Computational Linguistics and Beyond

Edited by

Chu-Ren Huang and Winfried Lenders

Institute of Linguistics, Academia Sinica
Whither Linguistics? — an editorial

In launching this new series of monographs, FRONTIERS IN LINGUISTICS, it is an appropriate moment to glance back on some of the influences that have shaped our field, and look ahead to where we may be going. Here are some thoughts that occur to me, necessarily a personal view.

Interest in language dates back to different times, varying according to culture and tradition. While these traditions may be traced to many distant parts of the ancient world, there is no doubt that the main-stream linguistics of today derives most directly from a European tradition, particularly from Indo-European studies and the comparative method. This tradition was stimulated at various times with the discovery of ancient inscriptions, and with the realization of how far-flung languages may have ‘sprung from some common source,’ in the words of William Jones of 1786.

Given such an orientation, it is not surprising that historical linguistics and evolutionary biology, which started within decades of each other as systematic sciences, should be concerned with each other’s methods and theories. Discovering ancient inscriptions is not unlike unearthing fossils, and comparing grammars is not unlike comparing anatomies. The exchanges between Darwin and Schleicher are well-known; cf. Schleicher’s famous monograph of 1863 relating Darwin’s theory to linguistics. Less known is the fact Darwin kept detailed notes of the language developments of his children, an early study in language acquisition.

Our field took a major turn in the 20th century, channeled by two Saussurean dichotomies: diachrony and synchrony, and langue and parole. The former dichotomy put into relief the internal organization of a language, injecting a balance to the exclusively historical orientation of earlier decades. The distinction was a timely and useful one for the field as thousands of indigenous languages in many parts of the world came into view around then, and these languages needed to be described in terms of their own synchronic structures, rather than with a ‘Standard Average
European’ type of template. So the linguistics with this kind of practical mission in mid-century was often called ‘descriptive’ or ‘structural’; and later called ‘taxonomic’ as a mark of condescension by the more theory-minded generativists.

The other Saussurean dichotomy of langue and parole, in its many variant forms, encourages the attribution of various abstract units and principles to langue, even when the evidence in the parole may not be clearly there. In positing these abstractions, linguistics may have been simply accepting the general climate of reductionism in research, which has led to many enviable successes in the physical sciences. Unfortunately, however, there has not been sufficient reflection on to what extent such reductionism can be profitably pursued in the science of language. Indeed, one linguist recently wrote that abstractions such as ‘the subjacency constraint, the empty category principle, and the binding principles are theory-internal affairs and simply do not exist in usage-based theories of language.’1 In this view, at least, the pendulum has swung too far in the direction of langue and formalism; it is time to take parole, or usage, a lot more seriously.

“Whereas speech is heterogeneous,” Saussure had said, “language, as defined, is homogeneous.”2 Careful attention to usage will reveal that ‘orderly heterogeneity’,3 to use the felicitous phrase of Weinreich et al., abounds at every level of language, for the idiolect as much as for various layers of the speech community. Answers to the deeper questions will always elude us unless we develop methods and models to understand this key aspect of language. The eminent geneticist Dobzhansky had observed that: “Nothing in biology makes sense except in the light of evolution.”4

2 de Saussure, Ferdinand. 1959. Course in General Linguistics. Quote is from p.15.
believe this observation is every bit as true for linguistics. Heterogeneity is a basic driving force that shapes languages into the objects they are, much as it is the force shaping biological organisms. Ignoring heterogeneity (or variation) to study an illusory ‘homogeneous language’ has deprived our field of much of its natural significance. It is encouraging that the past several decades have seen much infusion of evolutionary thinking into our field.5

20th century linguistics had a penchant to declare and emphasize its autonomy, particularly as the first departments of linguistics came into being around mid-century. Perhaps this was necessary for a new discipline to assert its individuality and identity, having been incubated in other more established departments, (be it anthropology, psychology, or some language department), and needing to justify the splitting off. But the emphasis on autonomy sometimes had the undesirable consequence of moving away from the healthy inter-disciplinary synergies of the Darwin-Schleicher type.

Fortunately, in spite of this anxiety for autonomy, linguistics never fully isolated itself. As the 21st century begins, we see vibrant interactions among investigators in many fields relevant to our understanding of language. I shall mention only two such frontiers here. On one frontier, geneticists, anthropologists, and computer scientists have joined forces with linguists to provide us with a panorama of human evolution, from the first hominids that stood up to the cultural achievements of recent millennia, including the great diasporas to all parts of the world. A leading contributors to this multi-disciplinary area, Cavalli-Sforza, has provided an authoritative overview of recent developments in his GENES, PEOPLE AND LANGUAGES.6 Going down a complementary path are the many computer studies on simulation and modeling, which are attracting a great deal of important pioneering research.7

7 A representative sample can be seen in Christiansen, Morten H. and Simon
There are at least two reasons why we must attend to this evolutionary panorama seriously. This is, after all, the macrohistory within which language transitioned via several crucial steps from crude gestures and simple calls into the wonderful mental instrument we all have today. Just as importantly, we must take their results into account when we reconstruct ancient language families, since these are inextricably interwoven with the ancient migrations. Regardless of whether it is the linguist, the geneticist or the anthropologist that is doing the reconstruction, ultimately there was only one past. Each discipline provides an additional window on this past, and they all need to be reconciled with each other to arrive at true knowledge.

On another frontier, psychologists, ethologists, and neuro-scientists have been comparing our communicative and cognitive behaviors with those of other animals to find out why language is possible only in our species. Since we share the great bulk of our DNA with other animals, particularly with the chimpanzee, this question is all the more intriguing: In what ways is our neuro-anatomy distinctive that we are language-ready and they are not? How do these distinct language-ready features in our bodies get expressed ontogenetically, and how did they emerge phylogenetically?

On this frontier, it is especially heartening to see the explosive growth of research which exploits the recent breakthroughs in brain imaging. Whereas earlier linguistic investigations on the brain were largely restricted to the more accessible cortical surface, the new technology allows remarkable 3-D viewing of the entire brain volume during linguistic tasks, with refinements in spatial and temporal resolution being achieved at an accelerating pace. Even at this early stage, we are already gaining a deeper understanding of how very diffusely and widely distributed linguistic

---


behavior shows up in the entire brain, giving no support to any hope of finding any special ‘language organ.’ To make the problem even more challenging, we are recognizing more and more how variable and plastic human brains are, even at the level of monozygotic twins. In any case, studying language with the powerful new tools of neuroscience will surely yield exciting results for many years to come.10

These two frontiers of evolutionary linguistics and neurolinguistics will inevitably converge, perhaps several decades down the road. The findings on language behavior using methods of neuroscience in the form of nerve tissues and proteins must eventually be explained in the form of DNA and genes, and the evolutionary trajectories which produced them only in our species. Recent claims of discovering ‘a language gene’ surely must not be taken literally. Something as complex and diffusely distributed as language cannot be anything but multigenic and heterogeneous. We are only at the beginning of an exciting journey to discover the brain mechanisms that make language possible, and the phylogenetic events that brought about these mechanisms.

There are other important frontiers for linguistic research, interdependent and interactive with the two new ones I just mentioned. Historical linguistics continues to make steady progress, gradually sorting out the effects of vertical and horizontal transmission, including unusual contexts which produce creolization. Studies on language acquisition are getting streamlined with large shared data-bases, drawing upon an increasing range of languages. Increasing attention is being given to the various forms of selective language impairment, for both spoken and written language. And the description and analysis of linguistic patterns from a great diversity of languages, which forms the bedrock of linguistics, continues to grow robustly along complementary lines, even if the labels emphasized may be different; such as construction grammar, functionalist grammar,

10 In this connection it is gratifying to note that some results on Chinese are already being reported in the comprehensive volume, Handbook of East Asian Psycholinguistics (vol.1: Chinese), ed. by Ping Li, Lihai Tan, Elizabeth Bates, and Ovid J. L.Tzeng, forthcoming from Cambridge University Press.
cognitive grammar, etc. As long as the work produces results on language which are objective and cumulative, our horizons will be extended.

None of the research alluded to above can be carried out effectively anymore without the computer. There is a Chinese saying, dating back to the Analects of Confucius: 工欲善其事，必先利其器. To do one’s work well, one must sharpen his tools. The tool that has become as commonplace as pencil and paper for any science is the computer. It is auspicious that the first volume of this series should be an anthology on COMPUTATIONAL LINGUISTICS AND BEYOND.¹¹ This anthology gives a good perspective on how the computer is currently used in linguistics. It serves well as the starting point of this series of monographs. Let us hope that future monographs will emulate this tradition as we continue to push forward the many frontiers of our science beyond their present boundaries.

William S-Y. Wang, Editor
June 2004

Contents

Whither Linguistics? — an editorial
William S-Y. Wang................................................................. i

Computational Linguistics and Beyond: An Introduction
Chu-Ren Huang & Winfried Lenders................................. 1
   Biographic Sketch............................................................. 15

FrameNet and Representing the Link between Semantic and
Syntactic Relations
Charles J. Fillmore, Josef Ruppenhofer & Collin F. Baker
   Introduction.................................................. Winfried Lenders 17
   Article................................................................. 19
   Biographic Sketch.................................................... 60

Computational Studies of Language Evolution
William S-Y. Wang, Jinyun Ke & James W. Minett
   Introduction............................................. Chu-Ren Huang 63
   Article................................................................. 65
   Biographic Sketch.................................................. 105

New Chances for Deep Linguistic Processing
Hans Uszkoreit
   Introduction.................................................. Winfried Lenders 109
   Article................................................................. 111
   Biographic Sketch.................................................. 135
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
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
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 reëlected 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/
The semantic gap in current natural language programs attracts immediate attention when one considers optimizing natural language processing tasks, e.g., to get better retrieval results in a future semantic web. Several large projects were set up in the past decade to fill this semantic gap. One of these is the FrameNet project, explained in detail in the paper that follows. At COLING2000, the father of case grammar himself, Charles Fillmore, presented this paper, co-authored with Josef Rippenhofer and Collin F. Baker. Fillmore proposed Case Grammar and elaborated on it in the late 'sixties, and in the following decade developed it within his so-called frame semantics, and in the related theory of Construction Grammar. Here Fillmore specifies the exhaustive lexicographical basis of this theory, to be developed in a multiple-year research project funded by the U.S. National Science Foundation.

Although FrameNet looks like a lexicon—or more aptly a lexical knowledge base—it is much more than that: its basic units are not individual words (such as provided by traditional lexica), but conceptual structures called frames, which consist of configurations of frame elements. The frame elements represent the semantic roles or functions that are associated with particular words in particular sentences or sets of sentences. The paper provides a description of these fundamental components of the system and illustrates the methods and tools by which FrameNet is constructed and used.

One of these tools is the annotation tool by which sentences taken from large corpora are annotated with sets of frame elements like agent, degree, instrument, or purpose. Patterns of these frame elements represent the frames, which are the real and essential units of the system. Since the
frame elements in this annotation process are associated with lexical units, the involved frames together with the different syntactic realizations and valence patterns in which they occur are retrievable by these units. The frames are interrelated by frame-to-frame relations, like subordination and inheritance, which give the system the characteristics of a hierarchical network in which the transfer of properties from one node to subordinate nodes is possible.

In addition to these features the FrameNet database allows the derivation of what Fillmore calls Kernel Dependency Graphs (KDG). These are graphical representations of particular frames that are evoked by certain possible predicators in sentences (e.g., verb, noun, or adjective). These KDGs show the essential frame elements and the combinatorial properties of lexical units when they occur in a complex sentence. The KDG are instantiated by head lexicalization and gap filling processes. For every simple predication a simple KDG is derived and for complex sentences there are multiple KGD cross-references linking them to each other.

