PROCEEDINGS OF

THE 27TH PACIFIC ASIA CONFERENCE ON LANGUAGE, INFORMATION, AND COMPUTATION

(PACLIC 27)

TAIPEI, TAIWAN

Conference Dates: November 21-24, 2013
Conference Venue: Administration Building, National Chengchi University
Organizer: Department of English, National Chengchi University
Co-organizers: Institute of Linguistics, Academia Sinica
Graduate Institute of Linguistics, National Chengchi University
Department of Computer Science, National Chengchi University
Supporting Organizations: Linguistic Society of Taiwan (LST)
Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

Sponsors:
National Science Council, Executive Yuan, R.O.C.
Institute of Linguistics, Academia Sinica
NCCU Office of Research and Development

Welcome Message from Conference Honorary Chair

On behalf of ILAS, a co-host of this conference with National Chengchi University, I would like to welcome you all to the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC27).

The PACLIC conference has a long history,dating back to 1982 where the first conference of this series was organized with the original name “Korea-Japan Joint Conference on Formal Linguistics”. It was the consensus of the organizer of the 1994 Joint Conference of the Asian Conference on Language, Information and Computation (ACLIC) and the Pacific Asia Conference on Formal and Computational Linguistics (PACFoCoL) that the two conferences would continue to be held jointly in the future as the Pacific Asia Conference on Language, Information and Computation, with the 1995 conference being numbered the 10th. Over the years the conference series has developed into one of the leading conferences in the Pacific-Asia region. Like the previous PACLIC conferences, PACLIC27 has received 123 submissions (workshop and main conference included) in the fields of theoretical and computational linguistics, and participants coming from 27 countries.

The long tradition of the conference has been the keynote and invited speakers, and this year is no exception. The 5 eminent scholars who kindly agreed to deliver keynote speeches for this year are Professor Alec Marantz (New York University, USA), Professor Junichi Tsujii (Microsoft Research Asia, Beijing), Professor Wen-Lian Hsu (Academia Sinica, Taiwan), Professor Yukio Tono (Tokyo University of Foreign Studies, Japan),and Professor Stefan Th. Gries (University of California, Santa Barbara, USA). Furthermore the 3 distinguished scholars for the invited talks are Professor Chengqing Zong (Chinese Academy of Sciences, China), Professor KingkarnThepkanjana (Chulalongkorn University, Thailand) and Professor Aesun Yoon (Pusan National University, Korea). I have no doubt that in the three days there will be many opportunities for you to explore the intellectual fascination of theoretical and computational linguistics with these internationally renowned scholars and the other participants as well. It is my sincere hope that some of these interactions will lead to possible collaborations in the future or ring a bell in your memory in the years to come.

Thank you!

Chiu-yu Tseng (Conference Honorary Chair)
Director, Institute of Linguistics, Academia Sinica

Welcome Message

The 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27) is being held at National Chengchi University in Taipei, Taiwan from 21-24 November 2013. PACLIC is hosted annually by different academic institutions in the Asia-Pacific region. It has been nine years since the conference was first held in Taiwan, when PACLIC 19 was hosted by the Academia Sinica in 2005, and we are truly honored that National Chengchi University has the opportunity to take up the task this time.

For the past years, PACLIC has provided platforms for scholars to share new ideas about language, information, and computation, and, as such, has developed into one of leading conferences on the synergy of language studies and computational analysis. PACLIC 27 in Taiwan aims to carry on the mission of providing a great opportunity for linguists and computer scientists to gain stimulation from the exchange of the most up-to-date knowledge. A pre-conference workshop on Computer-Assisted Language Learning that represents an exemplary synergy of language, information, and computation is organized to address the study of computers and information technology in language teaching and learning. The theme is ‘Corpora and Language Learning’. Together with the main conference, PACLIC 27 provides the best access to the current trends in both linguistics and computational linguistics among the international research community, and most importantly, allows for the generation of synergies among research approaches and findings.

We received paper submissions representing enormous diversity, with authors from 27 countries or regions, namely, Canada, China, the Czech Republic, Denmark, France, Germany, Hong Kong, India, Iran, Ireland, Japan, Korea, Libya, Macao, Malaysia, Morocco, the Netherlands, the Philippines, Portugal, Singapore, Switzerland, Taiwan, Thailand, Turkey, the United Kingdom, the United States, and Vietnam. All submissions were rigorously reviewed by three reviewers to ensure the quality of all of the accepted papers. Of the 114 submissions, 39 papers (34%) were accepted for oral presentations, and another 17 papers (15%) for poster presentations. The research topics this year include grammar and syntax, language generation, discourse and pragmatics, lexical knowledge learning, speech perception, language learning, language acquisition, corpus compilation and analysis, machine translation, phonetics, lexical semantics, morphology and syntax, and sentiment analysis. The turnout reflects a diverse, inspiring and high-quality collection of research.

The key to the guarantee of high-quality results lies in the tremendous efforts and professional contributions of the program committee members from 18 countries, to whom we must extend our greatest gratitude, and the conference is enriched by the resulting combination of keynote speeches, invited talks and oral and poster presentations. The five keynote speeches for the main conference are given by internationally well-known scholars—Professor Alec Marantz from New York University, Professor Wen-Lian Hsu from Academia Sinica, Professor Yukio Tono from the Tokyo University of Foreign Studies, Professor Stefan Th. Gries from the University of California, Santa Barbara, and Professor Junichi Tsujii from Microsoft Research Asia in Beijing. The three invited talks are given by Professor Chengqing Zong from the Chinese Academy of Sciences, Professor Aesun Yoon from Pusan National University, and Professor Kingkarn Thepkanjana from Chulalongkorn University. Professor Yukio Tono and Professor Jason S. Chang from National Tsing Hua Univeristy present their keynote speeches in the workshop. The chance to hear first hand of their respective expertise definitely provides us with inspiring insights for research. On behalf of the organizing committee, we express our wholehearted appreciation to them. We would also like to thank the steering committee for their supervision, to Professor Zhao-Ming Gao from National Taiwan University and Professor Jyi-Shane Liu from National Chengchi University for organizing the workshop, to Professor Siaw-Fong Chung from National Chengchi University, Professor Jing-Shin Chang from National Chi Nan University and Liang-Chih Yu from Yuan Ze University for their efforts of compiling the proceedings, and to the local staff members at National Chengchi University for their exceptional dedication and coordination in their work.

Finally, we hope that you will enjoy the conference, and take advantage of this special occasion to renew contacts, and exchange ideas and the results of the latest developments. More importantly, you cannot miss the chance to explore and discover the beauty of Formosa, and to experience the great hospitality of this island.

Conference Chair and Co-Chair:
Huei-ling Lai and Kawai Chui (National Chengchi University)

Program Committee Chairs:
Chao-Lin Liu (National Chengchi University)
Shu-Chuan Tseng (Academia Sinica)

Steering Committee:

Standing Members:

Chae, Hee-Rahk, Hankuk University of Foreign Studies, Seoul

Huang, Chu-Ren, The Hong Kong Polytechnic University, Hong Kong

Roxas, Rachel, De La Salle University-Manila, Manila

Sun, Maosong, Tsinghua University, Beijing

Tsou, Benjamin, City University of Hong Kong, Hong Kong

Yoshimoto, Kei, Tohoku University, Sendai

Zhang, Min, Institute for Inforcomm Research, A-STAR, Singapore

Ex Officio Members:

Harada, Yasunari, Waseda University, Tokyo (Digital Archivist)

Lai, Huei-ling, National Chengchi University, Taipei (PACLIC 27 Local Organizer)

Manurung, Ruli, University of Indonesia, Depok (PACLIC 26 Local Organizer)

Otoguro, Ryo, Waseda University, Tokyo (Associate Digital Archivist)

Honorary Chair:

Tseng, Chiu-yu (Academic Sinica)

Conference Chair:

Lai, Huei-ling (National Chengchi University)

Conference Co-Chair:

Chui, Kawai (National Chengchi University)

Program Committee:

Chairs:

Liu, Chao-Lin (National Chengchi University)

Tseng, Shu-Chuan (Academic Sinica)

Co-Chairs:

Bond, Francis (Nanyang Technology University)

Ji, Donghong (Wuhan University)

Kwong, Olivia (City University of Hong Kong)

Manurung, Ruli (University of Indonesia)

Otoguro, Ryo (Waseda University)

Roxas, Rachel (De La Salle University-Manila)

Yeom, Jae-Il (Hongik University)

PC Members:

