Computational Linguistics and Beyond:  
An Introduction

Chu-Ren Huang and Winfried Lenders

1. Background: COLING and Computational Linguistics

In the preface to Proceedings of COLING 2000, Martin Kay notes that when David Hays coined the expression Computational Linguistics back in the ’sixties, his intention was “to provide a more solid theoretical foundation for work on machine translation.” And, although this meaning has always lurked in the background, the term right from the beginning could not be restricted solely to machine translation—fortunately. Scholars from diverse disciplines—mathematics, linguistics, logic, information theory, statistics, the humanities—came together to pool their knowledge to attain a common goal: to create a machine that could translate from one language to another; or even more ambitiously: to enable a computer to behave like a human language user. The first COLING conference, in 1965 in New York, revealed this idea of an international, interdisciplinary forum for discussion. In following years, COLING became a truly international event, carefully orchestrated under the auspices of the International Committee on Computational Linguistics (ICCL), which “exists for the sole purpose of arranging” the conference every two years. “The members of the ICCL represent only themselves; certainly not the countries or institutions they come from” and they are independent from any particular scientific organization. (See Martin Kay at www.dcs.shef.ac.uk/research/ilash/iccl/.)

On that basis, COLING conferences have become a well-established occasion for the exchange of ideas and experiences in the field. COLING-conferences have tried to cover all aspects of the field and to reflect the main directions of its participants, such as machine translation, machine-aided translation, information retrieval and information extraction, grammar formalisms, parsing, semantic models, summarization, generation, natural
Computational Linguistics and Beyond

language understanding, and question/answering systems. Innovations and extraordinary ideas were always welcomed.

Over the years, Computational Linguistics, partly supplemented by Artificial Intelligence, has become a regular academic discipline in many countries all over the world. Computational Linguistic research has also taken a leading role in language engineering and language technology. The world-wide evolution of Computational Linguistics, on the other hand, has been influenced by several more or less ‘external’ conditions:

First of all, the rapid development of computer power has disposed of the most serious initial obstacle to natural language processing: Since speed of processing and memory size are no longer insurmountable hindrances, researchers are no longer forced to work with restricted small-scale vocabularies or knowledge bases (i.e., the so-called ‘toy’ systems). Both rule-based processing assisted by large lexica and the tour-de-force application of statistical processes and learning algorithms based on large-scale corpora have become widely available. Many research issues previously considered to be nearly impossible for machines to resolve (such as POS assignment) have received robust and replicable solutions with the stochastic methods supported by large corpora. With the past decade dominated by stochastic approaches and their impressive results, computational linguists are now looking back to integrating rule-based knowledge. Combinations of rule-directed deep analysis with robust statistically based shallow analysis systems are now under consideration and may help to bring language processing systems to previously unforeseen levels of efficiency and application.

A second development that has had an enormous and persistent influence on the development of Computational Linguistics concerns the availability and re-usability of language resources. Most importantly, affordable, fast computing power is a prerequisite to the wide application of large-scale language resources in Computational Linguistics. The first step has been the PC revolution, which facilitated the construction and availability of language resources in languages other than English. Once computing became available and accessible for many different languages, the collection of language resources followed. These resources in turn have provided the basis for natural language processing in those languages. The
The next step was a rising demand for language resources, when the issue of re-usability became central. On the one hand, the building of language resources is very labor-intensive, hence re-using language resources would save time and money. On the other hand, carefully constructed language resources can be used in different applications, and subsets of language resources can be combined under different criteria to form a different language resource. In other words, re-usable language resources can create added value. It is also crucial to note that linguistic acts are transient and that language varies with time and location. Hence some language resources are no longer replicable because of the change of contexts or the loss of native speakers. Thus, re-using language resources makes both budgetary and theoretical sense.

