世界における言語資源・言語習得研究の動向 : 第24回太平洋アジア言語・情報・計算会議の成果から

報　告

世界における言語資源・言語習得研究の動向

―第24回太平洋アジア言語・情報・計算会議の成果から―

吉本　啓[1)]*，北原良夫[1)]

1）東北大学高等教育開発推進センター

1．はじめに

太平洋アジア言語・情報・計算会議（Pacific Asia Conference on Language, Information and Computation; 略称 PACLIC）は理論言語学およびコンピュータ言語処理研究を軸として，言語資源や言語習得もトピックとするテーマの幅の広い学会であり，東アジア持ち回りで毎年開かれている．平成22年11月４～７日の間，第24回大会（PACLIC 24）が東北大学川内キャンパスで，東北大学高等教育開発推進センターおよび日本論理文法研究会の主催により開催された．世界の21の国および地域から126人が参加して発表論文数は102本にのぼり，これまでで最大の学会となった．今回の大会では大会初日（11月４日）をワークショップの日としてコーパスおよび言語習得に関する研究発表に当て，また２日目（11月５日）午後に言語資源をテーマとするシンポジウムを開いた．この他，通常セッションでも言語資源や言語習得に直接・間接に関係する研究が多数発表された．それらの内容は報告者らが高等教育開発推進センターで取り組んでいる，言語情報処理技術を用いた外国語教育の高度化というテーマとも密接に関係する．そこで本稿では，同学会での講演および研究発表の中から言語資源および言語習得に関連するものについて報告し，世界における最新の研究動向を探ることにする．

2．ワークショップ

上記のように，初日の11月４日はワークショップに当てられ，コーパスと言語習得をテーマとして２つのワークショップが開催された．そのうち，Janne B. Johannessen（オスロ大学）の企画，チェアによる Workshop on Advanced Corpus Solutions では，主として人文系の言語研究者の観点から，コーパスの開発や利用に関わる諸問題について議論がかわされた（表１を参照のこと）．人文系の言語研究者は当然のこととして情報処理技術に詳しくない者が多いが，それにもかかわらずコーパスを利用した研究を進めるにあたっては高度の情報処理技術を必要とすることが多い．また，そのような言語研究者の側の様々なニーズがコーパスの開発・研究の重要な推進力となってきており，その事情は今後も変わらないであろう．本ワークショップはそのような観点から，コーパスの様々なツールやタイプに関わる問題を議論するために開かれた（Johannessen 2010）．

発表論文のうち，Wilson らは主要なヨーロッパ諸語および日本語・中国語を含む12か国語について，外国語学習者，外国語教員，言語学者や翻訳者が使用できるコーパス・インタフェースについて発表した．Jakubicek らは同様に言語研究のツールとして，柔軟かつ高速に例文を検索するための多言語対応システムについて発表した．また Goller は，複雑な検索の高度の処理を可能にするデータ構造である「並列接尾辞配列（parallel suffix array）」の開発について報告した．

コーパスの外国語教育への応用例として Yamura-Takei らは日本語・英語の外国人学習者と母語話者による作文コーパス中の指示表現の結束性について調査した．Centering Theory にもとづいて分析した結果，

*）連絡先：〒980－8576　宮城県仙台市青葉区川内41　高等教育開発推進センター　kei@compling.jp

外国語学習者の間では母語の影響が認められた.

また，Johannessen らはスカンディナビアの 6 言語について構築したマルチ・モーダル対応コーパスについて発表した．Kunst and Wesseling は，オランダ語方言の構文分布地図コーパスである SAND にもとづいて，他の言語との比較や構文以外の言語レベルの処理への応用について報告した．さらに，Bick はチャットおよび電子メールのコーパスの文章の「話し言葉性」を測定する方法について述べた.