Aroonmanakun, Wirote (Chulalongkorn University)
Baldwin, Tim (University of Melbourne)
Bressan, Stephane (National University of Singapore)
Chae, Hee-Rahk (Hankuk University of Foreign Studies)
Chang, Chia-Hui (National Central University)
Chang, Claire H.H. (National Chengchi University)
Chang, Jung-Hsing (National Chung Cheng University)
Chang, Tao-Hsing (National Kaohsiung University of Applied Sciences)
Chang, Yungli (Institute of Linguistics, Academia Sinica)
Chen, Chun-Yin Doris (National Taiwan Normal University)
Chen, Hsin-Hsi (National Taiwan University)
Chen, Kuang-Hua (National Taiwan University)
Cheng, Pu-Jen (National Taiwan University)
Chng, Eng-Siong (Nanyang Technological University)
Daille, Beatrice (University of Nantes)
Dalrymple, Mary (Oxford University)
De Busser, Rik (National Chengchi University)
Dong, Minghui (Institute for Infocomm Research)
Fu, Guohong (Heilongjiang University)
Harada, Yasunari (Waseda University)
Her, One-Soon (National Chengchi University)
Hong, Munpyo (Sungkyunkwan University)
Hsiao, Yu-Chau E. (National Chengchi University)
Hsieh, Shelley Ching-Yu (National Cheng Kung University)
Hsieh, Shu-Kai (National Taiwan University)
Hsu, Dong-Bo (National Taiwan Normal University)
Huang, Chiung-Chih (National Chengchi University)
Huang, Meei-Jin (Shih Chien University)
Huang, Xuanjing (Fudan University)
Inui, Kentaro (Tohoku University)
Ji, Donghong (Wuhan University)
Kim, Jong-Bok (Kyung Hee University)
Kordoni, Valia (Saarland University and DFKI GmbH)
Kwong, Oliver (City University of Hong Kong)
Lai, Bong Yeung Tom (City University of Hong Kong)
Law, Paul (City University of Hong Kong)

Lee, Yae-Sheik (Kyungpook National University)
Lenci, Alessandro (University of Pisa)
Levow, Gina-Anne (University of Washington)
Li, Haizhou (Institute for Infocomm Research)
Lin, Chuan-Jie (National Taiwan Ocean University)
Lin, Jo-Wang (National Chiao Tung University)
Lin, Shou-De (National Taiwan University)
Liu, Qun (Dublin City University & ICT Chinese Academy of Sciences)
Lu, Wen-Hsiang (National Cheng Kung University)
Ma, Qing (Ryukoku University)
Ma, Yanjun (Baidu)
Maekawa, Takafumi (Hokusei Gakuen University Junior College)
Matsumoto, Yuji (Nara Institute of Science and Technology)
Matsushita, Mitsunori (Kansai University)
Morey, Mathieu (Nanyang Technological University)
Ng, Vincent (University of Texas at Dallas)
Nie, Jian-Yun (Universit de Montral)
Ogihara, Toshiyuki (University of Washington)
Otoguro, Ryo (Waseda University)
Paris, Cecile (CSIRO - ICT Centre)
Park, Jong C. (KAIST)
Prévot, Laurent (Aix-Marseille Université)
Qi, Haoliang (Heilongjiang Institute of Technology)
Qiu, Long (Institute for Infocomm Research)
Ranaivo-Malançon, Bali (Universiti Malaysia Sarawak)
Roxas, Rachel (De La Salle University-Manila)
Sah, Wen-Hui (National Chengchi University)
Shaikh, Samira (State University of New York - University at Albany)
Shyu, Shu-Ing (National Sun Yat-sen University)
Siegel, Melanie (Hochschule Darmstadt)
Singhapreecha, Pornsiri (Thammasat University)
Smith, Simon (Coventry University)
Sornlertlamvanich, Virach (National Electronics and Computer Technology Center)
Su, Keh-Yih (Behavior Design Corporation)
Su, Lily I-Wen (National Taiwan University)
Su, Yi-Ching (National Tsing Hua University)
Sung, Li-May (National Taiwan University)

Tabata, Tomoji (The University of Osaka)
Thompson, Henry S. (University of Edinburgh)
Tsai, Ming-Feng (National Chengchi University)
Tsai, Richard Tzong-Han (Yuan Ze University)
Tseng, Yuen-Hsien (National Taiwan Normal University)
Van Genabith, Josef (Dublin City University)
Villavicencio, Aline (Federal University of Rio Grande do Sul, University of Bath)
Wan, I-Ping (National Chengchi University)
Wang, Haifeng (Baidu)
Wang, Houfeng (Peking University)
Wang, Hsu (Yuan Ze University)
Wang, Hui (National University of Singapore)
Wang, Yu-Fang (National Kaohsiung Normal University)
Wu, Jing-Lan Joy (National Taiwan Normal University)
Wu, Jiun-Shiung (National Chung Cheng University)
Yang, Cheng-Zen (Yuan Ze University)
Yeh, Jui-Feng (National Chia-Yi Universty)
Yeom, Jae-Il (Hongik University)
Yokoyama, Satoru (Tohoku University)
Zhang, Jiajun (Chinese Academy of Sciences)
Zhang, Min (I2R)
Zhao, Hai (Shanghai Jiao Tong University)
Zock, Michael (CNRS-LIF)
Zong, Chengqing (Chinese Academy of Sciences)

Workshop Chairs:

Liu, Jyi-Shane (National Chengchi University)
Gao, Zhao-Ming (National Taiwan University)

Publication Chairs:

Chung, Siaw-Fong (National Chengchi University)
Chang, Jing-Shin (National Chi Nan University)
Yu, Liang-Chih (Yuan Ze University)

Table of Contents

Main Conference Keynote Speeches

Workshop Keynote Speeches

Main Conference Invited Talks

PACLIC 27 Papers

Oral Presentation Session 1A: Grammar and Syntax

Oral Presentation Session 1B: Language Generation

Oral Presentation Session 2A: Discourse and Pragmatics

Oral Presentation Session 2B: Lexical Knowledge Learning

Oral Presentation Session 3A: Speech Perception

Oral Presentation Session 3B: Language Learning

Oral Presentation Session 4A: Language Acquisition

Oral Presentation Session 4B: Corpus Compilation and Analysis

Oral Presentation Session 5A: Grammar and Syntax

Oral Presentation Session 5B: Machine Translation

Oral Presentation Session 6A: Morphology and Syntax

Oral Presentation Session 6B: Phonetics

Oral Presentation Session 7A: Semantics

Oral Presentation Session 7B: Sentiment Analysis

Poster Session 1

Poster Session 2

Workshop Session 1

Workshop Session 2

Workshop Session 3

Words and Rules Revisited: Reassessing the Role of Construction and Memory in Language

Alec Marantz
New York University, USA
marantz@nyu.edu

Abstract

Pinker's influential presentation of the distinction between the combinatoric units of language (the "words") and the mechanisms that organize the units into linguistic constituents (the "rules") rested on a strong, but ultimately incorrect, theory about the connection between a speaker's internalized grammar and his/her use of language: that what is linguistically complex, and thus constructed by the grammar, is not memorized; thus experience with complex constituents (as measured in corpus frequency, for example) would have no effect on processing such complex constituents. I argue that recent results within linguistics and within psycho- and neuro-linguistics show instead that memory and frequency effects are irrelevant to the linguistic analysis of language but always influence processing, across simple and complex constituents. Phrases and words can be shown always to decompose down to the level of morphemes both in representations and in processing, and, contrary to Pinker's claim, the "memorized" status of a complex structure holds no import for its linguistic analysis. On the other hand, speakers' experience with language is always reflected in their use of language, so frequency effects are always relevant to processing, even for completely regular combinations of words and morphemes. I will present neurolinguistic evidence for full decomposition of irregular forms (such as English irregular verbs), as well as evidence for frequency effects for regular combinations of morphemes and words.

A Principle-Based Approach for Natural Language Processing:
Eliminating the Shortcomings of Rule-Based Systems with Statistics

Wen-Lian Hsu
Institute of Information Science, Academia Sinica
hsu@iis.sinica.edu.tw

Abstract

In natural language processing, an important task is to recognize various linguistic expressions. Many such expressions can be represented as rules or templates. These templates are matched by computer to identify those linguistic objects in text. However, in real world, there always seem to be many exceptions or variations not covered by rules or templates. A typical approach to cope with this situation is either to produce more templates or to relax the constraints of the templates (e.g., by inserting options or wild cards). But the former could create many similar case-by-case templates with no end in sight; and the latter could lead to lots of false positives, namely, matched but undesired linguistic expressions. Thus, the flexibility of rule matching has troubled the natural language processing (NLP) as well as the artificial intelligence (AI) community for years so as to make people believe that rule-based approach is not suitable for NLP or AI in general. On the other hand, fine-grained linguistic knowledge cannot be easily captured by current machine learning models, which resulted in mediocre recognition accuracy. Therefore, how to make the best out of rule-based and statistical approaches has been a very challenging task in natural language processing.

This paper describes a partial matching scheme that enables a single template to match a lot of semantically similar expressions with high accuracy, which we refer to as the Principle-Based Approach (PBA).

In PBA, we use a collection of frames to represent linguistic concepts or rules. Each frame is a collection of slots (also called components) with relations specified among them. A slot can be a word, phrase, semantic category, or another frame concept. One can specify position relations, collocation relations, and agreement relations and others among its slots. Unlike normal templates that involve mostly left-right relations among its components in a sentence, relations within frames can be multi-dimensional. For example, one slot could be a variable indicating the topic which other slots belong to.