The re-usability of language resources depends crucially on the standardization of formats and tools. The worldwide endeavors to standardize the format of data and annotations started from the very beginning of Computational Linguistic research. Success came with SGML in the late ’eighties and—for Computational Linguistics— with the Textual Encoding Initiative (TEI) in the early ’nineties. But the implications of standardization go far beyond facilitation of information exchange. Standardization also makes international resource sharing possible. Textual corpora, lexica, grammars, and tools can all be shared on an international level nowadays regardless of the languages involved. Thus, it is no longer necessary to construct separate dictionaries (or grammars, or textual resources) for every new project. Instead, linguistic knowledge is provided in a theory-independent manner and in an international distributable standard. Standardization, in connection with the development of the World Wide Web, promoted these efforts and led to more international cooperation than ever before. In addition to this, network-based projects like WordNet and the recent Semantic Web Vision demonstrate that there are common structures between languages and cultures that may allow computerization of language-independent representation of knowledge, which is in turn accessible in a multi-lingual environment.

The third point concerns the introduction of multimodality into the scope of Computational Linguistics. The rapid technological progress of the last fifteen years has made it possible to incorporate communication
channels other than written text, i.e., spoken and visual channels. Computational Linguistics for more than 40 years has remained focused on orthographically rendered language. Indeed, there was significant research in speech recognition and speech production, called Digital Signal Processing even in the ’sixties. However, as the name suggests, these studies typically do not attempt to use linguistic structure and hence there was no real interaction with Computational Linguistics, although there were a few studies concerning the relationship between written and spoken language, such as phonological programs simulating the idea of minimal pair analysis. However, these studies all started with and focused on annotated phonetic input/output. Nowadays, since modern computers have multimedia capacities—including audio and visual—it is possible to integrate digital signal processing with the so-called ‘higher’ levels of understanding, syntax and semantics. We now are able to use parameters of spoken language like word stress and intonation patterns for the description and disambiguation of sentence and discourse structures. In the near future, the visual channel will become available for transcribing gestures and face expressions. This information will be used for structural analysis and for generation.

Last but not least is the wide accessibility of the web. Before the World-Wide Web became the dominant medium of information exchange, the applications of natural language technology were limited. It was difficult to imagine that Computational Linguistic technology would apply to many languages in the world and could be used daily by the general public. The Web as an information infrastructure drew users with no technological background to the computer. These users had common needs to access information using daily language, which in turn, redirected the focus of human language technology back to natural language processing. Although Computational Linguistics has yet to prove that it delivers the optimal solution for cross-lingual and cross-modal information access problems, its does clearly identify and define the relevant issues. The dominance of the web also brings a renewed focus to semantics and the relation between language and information in Computational Linguistics.

COLING Conferences are organized to encourage new developments and ideas, regardless of theoretical, geographic, political, or commercial
interest. Conference presentations (papers and project notes screened by a rigorous formal reviewing system) are published in the conference proceedings. In addition to the presentations, COLING conferences have fostered a pleasant, yet intellectually stimulating environment that allows for lively discussion. Important in this context are plenary lectures and panel discussions where new ideas and developments may be presented. A more recent tradition (just a little more than two decades old) is the sponsoring of pre-COLING and post-COLING events, such as tutorials and workshops. At these events, particular problems and trends within the field can be discussed at greater leisure.

2. Computational Linguistics at the start of the new century: New frontiers

Held at the beginning of a new century, COLING2002 was the first COLING conference not held in Europe, America, or Japan, but outside the industrialized West. The conference setting thus symbolized in a way the unifying theme of the papers herein: to boldly explore the frontiers of computational linguistics. The present volume consists of the plenary lectures and the results of two workshop panels. (It thus complements the already published conference proceedings of the 19th COLING.) The following surveys these papers from the frontier of computational linguistics.

2.1 From Structure to Meaning: Building upon foundations in Computational Linguistics

“If I have seen farther than others, it is because I was standing on the shoulders of giants.” As Sir Isaac Newton acknowledged, scientific progress is built upon the knowledge accumulated by previous research. Computational Linguistics as a mature discipline is no exception. It is appropriate that some of the frontiers explored in this volume are solidly based on a historically strong foundation in Computational Linguistics: the mapping from structure to meaning.
The initial development of Computational Linguistics was stimulated by the observation that linguistic facts can be described by structural rules, and the claimed implication was that grammatical structures correspond to meaning representation. However, this foundational methodology gradually yielded to stochastic approaches in the ’eighties. As the system, size of data, and applications scaled up, it was discovered that, despite Chomsky’s earlier criticism, the ‘quick and dirty’ Markov models, in general, obtained better empirical results than the so-called rule-based approaches. Research on mapping from structure to meaning continued, but the emphasis was no longer on accounting for the mapping rules, but rather on how structures as objects could be statistically manipulated. It was not until the beginning of this century that computational linguists realized that there were limitations to stochastic models, and that linguistic processing paradigms might offer some answers.