Clearly the aim of FrameNet is not to cover a broad variety of different sentences or texts, but to collect as completely as possible the underlying frames that represent the syntagmatic relations by which linguistic units are connected and form the semantic structure of sentences. Although it is neither a lexicon with a broad coverage of lexical units nor an ontology which covers the knowledge of a domain, FramNet is an important instrument for further semantic processes as are needed in the semantic web.
FrameNet and Representing the Link between Semantic and Syntactic Relations

Charles J. Fillmore, Josef Ruppenhofer, and Collin F. Baker

1. Abstract

FrameNet is a corpus-grounded frame-based computational lexicography project funded by the US National Science Foundation. Now in its sixth year, the project generates information about the articulation of the semantic and syntactic requirements of English lexical items and presents this information in a variety of web-based reports. The name “FrameNet”, inspired by “WordNet”, reflects the twin facts that the project exploits the theory of Frame Semantics and that it is concerned with the semantic networks through which word meanings are connected with each other.

The present paper has two functions. The first is to explain to the linguistic and NLP communities the goals, the procedures, and the intended final products of FrameNet. The second is to show how the FrameNet data can serve as a research tool for a particular derivative task; this secondary task involves extracting from the data developed in the main project a new resource, called kernel dependency graphs (KDGs). KDGs consist of structured clusters of lexical items, each such cluster containing a governor and the lexical heads of all of its dependents, with each of the latter marked for the semantic role it bears to the governor. KDGs can be displayed as modified dependency trees where nodes are lexical items,

---

1 The National Science Foundation grant NSF ITR/HCI #0086132: “FrameNet++: An On-Line Lexical Semantic Resource and its Application to Speech and Language Technology”, September 2000–August 2003. PI is Charles Fillmore; co-PIs are Srin Narayanan, Dan Jurafsky and Mark Gawron; Project Manager is Collin Baker; lexicographic consultant is Sue Atkins. Detailed information is available at http://framenet.icsi.berkeley.edu/~framenet.

2 By governor we mean the lexical head of a headed phrase: the head verb of a VP, the head noun of a NP, the head adjective of an AP, etc.
Computational Linguistics and Beyond

and the semantic roles connecting governors to their dependents are indicated as labels on the branches. Prepositional and other function-word indicators of semantic roles are also represented. Put in “slot-filler” terminology, the head of a KDG evokes a configuration of slots named by the labels on the branches; the fillers of these slots are the lexical heads of the dependents.

KDGs can be generated automatically from existing FrameNet annotations yielding information about collocations and semantic-syntactic linking for sentences that FrameNet annotators chose to annotate. We look forward to the development of systems to assign FrameNet-like annotations to new texts (Gildea & Jurafsky 2002, Manning et al. forthcoming). KDGs derived from a large corpus will provide a database offering reliable information about frequencies and collocations in the kind of corpus being used, and KDGs recognized in specific documents can be read off as indications of the subject matter and basic claims of given passages in the documents.

2. What should we expect from a lexical resource?

The types of information needed for any lexical resource capable of serving NLP applications should include at least these:

(1) representations of the meaning of each lexical unit (LU),

---

3 One of the settings in the FrameSQL FrameNet viewer designed by Hiroaki Sato, accessible through the FrameNet public website, allows the display of KDGs.

4 These choices do not necessarily reflect frequency, but annotators are instructed to prefer sentences with frequent or significant collocates.

5 Briefly, a lexical unit is a word paired with one of its senses, that is, a word in a single meaning (Cruse 1986). Given this usage, therefore, what we think of as a polysemous word is really a set of separate lexical units that use a single lexeme. The full set of distinctions required in FrameNet is fairly complex, but the concept of lexeme covers a set of word forms that as a group can be associated with a single meaning. Thus, bring - brought - brings - bringing are different forms of a single lexeme. The full set of contrasts we need are expressed as word-form, lexeme, lemma (included multi-lexemic lemmas), and lexical unit,
(2) various types of LU-to-LU relations,
(3) information about a word’s capacity to combine with other linguistic units to form phrases and sentences,
(4) semantic information associated with individual words that allows us to interpret the phrases that contain them.

These four types of information are directly relevant to most non-speech NLP applications, and each will be discussed in some detail below. All of them find a place in FrameNet descriptions. A complete lexicon, of course, will also have information about pronunciation, regional variations in vocabulary, etymology, register and attitudinal values and the like. A lexicon based on corpus evidence might also be expected to carry metadata about author/speaker, region, occasion, and the like, but these will not be considered here.

2.1 Definitions

Definitions or other meaning representations are of course a standard feature of familiar print dictionaries, and of the machine-readable dictionaries derived from them. In general the language of dictionary definitions cannot be counted on to provide directly usable information for NLP in a systematic way, mainly because dictionary style manuals have not been designed with such capabilities in mind. In fortunate cases, however, regularities in definition formats make it possible to derive other sorts of information about words (Wilks et al. 1996): words with the same definition are assumed to be synonyms, and noun taxonomies and related lexical networks are constructed based on certain phrases in definitions. For example, anything defined as “a type of fish” can be taken as labeling a subtype of the category fish; word lists followed by “etc.” can generally be seen as indicating coördinate concepts on some level; and definitions containing cross-references can suggest other word-to-word relations.

The definitions included in FrameNet are either copied from the
Concise Oxford Dictionary (COD) or, where no appropriate COD sense is available, are composed by FrameNet lexicographers. While the main function of current FrameNet definitions is to serve as a quick guide to the human user on the intended meaning of the LU, efforts are under way to devise more structured definitional formats for including the kinds of NLP-relevant information not discoverable in the FrameNet annotations themselves, such as morphological derivatives, domain names (Zoology, Psychoanalysis, etc.), semantic incorporation, etc.

2.2 Paradigmatic relations between LU types

Though dictionary definitions also specify paradigmatic relations of various kinds between LUs, such information is more systematically organized in thesauruses, such as the online WordNet, which includes relations such as those of “sameness” (synonymy), “oppositeness” (the varieties of antonymy), the “is-a-kind-of” relation and its inverse (hyponymy, hyperonymy), the “is-a-part-of” relation and its inverse (meronymy, holonymy), and a few others. It is also important to register morphological relatedness associating LUs with other LUs by virtue of shared stems. Tables of stem-relatedness classified by derivational process, for example, will show the adjective-noun relation of brief to brevity, verb-noun relations such as that of expand to expansion, or adjective-to-verb relations such as clean to cleanse, and so on.


7 For most cases it would make sense to state that stem-relatedness involves “words” (lexemes) rather than LUs, but there are a large number of cases in which ambiguous words support frame-sharing derivatives only in one sense. Thus, while the adjective curious has two senses—one attributing an attitude to an experiencer (I’ve become curious about the lives of insects) and one assigning some property to an object or phenomenon (that person’s behavior is quite curious)—the relevant sense of derived noun curiosity, which is itself ambiguous—but in different ways—belongs only to the former sense. (Compare my curiosity about insects with *the curiosity of his behavior). Similarly, two of the senses of the verb observe (‘follow the rules or practices of’ versus ‘be
Because this sort of information can be obtained by some combination of other available lexical resources, FrameNet has not felt it necessary to duplicate such efforts. In many cases morphologically related words are included: *possess* and *possession* are in the same frame, as are *replace* and *replacement*, or *eager* and *eagerness*. But FrameNet does not identify such relationships explicitly.

### 2.3 Syntagmatic relations between heads and dependents

The combinatorial potential of the lexical head of a grammatical construction is expressible in both semantic and syntactic terms; both of these can be referred to as *valence*. The **semantic valence** of a governing word is an account of the semantic roles (agent, patient, etc.) of its dependents as determined by its meaning. The **syntactic valence** is an account of the grammatical categories of the phrasal constituents that depend on a governor. FrameNet provides a complete valence for each LU which combines the semantic and syntactic aspects thus showing the syntactic patterns by which the semantic roles of the head’s dependents can be expressed.

FrameNet provides valence descriptions of each frame-bearing word by listing and exemplifying all of the major combinations of semantically and syntactically defined dependents that are found in the corpus. This means that for any word that is not richly attested in the corpus there will be “accidental gaps”—possibilities that native speakers would accept as part of the language, but which for some reason do not appear in the data. This is in contrast with the kind of lexicon built for a generative grammar, where linking rules are created to capture generalizations about the combinatorial possibilities, on the basis of semantic patterns wherever such relations are regular. This is not to say that FrameNet data could not

---

8 The semantic valence is analogous to the notions of *case frame* or *thematic grid* in other frameworks (Fillmore 1968).

9 The syntactic valence is analogous to the notion of subcategorization frame (Chomsky 1965) except that we always include the subject as one of the elements, unlike COMLEX (see http://nlp.cs.nyu.edu/comlex/).
constitute the basis for inducing such rules.

In many cases the valence descriptions of a governor need to be amplified by lexically specific information; for example, sometimes a verb or other governor shows a strong preference for a particular word or a narrow class of words within their dependents. Such LU-to-LU relations (co-occurrence relations) are referred to as collocations, ranging from such high-strength lexical associations as *cast aspersions* or *blithering idiot* through medium-strength associations like *brandish a weapon* or *precious cargo* to low-strength co-occurrences that are trivially associated with the general meanings of each partner in the relationship, for example that between verb and object in the VP *eat a sandwich* or *read the newspaper*. At this more general end we find lexically dependent cases of semantic prosody, such as the fact that the objects of the English verb *cause* more often than not designate negatively evaluated objects—*disaster*, *infection*, *failure* or the like (See Stubbs 1995). The higher-strength relations need to be included in an inventory of multi-word expressions and in the lexicon proper; in an NLP-relevant lexicon, generalizations about the lower-strength relations also need to be captured.

In addition, certain kinds of verb-to-noun syntagmatic relations are captured in FrameNet with the concept of support verb: this would capture the relationship in *cast aspersions*, allowing us to say that the frame is evoked by the noun itself, and the verb allows the concept evoked by the noun to be rendered as a finite clause. (We discuss support verbs in more detail in §4.3.) In the case of the relation between *brandish* and its object nouns, this should be derivable from FrameNet data by noting the class of nouns occurring in object positions with the verb.  

Identifying the class of nouns occurring in such a context need not be thought of as a constraint on “lexical selection”, that is, as a requirement to fill the slot with words from some list; instead we can think of the frame-bearing word (here the verb *brandish*) as inviting the reader/hearer to construe an object noun as designating something that can be used as a weapon. Thus, expressions about brandishing an umbrella or a fork need not require the belief that these nouns have linguistic properties identifying them as the names of potential weapons. It is of course still likely that the majority of nouns to occupy this position will in fact be names of weapons.
2.4 Clues for Semantic Integration

For a lexicon to have the potential to serve NLP applications, the information it provides about individual words must be structured in a way that contributes to the interpretation of phrases and sentences containing the word. In short, information is needed that allows the meanings of individual LUs to be combined to form the larger semantic structures that constitute the meanings of the phrases that contain them. In the simplest cases, a verb evokes a frame, particular semantic roles are associated with components of the frame, phrases in particular syntactic positions semantically identify the fillers of those roles, and inferences associated with the frame become linked to the referents of the syntactic constituents. Thus in \{Officer Clancy arrested my daughter yesterday for burning a flag\}, an instance of the frame associated with the verb *arrest* becomes populated with details provided by neighboring phrases inside the sentence, and this completed scenario affords inferences about past events in the lives of the participants, as well as predictions about future events in their lives, which can be integrated into a single narrative. From the sentence we learn the identity of the Arresting Officer, the Suspect, the Offense, and the Time of the event; from the location of the Arresting frame within the larger frame of judicial procedures, we infer a previous event of the commission of a crime, and we anticipate a series of future events, beginning with an Arraignment.