To illustrate our partial matching scheme, consider a simple frame concept involving 5 components such that their relations in a sentence are arranged as 1, 2, 3, 4, 5 from left to right. Suppose in a sentence we can identify components 2, 3, and 5 in that order. So 1 and 4 are missing (deletion), and there maybe words inserting

between 2 and 3 (insertion), and also between 3 and 5. Furthermore, a match for slot 5 could be on word-sense rather than on the word themselves (substitution). Our partial matching scheme allows for insertion, deletion and substitution. An insertion is given a positive score if it tends to collocate with its left or right matched components in general (otherwise, negative). A deletion can be harmless if slots 2, 3, and 5 contain a key combination for the frame. Note that many such key combinations can be pre-specified as indices of the frame. Collocation and bigram statistics can be incorporated in such score estimation. A substitution is given a lower score depending on their closeness in a semantic tree. After all these scores are determined, we can use an alignment algorithm to measure the fitness score and to decide how well the frame matches with the sentence.

PBA is inspired by the fact that when one studies a foreign language, he or she is usually presented with a collection of rules. These rules and their possible extensions and variations are practiced over and over again in real life to be mastered by the learner. PBA is flexible in that, it tends to relieve the burden of having to match with something “exactly” as specified and fine-grained linguistic knowledge can be more easily adopted to help estimate the scores of insertion, deletion and substitution in a PBA frame match.

We believe PBA can model more linguistic phenomena than current machine learning models, and is more suitable for NLP and AI in general. More details and examples of PBA will be covered in the talk.

Extracting “Criterial Features” for the CEFR Levels Using Corpora of EFL Learners’ Written Essays

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

Abstract

In this talk, I will report on the on-going project on systematic extraction of criterial features from multiple source corpora based on the Common European Framework of Reference for Languages (CEFR). First, a brief description of the CEFR itself, the project and the design of several different corpora newly compiled for the project will be given, followed by methodological issues regarding how to extract criterial features from CEFR-based corpora using machine learning techniques.

The CEFR-J and Reference Level Descriptions

The project aims to support the implementation of the CEFR-J, an adaptation of the CEFR into English language teaching in Japan (Tono & Negishi 2012). After the release of Version 1 of the CEFR-J in March, 2012, we launched a new government-funded project called the “CEFR-J Reference Level Description (CEFR-J RLD)” Project. RLD is a term used for the CEFR to prepare an inventory of language (lexis and grammar) for each individual language for the purpose of level specification.

Table 1 shows the list of corpora to be used for the project:

Type of Corpora	Name	Features
Input corpus	ELT materials corpus (to be completed)	ELT course books Major textbooks that claim to be CEFR-based
Interaction corpus	Classroom observation data	30 hours secondary school ELT classes
Output corpus	JEFLL Corpus (0.7 million)	Written, secondary school, CEFR level
	NICT JLE Corpus (2 million)	Spoken, interview test scripts, 1,280 participants, CEFR level
	ICCI (0.6 million)	Written, primary & secondary school, 9000 samples, CEFR level
	GTECfS Corpus (to be comleted)	Written, exam scripts, 30,000 samples, CEFR level
	MEXT Corpus	S/W 2000 students randomly

	MEXT Corpus (S: 8,000 words) (W:3,0000 words)	S/W 2000 students randomly selected from all over Japan

Table1: Corpora used for the project

Three types of corpora have been either newly compiled or re-organised: input, interaction, and output corpora. For input corpora, major ELT publishers' CEFR-based course materials have been scanned and processed by OCR. For output corpora, major learner corpora for Japanese EFL learners, the JEFLL Corpus and the NICT JLE Corpus, have been selected, but for our project, the essays originally classified according to the school grades or oral proficiency test scores, have been re-classified according to the estimated CEFR levels assigned by trained raters based on their holistic scorings. Two additional corpora have been made available. One is an exam-based corpus called the GTEC for STUDENTS Writing Corpus, provided by the Benesse Corporation. It consists of more than 30,000 students essay data with approximately 5,000 samples aligned with correction data. The other is the data collected by Ministry of Education (MEXT), in which more than 2,000 students were randomly selected from all over Japan. They were given written and oral proficiency exams in English. This data shows the average performance of EFL learners in Japan, after the three year instructions in secondary school.

Finally, a corpus of classroom interaction between teachers and students has been added to the resource. This is an on-going project and the size is relatively small, but I hope that it will shed light on the understanding of what is happening in the classroom.

Our aim is to identify criterial features by looking at input and output corpora across CEFR levels. The language presented in the input corpora may not be produced in the output corpora. By examining both input and output, descriptions of criterial features will become more systematic. The interaction corpus also helps better understand the learning/acquisition process in the classroom. Input from textbooks as well as input and interactions in the actual classroom will play an important role in learning a target language. The major goal is to find out criterial features for the levels specified in the CEFR-J and complete the inventory of grammar and vocabulary for teaching and assessment, with a special reference to teaching and learning contexts in Japan.

In the past few years, various linguistic criteria have been proposed as "criterial", but they need to be validated against a particular learner group like Japanese EFL learners because the data used in Europe are very different from our learner group. Also each proposed criterial feature should be evaluated and weighed in terms of usefulness as CEFR-level “classifiers”. Then a bundle of criterial features have to be tested and validated to find out which combinations of criterial features work best to predict the CEFR-levels. In a way, for assessment purposes, it is sufficient to identify the most salient criterial feature that can distinguish all the levels clearly. For teaching purposes, however, all the learning items need to be somehow evaluated against their ‘criteriality.’

There are various ways of extracting criterial features from the data. Machine learning techniques such as random forest seem to be very promising for this purpose. For instance, random forest is very useful in that it gives estimates of what variables are important in the classification. Table 2 shows the results of variable importance measure by Gini impurity criterion. Basically, the higher the score is, the more important the variable is. By using this kind of information, one can profile which linguistic feature will be most effective in classifying texts into CEFR levels. The major aim of the project is to decide on which machine learning algorithms to take, and evaluate a range of criterial features for its effectiveness as assessment and teaching points.

Linguistic features	MeanDecreaseGini
Total n. of words	440.3
Total n. of sentences	134.8
N. of VPs	277.2
N. of clauses	182.4
N. of T-units	121.3
N. of dependent clauses	102.6
N. of complex T-units	114.6
N. of complex nominals	210.2

Table2: Variable importance measured by Mean Decrease of Gini

In this paper, I will report on the performance of different machine learning techniques, including random forest, support vector machine, decision tree (C4.5), and naïve Bayes over CEFR-level classified texts and compare which programs produce the best result and useful additional information to evaluate the importance of criterial features.

References

Hawkins, J.A. & Filipović, L. (2012). Criterial Features in L2 English. Cambridge: Cambridge University Press.

Tono, Y. 2012a. Developing corpus-based word lists for English language learning and teaching: A critical appraisal of the English Vocabulary Profile. In J. Thomas & A. Boulton (eds). Input, Process and Product: Developments in Teaching and Language Corpora (pp.314-328). Brno: Masaryk University Press.

Tono, Y. 2012b. International Corpus of Crosslinguistic Interlanguage: Project overview and a case study on the acquisition of new verb co-occurrence patterns. In Y. Tono, Y. Kawaguchi & M. Minegishi (eds.) Developmental and Crosslinguistic Perspectives in Learner Corpus Research (pp.27-46). Amsterdam: John Benjamins.

Tono, Y. & Negishi, M. 2012. The CEFR-J: Adapting the CEFR for English language teaching in Japan. JALT Framework & Language Portfolio SIG Newsletter No.8 (September, 2012), pp. 5-12.

It's about Time: More and More Sophisticated Statistical Methods in Corpus Linguistics

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Abstract

By its very nature, corpus linguistics is a discipline not just concerned with, but ultimately based on, the distributions and frequencies of linguistic forms in and across corpora. This undisputed fact notwithstanding, for many years, corpus linguistics has been dominated by work that was limited in both computational and statistical ways. As for the former, a lot of work is based on a small number of ready-made proprietary software packages that provide some major functions but can of course not provide the functionality that, for instance, programming languages provide. As for the latter, a lot of work is very unstatistical in nature by relying on little more than observed frequencies or percentages/conditional probabilities of linguistic elements.

However, over the last 10 years or so, this picture has changed and corpus linguistics has evolved considerably to a state where more diverse descriptive statistics and association measures as well as multifactorial regression modeling, other statistical classification techniques, and multivariate exploratory statistics have become quite common. In this talk, I will survey a variety of recent studies that showcase this new-developed methodological variety in both synchronic and diachronic corpus linguistics; examples will include applications of generalized linear (mixed-effects) models, different types of cluster-analytic algorithms, principal components analysis and other dimension-reduction tools, and others.

Linking Text with Data and Knowledge Bases

Junichi Tsujii
Microsoft Research Asia
Beijing, China
jtsujii@microsoft.com

Abstract

In the last two decades, we have witnessed the rapid development of techniques in statistical modeling of language, which exploit large collections of text to reveal statistical regularities in language uses. However, the statistics-based approach to language,which tends to ignore or deemphasize structural issues of language, has shown its own limitations. The approach in its strictest form, for example, fails to treat the systematic mapping between syntax and semantics of language (i.e. the compositional aspect of meaning).An increasing number of researchers have become interested in combining linguistic theories, which treat the compositional aspect of meaning, with statistical modeling of language.