Uszkoreit’s keynote paper ‘New Chances for Deep Linguistic Processing’ explicates this position as well as charts a model for future development in Computational Linguistics. He envisions that Information Extraction (IE) will be the dominant research paradigm for computational linguistics in the future. And since IE has an increasing demand for high quality and precise information, deep linguistic processing must be used to refine the coarser results obtained by the stochastic approaches. He also envisions a century of large-scale international collaboration. For the international collaboration to work, researchers from different sites must share tools and resources developed and written in different languages without loss of information. This requires highly sophisticated linguistic annotation on tools and resources. And the only way that such information can be provided is with deep linguistic processing.

Fillmore’s work on FrameNet, and his keynote paper entitled ‘Linking Sense to Syntax in FrameNet’, may be a case of “a giant standing on his own shoulders”. Fillmore’s earlier work on Case Grammar literally generated a whole paradigm of research on argument roles in Computational Linguistics and is still influential. His more recent work on Construction Grammar brought focus back to how meaning interacts with syntactic structures. And in the past ten years, his focus has been on FrameNet theory and the lexical knowledgebase which aims to specify both the
semantic information (in terms of event frames) that is lexically encoded and predict how this information is projected syntactically. In this paper, he and his co-authors describe how the frame information is linked to syntactic structures.

Uszkoreit and Fillmore approach the structure-meaning link from different directions. Uszkoreit focuses on how deep linguistic (and structural) processing leads to fine-grained linguistic information. And Fillmore underlines how fine-grained lexical semantic knowledge can be used to predict linguistic structure.

T’sou’s summary paper on the SigHAN (ACL Special Interest Group on Chinese Language Processing) panel entitled ‘Chinese Language Processing at the Dawn of the 21st Century’ also fits in with the theme of building on previous computational linguistic research. While Chinese is a language with the most speakers in the world, its computational linguistic research was developed much later and less fully than English. Since this was the first time that COLING had been held in a Chinese speaking environment as well as the inaugural meeting of SigHAN, a panel was convened to examine the development of Chinese language processing over the past ten years as well as to predict its developments in the next 10 years. It would be fair to characterize the lack of progress in the early stages of Chinese computational linguistics as highly correlated with the field’s original reliance on processing orthographically represented texts. Since the non-alphabetical orthography of Chinese characters was recalcitrant to early computer technology, no real progress in Chinese language processing could be made. The dawn of a new century brought on several breakthroughs to overcome the orthographic problems, such as intelligent character encoding and robust word segmentation. These breakthroughs coincided with the new emphasis on multi-linguality and multi-modality and returned research focus to knowledge content in computational linguistics. It is interesting to observe that one of the central issues discussed was the need to focus on semantic processing, as well as to study specific Chinese linguistic structures that would shed light on semantic content.
2.2 From World-Wide Web to Semantic Web: Challenges to Computational Linguistics

The Web had a profound and positive impact on computational linguistics in the past decade as the popular use of the Web defined some of the hottest research topics in computational linguistics, such as Information Retrieval and Question and Answering. Computational linguistics also gained an important media on which to test and apply its research results. But will there be continued synergy between the future Web and the development of computational linguistics? This issue gained some urgency after the bold prediction by Tim Berners-Lee and colleagues (in the May 2001 issue of *Scientific American*) that the Semantic Web will be the new Web. On the one hand, the initiatives to convert unstructured web-based digital information into a semantically structured knowledge base seem well-suited for computational linguistic research. On the other hand, however, this claim emphasizes that semantics is for computers—not humans—to comprehend. This emphasis may be orthogonal to important premises of computational linguistic research.

The urgency and potential long-term impact of this issue was brought into focus by two panels during COLING2002. The first plenary panel, organized by Hans Uszkoreit, was entitled ‘Semantic Web: A New Challenge for Language Technology.’ The second panel, entitled ‘The Roles of Natural Language and XML in the Semantic Web,’ was organized by Nancy Ide, Laurent Romary, and Graham Wilcock for the post conference workshop NLPXML-2002. Unfortunately, we are only able to include the summary paper from the second panel in this volume. However, the content of the plenary panel is summarized and discussed here since it deals with the basic issues involving computational linguistics and the Semantic Web.