The organization of frames in FrameNet, and the system of frame-to-frame relations, whereby some frames are seen as components or subparts of other frames, are designed to enable such semantic integration. The central work of the FrameNet project does not concern itself with the principles of compositional semantics *per se*, but such principles are being developed in associated projects dealing with the use of FrameNet data for information extraction.\(^\text{11}\)

---

\(^{11}\) See Mohit and Narayanan 2003.
3. What FrameNet does

Though FrameNet is a lexicographic project, it differs from traditional lexicography in being explicitly based on a particular theory of semantics. **Frame Semantics** is an approach to the study of lexical meaning based on work by Fillmore and his collaborators over the past thirty-odd years (1975, 1977a, 1977b, 1982, 1985, among others). The central idea of Frame Semantics is that word meanings must be described relative to semantic frames. Frames are schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community.

The tasks we set for ourselves in the FrameNet project are the following:

1. to describe the conceptual structures, or frames, that given lexical units belong to;
2. to extract sentences from the corpus that contain the word and to select sentences that exemplify the intended sense of each of the LUs that we analyze;
3. to annotate selected sentences by assigning frame-relevant labels to the phrases found in the sentences containing the LU; and
4. to prepare reports that summarize the resulting annotations, showing succinctly the combinatorial possibilities of each LU; these are called **valence descriptions**.

3.1 The semantic basis

An understanding of our approach to semantic frames is perhaps best achieved by working through a simple example. Here we shall look at a group of words connected with a frame we call the Revenge frame. Lexical units that evoke the idea of Revenge include *avenge, avenger, get back (at), get_even (with), retaliate, retribution, revenge.n, revenge.v*, and *vengeance*. Revenge has to do with the infliction of punishment in return for a wrong **

---

12 The criteria we follow in selecting sentences are discussed in section 3.3.
suffered. An AVENGER inflicts a PUNISHMENT on an OFFENDER as a consequence of an earlier action perpetrated by the offender, the INJURY. The AVENGER may or may not be the same individual as the INJURED PARTY, the person who suffered the INJURY. The judgment that the OFFENDER had inflicted an INJURY is made independent of the law; this requirement distinguishes the concept of revenge from legally sanctioned punishment. The events and participants involved in instances of the revenge scenario, e.g. the AVENGER and the PUNISHMENT, are called the frame elements (FEs).  

\[ 1 \]

Something should be said about the naming of FEs. In defining a frame, we are satisfied when we find that the posited collection of frame elements is adequate for tagging those dependents of the frame-bearing word which are clearly related to our understanding of the frame. In choosing FE names we are more concerned with choosing names whose meanings are clear to the annotator than names selected from some pre-established list. In particular, the FE names which make up the semantic part of the valence descriptions are not limited to any fixed set of semantic roles like agent, patient, theme, experiencer, instrument, source, goal, path, location, and so on. (See Fillmore 1977, 65 for a “standard” list within the case grammar tradition; for a differently structured list, see Somers 1987, 206.) We find that the “standard” lists do not cover all of the semantic roles needed for the description of our frames without distorting the interpretation of the standard roles. For example, in the Revenge frame, the INJURED PARTY could not easily be described in terms of the standard role set; though the INJURED PARTY is a patient in the earlier offensive action, it is not a direct participant in the act of taking revenge. Similarly, in a sentence like \{You risked death\}, it is hard to imagine assigning any of the standard roles to either you or death. We create frame-specific role names without trying to explain their commonalities with roles that might have received the same names in other frames.

Nonetheless, we often do choose the same names for FEs in different frames. This is partly because we simply do not have enough distinct and memorable English words to cover all the types we need; we are in principle satisfied to recycle names without claiming identity. In many cases, however, we want to assume that the situations in which identical FE names are used are cases of frames that are bound to particular roles from some more abstract frame which is simultaneously active in the given frame. The relevant mechanisms are briefly discussed in section 3.5.
Consider the following marked-up example sentences for lexical units in the Revenge frame:

1. [Ethel \textit{avenger}] eventually \textit{got even} [with Mildred \textit{offender}] [for the insult to Ethel’s family \textit{injury}]

2. Why hadn’t [he \textit{avenger}] sought to \textit{avenge} [his child \textit{injured party}]?

3. Yesterday [the Cowboys \textit{avenger}] \textit{avenged} [their only defeat of the season \textit{injury}] [by beating Philadelphia Eagles 20-10 \textit{punishment}].

4. The Old Bailey was told [he \textit{avenger}] was desperately in love and wanted to \textit{get back} [at the woman \textit{offender}] [“for ending their relationship” \textit{injury}]

5. [The USA \textit{avenger}] \textit{retaliated} [against the harassment of its diplomats \textit{injury}] [by expelling 36 staff from the Iraqi embassy in Washington on Aug. 27 \textit{punishment}]

From these and other examples it seems clear that we have the FE distinctions needed for covering the central participants. We can now consider how the various frame elements are linguistically realized, i.e., how FEs are linked to syntactic constituents. Different LUs sometimes offer different possibilities.

In the case of the verbs in this frame, we find that in active-voice sentences, the \textit{avenger} is the subject. The \textit{offender} typically occurs in a prepositional phrase. The identity of the preposition is lexically specific; that is, it depends on the lexical unit: with \textit{get even} it is \textit{with} as shown in

\footnotesize

\begin{itemize}
\item \textit{got even} with \textit{injury}
\item \textit{avenge} his child \textit{party}
\item \textit{avenged} their only defeat of the \textit{season}
\item \textit{got back} at the woman \textit{offender}
\item \textit{retaliated} against the harassment of its diplomats \textit{injury}
\end{itemize}

\normalsize

\textsuperscript{14} It should be noticed that while we think of the FEs as identifying the persons, artifacts, events, utterances, propositions, etc., that participate in frames, the actual phrases that receive FE labeling include the function words that mark their syntactic position in the sentence - prepositions, complementizers, and the like. The annotation is to show the phrases that occur as syntactic dependents of a governor, and to label them in terms of the frame elements that they introduce.

\textsuperscript{15} As the \textit{injury} frame element demonstrates, what we initially called “participants” in a frame need not necessarily be objects or people. They can be events or properties.
(1), but in the case of get back it is at, as shown in (4). The INJURY is mostly realized in prepositional phrases with for but in the case of the verbs revenge and avenge it can occur as a direct object. Expressions of the INJURY can either be understood from the point of view of the original act (my brother’s murder) or the effect on the injured party (my brothers’ death). The PUNISHMENT is typically expressed as a by-phrase with a gerundial complement. Finally, the INJURED PARTY can sometimes be realized as a separate constituent, in particular as the direct object of avenge as in (2).

Some of the constituents that are in syntactic construction with a verbal head bear meanings that are more specifically connected to the verb’s frame than others. Accordingly we distinguish core FEs from non-core FEs. Though there is a considerable overlap, this distinction is not the same as the one syntacticians have traditionally made between arguments and adjuncts. The latter distinction is based mostly on assumptions about syntactic configurations and syntactic phenomena such as extraction. Our notion is primarily semantic, concerned with whether some notion is conceptually necessary to the understanding of the frame. The valence patterns that we consider most relevant to the description of the verb are based only on the core elements. The non-core elements include various kinds of circumstantial adjuncts that are more or less compatible with any kind of event or state of affairs. An example of a non-core element is the temporal adverb yesterday occurring in example (3) above. While it is clear that any act of Revenge has spatial and temporal properties, the temporal modifier yesterday has no specific relevance to the meaning of the verb avenge. Since a secondary goal of our work is to provide at least partial semantic parses of the sentences we examine, the annotators often tag such constituents with appropriate FE labels (TIME, PLACE, etc.), but the basic valence descriptions for the relevant verbs include only the annotations of the core FEs.

---

16 The latter does not need to be seen as a separate FE since it can be interpreted as an instance of coercion in the sense of Pustejovsky (1995). One avenges a hurtful act, and in this context the simple event death is coerced as meaning an intentional action which produced the death.
3.2 The Unit of Analysis

It should be clear by now that FrameNet’s unit of analysis is not the individual word in all its senses but the frame. Traditional lexicography considers each word in turn, trying to cover all of its senses. FrameNet proceeds from frame to frame and tries to include in each frame all the lexical units that evoke it. The resulting pattern of growth of the FrameNet lexicon could be called “organic”: new frames are typically suggested by considering other meanings of the polysemous words in the frames we’ve already treated. This means that the presence of a word in one or more frames in our database does not amount to a claim that we have covered all of the senses of the word. In traditional lexicography, a dictionary in progress is likely to have finished all the words up through, say, the letter M, and the ones from N on have not been done. In our case, stopping in the middle will necessarily mean that many of the senses of individual words will not have been included since the relevant frames have not yet been defined; in many cases, to the surprise of visiting lexicographers, a missing sense might be the main sense of the word.

Another result—though not a necessary one—of our focus on frames rather than lexical units is the absence of indications of paradigmatic relations between LUs. For instance, morphological relatedness is not explicitly indicated, and neither are such semantic relations as synonymy and antonymy. In some cases of antonymy, we treat them as parts of different frames; this is so for LUs like place and remove, which represent the Placing frame (moving things to a place) and the Removing frame (moving things away from a place) respectively. On the other hand, the opposites reward and punishment are included in a single frame; and scalar opposites are in principle included in the same frame, the frame that covers the scale itself (old, young; near, far; tall, short, etc.). Human readers will recognize LU-to-LU relations in FrameNet definitions, but such relations are not made explicit in the database.

3.3 The Annotation Work

Without going into too much detail, we would like to point out several
important characteristics about the way FrameNet annotation is done, which have a bearing on what can be concluded from the resultant data.

Our annotation is always done relative to one particular lexical unit, the target, which is most often a single word but can also be a multi-word expression such as a phrasal verb (e.g., give in) or an idiom (e.g., take into account). Most sentences require evoking more than one frame and a full understanding of a sentence would require annotating it from the point of view of each frame-evoking unit. Through a “layering” facility in the annotation software we have the ability to annotate all the words in a sentence, but the database consists almost entirely of annotation of sentences with respect to single lexical targets.

We annotate whole constituents rather than just the head words of the target’s syntactic dependents. For instance, when annotating sentences with the lexical unit get back in the Revenge frame, we include the preposition at in the constituent expressing the VICTIM frame element.

Sentences are pre-selected because they contain a predetermined target word. They are not just a random sample of sentences in the corpus.

We annotate individual sentences rather than running text. This means that (1) we are not in a position to deal with anaphora resolution and (2) we do not annotate constituents whose denotata we only infer to be frame elements, but which are not in syntactic construction with the target. For instance, if we annotate a sentence like {Sue had stolen my watch and so I took revenge.} with revenge as a target, we would annotate neither Sue nor any other element in the first of the two conjoined clauses. In other words, annotation for lexicographic purposes differs from annotation for text understanding purposes.

The sentences we annotate are not intended to be either quantitatively or qualitatively representative of the average sentence with a particular valence pattern. Thus, if our annotation includes two sentences where avenge occurs with INJURED PARTY as its direct object and five sentences where it occurs with INJURY as a direct object, this should not be taken to mean that the latter pattern is 2.5 times more common in the corpus than

---

17 We do, however, regard subjects of controlling verbs as available for annotation.
the former. Likewise, if our annotated sentences on average contain more full lexical noun phrases than sentences of spoken conversation or even of writing, then this is so because annotators prefer sentences with such NPs because they are more informative with respect to selectional restrictions than sentences containing personal pronouns or zero forms. Our criteria for selecting sentences for annotation can be summarized as follows:

- a. they must exemplify the LU in the target frame
- b. all else being equal, structurally simpler sentences are to be preferred
- c. the sentences we choose should contain typical collocations
- d. the set of annotated sentences for a given LU should cover all valence patterns that the LU exhibits in the target frame but avoid repetition of identical structures
- e. sentences should exemplify World English rather than any particular variety

3.4 Viewing the Results of Annotation

Our main tools for studying our annotated sentences and the valence patterns they manifest are two automatically generated, web-based reports, which are available via the public web page. These are called the Annotation by LexUnit Report and Lexical Entry Report. The Annotation by LexUnit Report shown in Figure 1 for the LU avenge in the Revenge frame displays all the annotated sentences for a given Lexical Unit. At the top of the report is a list of all the frame elements found within the frame; this is called the Frame Element Table. Though it may not be clear from the reproduction of the report, the annotated sentences are displayed with the words instantiating Frame Elements highlighted in the same color in which the Frame Elements appear in the Frame Element Table.
Figure 1: Annotation Report for avenge

The Lexical Entry Report summarizes the syntactic realization of the Frame Elements and the valence patterns of the Lexical Unit in two tables. The realization table for the LU avenge in Figure 2 shows all of the core Frame Elements, the number of annotated examples and their syntactic realization.
Figure 2: Lexical Entry Report for avenge, Frame Element Table

The second table, shown in Figure 3 for the LU *avenge*, illustrates the valence possibilities of a given LU.
Figure 3: Lexical Entry report for avenge, Valence Patterns

In both tables of the Lexical Entry Report, each entry is hyperlinked to the sentences which manifest that particular syntactic realization or valence possibility.