On the other hand, the community of knowledge-mining and semantic search has constructed large knowledge bases such as Freebase, Yago and Wikipedia. Although these knowledge-bases have been constructed independently of the interests in the NLP research community, they provide essential resources for research on Natural Language Understanding, which aims to develop a system which understands language as human being does. The first step of such an understanding system is to relate surface forms of language with corresponding units in knowledge-bases. Once text is mapped to representation in the knowledge domain, one can perform inferences of various sorts by combining it with knowledge in the knowledge base. Inferences, which combine information embedded within text with human knowledge which is external to text, are deemed essential in text understanding.

The two streams of research in the above seem to be tackling the same problem of how surface expressions in text can be linked with extra-linguistic representation in the knowledge domain, and what roles the structure of language plays in such a linking process.

With this broad perspective in mind, I will address the following research topics which I have been involved in:

(1) Parsing and Semantics: While the performance of a syntactic parser has been improved substantially of late, it still fails to treat semantically crucial constructions. In order to resolve the difficulties which remain in parsing, we have to treat semantics of language more systematically than the current state of the arts parsers do. I would argue that we cannot resolve the difficulties without referring

to proper theories of syntax.

(2) Entity linking: Disambiguation in entity-linking has been carried out by using characteristics specific to individual entities. However, in order to treat long-tail problems in entity-linking, not only properties of individual entities but also classes of entities and their properties in knowledge bases have to be exploited. The results of our recent experiments will be presented, in order to illustrate how structures in knowledge bases can be used for interpretation of expressions in language.

(3) Relation linking: The same relation in the knowledge domain can be expressed by diverse surface expressions in language. To gather surface relation expressions for a given set of relations in the knowledge domain is a crucial step of linking text with knowledge. Some of our recent studies in relation extraction will be presented as the next step of linking text with knowledge bases.

(4) Paraphrasing and structures of sentences: While semantics of words have been studied extensively both in distributional semantics and traditional linguistics (e.g. synonyms, antonyms, etc.), semantics of larger units such as phrases and clauses have not been studied with similar degrees of details. Paraphrase recognition by structure alignment will provide a framework to capture semantics of larger units in language than words. We discuss how structures of sentences together with inferences based on meaning can give fine grained explanation of paraphrases, and how such research will contribute to the task of linking text with knowledge.

Mining Language Learners' Production Data for Understanding of L2 Learning Systems

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

Abstract

In this workshop, I will share my experience in the field of learner corpus research (LCR). First I will define learner corpora in terms of its design criteria. Second, I will show how L2 learners' production data as corpora can be exploited to find linguistic features that characterize the progress in L2 learning systems. Third, such transitional competence should be explained by various internal and external factors such as cognitive, affective, and instructional effects. I would like to discuss with the participants how to model L2 learning systems by showing various examples of features marking different stages of learning in English as a foreign language.

Introducing Linggle: From Concordance to Linguistic Search Engine

Jason S. Chang
Department of Computer Science
National Tsing Hua University
son.jschang@gmail.com

Abstract

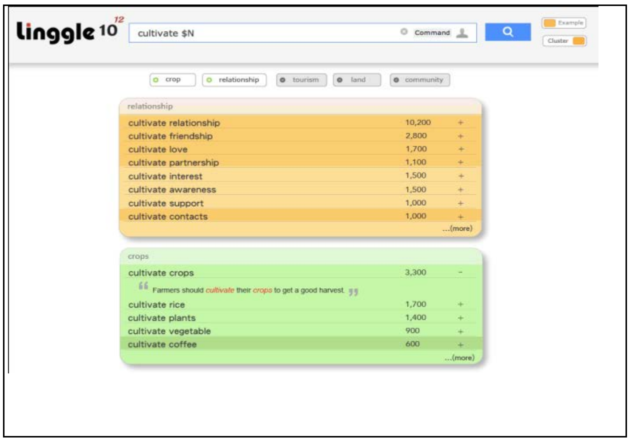
We introduce a Web-scale linguistics search engine, *Linggle*, that retrieves lexical bundles in response to a given query. Unlike a typical concordance, *Linggle* accepts queries with keywords, wildcard, wild part of speech (PoS), synonymous words, and additional regular expression (RE) operators, and returns bundles with frequency counts. In our approach, we argument Google Web 1T corpus with inverted file indexing, PoS information from BNC, and semantic indexing based on Latent Dirichlet Allocation. The method involves parsing the query to transforming the query to several keyword retrieval commands, retrieving word chunks with counts, filtering the chunks again the query as a RE, and finally displaying the results according the count, similarity, and topic. Clusters of synonymous or conceptually related words are also provided. In addition, *Linggle* provide example sentences from *The New York Times* on demand. The current implementation of *Linggle* is the most comprehensive functionally, and is in principle language and dataset independent. We plan to extend *Linggle* to provide a fast and convenient access to a wealth of linguistic information embodied in Web scale datasets including *Google Web 1T* and *Google Books Ngram* for many major languages in the World.

For non-native speakers, doubts concerning the usage of a preposition, the mandatory presence of a determiner, the correctness of the association of a verb with an object or the need for synonyms of a term in a given context are problems that arise frequently when writing in English. Printed collocation dictionaries and reference tools based on compiled corpora offer limited coverage of word usage while knowledge of collocations is vital for the competent use of a language. We propose to address these limitations with a comprehensive system that truly aims at letting learners "know a word by the company it keeps". *Linggle* (**linggle.com**) is a broad coverage language reference tool for English as Second Language learners (ESL). The system is designed to access words in context under various forms.

First, we build inverted file index for the *Google Web 1T Ngram* to support queries with RE-like patterns including PoS and synonym matches. For example, for the query "$V $D +important role", *Linggle* retrieve 4-gram chunks that start with a

verb and a determiner followed by a *important* synonym and the keyword *role* (e. g., *play a key part* 15,900). A natural language interface is also available for users that would be less familiar to pattern based search. For example the question "*How can I describe a beach?*" would retrieve two word chunks with count such as "*sandy beach* 413,300" and "*rocky beach* 16,800". The n-gram search implementation is achieved through filtering, re-indexing, and populating Web 1T ngram in a HBase database and augmenting them with the most frequent PoS for words (without disambiguation) derived from the British National Corpus.

The n-grams resulting from the queries can then be linked to examples extracted from the New York Times Corpus in order to provide full sentential context for more effective learning. In some situations, users might need to search for words in a specific syntactic relation (i. e., *collocates*). Let's consider the example "absorb $N" that queries all the objects of the verb *absorb*. In this case, grouping the words that belong to similar domains together offers a better overview of the usage of the verb than a list of objects ordered by frequency. For example the verb *absorb* takes clusters of objects related to the topic *liquid/energy*, but also to the topics *money, knowledge* or *population*.



relationship

cultivate relationship	10,200	+
cultivate friendship	2,800	+
cultivate love	1,700	+
cultivate partnership	1,100	+
cultivate interest	1,500	+
cultivate awareness	1,500	+
cultivate support	1,000	+
cultivate contacts	1,000	+

...(more)

crops

cultivate crops	3,300	-
"Farmers should cultivate their crops to get a good harvest."		
cultivate rice	1,700	+
cultivate plants	1,400	+
cultivate vegetable	900	+
cultivate coffee	600	+

...(more)

This tendency of predicates to prefer certain classes is defined by Wilks (1978) as selectional preference and widely reported in the literature. *Linggle* proposes *preferred* clusters of synonymous query arguments of adjectives, nouns and verbs. The clustering is achieved by building on Lin and Pantel (2002)'s large-scale repository of dependencies and word similarity scores and on an existing method for selectional preference induction with a Latent Dirichlet Allocation (LDA) model.

References

Chang, Jason. S. 2008. Linggle: a web-scale language reference search engine. Unpublished manuscript.

Fletcher, William H. 2012. Corpus analysis of the world wide web." In *The Encyclopedia of Applied Linguistics*.

Kilgarriff, Adam, and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of COLLOCTION: Computational Extraction, Analysis and Exploitation workshop*, pp. 32-38.

Kilgarriff, Adam. 2007. Googleology is bad science." *Computational linguistics* 33(1), pp. 147-151.

Lin, Dekang, and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of COLING*.

Lin, Dekang, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil et al. 2010. "New tools for web-scale n-grams." *Proceedings of LREC*.

Potthast, Martin, Martin Trenkmann, and Benno Stein. Using Web N-Grams to Help Second-Language Speakers. 2010. In *Proceedings of SIGIR Web N-Gram Workshop*, pages 49-49.

Wu, Shaoqun, Ian H. Witten, and Margaret Franken. 2010. Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge." *ReCALL* 22(1), pp. 83-102.

Wilks, Yorick. 1978. Making preferences more active. *Artificial Intelligence* 11(3), pp. 197-223.