The COLING2002 Semantic Web panel, which consisted of Paul Buitelaar, Eduard Hovy, Chu-Ren Huang, and Hans Uszkoreit, identified six fundamental challenges posed by the Semantic Web to language technology:
1. The employment of language technology (LT) for the construction of useful ontologies.
2. The exploitation of Semantic Web ontologies for LT such applications as information extraction.
3. The challenge of (partially) automating the detection and annotation of concepts.
4. The utilization of the Semantic Web as a resource for machine learning in NLP.
6. The Semantic Web and language variation.

What emerged from the panel, including active participation from the audience, was a very dynamic vision of the Web and computational linguistics. First, even though researchers do not have a clear consensus as to whether the Semantic Web will replace the current Web, they do agree that content with richer semantic structure is definitely in the future of the Web. Second, they also agreed that multi-linguality and multi-modality will become a defining feature of web-based information. Hence, a conceptual infrastructure for exchanging knowledge among different models or languages is definitely needed, though it may not necessarily end up being in the form of ontology as proposed in the Semantic Web. Their reasons vary, but they are mainly concerned with the growing demand for more precise content knowledge, as well as the trend towards a multilingual web. One observation worthy of special note was made by Hovy. He cited current web content statistics and predicted that by the end of 2005, there will be over 2,900 billion words on the web and the non-English resources combined will out-number English resources for the first time in the short history of the Web. In addition, Huang made the observation that with a significant number of new web users expected to come online with little or no English knowledge, the use of the Web will also undergo a qualitative change towards real multi-linguality. The most salient case is an ambitious program to give web access to five billion farmers in China. These complementing perspectives made it clear that the simplistic view that the Semantic Web ontology only needs to deal with English is a fallacy. Once multi-lingual texts as well as cross-lingual
information extraction are both involved, the usefulness of computational linguistics cannot be underestimated.

Given the issues and challenges, the discussion in the set of papers on ‘The Roles of Natural Language and XML in the Semantic Web’ can be appropriately highlighted. In addition to the introduction (by Wilcox) and the conclusion (by Ide), there are four position papers included in this summary: Buitelaar introduces the Semantic Web vision and implementation, Pareja-Lora discusses the experience of using RDF for hybrid web page annotation, Bryant discusses how domain knowledge can be represented with DAML, and Lin makes several proposals regarding how to bring natural language to the Semantic Web. These papers show how the combination of natural language processing technology with higher level annotation will positively affect our use of the web in the future. The Web with semantically annotated content is here to stay.

2.3 Language and Biology

Although the juxtaposition of linguistics and biology seems oxymoronic, they closely relate to each other and present new and exciting frontiers in Computational Linguistics.

The synergy between linguistics and biology is based on the human perspective that cognition and genetics are two areas that define our species and that these are the areas where many scientific questions remain unanswered. Merging these two main concerns, bioinformatics aims to solve the functional and informational structure of genomes. On the other hand, the study on Language and Evolution specifically integrates results from genetic studies to account for how language comes into being and how languages change. Interestingly, both fields adopt Computational Linguistic techniques. They were also both explored during COLING2002.

Bioinformatics, appropriately, was also one of the first areas with major scientific advances in the new millennium. The new century saw the complete solution of genetic sequencing. The next unsolved mystery is the explanation and prediction of the functions of genomes. As discussed in §2.1 above, the mapping from structure to meaning and function is a basic tenet of computational linguistics. Hence, many see bioinformatics as the
new area where these lines of inquiry will be richly rewarded with new applications and new breakthroughs. Bioinformatics has become the new hottest topic at computational linguistics conferences since 2001. COLING 2002 included a tutorial on bioinformatics. Since the tutorial contains mostly non-original materials, it is not included in this volume.