In addition to these web-based reports there is also a powerful web-based database querying tool, FrameSQL, developed by Prof. Hiroaki Sato of Senshu University, which is available by link through the FrameNet web page. This tool allows the user to query the database with multiple search parameters such as frame name, frame element name, grammatical function, etc. One could, for instance, find all the sentences in any frame where a Frame Element called REASON is expressed by a constituent whose phrase type is Prepositional Phrase.
3.5 Frame-to-Frame Relations

One of the ways in which we claim a “net”-like system for FrameNet is through a system of frame-to-frame relations, the simplest of which is Inheritance. In the case of the Inheritance relation between a more general and a more specific frame, all of the properties assigned to the parent frame correspond to something in the child frame. Most of the instances of pure inheritance are the lexically specific frames, and these are of a more detailed granularity than we generally deal with; bequeathing is a subtype of giving; gossiping is a subtype of speaking; swaggering is a subtype of walking. Each of the “child” concepts in these pairs is an elaboration of the meaning of the “parent” concept.

A more general Using relation is posited to obtain between two frames when one frame makes reference to the structure of a more abstract, schematic frame, though it may not provide dependents for each of the FEs of the more inclusive frame. Typically, this occurs in cases where the specific frame provides a perspective on the schematic frame. For these, we say that the specific frame has a Using relationship with the schematic frame, and bindings between the Frame Elements and Subframes may be specified. For example, the Commercial_transaction frame specifies a complex schema involving an exchange of multiple Themes (the MONEY and GOODS) between a BUYER and a SELLER. The more specific Selling frame has a Using relationship with the Commercial transaction frame in which the SELLER and the GOODS have core status. The perspectivalizing is expressed by seeing the Selling frame as inheriting the Giving frame and by seeing the Buying frame as inheriting the Getting frame, where the FE THEME is bound to the GOODS in both cases and the FE AGENT is bound to the SELLER in one case and the BUYER in the other.

A relation we refer to as the Subframe relation is needed for frames that are temporally complex: they represent sequences of states of affairs and transitions between them, each of which can itself be separately described as a frame. An example we are currently working with is the Criminal Process frame, which has at least Arrest, Arraignment, Trial, and Sentencing among subframes. The separate frames are related to the complex frames via the Subframe relation. In such cases, frame elements
of the complex frame may be identified with (mapped to) the frame elements of the subparts, allowing us, for instance, to capture the fact that the SUSPECT of the Arrest frame is the DEFENDANT of the Trial frame. However, it is not necessary that all frame elements of one need have any relation to the other. In this respect, Subframe relations contrast with inheritance relations, where this requirement does obtain. With Subframe relations, we also indicate the full or partial ordering and other temporal relationships of the subframes. Notice that subframes may themselves be complex frames with their own subframe structures, as is clearly true in the case of the separate phases of the Criminal Process frame.

3.6 The Treatment of Slot-Filler Nouns

Like most other lexical research,\(^\text{18}\) FrameNet focuses on predicates, that is, on intrinsically relational words that do not refer to abstract or concrete entities but to states of affairs and events. However, we have also done some annotation relative to slot fillers. That is, we have annotated referring expressions where they serve as frame elements of frames determined by their governors.

In the case of nouns that serve as the lexical heads of dependent phrases, we record information about their syntactic realization with various governors. For example, a noun like knife can occur in an instrumental phrase with verbs like cut, stab, slice, etc., in direct object position with sharpen, and in subject position with sharp, dull, etc. These nouns do not necessarily evoke frames by themselves,\(^\text{19}\) but mostly occur as slot fillers in frames evoked by verbs, adjectives, or other nouns. Nevertheless, we tag a governing verb or preposition as a governor for these slot filler nouns.

The reason we annotate the slot-fillers is that it is important to know about which frames they typically appear in as slot fillers, and conversely

\(^{18}\) Work based on Pustejovsky’s Generative Lexicon theory (1991, 1995) is an important exception.

\(^{19}\) In the case of artifact nouns, of course, we could say that they evoke the frames associated with the purpose for which they were manufactured, their telic quale in the sense of Pustejovsky.
what a typical slot filler for a particular frame element is. Consider the Cutting frame. We might wonder what kinds of things are usually sliced as opposed to simply cut. Conversely, thinking about tomatoes, we might wonder what kinds of events they are participants in. We may intuit that tomatoes are often mentioned as objects of slice but are there other kinds of cutting, for instance, that apply to tomatoes?

We could try to derive this kind of information automatically from our annotation relative to governors like slice, cut, chop, etc. However, for that to be feasible we would have to annotate many more sentences for each of the governors than we currently do. For lexicographic purposes alone, it is not necessary to document that in addition to onions and carrots, tomatoes can be sliced, so we have not chosen to represent such information. Thus, the easiest way to address questions like the ones above is to look for sentences with nouns such as tomato, onion etc. and treat them as targets. We can then record what syntactic governors (Gov) take phrases containing the artifact nouns as arguments and what kind of syntactic constituents contain the target nouns. (For lack of a better term, we call the constituent containing the target slot filler noun ‘X’.)

6. He [grabbed\textsubscript{gov}] [two tomatoes and the biggest apples\textsubscript{x}]
7. Skin and [quarter\textsubscript{gov}] [the tomatoes\textsubscript{x}]

4. Kernel Dependency Graphs

It is often difficult to recognize the central core of the combinatorial properties of a lexical item when it is embedded in a complex sentence with its parts interrupted or expanded by surrounding material. A type of notation has been defined which sees through all of this junk leaving only the essence. A \textit{kernel dependency graph} (KDG) is a partial analysis of a predication\textsuperscript{20} in some given sentence for which the following entities are identified: (1) the frame associated with the governor in a set of governor-dependent relations, (2) the FEs of that frame that are expressed

\textsuperscript{20} Predication in this sense is any relation and its relata.
or understood in the sentence, (3) the lexical heads of the constituents that express the FEs, and (4) the manner in which the FEs are syntactically presented. The term “kernel” is intended to indicate that these are provided for each predicator in a sentence: a complex sentence would generate multiple KDGs with cross-references linking them to each other.  

In principle, a compositional semantics for a dependency parsing of a sentence would assemble the KDGs through connections among their shared components. At present the elaboration of such compositional principles is not a part of FrameNet’s mission.

It is now possible to derive KDGs from existing annotations in the FrameNet database. With improvements in frame- and FE-identifying machine learning techniques it should be possible some day to extract KDGs from raw text. There are two reasons for wanting to do this. The first is purely linguistic, having to do with an interest in collocations for their own sake; a body of KDGs can be seen as a presentation of the lexico-syntactic collocations found in a corpus, that is, the (syntagmatic) relations between LUs that are in grammatical construction with each other in text. The second reason is to be able to find, by extracting the topmost salient KDG of a sentence, simple clues to the semantic content of the passage from which they were derived. This is because the elements of a KDG, in the typical case, are indications of the frame (associated with the head of the KDG) and essential information about the content of the associated FEs.

21 This is in analogy with the concept of kernel sentence in early transformational grammar. A surface sentence like \{I think they wanted to leave.\} is decomposable into three kernel sentences, one with think, one with want, and one with leave. See Chomsky 1957, Harris 1957.

22 The identification of the topmost “salient” KDG in a sentence may be sensitive to embeddings that exist mainly for the communication of epistemic or evidential qualifications: for example, in some newspaper crime reports, almost every sentence is qualified by “the police said” or “authorities have revealed”—but the actual news about muggings and thefts is found inside those statements and revelations.
4.1 Deriving Simple KDGs

In the simplest cases, the FrameNet annotations permit the derivation of KDGs by recognizing the lexical identity of the target predicate, the frame within which the target word is being annotated, the represented FEs, and the lexical heads of the constituents that instantiate them.23

The derivation of KDGs can be thought of in terms of converting phrase-structure representations to dependency representations by processes of head lexicalization (representing constituents as their heads—and doing this recursively), gap filling (getting the lexical material that is syntactically absent from the heads of the constituents that syntactically “control” their interpretation), and coordination deconstruction (by which, for example, \{We were slicing tomatoes and carrots.\} will yield two KDGs, not one, one having to do with slicing tomatoes, and one with slicing carrots). In addition, anaphora resolution is required when the purpose of collecting KDGs is to associate contentful lexical material with FEs of particular frames, as when studying selectional restrictions.

We begin by considering some KDGs that can be derived by using only the process of head lexicalization. A simple example of a KDG based on an intransitive verb is shown in Figure 4. The frame evoked in this sentence by the target die is the Death frame. In the example sentence given in the caption below the KDG, the target of the analysis is underlined.

![Figure 4: Death: “Truman died”](image)

23 The FrameSQL browser for FrameNet, created by Hiroaki Sato and accessible through the public FrameNet website, has a facility that in principle provides KDG information as an alternative representation of valence patterns. Fine-tuning the performance of this facility has not until the present been an emphasis in FrameNet, and unfortunately the British National Corpus we are using does not have dependency parsing.
A KDG based on a transitive verbal predicator is shown in Figure 5, where we are considering the sentence \( \text{The puppy drank the milk.} \). As with the intransitive example above, we omit information about grammatical functions, since in general that is predictable either by linking properties associated with the FEs or by the nature of the phrase type of the FE constituent.

![Diagram of Ingestion: "The puppy drank the milk"](image)

Figure 5: Ingestion: “The puppy drank the milk”

The name of the frame is Ingestion. The lexical head of the VP (and hence of the sentence) is \( \text{drank} \), lemmatized as \( \text{drink} \); the heads of its two complements (subject and object in this case) are \( \text{puppy} \) and \( \text{milk} \); and the FEs are \( \text{INGESTOR} \) and \( \text{INGESTIBLES} \).

For a verb that has prepositional or verbal complements we can indicate the grammatical ways in which they are realized with certain abbreviations. In the following KDG, the label “P-TO” represents the preposition \( \text{to} \) (rather than the infinitive marker \( \text{to} \), which would be labeled “V TO”).

![Diagram of Showing: "The professor demonstrated the proof to the class"](image)

Figure 6: Showing: “The professor demonstrated the proof to the class”

There are several different types of frame-evoking nouns. We shall here consider deriving a KDG for an event-denoting noun. Figure 7 shows the analysis of the noun phrase \( \text{The company’s purchase of its own stock} \).
The target noun *purchase* belongs to the Commerce_buy frame. The BUYER frame element is realized as a possessive and the GOODS frame element is realized as an of-prepositional phrase.

![Figure 7: Commerce_buy: “The company’s purchase of its own stock came as a surprise”](image)

Nominal predicators have different syntactic relationships to their frame elements than verbal ones. They do not, for instance, have a position comparable to subject and they cannot take nominal objects directly (*her purchase the book*). Additionally, a given FE can often be realized in multiple ways. Compare the *purchase by the company of its own stock* as well as the *stock purchase by the company* to the example in Figure 7, the *company’s purchase of its own stock*. Because of this flexibility with respect to argument realization it is necessary to keep track of the grammatical markings of all the frame elements.