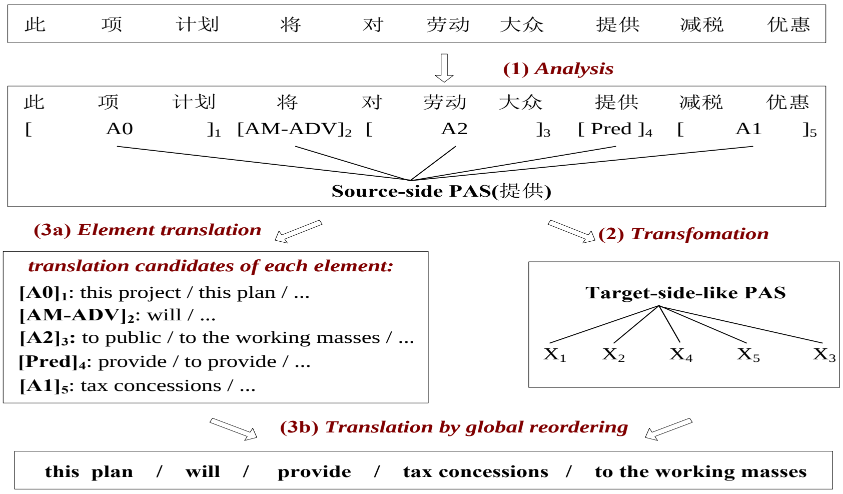
Statistical Machine Translation Based on Predicate-Argument Structure

Chengqing Zong
Institute of Automantion, Chinese Academy of Sciences
No. 95, Zhong Guan Cun East Road, Beijjing 100190, China
cqzong@lpr.ia.ac.cn

Abstract

As we have well known that it is always a basic requirement for statistical machine translation (SMT) to maintain semantic equivalence between a source sentence and its translation. However, nearly all of the existing translation models do not deal with the semantic structure between two languages at all. In this talk, I will present a novel translation method based on semantically-motivated framework, using predicate-argument structure (PAS). Generally, PAS depicts the semantic relation between a predicate and its associated arguments, and it always indicates the semantic frame and skeleton structure of a sentence. Thus, we believe the PAS would be much beneficial for machine translation in grasping the semantics of sentences. Furthermore, after analysis of the weakness of PAS representation during translation, I will propose a concept of syntax-complemented PAS (SC-PAS). It effectively overcomes the drawback of the prevalent gaps in PAS and provides more useful knowledge for SMT.

We also call the semantically-motivated framework as Analysis-Transformation-Translation (ATT) framework, which is just based on the PAS and SC-PAS. As the following figure shows, this framework divides the whole translation process into three steps: (1) Analysis: to analyze the source sentences and obtain their PASs (or SC-PASs) automatically; (2) Transformation: to convert the source-side PASs (or SC-PASs) to target side by predicate-aware transformation rules; (3) Translation: this step is further divided into two parts: (a)element translation is to translate each element of PAS (or SC-PAS); (b)translation by global reordering is to combine the resulting translation candidates to translate the entire structure. By taking advantage of PAS (or SC-PAS), the ATT framework can well keep the semantic structure consistency of the source language and the target language and consequently show the great potential to improve translation quality.


此 项 计划 将 对 劳动 大众 提供 减税 优惠
(1) Analysis
此 项 计划 将 对 劳动 大众 提供 减税 优惠
[A0]1 [AM-ADV]2 [A2]3 [Pred]4 [A1]5
Source-side PAS(提供)
(3a) Element translation
(2) Transfomation
translation candidates of each element:
[A0]1: this project / this plan / ...
[AM-ADV]2: will / ...
[A2]3: to public / to the working masses / ...
[Pred]4: provide / to provide / ...
[A1]5: tax concessions / ...
Target-side-like PAS
X1 X2 X4 X5 X3
(3b) Translation by global reordering
this plan / will / provide / tax concessions / to the working masses

Using Hierarchically Structured Lexicon as Key Clues Solving Data Sparseness Problems in Word Sense Disambiguation: a Case for Korean and Its Applications to English and Chinese

Aesun Yoon
Korean Language Processing Laboratory
Dept. of French
Pusan National University
Busan, 609-735, Rep of Korean
asyoon@pusan.ac.kr

Minho Kim, Hyuk-Chul Kwon
Korean Language Processing Laboratory
Dept. of French
Pusan National University
Busan, 609-735, Rep of Korean
{karma, hckwon}pusan.ac.kr

Abstract

Word sense disambiguation (WSD) determines the accuracy of almost all tasks in natural language processing. Korean Processing Laboratory of Pusan National University has been working on efficient automatic WSD methods, especially for Korean language. This paper presents our unsupervised model using hierarchically-structured lexicon, i.e. Korean WordNet (KorLex). KorLex can provide us with key clues for solving data sparseness problems, which are inherent in the unsupervised WSD. The proposed model shows 91.14% average accuracy, which is 26.95% higher than the best performance obtained by a supervised method (Lesk's dictionary-based WSD). Our model obtains also a higher accuracy for English and Chinese, using Princeton WordNet and HowNet.

Effects of Constituent Orders on Grammaticalization Patterns of the Serial Verbs for 'Give' in Thai and Mandarin Chinese

Kingkarn Thepkanjana
Chulalongkorn University
Phayathai Road, Bangkok 10330
THAILAND
Kingkarn.T@chula.ac.th

Satoshi Uehara
Tohoku University
41 Kawauchi, Aoba-ku, Sendai
980-8576 JAPAN
uehara@intcul.tohoku.ac.jp

Abstract

The verbs meaning 'give' across languages are known to be among the most highly grammaticalized verbs, which exhibit a high degree of polyfunctionality. This paper aims to (i) present commonalities and differences in the grammaticalization of the verbs for 'give' in Thai and Mandarin Chinese, namely, hây in Thai and gěi in Mandarin Chinese, and (ii) investigate how different constituent orders of the head vis-à-vis the modifier and complement in Thai and Mandarin Chinese bear on patterns of grammaticalization of the two verbs. It is found that the functions that hây in Thai and gěi in Mandarin Chinese share in common are (1) the ditransitive verb use, (2) the dative-marking use, (3) the benefactive-marking use, and (4) the causative-marking use. As for different functions of hây and gěi, hây exhibits the clause connective use, which is lacking in gěi, whereas gěi exhibit the passive-marking use, which is lacking in hây. It is argued that the head-modifier order in Thai seems to be compatible with postverbal grammaticalized morphemes whereas the modifier-head order in Mandarin Chinese seems to be compatible with preverbal grammaticalized ones.

1 Introduction

It is generally known that Thai and Mandarin Chinese are typologically similar in many respects. They are isolating, topic-prominent, serializing, have the SVO basic word order and rich with grammaticalized morphemes. However, there is one important difference between them, i.e. difference in constituent order. Mandarin Chinese has the modifier-head order whereas Thai has the head-modifier one. This paper investigates how the difference in constituent order in Thai and Mandarin Chinese bears on patterns of grammaticalization of serial verbs in the two languages. The serial verbs for 'give' in Thai and Mandarin Chinese, i.e. hây and gěi, are used as a case study. The verbs meaning 'give' across languages are known to be among the most highly grammaticalized verbs, which exhibit a high degree of polyfunctionality. The analysis in this paper is based on the findings of a synchronic contrastive study of hây and gěi presented in Thepkanjana and Uehara (2008).

2 Commonalities and differences

Thepkanjana and Uehara (2008) make a synchronic contrastive study of the polysemous morphemes hây and gěi in Thai and Mandarin Chinese. It is found in Thepkanjana and Uehara (2008) that hây and gěi share four main uses, namely, the ditransitive (main) verb use, the dative-marking use, the benefactive-marking use and the causative-marking use. As for differences between hây and gěi, one important use that is missing in hây is the passive-marking use whereas one that is missing in gěi is the clause connective function. The commonalities between the two verbs are discussed in section 2.1 and the differences in section 2.2. The examples provided are drawn from Thepkanjana and Uehara (2008).

2.1 Commonalities between hây and gěi

The first common function between hây and gěi is the ditransitive main verb use. Hây and gěi in this use co-occur with two NPs following each other in a row. The structural schemas of the ditransitive verbs hây and gĕi and some examples of this use are given below. Notice that the semantic roles of NP1 and NP2 in Thai and Mandarin Chinese are different.

Ditransitive verb use
Thai: [hây + NP1 + NP2]
(thing) (recipient)

(1) sǒmsàk hây ŋə*n* sǒmchay
Somsak give money Somchay
'Somsak gave Somchay some money.'

Mandarin Chinese: [gĕi + NP1 + NP2]
(recipient) (thing)

(2) Zhāngsān gĕi Lǐsì qián
Zhangsan give Lisi money
'Zhangsan gave Lisi some money.'

Dative-marking use
Thai: [V + NP1 +hây+ NP2]
(thing) (recipient)

(3) sǒmsàk sòŋ ŋə*n* hây sǒmchay
Somsak send money give Somchay
'Somsak sent some money to Somchay.'