Language and Evolution, although posing no less a mystery or challenge, has not received as much attention in computational linguistics because of its more esoteric nature. It is important to note, however, that the study of language and evolution and computational linguistics share two fundamental issues: what language is; how languages work. Computational linguistics is mainly concerned with accounting for the present state of contemporary languages even though there are no theoretical constraints against the computational processing of ancient languages. The study of language evolution, on the other hand, is concerned with the origin of language and how languages evolved over human history. However, if we take the view that macro changes are the sum of micro changes, and that a language is a living organism, we shall arrive at the conclusion that the same motivations and explanations should work both for language evolution and daily language use. In other words, language evolution is simply the accumulated sum of the language use of people over a long period of time; it does not necessarily apply with a different set of rules on a different cognitive level.

William S-Y. Wang is another giant standing on his own shoulders. The current keynote paper that he co-authored with two other colleagues stands on his early work, including the theory of lexical diffusion for language changes and early computational simulation of linguistic rules and dialect variation. Wang et al. discuss several mathematical models and their computational simulation of language change and language emergence in their paper ‘Evolutionary Linguistics and Computer Modeling.’ What it shares with prototypical computational linguistics is more than statistical models and computational simulation. It also runs experiments and verifies results based on large scale language resources. Its innovative combination of results from genetic studies and computational simulation of language data opens up a whole new line of research.
2.4 From Sharable Resources to International Standards

We observed in the first part of this paper that the availability and re-usability of language resources played a central role in the recent developments of Computational Linguistics. Work along this line toward sharable language resources is one of the leading frontiers of research in Computational Linguistics for the new century. It underlines research methodology mentioned specifically in three of the papers: Uszkoreit’s paper on deep linguistic processing, T’sou’s paper on the development of Chinese language processing, and Graham et al.’s paper on Semantic Web and XML. Large scale language resources also played a substantial role in the research on both FrameNet and language evolution. In other words, all the research reported in this volume used language resources. During the COLING2002 conference, there were also at least two workshops that were specifically devoted to language resources and standards: the 3rd Workshop on Asian Language Resources and International Standardization, and the 2nd Workshop on NLP and XML (NLPXML-2002). The consensus from these works for the field of computational linguistics is that both standards and infrastructures for resource-sharing must be the cornerstone of future development. Building on this consensus, there have been three significant recent breakthroughs in the frontier of language resources for Computational Linguistics. We briefly comment on three of the most important recent developments to underline the crucial role that language resources will continue to play in computational linguistics.

The most salient one is perhaps the founding of ISO/TC 37/SC 4 on Language Resources Management (http://www.tc37sc4.org/). The formation of this ISO subcommittee and the expected ratification of the drafted standards in the future is making a statement that language resources and related technology are ready to play a role in our daily life. There are five working groups under ISO/TC 37/SC 4: Basic descriptors and mechanisms for language resources, Representation schemes, Multilingual text representation, Lexical database, and Workflow of language resource management. These workgroups and the subcommittee will integrate research results for the field of computational linguistics.
The second breakthrough was the founding of OLAC (Open Language Archives Community, http://www.language-archives.org) in 2000. While the ISO subcommittee aims to formalize standards that can be followed internationally, OLAC aims to build a community-wide consensus among language archive builders and users. The consensus must be built upon the actual sharing of data and knowledge. In other words, OLAC is more than a set of standards, it is also an infrastructure for sharing language resources, tools, and know-how. It is significant that the OLAC community includes both language engineers and field linguists. A critical issue regarding our cultural heritage is the archiving and preservation of endangered and other less-studied languages. Such critical work cannot be successfully carried out without computational and field linguists working together as a community.

The third breakthrough involves the continuing productive international collaboration to develop standards as well as sharable resources. One of the pioneering projects is the EAGLES (Expert Advisory Group on Language Engineering Standards) project that was completed in 1996 under the guidance of Antonio Zampolli, an ICCL member. This was the first time that a set of standards was created by experts from different countries to cover a wide range of language engineering needs and languages. The initial success of EAGLES led to further developments such as ISLE (International Standards for Language Engineering, 2000-2002). ISLE included both American and European experts as formal partners, and experts from Asia as participants. From the above and on-going work, such as a series of workshops on an international roadmap for language resources organized by ELSENET, we can see the clear emergence of an international framework for collaboration. This framework and the resultant infrastructure for resource-sharing will lead to a genuine integration of international experts and open new frontiers by successfully dealing with common issues such as multi-linguality and cross-domain knowledge processing.
3. Conclusion: Computational Linguistics and Beyond