While predicative adjectives, unlike event nouns, can occur with subjects across copular verbs, they also are unable to take noun objects directly. Therefore, we need to record the prepositions that head the phrases containing the adjectives’ frame elements. This is shown in Figure 8, where we analyze *similar*, which belongs to the Similarity frame.

![Figure 8: Similarity: “Carving chisels are similar to carpentry chisels”](image)
4.2 KDGs involving Gap Filling

Having shown that we can derive KDGs for verbs, nouns, and adjectives by the process of head lexicalization, we now turn to the process of gap filling. Gap filling identifies the overt antecedents of zero forms in controlled structures. The graph in Figure 9 shows the same relations among the elements of the Ingestion frame, with the verb drink, as in Figure 5. Even though this time the NP the puppy serves directly as the object of expect, it is still construed as the subject of drink. Here the pseudo-grammatical-function tag “EXT” indicates that puppy is only in indirect construction with drink, mediated through the verb suspect: it is external to the constituent headed by drink.\(^4\)

![Figure 9: Ingestion: “Everybody suspected the puppy of having drunk the milk”](image)

Note that a constituent that is external to the syntactic domain governed by the frame-evoking predicator of one frame, is a dependent in the KDG of some other governor. Consequently, a second KDG can be derived in all such cases. The second KDG for the sentence in Figure 9 is centered on the verb suspect, as shown in Figure 10. The ability to identify the puppy as the INGESTOR of drink in this case comes from the grammar of the phrase built around the verb suspect: by virtue of instancing the SUSPECT of the Suspicion frame with the transitive verb suspect it stands as well for the INGESTOR of the drinking event expressed in its complement.\(^5\)

\(^4\) Note that the use of the tag EXT in FrameNet annotation is broader than our use in KDG diagrams, being applied to simple subjects of verbs as well.

\(^5\) The verb suspect exemplifies what in the linguistic literature has been called
Here we see that the content FE that goes with *suspect* is syntactically introduced by the preposition *of* and a VP in the gerundial form.

We also see in the sentence shown in Figure 10 that the verb which syntactically heads the content expression is not *have* but *drink*. This means that we need to recognize certain structures which are “transparent” to the process of identifying the elements of the KDG. The indication of perfect aspect in this sentence (the *have* of *have drunk*) is not relevant to the kinds of collocations we are interested in.

Figure 9 gave us an example of a context in which the object of the Raising verb *suspect* was construed as the subject of its complement; Figure 11 shows a case in which the subject of the Equi verb *try* serves also as subject of its complement. The target is *kill*, in the Killing frame; again, the “EXT” indicates that the constituent realizing the frame element is not in direct construction with the verb *kill*.

Equi-NP deletion, which together with the phenomenon of Raising, falls under the rubric of Control. Theoretical disagreements on the treatment of such structures have no bearing on our decisions about the Ext categories.
And of course there is a second KDG that can be built for the sentence (Figure 12), centered on the predicator try belonging to the Attempt frame.

![Figure 12: Attempt: “The man tried to kill my father”](image)

So far the controllers have been verbs, but adjectives and nouns can also serve as controllers. Here we shall only illustrate a noun controller. The KDG in Figure 13 is centered on the verb sign in the Hiring frame. Because of the grammar of desire, the possessor of the governing noun desire, the team, is necessarily the actor in the associated act. Thus we can build a complete KDG for sign, acknowledging the external syntactic status of the AGENT FE.

![Figure 13: Hiring: “the team’s desire to sign the player”](image)

Of course we can also build a KDG for the noun desire in the Desiring frame.

![Figure 14: Desiring: “the team’s desire to sign the player”](image)
Note that the frame element realization seen in Figure 14 is but one possibility. The experiencer FE for desire has an alternative realization as a preposition phrase, as in the desire [on the part of the team] to sign the player. By-prepositional phrases are also sometimes found with desire, as in the desire [by the team] to sign the player.

In other cases a semantic governor is not a syntactic governor of its “subject” but is in a secondary predicate position. In this next sentence, the puppy is the object of the verb find, but the “subject” of the adjective dead.

![Figure 15: Becoming aware: “The investigators found the puppy dead”](image)

With respect to the word dead, however, the argument is the puppy.

![Figure 16: Dead or alive: “The investigators found the puppy dead”](image)

In addition to the control constructions we have considered so far there exist extraction constructions such as topicalizations, wh-interrogatives, and relative clauses, etc., in which a constituent tagged to fill a frame element role is not found inside the maximal projection of the target word.

8. Kim I have always liked very much.
9. Who did you call?
10. The car that John bought makes a funny noise.

However, in these cases we do not treat the relevant constituents as external since they still have grammatical functions and case assigned to them by the target predicator.

It should be noted in this context that the FrameNet treatment of frame-bearing expressions inside relative clauses gives an “incorrect” constituent structure where an FE is represented by the relativized NP. This too is in the service of providing lexical heads for collocational purposes. Thus sentence 10 would be annotated as follows:

The [car that \textit{GOODS}] [John \textit{BUYER}] \textbf{bought} makes a funny noise.

In this case both the noun \textit{car} and the relative pronoun that are included in the label, and the noun would be selected as a member of the KDG. The discrepancy with the syntax is more extreme in that same sentence without the word \textit{that}: \textit{The car John bought makes a funny noise}. Here there is no surface realization of the “real” direct object of \textit{buy}. The collocational relation between \textit{John} and \textit{car} (through \textit{buy}) is still recoverable, though there is a zero representation of the direct object of \textit{buy}.

It is not always the case that there are semantically meaningful ways in which the syntactic head of a phrase is the most important frame-evoker or that the syntactic heads of dependent phrases are the most important indicators of the meanings of those phrases. In the sections below, we shall treat several such discrepancies, including: (1) support verbs, where the syntactic head of a verb phrase has a minor semantic role and the major frame introducer is the noun it is associated with; (2) null instantiation, where clearly present conceptual elements are not given expression in the sentence; (3) transparent nouns, where the syntactic head of an NP indicates a quantity or type or evaluation and its complement bears the

\footnote{For those who would not think of \textit{that} as a relative pronoun in the first place we can change it to \textit{which} in the former example.}
semantic core of the NP; and (4) frame element fusion, where the information associated with two frame elements is expressed in a single constituent.

4.3 Support Verbs

A special kind of gap-filling occurs when nouns occur with **support verbs** (SVs). As noted earlier, typically the major frame in a clause is evoked by a verb. However, there are cases where a noun is the major frame-evoker even though it is governed by a verb, as in *{The president will make an important announcement later today.}*}. In these cases, it seems the verb is present mainly to allow the formation of a tensed clause and to introduce one (or sometimes two) of the elements licensed by the frame structure of the noun. The combination of support verb and event noun has essentially the same meaning, though with different syntax, as the verb underlying the noun. Thus, *making an announcement* is an instance of *announcing something* rather than of *making something*. Likewise, when somebody *has an intention* to do something, that is because somebody *intends* something and not because somebody *has* something. The clearest examples of support verbs—so-called **light verbs** such as *do, give, have, make, and take*—are very frequent and combine with a large number of event nouns and make almost no semantic contribution to the scene that the noun evokes. In addition to the well-known light verbs, there are many other support verbs that collocate much more narrowly with event nouns: *attention is paid; prayers are said; lies are told; complaints are filed; and sins are committed.*

Because support verbs are specific to particular nouns—we might say that nouns select their governing support verbs—we want to record support verbs that occur with nouns for which we are building KDGs. Consider the KDG in Figure 17. In the analyzed sentence, the *team* is introduced as an external argument, by being the subject of *have*. As usual we record this by prefixing the head of the EXPERIENCER FE with “Ext”. Now the support verb *have* is included in the KDG as well given that the noun selects it. But with support verbs we reverse the direction of the arrow to show that syntactically the noun is the dependent of the verb.
Figure 17: Desiring: “The team has the desire to sign the player”

Note that while the light verbs discussed above represent the situation where a support verb is associated with multiple event nouns, the opposite situation also occurs: some nouns occur with several support verbs. In some cases, the support verbs distinguish separate senses of the noun: *have an argument* is concerned with quarreling, while *make an argument* is concerned with reasoning. In other cases, the different support verbs reflect subtler distinctions. For instance, the different SVs may pick out different participants along the lines of Igor Mel'cuk’s *lexical functions*. The prototype example is the difference between *perform an operation* (Oper1 in Mel'cuk’s classification) and *undergo an operation* (Mel'cuk’s Oper2), where the subject of *perform* is the active participant in a situation, and the subject of *undergo*, the passive participant. In other sets of SVs, the members differ in terms of register: *lodge a complaint* is more formal than *make a complaint*, *exact revenge* is more formal than *take revenge*; etc.

The SV *give* offers a special case: it can provide two arguments for the frame introduced by the semantic head. Consider *{The audience gave the speaker a standing ovation.}* , the KDG for which is diagrammed in Figure 18.
“The audience gave the speaker a standing ovation”

Here the frame-bearing word is a noun, and its FEs are both external to the NP that contains it, in this case as the subject and the (first) object of give.

4.4 Null Instantiated FEs

There are situations in which a core frame element is neither expressed as a dependent of the predicator nor can it be found through gap filling; we call this null instantiation of the frame element. We record the absence of such elements because it provides lexicographically relevant information regarding omissibility conditions. It is important to note that not all cases of frame element omission are alike. We make a three-way distinction between constructional, definite, and indefinite null instantiation (abbreviated CNI, DNI, and INI). The first category covers, among other things, the omitted subject of an imperative and the omitted by-phrase agent in a passive. Constructional null instantiation is not lexically specific: any verb that can occur in the construction allows the null instantiation in question. The latter two categories, by contrast, are lexically specific. Cases of definite (also called anaphoric) instantiation are those in which the missing element must be something that is already understood in the linguistic or discourse context, as in the following example where the place of departure must be accessible from the context.

11. [John _AGENT_] left. [DNI _SOURCE_]

Cases of indefinite (sometimes also referred to as existential) null instantiation are the missing objects of verbs like eat, sew, bake, drink, etc. where the nature or semantic type of the missing element can be understood given interpretational conventions, but there is no need to retrieve or construct a specific discourse referent.

12. [Sue and Peter _INGESTOR_] had eaten already. [INI _FOOD_]

The need to recognize FEs that are conceptually present in the frame
evoked by a given target was discussed earlier. It is possible to show such structures in KDGs by representing the terminal node of that branch with a symbol that indicates the kind of null instantiation exhibited by the sentence. In our examples we shall use the Replacement frame, in which an AGENT causes something NEW to occupy the PLACE formerly held by something OLD; this frame is evoked by the verbs replace and substitute in their agentive sense. If the verb replace is used with only a direct object, the missing element is understood to be the NEW entity and it has the INI interpretation: *We replaced the blue one*—with something but I’m not telling you what it is. Where substitute is used with only a direct object, the missing element is the OLD thing, and it has the DNI interpretation: *We had to substitute sugar*—for the stuff we’ve just been talking about. Some sample representations:

![Diagram of the Replacement frame with the sentence: The committee replaced Harry with Susan.](attachment:image)

**Figure 19:** Replacement: “The committee replaced Harry with Susan”

![Diagram of the Replacement frame with the sentence: The committee replaced Harry.](attachment:image)

**Figure 20:** Replacement: “The committee replaced Harry”

27 The FE labels are essential to KDGs. The sentence in Figure 20 is ambiguous. For the meaning that the committee has taken over Harry’s duties, committee would be labeled NEW and there would only be two dependents.
The possibility for FE omission with frame-bearing nouns is much greater than for verbs or adjectives. The nouns in such expressions as \{The decision was foolish.\}, \{The promotion was surprising.\}, and \{There was a quarrel.\} refer to events of the type introduced by the noun, but no information about the participants in those events is provided. This is generally not possible with verbs.  