Mandarin Chinese: 2 schemas
Schema 1: postverbal *g*ĕi
[V + NP1 + gĕi + NP2]
(thing) (recipient)

(4) Zhāngsān jì-le yì fēng
Zhangsan send-ASP one CLS
*x*ìn gĕi Lǐsì
letter give Lisi
'Zhangsan mailed a letter to Lisi.'

Schema 2: preverbal gĕi
[gĕi + NP1 + V + NP2]
(recipient) (thing)

(5) Zhāngsān gĕi Lǐsì mǎi
Zhangsan give Lisi buy
*y*ì bĕn shū
one CLS book
'Zhangsan bought a book for (and gave it to) Lisi'

Notice that the dative hây in Thai occurs postverbally whereas the dative gĕi occurs both preverbally and postverbally.

Newman (1993b) argues that an act of giving naturally results in some kind of benefit to the recipient. Even a non-giving action, such as driving, speaking and cleaning can also be done for the benefit of someone. The person who benefits from the agent's action is usually called a beneficiary. Therefore, it is natural that hây and gĕi can also function as benefactive markers. The notion of benefactive is more complicated than generally assumed. Three types of benefactive are postulated in this paper as below.

(a) Recipient benefactive: The beneficiary gains a benefit by virtue of being a recipient of a concrete entity, for example, J*ohn* bought a sweater *f*or Mary.
(b) Benefit benefactive: The beneficiary gains a more or less abstract benefit from somebody's action, for example, Jo*hn s*ang *a s*ong *f*or *M*ary.
(c) Behalf benefactive: The beneficiary gains a benefit from somebody who performs an action on his/her behalf, for example, J*ohn* drove *a* car for Mary *be*cause she *w*as drunk.

It is found that the Thai hây can be used to mark the three types of benefactive as shown below.

Recipient benefactive
(6) sǒmsàk sɯ́ɯ sɯ̂anǎaw hây
Somsak buy sweater give
sǒmchay
Somchay
'Somsak bought a sweater for Somchay.'

Benefit benefactive
(7) sǒmsàk tàt phǒm hây sǒmchay
Somsak cut hair give Somchay
'Somsak cut hair for Somchay.' Or
'Somsak cut Somchay's hair.'

Behalf benefactive
(8) sǒmsàk khàprót hây sǒmchay
Somsak drive a car give Somchay
'Somsak drove a car for Somchay.'

It is noted that the benefactive hây is ambiguous between the recipient benefactive and behalf benefactive readings if the main verb incorporates the sense of giving or involves the the manipulation of an entity as shown in (9) and (10).

(9) sǒmsàk sòŋ *c*òtmǎay hây sǒmchay
Somsak send letter give Somchay
'Somsak sent a letter to Somchay.' Or
'Somsak sent a letter on Somchay's behalf.'

(10) sǒmsàk suɯ́ɯ nǎŋsuɯ̌ɯ hây
Somsak buy book give
sǒmchay
Somchay
‘Somsak bought a book and gave it to Somchay.’ Or
‘Somchay bought a book on Somchay’s behalf.’

It is found that the Mandarin Chinese gěi can be used to mark the recipient benefactive and the benefit benefactive in some cases as shown below.

(11) Zhāngsān gěi Lǐsì mǎi
Zhangsan give Lisi buy
yì běn shū
one CLS book
‘Zhangsan bought a book for (and gave it to) Lisi’

(12) Zhāngsān gěi wǒme*n* chàng
Zhangsan give us sing
yì shǒu gē
one CLS song
‘Zhangsan sang a song for us.’

The structural schemas of the benefactive hây and gěi are given below.

Benefactive-marking use
Thai: [V + (NP1) + hây + NP2]
(beneficiary)

Mandarin Chinese: [gěi + NP1 + V+ (NP2)]
(beneficiary)

‘

Notice that the benefactive hây and gěi occur in different positions. The former occurs postverbally, i.e. after the main verb, whereas the latter occurs preverbally, i.e. before the main verb.

The third common use of hây and gěi is the causative use. The causative constructions with the causative-marking hây and gěi in Thai and Mandarin Chinese have the same syntactic schema as below.

Causative-marking use
Thai and Mandarin Chinese:
[NP1 + hây/gěi + NP2 + VP]
(causer) (causee)

(13) sǒmsàk hây sǒmchay ʔɔ̀ɔk pay
Somsak give Somchay exit go
‘Somsak had Somchay go out.’

(14) Zhāngsān gěi Lǐsì kàn
Zhangsan give Lisi look
‘Zhangsan let Lisi look.’

The NP1 in the schema above is the causer whereas the NP2 is the causee. The causer is typically human whereas the causee is typically animate. The causative verbs hây and gěi express an indirect causation in which the causer intentionally causes an event to take place by doing something to prompt the causer to act or by not doing something which prevents that event to take place. The causee is the person who directly causes the event to take place. Notice that the causative gěi occurs in the same position as the benefactive gěi in Mandarin Chinese, which results in ambiguity between the causative and benefactive readings in some cases as shown in (15), which is taken from Newman (1996:20).

(15) wǒ gěi nǐ kàn
I give you look
‘I let you look.’ (causative) Or
‘I look on your behalf.’ (benefactive)

According to Yap and Iwasaki (1998), native speakers of Mandarin Chinese tend to interpret gěi in (15) as the benefactive marker rather than the causative one as in (16).

(16) tā gěi *wǒ* *zào*-le
s/he give me build-ASP
yì dòng fángzi
one CLS house
‘S/he built a house for me.’ (preferred)
‘S/he had me build a house.’ (awkward)

Yap and Iwasaki (1998) note that Mandarin Chinese prefers the causative verbs ràng and jiào to the verb gěi in expressing indirect causation as in (17).

(17) tā *gěi/ràng/jiào háizi shuì-jiào
s/he CAUSE child sleep
‘She let the child sleep.

The use of ràng and jiào rather than gěi to express causation helps prevent the ambiguity between the causative and benefactive readings that can arise if gěi is used as the causative verb, which occurs in the same position as the benefactive gěi. It is therefore not surprising that the use of the causative gěi in Mandarin Chinese is much more restricted than the use of the causative hây in Thai because the latter does not create ambiguity as the former.

2.2 Differences between hây and gĕi

Hây and gĕi are different in two ways. There is one important use of hây which is missing in gĕi, namely, clause connective use, and one important use of gĕi which is missing in hây, namely passive-marking function. The clause connective use, which is missing in gĕi is discussed first.

The connective hây in Thai takes place in complex constructions in which hây functions as a subordinator which links two predicates or two clauses. The first clause in the complex construction is the matrix clause and the other is the subordinate one. The complex constructions in which *hâ*y functions as the subordinator can be classified into three types, namely, a purposive construction, a jussive construction and a complementation construction. The purposive construction is a complex construction in which the subordinate clause functions as a purpose of the performance of an action denoted by the matrix clause. The jussive construction expresses a command, request or demand made by one participant towards another in order for the latter to perform an action (Van Valin and LaPolla, 1997). The complementation construction is a complex construction in which the subordinate clause functions as a complement of the desiderative predicate of the matrix clause. The structural schema of the connective hây and some examples of the three types of complex constructions containing hây are given below.

Clause connective use
<u>Thai</u>: $_{\mathrm{S1}}$[NP1 + VP1] + hây+ $_{\mathrm{S2}}$[NP2 + VP2]

From Rangkuphan (1997:36)
Purposive construction
(18) nuan phlàk kæ̂æw hây
Nuan push Kaew give

*k*lîŋ pay rɯ̂ayrɯ̂ay
roll go continually
‘Nuan pushed the glass in order for it to keep rolling.’

(19) nuan khon námtaan hây lalaay
Nuan stir sugar give melt
‘Nuan stirred the sugar in order for it to melt.’

Jussive construction
(20) sŏmsàk bɔ̀ɔk hây sŏmchay maa
Somsak tell give Somchay come
‘Somsak told Somchay to come.’

(21) sŏmsàk sàŋ hây sŏmchay
Somsak order give Somchay
klàp bâan
return home
‘Somsak ordered that Somchay go home.’

Complementation
(22) sŏmsàk yàak hây sŏmchay
Somsak want give Somchay
m*aa* *hăa*
come see
‘Somsak wanted Somchay to come to see him.’

(23) sŏmsàk tɔ̂ŋk*aan* hây lûuk
Somsak want give child
rian phæ̂æt
study medicine
‘Somsak wanted his child to study medicine.’

Thepkanjana and Uehara (2008) argue that each type of complex construction results from a reanalysis of hây from the causative verb to the subordinator. In the reanalysis process, the causative hây is semantically bleached out and loses its verbal properties to varying degrees in the three types of complex construction. In other words, hây in the three types of complex construction has different degrees of function word properties. It is argued in Thepkanjana and Uehara (2008) that the connective hây in the complex constructions is derived, extended or grammaticalized from the causative hây. The hây’s in all of these cases are followed by a clause or a predicate. The causative hây functions as the main verb in the causative construction whereas the connective hây is preceded by a main verb and followed by a clause or a predicate. It is found that there is an intention that an event take place in the subject of the matrix clause in all of the three types of complex construction and in the subject of the causative hây. It is argued in Thepkanjana and Uehara (2008) that the notion of indirect causation has the highest degree of saliency in the causative hây but has decreasing degrees of saliency in the purposive, jussive and complementation constructions.