我が国ではコーパス言語学はまだそれ程盛んでなく，しかも外国語教育や日本語学等，旧来の領域の枠内で行われているのが現状である．これに対して本ワークショップは，そのような小さな専門ごとの垣根を超えて行われたこと，また人文系研究者や開発者のニーズという観点からのソフトウェアの開発，またそれらの研究への応用について発表が行われたとういう点で，大きな意義があったと言える.

Workshop on Model and Measurement of Meaning（Shu-Kai Hsieh，国立台湾大学，の企画およびチェアによる）は，台湾－フランス間の同名の国際共同研究

表 1 ：Workshop on Advanced Corpus Solutions のプログラム

発　表　者	タ イ ト ル	予稿集中のページ
J. Wilson, A. Hartley, S. Sharoff and P. Stephenson	Advanced Corpus Solutions for Humanities Researchers	769-778
M. Yamura-Takei, M. Fujiwara and E. Yoshida	Entity Coherence in Comparable Learner Corpora: Seeking Pedagogical Insights	779-788
J. B. Johannessen, J. Priestley and A. Nøklestad	A Multilingual Speech Resource: The Nordic Dialect Corpus	749-758
E. Bick	Degrees of Orality in Speech-like Corpora: Comparative Annotation of Chat and E-mail Corpora	721-729
J. P. Kunst and F. Wesseling	Dialect Corpora Taken Further: The DynaSAND Corpus and Its Application in Newer Tools	759-767
M. Jakubicek, A. Kilgarriff, D. McCarthy and P. Rychlý	Syntactic Searching in Very Large Corpora for Many Languages	741-747
J. Goller	Parallel Suffix Arrays for Corpus Exploration	731-740

表 2 ：Workshop on Model and Measurement of Meaning（M 3 ）のプログラム

発　表　者	タ イ ト ル	予稿集中のページ
L. Prévot, C.-H. Chang and Y. Desalle	Computational Modeling of Verb Acquisition, from a Monolingual to a Bilingual Study	841-851
B. Gaillard, Y. Chudy, P. Magistry, S.-K. Hsieh and E. Navarro	Graph Representation of Synonymy and Translation Resources for Crosslinguistic Modelisation of Meaning	819-830
Y. Desalle, S.-K. Hsieh, B. Gaume and H. Cheung	Towards an Automatic Measurement of Verbal Lexicon Acquisition: The Case for a Young Children-versus-Adults Classification in French and Mandarin	809-818
P. Šimon and C.-R. Huang	Cross-sortal Predication and Polysemy	853-861
C.-F. Pan	Exploring Chinese Verbal Lexicon Developmental Trend with Semantic Space	831-839
H. Cheung, Y. Desalle, K. Duvignau, B. Gaume, C. Chang and P. Magistry	The Use of a Cultural Protocol for Quantifying Cultural Variations in Verb Semantic between Chinese and French	791-798
T.-H. Wu	Verb Use in Chinese Children: Extensibility of Instrument	863-872
A. Das and S. Bandyopadhyay	Towards the Global SentiWordNet	799-808

表 3 ：Symposium on Language Resources のプログラム

発　表　者	タ イ ト ル	予稿集中のページ
Julia Hockenmaier	The Future Role of Language Resources for Natural Language Parsing（We Won't Be Able to Rely on Pierre Vinken Forever… or Will We Have to ?）	13
Thomas Hun-Tak Lee	The Acquisition of Word order in a Topic-prominent Language: Corpus Findings and Experimental Investigation	15
Masataka Goto	PodCastle: A Spoken Document Retrieval Service Improved by Anonymous User Contributions	3-11

プロジェクトに参加した研究者が主体となって開催された（表 2 を参照のこと）．同プロジェクトは，中国語（普通話）およびフランス語の動詞の意味について，言語心理学および計算科学の立場からアプローチしようとする先端的な研究であり，特に動詞の意味の習得に重点を置いている．