4.5 Transparent Nouns

Earlier (in Figure 10) we pointed out auxiliaries such as the have indicating perfect aspect as a case in which the syntactic head of a structure is not the semantic governor and should be treated as “transparent” for the purpose of building KDGs. A similar discrepancy between syntactic and semantic dependency structure also occurs with certain nouns. If we return to the verb drink and look for the head nouns in its direct objects, in

---

28 See Grimshaw 1990 and references therein for discussion of whether at least some event nouns do have obligatory arguments.
addition to champagne and whiskey and water, we also find some nouns that do not refer to liquids: pint, liter, cup, glass, bottle, and toast are among the frequent collocates. These tend to be interpreted as quantities. In each case the quantity is relevant to the meaning of the sentence (e.g., \{Drinking several pints of water should help relieve the symptoms.\}), but most of the important collocations with drink allow the quantifying word to be skipped. The words that have this “transparent” function are words that indicate quantifiers, containers, portions, aggregates, and types. Sato’s FrameSQL has a facility which, in building KDGs, skips past the nouns that are or that can be interpreted as transparent nouns. In the annotation of transparent nouns taken as LUs targeted themselves for annotation, care is given to choosing cases in which the annotators feel a collocational relation between the governor of the NP and the noun in the complement of the target noun rather than the noun itself.

Note that certain nouns which are often transparent need not always be transparent, as shown in the following examples. In the sentence \{The horse ate a number of poisoned apples.\}, the most relevant verb-object collocations connect eat and apple; on the other hand, the sentence \{The detective tried to estimate the number of apples the horse ate.\} shows a verb-object collocation connecting estimate and number. When the indefinite article occurs with aggregate, measure, or quantity nouns, the noun always seems to receive a transparent, quantificational interpretation: notice the unacceptability of \{*Each contestant had to guess a number of people in a room.\}. The definite article predominantly occurs with non-transparent uses of such nouns.

The importance of selecting the semantically relevant noun when it is not the same as the syntactic head of an NP, can be illustrated with a sentence like the following (modeled after examples in the Penn TreeBank). The KDG in Figure 23, with the syntactic heads, is much less informative than that in Figure 24, where we have seen “through” the transparent nouns to the semantic heads of the NPs.
Finally, note that certain transparent nouns are themselves involved in lexicographically interesting collocations. For instance, nouns denoting aggregates of animals such as *flock* or *herd* occur only with nouns denoting certain kinds of animals as their dependent. Similarly, the type nouns *strain* and *breed* have a much narrower range of complement than the type nouns *type* and *kind*.

### 4.6 Frame Element Fusion

In certain frames, pairs of frame elements are so closely connected with each other that the grammar permits the omission of one of them because it is inferable from the other, overtly realized FE. For instance, words in the Cure frame (related to medical treatment) typically allow the fused or separate expression of the PATIENT and the AFFLICTION:

13. Jesus cured [the leper PATIENT + AFFLICTION].
14. Jesus cured [the man PATIENT] [of lepros/ AFFLICTION].
Here, the leper denotes the PATIENT, but incorporates the AFFLICTION. Similarly, words having to do with hiring a person for a job or removing him allow the fused or separate expression of EMPLOYEE and POSITION.

15. I hired [her EMPLOYEE] [as my assistant POSITION].
16. I expect to hire [two new assistants EMPLOYEE+POSITION].

When building KDGs for sentences in which two frame elements are fused, we can simply associate the head of the constituent which expresses the fused FEs with the two relevant FEs. Compare Figure 25, the KDG for sentence 15, in which no fusion occurs, to Figure 26, the KDG for sentence 16, in which the POSITION and EMPLOYEE frame elements are fused.

![Figure 25: Hiring: “I hired her as my assistant”](image1)

![Figure 26: Hiring: “I expect to hire two new assistants”](image2)

5. Conclusion

We began this paper by describing what an ideal lexicon should provide and then outlined the information that is provided in the FrameNet database. Let us now consider the extent to which FrameNet satisfies the properties of our initial idealization.
Questions of the extent of coverage of the language’s lexicon have so far been left implicit in our account: every lexicon must have something to say about a significant portion of the lexical material that is likely to be encountered by its users (however that group is defined). From this point of view, FrameNet is still far from being complete. In terms of breadth of coverage, FrameNet is clearly more limited than even fairly small commercial dictionaries, much less WordNet or other large-scale digital lexica. As of this writing FrameNet has roughly 7,600 LUs in its lexicon, compared to the more than 190,000 “word-sense pairs” for WordNet. In terms of depth, however, FrameNet contains more detailed syntactico-semantic information than can be found in any of these lexica.

On matters of coverage, our working goal has not been to do whatever is necessary to increase the size of the lexicon, but to increase the coverage of familiar and frequent frames, which means that we emphasize particularly the “frame-bearing” LUs. Many function words are attached to or selected by the frame-bearing LUs, or participate in the creation of grammatical constructions headed by them and hence do not need lexical units of their own. In our work on text understanding, which so far is limited to newspaper accounts of crimes, we use a named entity recognizer (IdentiFinder, thanks to a license from BBN), which handles most proper nouns well, along with some time expressions, quantities, etc. The combination of these means that our coverage of the vocabulary of such crime stories is actually pretty good. If we consider other semantic domains, any serious application of the FrameNet data for a given topic or domain would require the lexicon to be enlarged with names and terminologies from that domain (e.g., names of diseases, machine parts, terrain features); the frame analysis for these LUs would need to be done by experts in those fields, although the initial term extraction can be partially automatic.

Because FrameNet focuses on syntagmatic relations connecting linguistic units in context, our lexicon may seem less complete than some other resources with regard to morphological or logic-semantic relation between LU types, or ontologically oriented structured definitions. On the other hand, the inductive, empirical approach we take to the task of building a lexicon results in rich, detailed information on argument realization. All our descriptions of the combinatory possibilities of predicators are based on
corpus evidence, which guards against overgeneralizations as well as against missing certain facts that do not readily come to mind when linguists work introspectively.

The FrameNet database also has a thesaurus-like quality, due to its grounding in frame semantics. However, at least in principle, our groupings are more structured than those found in a thesaurus and are related to each other as part of a hierarchy of frames. (In practice, much work remains to be done on defining the relations in this hierarchy.) In a thesaurus, the interrelation of concepts is either addressed by multiple listing or by pooling the related concepts at the most abstract level, neither of which is satisfactory. In FrameNet, however, the organization of frames and the system of relations between frames provides better clues for building up the meanings of phrases, sentences, and texts.

In the second part of the paper, we discussed the use of the FrameNet data in a proposed NLP task, i.e. deriving kernel dependency graphs (KDGs), both as a basis for studying collocations for their own sake, and as simple clues to the semantic content of a text passage from which they were derived. Taking a frame semantic approach to this task is helpful in two respects. First, because it is strongly semantically oriented it can comfortably deal with cases where syntactic and semantic headedness diverge: when support verbs and transparent nouns occur, the relevant collocations can still be recognized relative to the correct frame, evoked by the semantic head. Second, because our work is based on attested valence patterns, supplemented by a consistent annotation of core FEs, a KDG extraction tool based on FrameNet data should be better able to handle cases of non-canonical argument realization, such as null instantiation, FE fusion, etc. than one based on syntax alone, or even syntax combined with simple co-occurrence statistics.

As participants in the FrameNet project, we are keenly aware of the limitations of the lexicon with regard to issues such as breadth of coverage, completeness of frame-to-frame and LU-to-LU relations, etc. Nevertheless, we hope that its strengths will help to fill some critical gaps in existing lexical resources and thus improve the state of the art in many areas of NLP and information technology.
References

Mohit, Behrang, and Srini Narayanan. 2003. Semantic extraction with
wide-coverage lexical resources. Paper delivered at the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), Edmonton, Canada, May 2003.


Biographic Sketch and Research Outline of Authors

Charles J. Fillmore

Charles J. Fillmore is an Emeritus Professor in the Graduate School of UC Berkeley (lexicon, semantics, syntax, text comprehension, English, Japanese). He received his Ph.D. in Linguistics from the University of Michigan in 1960 and spent 5 years before that in Japan. He taught linguistics for 10 years at the Ohio State University before joining Berkeley’s Department of Linguistics in 1971. His research has concentrated mainly on questions of syntax and lexical semantics, and has emphasized the relationship between properties of linguistic form and matters of meaning and use. He collaborated with Paul Kay on a monograph on construction grammar, and has directed a research project to offer syntactically and lexically analyzed and translated Japanese texts on the Internet. He is now actively engaged in a project under NSF in computational lexicography, FrameNet, the fruit of which is a lexicon with word senses represented by semantic frames. For details of FrameNet, please see http://www.icsi.berkeley.edu/~framenet/. Prof. Fillmore was a Fellow at the Center for Advanced Study in the Behavioral Sciences (Stanford University) during 1970-71. He became a Fellow at the American Academy of Arts and Sciences in 1984 and the president of the Linguistic Society of America in 1990. He was awarded an Honorary Doctorate of Humane Letters by the University of Chicago in 2000.

List of Publications:
http://linguistics.berkeley.edu lingdept/Current/people/facpages/fillmore.html
http://www.icsi.berkeley.edu/~framenet/Papers.html

Josef Ruppenhofer

Josef Ruppenhofer is a lexicographic researcher at the FrameNet project and a doctoral candidate in the linguistics department at UC Berkeley. His research focuses on lexical semantics, argument realization, the discourse status of referents, and the interaction of the three.
With Laura A. Michaelis he has also worked on the treatment of argument alternations in Construction Grammar.

List of Publication:

**Collin F. Baker**

Collin Baker received a A.B. from Brown University, an M.A. from the University of South Carolina, and a Ph.D. from the University of California at Berkeley, all in linguistics. His dissertation research was on lexical semantics, specifically distinguishing and describing the senses of the English verb *see*.

He has taught and been an administrator in programs in English as a Second Language in Florida, Taiwan (Taipei Language Institute), the P.R.C. (Shanxi University), and Japan (Temple University, Japan). Since 1997, he has been working for the FrameNet Project (http://framenet.icsi.berkeley.edu/~framenet) at the International Computer Science Institute in Berkeley, California, serving as project manager since 2000.

List of Publications:


Introduction to
Computational Studies of Language Evolution

Chu-Ren Huang

How language came into being is one of the most fascinating scientific questions of our time. It touches upon both the ontological question of what is human and the epistemological question of how human beings express knowledge. In this scientific quest, two vastly different perspectives have been taken and both have led to fertile research ground: The perspective that language is the shared endowment of the human species has led to the study of language and evolution, and is discussed in the following paper. The perspective that the essence of language is what is innate in each individual human being’s brain has led to the study of language acquisition in the generative grammar paradigm. Both perspectives are paradigmatic and focus on the emergence of, and transitional stages within, language. The critical difference has to do with scale: whether the focus of interest is on the first years of an individual’s life or on the history of human beings.

The birth of computational linguistics was indeed closely related to generative linguistics. The generative mechanisms of syntactic structure, which accounted for the speed and completeness of language acquisition, were espoused by computational linguists because of their programmatic nature. There was genuine enthusiasm that language could finally be manipulated by rules, which, in turn, suggested that different languages could also be mediated by rules. Hence, early studies on computational linguistics and machine translation typically involved generation trees and transformational rules, including the pioneering work of the first author (Wang 1973).

The evolutionary view, on the other hand, has only recently received attention from computational linguists. Two important developments facilitated this interest. The first was advancement in genetic studies,
which shed light on the course of human evolution (Cavalli-Sforza 1994). Cavalli-Sforza’s research was inspired by Wang’s earlier work on lexical diffusion, and developed into a line of research whereby genetic evidence collaborates with evidence of language change. The second was the availability of large-scale language resources and computing power. These developments finally allowed for the simulation of language changes over a long period of time.