On the other hand, one important use of gĕi which is missing in hây is the passive-marking function. The passive-marking function is alternatively called the agentive-marking function. The structural schema of the passive-marking gĕi and some examples are given below.

Passive-marking use
<u>Mandarin Chinese</u>: [NP1 + gĕi+ NP2 + VP]
From Haspelmath (1990:48)

(24) Lĭsi gĕi Zhāngsān kànji*àn*-le
Lisi give Zhangsan see-ASP
'Lisi was seen by Zhangsan.'

From Newman (1993b:471)

(25) jīnyú gĕi m*āo* chī-le
goldfish give cat eat-ASP
'The goldfish was eaten by the cat.'

According to Xu (1994), the passive gĕi is used in colloquial speech whereas the other passive marker, bèi, is used in formal speech. In addition, a verb which co-occurs with the passive gĕi must be marked by the aspect marker le, otherwise the sentence with gĕi will not be interpreted as a passive sentence. Many works, such as Newman 1993a, b), Xu (1994), Yap and Iwasaki (1998, 2003) argue correspondingly that the passive gĕi is directly derived from the causative gĕi via the reflexive context. An important question is why the development from a causative use into a passive one does not take place in Thai. Yap and Iwasaki (1998) found out that hây in Thai takes only a volitional causer. Yap and Iwasaki (2003) argue that only nonvolitionality on the part of the causer can allow a passive interpretation to emerge. Therefore, the high degree of volitionality of the causer prevents hây from developing into a passive marker in Thai.

2.3 Summary

In summary, hây in Thai occurs in four constructions, namely, the ditransitive construction, the prepositional phrase, the causative construction and the complex construction. Hây functions as the ditransitive main verb, dative and benefactive markers, causative verb and clause connector or subordinator, respectively. Each of the four constructions has its own structural schema as below. The syntactic category of hây in each construction and function is specified under each structural schema in the rightmost column.

No.	Construction type Containing hây	Function of ḧây	Structural Schema
1	ditransitive construction	ditransitive (main) verb	hây+ NP1 + NP2 main verb
2	prepositional phrase	dative marker; benefactive marker	VP+$_{PP}$[hây +NP] preposition
3	causative construction	causative verb	NP1+hây+NP2+ VP causative verb
4.	complex sentence	clause connector	$_{S1}$[NP1+VP2] + hây + $_{S2}$[NP2+VP2] subordinator

Table 1. Functions and Structural Schemas of Hây

On the other hand, gĕi in Mandarin Chinese appears in four constructions, namely, the ditransitive construction, the prepositional phrase, the causative construction and the passive construction. Gĕi functions as the ditransitive main verb, dative and benefactive markers, causative verb and passive marker, respectively. The constructions in which gĕi appears, the functions and the structural schemas of all constructions containing gĕi appear in Table 2.

No.	Construction type Containing gĕi	Function of g̈ĕi	Structural Schema
1	ditransitive construction	ditransitive (main) verb	gĕi+ NP1 + NP2 main verb
2	prepositional phrase	dative marker	VP + $_{PP}$[gĕi + NP] preposition
			$_{PP}$[gĕi + NP] + VP preposition
		benefactive marker	$_{PP}$[gĕi + NP] + VP preposition
3	causative construction	causative verb	NP1+gĕi+NP2+ VP causative verb and passive marker
4	passive construction	passive marker	

Table 2. Functions and Structural Schemas of Gĕi

Some observations can be made regarding the functions, the structural schemas and the productivity of hây and gĕi in the functions specified in the tables above as follows.

(a) The clause connector use is possible for hây in Thai but is lacking for gĕi in Mandarin Chinese.
(b) The passive-marking use in possible for gĕi in Mandarin Chinese but is lacking for hây in Thai.
(c) The gĕi-marked dative PP in Mandarin Chinese can occur both before and after the main VP whereas the hây-marked dative PP can occur only after the main VP. That means there are two structural schemas of the dative gĕi whereas there is only one of the dative hây.
(d) Even though the gĕi-marked dative PP in Mandarin Chinese is claimed by many researchers to occur both before and after the main VP, only the preverbal gĕi-marked dative PPs, not the postverbal ones, are attested in a Beijing Mandarin speech corpus (Sanders and Uehara, 2012).
(e) The gĕi-marked benefactive PP in Mandarin Chinese can occur only before the main verb phrase.
(f) The postverbal [hây+NP] in Thai and the preverbal [gĕi+NP] in Mandarin Chinese can be ambiguous between the dative and benefactive interpretations if the main VP incorporates the sense of giving.

(g) The structural schemas of the causative and the passive gĕi are identical.
(h) The causative use of hây in Thai is productive but that of gĕi in Mandarin Chinese is not.

In section 3, we will argue for the relationship between constituent orders in Thai and Mandarin Chinese on the one hand and patterns of grammaticalization of hây and gĕi on the other.

3. Effects of constituent orders on patterns of grammaticalizationof hây and gĕi

In this section, we will point out how constituent orders in Thai and Chinese bear on patterns of grammaticalization of hây and gĕi in both languages. The constituent orders to be discussed in this section are those of a head vis-à-vis a modifier and those of a head vis-à-vis a complement. A complement is a syntactic category that is selected or subcategorized for by the head of a phrase. A complement is therefore semantically necessary for the head to become semantically complete. Some examples of complements are below.

(26) I *c*ut *a* tree.
(27) She *p*ut *a b*ook *on t*he table.

In (26) and (27), the direct object nominals *a t*ree and *a* book function as complements of the verbs cut and put respectively. In addition, the prepositional phrase *on* the table also functions as another complement of the verb put in (27) because the verb put is semantically incomplete without it. On the other hand, a modifier is an expression which limits or qualifies the meaning of a word, a phrase or a sentence. It is less semantically crucial to the meaning of a head than a complement. In other words, a modifier is more semantically peripheral than a complement. The underlined parts in (28) and (30) illustrate the modifiers in the sentences.

(28) The tree *i*s very *t*all.
(29) She *read* the *n*ewspaper *in* the *l*iving *r*oom.
(30) She *w*ent t*o s*ee *a m*ovie *a*fter *d*inner.

In (28), very modifies tall. In (29) and (30), the phrases i*n* the living room and after dinner modify the predicates in the clauses. The three sentences above are semantically complete without the modifiers. However, Langacker (1987) acknowledges that the demarcation between modification and complementation is sometimes hard to draw because the difference between them is a matter of degree.

It is generally known that the constituent orders in Thai and Mandarin Chinese are different in that Thai has the head-modifier constituent order whereas Mandarin Chinese has the modifier-head one. The difference in constituent order in the two languages is illustrated below. The adverbial modifiers in the examples are underlined.

Thai
(31) khun pay kɔ̀ɔn
you go first
'You go first.'

Mandarin Chinese
(32) nĭ xiān zǒu
you first go

'You go first.'

However, in case of the head and complement, the constituent orders in Thai and Mandarin Chinese are identical, that is, head-complement order. Therefore, in Mandarin Chinese, the modifier appears before the head whereas the complement appears after the head. On the other hand, in Thai, both the modifier and the complement appear after the head. In this section, we will point out that the constituent orders of the head and modifier and of the head and complement in Thai and Mandarin Chinese have some effects on patterns of grammaticalization of hây in Thai and gěi in Mandarin Chinese. To be specific, we will provide answers to the following questions in terms of different constituent orders in Thai and Mandarin Chinese.

1. Why does the benefactive [gěi+NP] occur only in the preverbal position, not the postverbal position, in Mandarin Chinese?
2. Unlike the benefactive [gěi+NP], the dative [gěi+NP] occurs both preverbally and postverbally in Mandarin Chinese. Why does the dative [gěi+NP] behave differently from the benefactive [gěi+NP]?
3. Why do the dative [hây+NP] and the benefactive [hây+NP] not occur in the preverbal position in Thai?
4. Why is the causative gěi not productive in Mandarin Chinese?
5. Why is gěi not used as a clause subordinator in Mandarin Chinese? In contrast, why is hây used as a clause subordinator in Thai? Moreover, why is the clause subordinator hây used highly productively in Thai?

The first question is why the benefactive [gěi+NP] occurs only in the preverbal position, not the postverbal position, in Mandarin Chinese. In order to answer this question, we have to understand the role of the benefactive PP in a sentence. The benefactive PP in a sentence serves as a modifier, rather than a complement, of the main VP because it is peripheral and can be omitted. It functions like an adverbial phrase modifying the main VP. It merely adds an extra piece of information regarding who benefits from the agent's action. Therefore, the preverbal benefactive [gěi+NP] matches the modifier-head constituent order in Mandarin Chinese. The postverbal benefactive [gěi+NP] would violate this constituent order in the language.

