し，そこに研究分野によって多彩な名詞・漢語動詞を入れ込むことによって研究内容を表していることがわかった．異なる分野においてこのような結果がみられることは，理工系口頭発表における文の形態，発表の構成が類似・標準化されていることを示唆しており，これらの語彙は工学教育基準の基本語彙の可能性があることを示している.」とした。表 2、3 には分野・品詞別の頻度数・割合を示す。また、表４～17　には品詞ごとの頻度数上位 10 語を掲げる。

表 2 分野・品詞別の頻度数

	化学	機械
提供された発表原稿数	40	69
協力 labo 数	13	22
形容詞	689	1287
格助詞-連語	1574	2179
接続詞	1069	1449
接頭詞	905	1232
動詞	8892	13517
名詞-サ変接続	9784	13683
副詞	855	979
名詞	13624	21503
名詞-形容動詞語幹	1045	1538
名詞-接尾	5735	6458
名詞-非自立	1927	1019
名詞-副詞可能	1746	2023
連帯詞	1155	1946
名詞-代名詞	867	1051

表 3 分野・品詞別の割合（%）

	化学	機械	知能機能創成工学
形容詞	1.4	1.8	2.1
格助詞–連語	3.2	3.1	3.3
接続詞	2.1	2.1	1.8
接頭詞	1.8	1.8	1.5
動詞	17.8	19.3	17.4
名詞–サ変接続	19.6	19.6	19.5
副詞	1.7	1.4	1.2
名詞	27.3	30.8	29.6
名詞–形容動詞語幹	2.1	2.2	2.5
名詞–接尾	11.5	9.2	9.4
名詞–非自立	3.9	1.5	3.9
名詞–副詞可能	3.5	2.9	4
連帯詞	2.3	2.8	2.6
名詞–代名詞	1.7	1.5	1.2
計	99.9	100	100

表 4 名詞の頻度数上位10語

化学	13624	機械	21503	知能機能創成工学	8639
錯体	358	粒子	341	表面	168
分子	340	条件	329	ロボット	167
触媒	294	図	302	温度	145
活性	228	熱	236	材料	138
光	205	システム	222	モデル	126
構造	198	方向	218	ナノ	122
炭素	181	軸	215	熱	118
基質	129	速度	197	形状	97
蛍光	128	工具	196	原子	91
金属	124	モデル	182	特性	85
10 語の割合	16%	10 語の割合	11%	10 語の割合	15%

表 5 形容詞の頻度数上位10語

化学	689	機械	1,287
高い	112	大きい	255
よい	87	小さい	134
大きい	64	高い	84
低い	47	長い	65
ない	32	ない	64
強い	31	多い	64
新しい	23	硬い	46
多い	23	厚い	40
小さい	21	よい	39
長い	19	少ない	37
10 語の割合	67%	10 語の割合	64%

表 6 名詞-形容動詞語幹の頻度数上位10語

化学	1,045	機械	1,538
可能	114	可能	127
同様	96	安定	86
明らか	84	必要	82
様々	65	同様	70
非常	58	困難	52
必要	54	十分	48
安定	49	主	42
重要	34	自由	41
新た	31	不安定	35
主	29	多様	32
10 語が品詞全体に占める割合	59%	10 語が品詞全体に占める割合	40%

表 7 接続詞の頻度数上位10語

化学	1069	機械	1449
また	276	また	338
および	178	および	150
そこで	108	そして	147
次に	94	そこで	143
一方	61	次に	142
しかし	52	しかし	91
つまり	36	それでは	64
そして	35	つまり	48
たとえば	23	一方	38
あるいは	18	なお	31
10 語が品詞全体に占める割合	82%	10 語が品詞全体に占める割合	82%

表 8 副詞の頻度数上位10語

化学	855	機械	979
さらに	137	まず	295
まず	116	さらに	106
ほとんど	57	次に	85
もっとも/最も	39	実際	62
ほぼ	34	ほぼ	52
より	33	もっとも/最も	38
全く	31	特に	33
特に	26	ほとんど	25
既に	20	同時に	23
よく	19	常に	22
10 語が品詞全体に占める割合	60%	10 語が品詞全体に占める割合	76%