This macro view of language has the potential of complementing the micro view, and to move computational linguistics out of its current rule-based and statistics-based polarity and into a synergized dichotomy where collective tendency and individual motivation are both accounted for.

References


Computational Studies of Language Evolution*

William S-Y. Wang, Jinyun Ke, and James W. Minett

The study of language evolution has been recently revitalized because of converging interests and endeavors from across different disciplines. Computational modeling is one such fruitful area. Various aspects of language evolution have been studied using mathematical modeling and simulation. In this paper we discuss several computational studies in language change and language emergence.

1. Introduction

Questions about the origin and evolution of language, our species’ most distinctive attribute, have piqued human curiosity since ancient times. Earlier opinions on the subject we adjudge to have been fruitless speculations, inasmuch as they had no empirical foundations upon which to build. And in fact, in utter dismay, the linguistic societies of Paris and London banned all such discussion in the 19th century. Only by the mid-20th century had our ability to deal with these matters improved dramatically enough so that specialists from a variety of disciplines could pool their talents in taking up the challenge scientifically.

These disciplines literally range from A to Z, from anthropological concern with the physical development of our remote ancestors, to

* This paper was presented by the first author at the 2002 COLING in Taipei. We thank the organizers of the conference, particularly Chu-Ren Huang, for their kind invitation. Our research is supported in part by grants from the City University of Hong Kong [#9010001], and from the Hong Kong Research Grants Council [#9040781].

1 Some good landmarks for the return to respectability of the discussion of these issues include the well-known paper by Hockett (1963), and the large conference anthologized by Harnad et al. (1976).
zoological interest in animal communication and culture. More central here are the discoveries by linguists of universal tendencies found in all languages (Greenberg), by psycholinguists of the dynamics of language acquisition and loss (Jakobson), and by neuroscientists of how language is organized in the brain (Deacon).

Over the last several decades, the range of disciplines has broadened in two major steps. First, genetics has come on board with important hypotheses regarding the age of our Most Recent Common Ancestor, and regarding the correlation between groups of peoples and groups of languages. This development started with the so-called classical markers, and has been successively refined to gender-specific materials, first mitochondrial DNA for the maternal line and then the Y-chromosome for the paternal line. A consensus is gradually emerging that although anatomically modern humans first appeared over 100,000 years ago, our Most Recent Common Ancestor may date to only some 50,000 years ago. Such a date correlates well with the sudden burst of cultural achievements at many sites in the world, including the navigational skills to sail across large expanses of water.

It is reasonable to associate the origin of language with this date, since it is most likely that the power of language facilitated these cultural achievements. A more recent emergence date certainly makes the question of emergence more tractable, since there has been less time to obscure the traces of our primordial language(s). Indeed, some bolder scholars have been prospecting for words that may have existed in the primordial language which have been preserved in most branches of the world’s languages today. And other scholars have been exploring the possibility that the unique click consonants still extant in South Africa were indeed part of the primordial language phonology which had become lost in the branch of humans that left Africa to populate the rest of the world.

\[2\text{ It is still an open question whether language was invented only once (monogenesis), or several times independently (polygenesis). Freedman and Wang (1996) present some arguments to support the latter view.}\]
Fascinating as these explorations are, the fact remains that most of the pieces of evidence collected from the various disciplines are circumstantial, and that it is not possible to directly reconstruct the stages whereby our ancestors invented language dozens of millennia ago. This leads us to the second major step after genetics—the use of computational linguistics in the study of language evolution, which for convenience we shall refer to as CSLE: the computational study of language evolution. This is an area that burst upon the scene with great vitality, attracting exciting research from a variety of viewpoints. This vitality can be seen from the many anthologies available since 1998, including those by Hurford et al. (1998), Knight et al. (2000), Cangelosi & Parisi (2001), Briscoe (2001) & Wray (2002).

Few proponents of CSLE take the innatist position that there is literally an autonomous organ for language, that language requires a special bioprogram, or that language is based on any instinct exclusive to it. Obviously, a very wide array of abilities must have been in place before our ancestors could have been ready for language, ranging over sensory, motor, memory, and cognitive dimensions, as well as social skills in courtship, forming alliances, collaborating in group activities, and strategizing against enemies. Many of these abilities are present to various extents in our ape relatives, even though it is clear our ancestors must have had more language readiness than the former now have. It is encouraging that some recent studies are beginning to give us hints on the neurobiological bases of some of these abilities, such as the discovery of the so-called mirror neurons and their implications for the ability to imitate.

The basic assumptions that CSLE makes are that numerous interactions among members of a community, as well as among members across communities, over a long span of time can result in behaviors and structures that are quite complex. Furthermore, the bottom-up paths leading to such complex structures often involve phase transitions, points in time at which there are abrupt non-linearities where the change seems to be more qualitative than quantitative.

3 Writing was invented much later, after the advent of agriculture some 10,000 years ago.
We see such phase transitions in the physical world, for instance, when ice changes abruptly to water, and then abruptly to steam, even when heat is added gradually and by a constant amount. Similarly, we can perhaps identify some phase transitions in the cultural evolution of language, as in the emergence of segmental phonology, the invention of hierarchic morphology and syntax, the use of recursion in sentence construction, etc. The points in time for such non-linearities and the driving forces for change are not nearly as well-defined and uniform as in physical systems, of course.

The linguistic analog to the addition of heat that drives the change in water would be the set of communicative needs the early hominids felt as their world became increasingly complex, often a result of their own expanding consciousness as they interacted with the environment (Schoenemann 1999). Furthermore, given that by 50,000 years ago there were numerous communities scattered in many parts of the Old World in diverse environmental niches (Klein 1989), it is very likely that the evolution of language proceeded at different rates in these communities, each community crossing the various linguistic thresholds in its own way.

We shall now consider three distinct approaches to computational studies of language evolution.

2. Modeling lexical diffusion and the snowball effect

In the middle of the 20th century, the dominant view of sound change was that the unit of change was the phoneme. This was a view that linguistics had essentially inherited from the influential Neogrammarians of the 19th century, who emphasized the doctrine that sound changes can have no exceptions. Taking cues from evolutionary theory in biology, the counter-proposal was that the unit of change is the word (Wang 1969). Wang suggested that a change proceeds by variation, often partitioning the relevant words into three classes: unchanged (U), variation (V) and changed (C). This view of language change was termed lexical diffusion, since the change diffuses itself across the population one word at a time.
An early study of lexical diffusion was conducted by Don Sherman (1975), who investigated the growth of diatones in the history of English, that is, the increase in the number of noun-verb pairs like permit/permit, contract/contract, etc. The earliest pronouncing dictionary he could find, that of 1570, listed only 3 such pairs, the next dictionary listed 8 such pairs, and so on, up to 1934. Plotting the growth in numbers of diatones against time, the graph that results, shown in Figure 1, suggests that we may have the beginning of an S-curve. Apparently, such curves are widely found in diffusions, both for cultural events and biological events. The biologists Luca Cavalli-Sforza & Marc Feldman (1981:29-30) wrote:

The new word that becomes part of a language… is an innovation and can be considered as an analog of mutation in biology… When the process of diffusion of an innovation is followed for a sufficiently long time, the frequency of use of the innovation almost always follows an S-shaped curve. At the beginning the number of acceptances rapidly increases. Then follows an approximately linear increase, and finally the increase slows down and is barely perceptible.

Figure 1: The chronological profile of diatone formation in English, after Sherman (1975).
Two centuries earlier, the poet Alexander Pope expressed the same idea in more social terms:

In words, as fashions, the same rule will hold
Alike fantastic, if too new or old
Be not the first by whom the new are tried
Nor yet the last to lay the old aside

With the realization that lexical diffusion of a single word has an S-shaped trajectory, the question naturally follows as to whether the words that are cohorts in a given change influence each other, and on the nature of this influence. One answer to these questions may be summarized in the term snowball effect. The term takes its name metaphorically from a snowball rolling down a snowy mountainside. The further down it rolls, the faster it goes, and the more snow it picks up along the way. So if the S-curve for the first word has a particular gradient, then the curve for the second word has a steeper gradient. Furthermore, the time delay between the first pair of words will be greater than that between the second pair of words, and so on.

So far, there have been two empirical studies on the snowball effect. The study by Ogura & Wang (1996) deals with the development of the -s suffix in the third person singular present indicative in the history of English, starting from the Early Modern English of the mid-15th century. The -s suffix competed with the -th suffix in this function for several centuries until it completely replaced the latter in the end.

The other, more detailed study is by Shen Zhongwei (1997) on the merger of two nasalized vowels in modern Shanghai. Shen used the ages of his informants as virtual time. Assuming that a person’s habits of pronunciation are largely fixed by age 15, say, then an informant who is 60 years old may reflect the speech of 45 years ago. Such a method is far from fail-safe, of course. But it is nevertheless very useful for shorter term changes that run their courses over several decades. Shen’s study has the merit of being based on a large number of informants—almost 400. There are 28 relevant words in Shen’s list, each of which is pronounced with a nasalized low front vowel by some older speakers. One by one, the vowel
in these relevant words moves back in its articulation, and the words become homophones with words that have back vowels. In other words, this is a classic case of vowel merger, of which there are numerous examples in language change. Our questions have to do with how the relevant words influence each other in the process.

2.1 A dynamical system model of lexical diffusion—one word

We begin our discussion of modeling lexical diffusion and the snowball effect by describing a dynamical system, first derived by Shen (1997), that models a sound change that affects a single word only.4 The model applies to a group of homogeneous language users who can each adopt one of two possible forms for the word that is undergoing the change, either the unchanged form $u$, or the changed form $c$; the model does not allow for free variation $v$. The state of this system at any time instant $t$, can be described in terms of the proportion, or frequency, of individuals who use the unchanged form $u(t)$, and the frequency of individuals who use the changed form $c(t)$. Note that since each individual must adopt either $u$ or $c$, $u(t) + c(t)$ must sum to 1 for all $t$. It is assumed that the frequencies $u$ and $c$ at some time instant can be calculated from the frequencies at an earlier time instant. In particular, it is assumed that use of the changed form is propagated by contact between pairs of speakers, one of whom uses the unchanged form, the other uses the changed form; thus the increase in the frequency of changed forms is proportional to the product $c(t) \times u(t)$. The increase in the frequency is also proportional to the rate of affective contact $\alpha$ (Shen 1997), and the length of time over which the sound change is observed $\delta t$. Hence the frequencies of changed and unchanged forms at time $t + \delta t$ can be written in terms of the frequencies at time $t$ as

4 We would particularly like to thank Jeff Chasnov of the Mathematics Department of the Hong Kong University of Science of Technology for his assistance in re-deriving Shen’s model for lexical diffusion of a sound change affecting a single word and deriving the models for a sound change affecting an arbitrary number of words—any errors in the presentation of these models are our own.
\begin{align*}
(1) \quad & c(t + \delta t) = c(t) + \alpha c(t)u(t)\delta t, \\
(2) \quad & u(t + \delta t) = u(t) - \alpha c(t)u(t)\delta t. \\
\end{align*}

The parameter $\alpha$ can also be interpreted as representing phonetic, social, or other pressures on individuals to adopt the changed form.

Figure 2 summarizes the rates at which the frequencies of changed and unchanged forms vary over a time interval of duration $\delta t$. Taking the difference in the values of $c$ at time $t$ and time $t + \delta t$, and dividing by the duration $\delta t$ gives the rate of change of the frequency of changed forms:

\begin{align*}
\frac{c(t + \delta t) - c(t)}{\delta t} = \alpha c(t)u(t).
\end{align*}

Taking the limit as $\delta t \to 0$ produces the differential equation

\begin{align*}
(4) \quad \frac{dc}{dt} = \alpha c(t)u(t).
\end{align*}

Recalling that $u + c = 1$, we obtain the differential equation

\begin{align*}
(5) \quad \frac{dc}{dt} = \alpha c(t)[1 - c(t)].
\end{align*}
The general solution to (5) is the well-known Logistic equation,

\[ c(t) = \varepsilon \frac{\exp(\alpha t)}{1 + \varepsilon \left[ \exp(\alpha t) - 1 \right]}, \]

where \( \varepsilon \) is the initial frequency of changed forms at time \( t = 0 \).