The second question is why the dative [gěi+NP] behaves differently from the benefactive [gěi+NP] in Mandarin Chinese. That is, the dative [gěi+NP] occurs both preverbally and postverbally whereas the benefactive [gěi+NP] occurs only preverbally. We argue that a dative constituent, which expresses a participant receiving a thing in a transfer event, is located somewhere on a continuum between a complement and a modifier. A recipient is sometimes analyzed as a semantically crucial participant for a transfer event to be semantically complete. This is because the transfer event is usually analyzed as consisting of three crucial participants, namely, a giver, a thing given and a recipient. However, the recipient is in some contexts perceived as not as semantically crucial as the other two participants as in Jo*hn* donates bl*ood* every month. On the other hand, the recipient in Jo*hn* gave a*n* expensive birthday present t*o* his mother, can be perceived to be a semantically crucial participant. That means the recipient can be perceived as a complement in some contexts and as a modifier in some others. Since the dative PP denoting a recipient fluctuates on the complement-modifier continuum, it is not surprising that the dative PP in Mandarin Chinese can occur both preverbally and postverbally according to the head-complement and modifier-head constituent orders in Mandarin Chinese. However, Sanders and Uehara (2012) found that the dative [gěi+NP] occur only preverbally in a speech corpus of Beijing Mandarin Chinese. This fact may suggest that the dative [gěi+NP] in spoken Beijing Mandarin Chinese is perceived to be modifier-like rather than complement-like. The examples below illustrate the preverbal dative [gěi+NP] in spoken Beijing Mandarin Chinese.

Data from Sanders' and Uehara's personal communication

(33) méi gěi nǐ xiě
not give you write
'I haven't written to you.'

(34) wǒ gěi nǐmen shuō ya
I give you (pl.) say PART.
'Let me tell you.'

The third question is why the dative and benefactive [hây+NP] do not occur preverbally in Thai. In the grammaticalization process, a string of [V1+NP1] + [V2+NP2] is reanalyzed into [V+NP1] + [P+NP2]. That is, the second verb is grammaticalized into a preposition marking a dative and benefactive NP. The PP functioning as a complement and a modifier occurs after the main VP. Therefore, the fact that the dative and benefactive [hây+NP] constituents do not occur preverbally matches the predominant head-complement/modifier constituent order in Thai.

The fourth question is why the causative gĕi is not productive in Mandarin Chinese. Unlike the benefactive gĕi and the dative gĕi, which are grammaticalized into prepositions, the causative gĕi is more verb-like in that it can be negated. Notice that the causative gĕi appears in the same position as the benefactive gĕi, i.e. the preverbal position, which bears two consequences. The first consequence is that the preverbal gĕi tends to be analyzed as the benefactive marker functioning as a modifier of the main VP, which corresponds to the predominant modifier-head constituent order in Mandarin Chinese, rather than as the causative verb. The second consequence is that the preverbal gĕi in some cases can give rise to ambiguity between the causative and the benefactive readings. It is found that the other causative verbs ràng and jiào are used more frequently than gĕi in order to avoid ambiguity as stated earlier in the paper.

The last question is why gĕi is not used as a clause subordinator in Mandarin Chinese but hây is in Thai? Moreover, why is the clause subordinator hây used highly productively in Thai? A complex construction consists of a matrix clause and a subordinating clause. Most subordinating clauses function as modifiers of the matrix VPs. In Mandarin Chinese, modifiers precede heads. Therefore, the postverbal position is not a perfect site for a verb to be grammaticalized into a subordinator in Mandarin Chinese. This is the reason why we do not find the postverbal subordinator gĕi in Mandarin Chinese. In contrast, the postverbal position is a perfect site for a verb to be grammaticalized into a subordinator introducing a subordinating clause in Thai because it matches the head-modifier constituent order in the language. That is why hây is used as subordinator with a high degree of productivity in Thai.

However, it is noted in some previous works that gĕi is used as a subordinator to introduce an adverbial clause occurring after a matrix clause in the head-adverbial clause order. This use of gĕi is exemplified by (35).

(35) Zhāngsān   chan̂g   gē   gěi
Zhangsan   sing   song   give
tā   tīng
he/she   hear
‘Zhangsan sang a song for him/her to hear.’

However, this construction is not attested in a Beijing Mandarin speech corpus according to Sanders and Uehara (2012). To express this meaning, the benefactive gěi is used instead as in (36).

(36) Zhāngsān   gěi   tā   chan̂g
Zhangsan   give   he/she   sing
gē
song
‘Zhangsan sang a song for him/her.’

The fact that the subordinator gěi is not found in spoken Beijing Mandarin Chinese confirms our hypothesis that the postverbal position is not a perfect site for gěi to be grammaticalized into a subordinator.

Another observation can be made regarding the grammaticalized passive marker gĕi in Mandarin Chinese. It is noted in Thepkanjana and Uehara (2008) that the passive gĕi in the structural schema [gĕi + NP + VP] has been developed into what Newman (1993b: 477) calls “the prefixal gĕi in passive constructions” as in (35).

From Newman (1993b: 477)
(37) tā   gĕi-mà-le
he   PASSIVE-scold-ASP
‘He/She was scolded.’

This phenomenon, which indicates that the second verb becomes the head which the prefix gĕi is attached to, corresponds with the modifier-head pattern constituent order in Mandarin Chinese.

4 Conclusion

This paper presents commonalities and differences in the grammaticalization of hây in Thai and gĕi in Mandarin Chinese and argues how different constituent orders in Thai and Mandarin Chinese bear on patterns of

Grammaticalization of the two verbs in the two languages. It is found that the common functions shared by hây and gĕi are (1) the ditransitive main use, (2) the dative-marking use, (3) the benefactive-marking use and (4) the causative-marking use. As for differences, hây, not gĕi, is used as a subordinator connecting two clauses in a complex construction whereas gĕi, not hây, is used as a passive marker. Five questions are posed regarding different patterns of grammaticalization of hây and gĕi in Thai and Mandarin Chinese. Facts about different patterns of grammaticalization of the two morphemes under discussion are accounted for in terms of different constituent orders in Thai and Mandarin Chinese, i.e. head-modifier/complement in Thai, modifier-head and head-complement in Mandarin Chinese. It is argued that the head-modifier constituent order in Thai seems to be compatible with postverbal grammaticalized morphemes whereas the modifier-head order in Mandarin Chinese seems to be compatible with preverbal grammaticalized ones.

Acknowledgments

This research work is partially supported by the Ratchadaphiseksomphot Endowment Fund of Chulalongkorn University (RES560530179-HS) awarded to the first author and a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (No. 24520416) awarded to the second author.

References

Dan Xu. 1994. The Status of Marker Gei in Mandarin Chinese. Journal *of* Chinese Linguistics, 22(2): 363-394.

Foong Ha Yap and Shoichi Iwasaki. 1998. 'Give' Constructions in Malay, Thai and Mandarin Chinese: A Polygrammaticization Perspective. Proceedings *of* the 34th Annual Meeting *of* the Chicago Linguistic Society, 421-438.

Foong Ha Yap and Shoichi Iwasaki. 2003. From Causative to Passive: A Passage in Some East and Southeast Asian Languages. In Eugene H. Casad and Gary B. Palmer (eds.) Cognitive Linguistics and Non-Indo-European Languages. Mouton de Gruyter, Berlin & New York, 419-445.

John Newman. 1993a. A Cognitive Grammar Approach to Mandarin Gei. Journal *of* Chinese Linguistics, 21(2): 313-336.

John Newman. 1993b. The Semantics of Giving in Mandarin. In Richard A. Geiger and Brygida Rudzka-Ostyn (eds.) Conceptualizations and Mental Processing in Language. Mouton de Gruyter, Berlin & New York. 433-485.

John Newman. 1996. Give: A Cognitive Linguistic Study. Mouton de Gruyter, Berlin & New York.

Kingkarn Thepkanjana and Satoshi Uehara. 2008. The Verb of Giving in Thai and Mandarin Chinese as a Case Study of Polysemy: A Comparative Study. Language Sciences, 30(6): 621-651.

Martin Haspelmath. 1990. The Grammaticalization of Passive Morphology. Studies in Language, 14(1): 25-72.

Robert D. van Valin, Jr. and Randy J. LaPolla. 1997. Syntax: Structure, Meaning and Function. Cambridge University Press, Cambridge.

Robert M. Sanders and Satoshi Uehara. 2012. A Syntactic Classification of the Synchronic Use of Gĕi in Beijing Mandarin: A Spoken Corpus-based Case Study of its Polyfunctionality. Chinese Language and Discourse, 3(2): 167-199.

Ronald W. Langacker. 1987. Foundations of Cognitive Grammar, volume I. Theoretical Prerequisites. Stanford University Press, Stanford, California.

Suda Rangkupan. 1997. An investigation of haŷ complex constructions in Thai. Available from: <http://www.buffalo.edu/soc-sci/linguistics/people/students/ma_theses/rangkupan/RANGKMA.PDF>