表 9 接頭詞の頻度数上位10語

化学	905	機械	1232
本	142	本	333
脱	100	各	113
超	49	高	84
当	43	非	71
環	42	超	48
再	42	低	40
約	42	被	37
重	38	約	36
単	35	再	34
不	33	第	31
10 語が品詞全体に占める割合	63%	10 語が品詞全体に占める割合	67%

表 10 名詞-接尾の頻度数上位10語

化学	5735	機械	6458
化	544	的	376
的	439	率	279
性	394	化	259
物	343	部	251
位	300	法	221
体	229	物	196
基	215	性	189
剤	204	値	182
率	166	流	182
量	163	面	174
10 語が品詞全体に占める割合	52%	10 語が品詞全体に占める割合	36%

表 11 格助詞 - 連語の頻度数上位10語

化学	1574	機械	2179
により/よって/よる	491	により/よって/よる	504
として/しまして	288	について	486
について	235	において/おける	446
において/おける	228	として/しまして	335
に対し/対して	142	に対し/対して	197
という/といった	91	という/といった	89
に関して/関する	36	に関して/関する	89
とともに	33	とともに	17
に従い/従って	12	を通して	4
につれ	8	に従い/従って	3
10 語が品詞全体に占める割合	99%	10 語が品詞全体に占める割合	99%

表 12 名詞-副詞可能の頻度数上位10語

化学	1746	機械	2023
結果	351	結果	378
ため	234	時間	188
場合	177	場合	149
ところ	99	以上	136
以上	89	それぞれ	88
それぞれ	77	以下	82
時間	60	ため	80
とき/時	56	とき/時	67
今回	43	今回	53
中	37	従来	46
10 語が品詞全体に占める割合	70%	10 語が品詞全体に占める割合	63%

表 13 名詞-非自立の頻度数上位 7 語

化学	1927	機械	1019
こと	1164	よう/様	752
の	102	もの	105
もの	101	こと	72
よう/様	424	ほう/方	44
ン	103	ン	17
ほう/方	21	点	17
点	12	の	12
8語が品詞全体に占める割合	100%	8語が品詞全体に占める割合	100%

表 14 連体詞の頻度数上位10語

化学	1155	機械	1946
この	790	この	1261
その	263	その	370
同じ	35	同じ	98
大きな	30	大きな	69
どの	15	小さな	62
さらなる	7	どの	50
なんらかの	3	ある	16
ある	2	なんらかの	5
いわゆる	2	いかなる	4
どういう	2	こうした	4
10 語が品詞全体に占める割合	99%	10 語が品詞全体に占める割合	99%

表 15 名詞-代名詞の頻度数上位10語

化学	881	機械	1051
これ	215	これ	266
こちら	183	こちら	260
これら	154	これら	161
それ	105	ここ	147
ここ	55	それ	88
それら	36	それら	32
いずれ	34	どちら	19
私	31	我々	14
そこ	12	そこ	12
どちら	9	その他	12
10 語が品詞全体に占める割合	95%	10 語が品詞全体に占める割合	96%

表 16 動詞の頻度数上位12 語

化学	8892	機械	13517
する	3526	する	4360
用いる	447	なる	671
行う	388	示す	513
なる	383	用いる	480
示す	376	行う	447
考える	313	できる	451
得る	305	わかる	383
できる	254	考える	288
よる	192	ある	281
わかる	175	みる	195
ある	143	得る	182
有する/有す	136	求まる	167
10 語が品詞全体に占める割合	75%	10 語が品詞全体に占める割合	62%

表 17 名詞-サ変接続の頻度数上位10語

化学	9784	機械	13683	知能機能創成工学	5701
反応	1015	研究	451	学習	164
生成	367	計算	353	研究	156
酸化	299	加工	331	結晶	151
進行	223	実験	321	計算	136
合成	220	制御	220	組織	132
検討	213	説明	211	溶融	118
研究	209	変化	199	変化	103
結合	208	発生	184	実験	95
選択	202	影響	173	解析	83
配	178	解析	168	凝固	80
10 語が品詞全体に占める割合	32%	10 語が品詞全体に占める割合	19%	10 語が品詞全体に占める割合	21%