Prévot らは上記のプロジェクトの中核をなしている，フランス語と中国語のビデオ・クリップの児童への視聴実験の大枠について解説している．これにもとづいて，Desalle らは，動詞語彙の習得の度合いを測定するための新しい統計的手法を提案している．また Gaillard らは，フランス語と中国語の同義語間の構造を比較するためのグラフ表示について発表した．また，Cheung らは同様のビデオ・クリップを用いて，フランス語・中国語母語話者間の文化の違いによる動詞の意味の理解に対する影響を調べた．その結果，視覚提示された映像に対する親近性の違いによって，動詞の選択に差が生じることが分った．さらに，Pan は動詞語彙習得データにもとづいて，意味空間の影響を評価した．子供の言語習得につれて，被験者間の語彙のバリエーションは減少するのに対して，特殊性の強い動詞の数は増加することが判明した．

上記のプロジェクトには含まれないが志向を同じくする研究がさらに 3 本発表された．Wu は中国語を母語とする子供における，動作に用いられる道具の親近性が習得に与える影響を調べた．Das and Bandyopadhyay は，インド系の言語を中心とする多数の言語について，テクストの書き手の意見・感想を自動的に抽出するための枠組みについて述べた．また，Šimon and C.-R. Huang は形式意味論を用いた文の意味 の表示におけるタイプの不一致の問題について，語彙の意味を構造化してより柔軟に扱うことによる解決法 を提案した．第一および第二言語習得に関して，意味の習得がどのように行われるのかについては不明な点が多い．本ワークショップでは，上記国際共同研究プロジェクトの言語心理学的手法と統計学・情報論的方法とを結合するアプローチによる研究発表が多数を占めた．このようなチャレンジングな研究により，将来における言語習得研究の新しい道が切り開かれることが期待される．

3．シンポジウム

Symposium on Language Resources は 5 日の午後，コーパスを初めとする言語資源の言語学・コンピュータ言語処理に関わる課題を討議するために，吉本および Alastair Butler のチェアにより開かれた（表 3 を参照のこと）．

Hockenmaier（Illinois 大学）は，無制限の大量の英語テクスト・データに対して形式統語理論である Combinatory Categorial Grammar(組合わせ範疇文法) を実装したシステムを適用して，高精度の統語解析木付きコーパスを実現させた Penn Treebank の開発者として知られる．本講演では，統語情報付きコーパスが汎用性や表示の豊かさの点で不十分なものであるにもかかわらず開発上の問題を抱えていることを指摘し，言語処理の進展のために必要な言語資源のあり方について提案した．

次の講演者の Thomas Hun-Tak Lee(香港中文大学) は，中国語普通話（共通語）や広東語に関する第一言語習得の研究，特に幼児による言語習得のコーパスの構築やその論理言語学的・認知科学的分析によって有名である．本シンポジウムでは，幼児による中国語の

語順の習得について講演を行った．中国語の基本語順はSVO であるが，他方，中国語は主題優位（topic-prominent）の言語であるとされている．後者の性質からは，目的語が主題化された OSV や SOV の語順が第一言語習得の初期の段階から表われるのではないかと予測される．しかし，第一言語習得コーパスを調査した結果，実際には後者のような語順は初期にはほとんど表われなかった．このことから，言語の習得の初期にまず語順と主題役割（thematic roles）とのマッピングが確立され，主題化のパラメータはそれよりも後に習得されることが分かった．

最後に，後藤真孝（産総研）は，自動発話分析の精度を上げるために匿名のユーザの協力を得て解析結果を訂正するためのシステム PodCastle について講演した．PodCastle はウェブ上のシステムであり，協力者は出力された音声とテクストとを比較し，音声認識結果の候補の中から正しいものを選択することを通じて認識率の向上に貢献する．実際に過去46か月の実験で音声認識システムが目覚ましい向上を遂げていることが示された．

高度の文解析情報をともなうコーパス構築，認知科学的言語習得研究，および音声発話自動解析の第一線の研究者による講演は，学会参加者から非常な好評を得た．講演を通じて，将来にわたるコーパスの研究や開発について刺激を受けたとの声が多く寄せられた．