In this introductory chapter, we laid out the background of the COLING conferences as well as the recent developments that facilitated the advance of the field of computational linguistics. Anchored on these developments, we introduced the frontiers of computational linguistic research as illustrated by the leading research papers collected in this volume. We see that even though the vision of the Semantic Web as well as the creative synergy between language and biology give tremendous momentum to new research in computational linguistics, the sustained growth of the field still must be rooted in the basic framework of the mapping from structure to meaning, as well as the essential infrastructure of re-usable language resources. The central issue of how to convert digital information to structured knowledge goes beyond computational linguistics. It will be a theme that technology and science will come back to again and again in the 21st century.
Chu-Ren Huang’s Biographic Sketch and Research Outline

Chu-Ren Huang is a research fellow at the Institute of Linguistics, Academia Sinica. He is the vice president of the Linguistic Society of Taiwan, founding council member of the Open Language Archives Community (OLAC), and founding co-chair of the Asian Language Committee. He currently serves on the editorial boards of Journal of Chinese Linguistics, Language and Linguistics, and Computational Linguistics and Chinese Language Processing.

He received his Ph.D. in linguistics from Cornell University in January 1987 and started his research position at Academia Sinica in the same year. He has held adjunct or visiting positions in various international academic institutions, including City University of Hong Kong, CNRS, CSLI, Peking University, University of Pennsylvania, and UCSB. He served as president and secretary of ROCLING/ACLCLP and was a founding executive board member of both IAACL and LST.

Over the past 17 years, he has played an active role to promote research on Chinese computational and corpus linguistics. He has directed or co-directed the successful construction of the following Chinese language resources: CKIP lexicon, Sinica Corpus, Classical Chinese Corpora, Sinica Treebank, and Academia Sinica Bilingual Ontological Wordnet. His linguistic research focus shifted from earlier work on GPSG and LFG to recent emphasis on lexical semantics, which led to the development of the MARVS theory. Both lines of research led to his current work on Chinese WordNet as well as merging CWN with English WordNet and Upper Ontology. His research vision is to develop lexical knowledge framework and infrastructure that will both anchor web-based knowledge engineering as well as shed light on the internal representation of human knowledge.

List of Publications:
http://corpus.ling.sinica.edu.tw/member/churen/
Research Accomplishments:
1. Academia Sinica Bilingual Ontological Wordnet, Sinica BOW  
   http://BOW.sinica.edu.tw  
2. Sinica Corpus 研究院語料庫(現代漢語平衡語料庫)  
   http://www.sinica.edu.tw/SincaCorpus/  
3. SouWenJieZi: A Linguistic KnowledgeNet 搜文解字(漢語語文知識網路)  
   http://www.sinica.edu.tw/~tibe/2-words.html  
4. Adventures in Wen-Land: 文國尋寶記(搜文解字 II: 中小學語文知識網路)  
   http://www.sinica.edu.tw/wen/  
5. Early Mandarin Corpus 近代漢語標記語料庫  
   http://www.sinica.edu.tw/Early_Mandarin/  
6. Language Archives and Linguistic Anchoring 語言典藏及語言座標  
   http://corpus.ling.sinica.edu.tw/project/LanguageArchive/

Winfried Lenders’ Biographic Sketch and Research Outline

Winfried Lenders, born in 1943, received his Ph.D in philosophy in 1970. Since 1974, he has served as Professor of Linguistic Data Processing/Computational Linguistics at the Institut für Kommunikationsforschung und Phonetik of the University of Bonn. From 1982 to 2003, he was Director of the Studium Universale at the University of Bonn. In 1986, Lenders was the local organizer of the 11th International Conference on Computational Linguistics (COLING ’86) and has stayed a member of the International Committee on Computational Linguistics since. Between 1993 to 1997, he was President of the Gesellschaft für Linguistische Datenverarbeitung (Society for Computational Linguistics and Language Technology) and also served as a Senator of the University of Bonn. He was re-elected as Senator in both 1998 and 2000. In 1997, Lenders was a Guest Professor at the Language Institute of Waseda University in Tokyo and in 2002, he acted as Program Chairman for the 19th International Conference on Computational Linguistics (COLING2002) held in Taipei.

List of Publications:
http://www.ikp.uni-bonn.de/~wle/