これらは「工学教育」に採択された論文に載せたデータであるが、論文の採択によりこの語彙リストは工学教育のエキスパートの承認を得たと考えている。

４.２　ムーヴ（表現意図）の特定

専門的な職業・学問に携わる人々の集団（ディスコース・コミュニティ）はその共通目的を達成するためにコミュニケーションを行うが、それが繰り返されることによりパターン化してジャンルが形成される。ジャンルは文全体の構造・文法要素・単語・フォーマットなどの統合体であり、書き手や読み手など当該ディスコース・コミュニティ内のメンバーに共通して利用されている。Swales は Genre Analysis (1990)においてジャンルの重要性について指摘し、理工系論文のジャンルにおけるムーヴ（move）解析を行った。ムーヴ解析は「その表現形式は，ディスコース・コミュニティのどのような表現意図を示すものか」を探る。しかし、Swales のムーヴ解析は Introduction、Method の一部に限定されていた。我々は本コーパスを解析し、英語による発表より日本語による発表の方がより形式化されておりムーヴ解析が容易であることを見出した。ムーヴのリストについては林ら(2009)において明らかにしたが、現在 JECPRESE の改良にあたり、リストも見直している。

５．まとめ

日英語による口頭発表の信頼できるデータを集めたコーパスを構築し、コーパスの一部を解析し頻度数の高い語彙・表現を抽出し、ムーヴ解析も行った。また、研究・口頭発表に用いられる表現を容易に検索できるサイト JECPRESE を開発した。本コーパスは今後も拡大していく可能性が高い。

６．今後の課題

我々のコーパスは理工系の発表に特化した信頼できるデータと考えられるため、今後さらに詳細な解析を行いたいと考えている。すでに林(2004)、林ら(2010)で指摘したように、形容詞などの各語彙・条件表現などに特徴的な傾向がみられ、また、Theme/Rheme，Given/New 概念を含む Information structure（情報構造）の日英語における違いが明らかになった。今後は対象を広げ日英語の比較を考慮しながら精密な解析を行いたいと考えている。

謝　辞

貴重なデータをご提供くださった皆様に感謝申し上げます。また、本研究は平成 21 年度より科学研究費基盤研究（C）補助金を受けています。

文　献

林洋子(1999)「大阪大学工学部教官の認識に関する調査」『専門日本語教育教材作成に向けてー教官へのアンケート調査からー』，大阪大学工学部国際交流委員会

林洋子(2001)『日本語で考える：理工系専門日本語　基礎コースⅠ』

林洋子(2002)「考える」プロセスを重視して-多文化クラスの試み-，専門日本語教育研究，第 4 号, pp. 37

米田由喜代・林洋子(2003)「口頭発表の序論部の談話構造と語彙・表現-農学部卒業論文発表の分析から-」，専門日本語教育研究，第 5 号, pp. 37-43

林洋子(2004)「工学系修士論文口頭発表に用いられた語彙・表現」専門日本語教育研究,第 6 号、pp. 25～32

野口ジュディ・国吉ニルソン (2005)‘ESP education based on JSP research’JACET Kansai Chapter 2005 Spring Conference, June 4, 2005, Wakayama University

野口ジュディー，林　洋子，国吉ニルソン，東條加寿子(2007)理工系日本語・英語口頭発表における move・表現が検索可能なオンラインコーパスの開発，言語処理学会第 14 回年次大会発表論文集,pp.516−519

林　洋子，国吉ニルソン，野口ジュディ，東條加寿子(2008)若い研究者の言語獲得，電子情報通信学会、技術研究報告, IECE Technical Report, TL2008-3,2008-05, pp.11−16, 2008

林　洋子, 国吉ニルソン, 野口ジュディ(2009)工学系修士論文口頭発表のムーヴ解析, 工学教育, 57−6,pp.137−143, 2009
林　洋子, 国吉ニルソン, 野口ジュディ(2010)化学系と機械系の基本語彙, 工学教育, 58−6,pp.130−136, 2010
国吉ニルソン, 野口ジュディ、東條加寿子、林　洋子(2011) “Building a bilingual corpus of presentations in science and engineering: Purpose, issues and procedures”　The 16th World Congress of Applied Linguistics, August 27, Beijin, China.
東條加寿子 (2011) “Analysis of rhetorical strategies to identify moves in English research presentations in science and engineering fields”　The JACET 50th Commemorative International Convention, September 2, Fukuoka, Japan.

関連 URL

JECPRESE, The Japanese—English Corpus of Presentations in Science and Engineering, http://www.jecprese.sci.waseda.ac.jp