4．通常セッション

通常セッション（口頭発表およびポスター発表）で直接コーパスや言語習得に関連する発表は，表 4 に見るように 9 本にのぼった．

これらのうち，Humayoun and Ranta は Punjabi 語のコーパス・辞書構築の現状について報告し，Manurung らはインドネシア語コーパスのオンライン貯蔵庫の開発について発表した．また，Buhay らは，タガログ語を代表とする，言語資源の乏しいマイノリティ言語について自動的にレキシコン等の言語資源を構築するための方法について考察した．さらに，Chen らはコーパスから言語学習者や辞書開発者らが連語や成句を抽出して学習や開発に利用するためのシステムについて述べた．

Fang and Cao は，言語学的に正確・詳細な品詞情報をコーパスにタグ付けすることにより，テクストの自動ジャンル分類がより効率的に行えることを示した．Xu らは大量のラベルなしコーパスから意見，感

表 4 ：通常セッション中の関連発表

発　表　者	タ イ ト ル	予稿集中のページ
A. Chengyu Fang and J. Cao	Enhanced Genre Classification through Linguistically Fine-Grained POS Tags	85-94
M. Humayoun and A. Ranta	Developing Punjabi Morphology, Corpus and Lexicon	163-172
R. Manurung, B. Distiawan and D. D. Putra	Developing an Online Indonesian Corpora Repository	243-249
H. Xu, K. Zhao, L. Qiu and C. Hu	Expanding Chinese Sentiment Dictionaries from Large scale Unlabeled Corpus	301-310
M. Chen, C. Huang, S. Huang and J. S. Chang	GRASP: Grammar- and Syntax-based Pattern-Finder for Collocation and Phrase Learning	357-364
J.-F. Hong, S.-J. Ker, C.-R. Huang and K. Ahrens	Using Corpus-based Linguistic Approaches in Sense Prediction Study	399-407
W. Kashino and M. Okumura	An Approach toward Register Classification of Book Samples in the Balanced Corpus of Contemporary Written Japanese	433-438
R. Spring	A Look into the Acquisition of English Motion Event Conflation by Native Speakers of Chinese and Japanese	563-572
E. L. C. Buhay, M. J. P. Evardone, H. B. Nocon, D. M. Dimalen and R. E. Roxas	AUTOLEX: An Automatic Lexicon Builder for Minority Languages Using an Open Corpus	603-611

情等の主観表現を抽出し，さらに受け手に与える影響の強弱の程度を付した辞書を自動的に作成する手法の開発について述べた．また，Hong らは，中国語の未知語の意味を大規模コーパスに基づいて推測する方法について発表した．Kashino and Okumura は，日本語で書かれた本の内容にもとづいて書誌情報として自動的に分類するシステムについて講演した．

また，Spring は，中国語および日本語を母語とする英語学習者が英語で動作を表現する場合に，母語の動詞の意味的分節化の違いにどのように影響を受けるかについて発表した．

5．結論

初めに述べたように，PACLIC 24 ではワークショップやシンポジウムを通じて，コーパスを中心とする言語資源研究およびその成果の外国語学習や言語習得研究を初めとする多様な目的への応用に重点を置いた．その結果，すでに報告したように，多彩な分野について最先端の研究発表がなされ，国境と分野の壁を真に超えた研究交流という点で成果を挙げることが出来た．採択論文は予稿集（Otoguro et al. 2010）として出版された．

謝辞

PACLIC 24 は東北大学高等教育開発推進センターの平成22年度高等教育の開発推進に関する調査・研究経費，東北大学大学院国際文化研究科言語脳認知総合科学研究センターよりの補助金，および平成22年度日本学術振興会国際交流事業による国際研究集会としての補助金を得て行われた．

参考文献

Johannessen, Janne Bondi.（2010）Workshop on Advanced Corpus Solutions. In: Otoguro et al.（2010）, 717-719.

Otoguro, Ryo, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto and Yasunari Harada, eds.（2010）*Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Tohoku University, 4-7 November.