A plot of \( c(t) \) is given in Figure 3 for an initial value of \( \varepsilon = 1\% \) and the rate of affective contact \( \alpha = 0.1 \). The plot exhibits the characteristic feature of the S-shaped logistic curve—a slow initial increase, followed by a period of more rapid, almost linear increase, that quickly drops off again as the frequency approaches 100%—and indicates the gradual diffusion of the changed form throughout the entire population.

![Figure 3: The logistic curve. (\( \varepsilon = 1\%, \alpha = 0.1 \))](image)

### 2.2 A dynamical system model of lexical diffusion—multiple words

Having derived an expression equivalent to (6), Shen (1997) went on to apply the model to the diffusion of a sound change across a group of words. He assumed that the frequency of changed forms for each word could be described independently by the model just discussed, determining parameter values with the best fit to data collected for 28 words in Shanghai Chinese that exhibit the merger of /\( \ddot{a} \)/ and /\( \tilde{a} \)/. We take a different approach, however, extending the model just described by explicitly accounting for coupling among the words themselves, in addition to the
coupling between speakers that has already been modeled; that is, we assume that the rate of diffusion of the sound change in one word may affect the rate of diffusion in other words.

Given a group of \( n \) words that are affected by a sound change, we denote the frequency of unchanged forms of word \( i \) at time \( t \) by \( u_i(t) \) and the frequency of changed forms of that word by \( c_i(t) \), where \( u_i(t) + c_i(t) = 1 \). We extend the definition of Shen’s rate of affective contact (\( \alpha \) above) by specifying for each pair of words the rate \( \alpha_{ij} \), at which adoption of the changed form of word \( i \) is induced by the frequency of changed forms of word \( j \)—we call this the coupling rate of word \( j \) on word \( i \), referring to \( \alpha_{ij} \) for distinct \( i \) and \( j \) as cross-coupling, and to \( \alpha_{ii} \) as self-coupling.

![Figure 4: Diffusion of unchanged and changed forms over a time interval of duration \( \delta t \)—multiple words.](image)

The rate of increase of changed forms of word \( i \) is assumed to depend on \( u_i(t) \), as in Shen’s model. However, due to the coupling that we assume to exist between words, we propose that the rate of increase is proportional to the combined affect on word \( i \) of the frequencies of changed forms of all the words participating in the sound change. The frequencies of changed and unchanged forms at time \( t + \delta t \) can therefore be written in terms of the frequencies at time \( t \) as

\[
(7a) \quad c_i(t+\delta t) = c_i(t) + u_i(t) \sum_{j=1}^{n} \alpha_{ij} c_j(t),
\]
\[(7b) \quad u_i(t + \delta t) = u_i(t) - u_i(t) \sum_{j=1}^{n} \alpha_{ij} c_j(t).\]

Figure 4 summarizes the rates at which the frequencies of changed and unchanged forms change over some time interval \(\delta t\).

Letting \(\delta t\) tend to zero, as previously, we obtain the following system of differential equations for the frequencies of changed forms:

\[(8) \quad \frac{dc_i}{dt} = (1 - c_i) \sum_{j=1}^{n} \alpha_{ij} c_j.\]

We have found no analytic general solution to (8), but we note in particular that the frequency of changed forms of any word does not follow a logistic curve, as assumed by Shen, unless the cross-coupling is zero, i.e., \(a_{ij} = 0\) for all \(i \neq j\). We can however characterize the behavior of the system using numerical methods.

Figure 5 shows the frequencies of changed forms predicted by the model for four words with the coupling rates and initial values given in Table 1. The system is initiated with only a single word having undergone any change. The other three words are distinguished by their having different coupling rates. In order to simplify the system somewhat, we have set the cross-coupling rate of each word with respect to all other words to a constant, although the value of the constant differs for each word.

### Table 1: Coupling rates and initial frequencies of changed forms of four words undergoing lexical diffusion.

<table>
<thead>
<tr>
<th>Coupling ((\omega_i))</th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word 3</th>
<th>Word 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>20%</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Word 2</td>
<td>4%</td>
<td>14%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>Word 3</td>
<td>3%</td>
<td>3%</td>
<td>16%</td>
<td>3%</td>
</tr>
<tr>
<td>Word 4</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>18%</td>
</tr>
<tr>
<td>(c_i(0))</td>
<td>1%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

At first, Word 1 grows at a faster rate than the other words. This is due to the frequency of changed forms of Word 1 far exceeding those of the other words—growth due to the self-coupling of Word 1 therefore
exceeds that due to cross-coupling. Because the frequencies of changed forms of the other words are initially zero, their growth is initiated by cross-coupling between themselves and Word 1. As the sound change progresses, the growth rates gradually increase as self-coupling is strengthened, although the rate of increase is initially greater for those words with higher cross-coupling. For example, at about $t = 10$, the rate of increase of the frequency of changed forms of Word 2 exceeds that of either Word 3 or Word 4. The rate of increase of the earlier words then begins to fall, allowing the later words to catch up. Eventually, the words that commenced the sound change later overtake the words that preceded them. For example, soon after time $t = 30$, Word 2 has attained a higher frequency of changed forms than Word 1; some time later, Word 3 has also changed more than Word 1 (although this is not visible in the figure). This feature is not unreasonable: as Ogura & Wang (1996) observed in their study of the sound change *th* to *s* in English, most of the words that commenced the sound change later actually completed the change before the earlier words. In our example, the frequency of changed forms in each language eventually approaches 100%.

Figure 5: Predicted frequencies of changed forms for four words. (parameter values given in Table 1)
Even for this simplified system, the relative progress of the sound change in each word does not behave in a transparent manner. For example, we do not know what combination of values of self-coupling and cross-coupling for each word cause certain words to attain a higher frequency of changed forms despite having commenced changing more slowly. We therefore simplify the system further such that the behavior becomes more transparent. We do this by setting both the self-coupling and the cross-coupling for each word to be constant. The coupling rates for a set of $n$ words can therefore be represented by two parameters: the self-coupling, denoted by $\beta$, and the cross-coupling, denoted by $\gamma$, i.e.

$$
\gamma_{ij} = \begin{cases} 
\beta : & i = j \\
\gamma : & i \neq j
\end{cases}
$$

Note that we require that $\beta \geq \gamma$ since we expect the influence of a word on itself to be at least that of other words. The differential equation that describes the evolution of the frequency of changed forms of each word (8) reduces to

$$
\frac{dc_i}{dt} = \left(1 - c_i\right) \left[\beta c_i + \gamma \sum_{j \neq i}^{n} c_j\right].
$$

Figure 6 shows the frequencies of changed forms predicted by this simplified model for four words with the coupling rates and initial values given in Table 2. Note that, since each word has the same values of self-coupling and cross-coupling, only different initial values of the frequency of changed forms, $c_i(0)$, generate different curves. For this reason, we have encoded each word with a different value of $c$; this emulates the words’ starting to participate in the sound change at slightly different times. The figure clearly shows that initially the frequencies of changed forms of words that participate in the sound change earlier grow at a faster rate than that of words that participate later. At about time $t = 10$, however, the rates of growth of the words are approximately equal. Later, the words that participate in the change later progress at a faster rate than the earlier words; in other words, we observe a snowball effect. Eventually, the frequency of changed forms in each language approaches 100%. At no time, however,
does a later word attain a higher frequency of changed forms than an earlier word.

Table 2: Coupling rates and initial frequencies of changed forms of four words undergoing lexical diffusion—simplified model.

<table>
<thead>
<tr>
<th>Coupling:</th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word 3</th>
<th>Word 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>All words</td>
<td>... $\beta = 20%$ $\gamma = 2%$ ...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_i(0)$</td>
<td>5%</td>
<td>2%</td>
<td>0.5%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

Figure 6: Predicted frequencies of changed forms for four words—simplified model. (Parameter values given in Table 2)

We now prove that a snowball effect is inevitable under the simplified model (10). Suppose that a sound change affects a group of $n$ words. To demonstrate the snowball effect we must show that when the frequency of changed forms of one word exceeds that of another word, the rate of change of the frequency of the latter word exceeds that of the former. We therefore specify the constraint

$$(11) \ c_1(t) > c_2(t),$$
at some time instant $t$ and determine the condition under which

$$\frac{dc_2}{dt} > \frac{dc_1}{dt}.$$  

(12) By equation (10), condition (12) holds when

$$\left(1 - c_2\right)(\beta c_2 + \gamma c_1 + \gamma k) > \left(1 - c_1\right)(\beta c_1 + \gamma c_2 + \gamma k),$$  

(13) where $k$ is defined in terms of the frequencies of changed forms of the other words:

$$k = \sum_{j=3}^{n} c_j.$$  

(14) Thus condition (12) holds when

$$\beta(c_1^2 - c_2^2) - \beta(c_1 - c_2) + \gamma(c_1 - c_2) + \gamma k(c_1 - c_2) > 0$$

$$\Rightarrow \beta(c_1 + c_2) - \beta + \gamma + \gamma k > 0$$

$$\Rightarrow \beta c_1 + \beta c_2 + \gamma k > \beta - \gamma$$

Equation (15) indicates that when the frequencies of changed forms of the two words, $c_1$ and $c_2$, are large enough, the frequency of the later word grows more quickly than that of the earlier word, thereby demonstrating a snowball effect. Note also that since the rates of increase of the frequencies of later words eventually exceed those of earlier words, the tendency is for the curves to converge over time. Whether or not the distance between successive curves is less for later words than for earlier words will depend on the initial values of $c_i$.

In Figure 7, we show the snowball effect for the four words with the parameter values specified in Table 2, drawing attention to the time instant at which later words adopt the changed form at a higher rate than the word immediately preceding them. The first such transition occurs at about time $t = 13$, when Word 4 begins to adopt the sound change at a faster rate than
Word 3. Soon after, the rate of growth of Word 3 overtakes that of Word 2, followed by Word 2 overtaking Word 1. The actual order of the transition depends on the initial values of $c_i$.

2.3 Discussion

The dynamical system just described appears able to capture a number of features that are typically observed among a group of words that undergo a sound change. However, its utility as a realistic model of the diffusion of a sound change is yet to be established. In order to make predictions about the behavior of the frequencies of changed forms for a sound change in progress, the values of the frequencies of changed forms of each word and the coupling rate between each pair of words would have to be known. While the former may be estimated (with some difficulty) by polling speakers of the dialect undergoing the sound change, the latter are not directly measurable. We must therefore perform a test of the model, much like that performed by Shen (1997), by collecting data for various sound changes that have run their course and estimating the values of coupling and initial frequency that produce the best fit between predicted behavior and observed behavior. We intend to test our multiple-word model using the two data sets used by Shen (1997), allowing the two models to be compared directly, as well as locating other data (such as can be found in Lee 2002) and collecting new data sets.
Both the dynamical system developed by Shen and the extension of it that we present here model lexical diffusion under highly regular conditions: a sound change within an isolated, static population of language users, each behaving identically. While such models may capture very well the expected evolution of a sound change in isolation, the reality is nowhere near so regular. As Ogura (1993) shows graphically, diffusion often proceeds in fits and starts; Figure 8 shows the progress of a syntactic change (periphrastic do in English) in various sentence types. Clearly the change was not regular; nor was it complete, the change being reversed in one of the sentence types analyzed. Realistic models of lexical diffusion should be able to capture such kinds of behavior. One way to extend the models discussed earlier to achieve this is to allow the coupling rates to vary with time. This should allow dynamic social and phonological pressures to be modeled—diffusion can be “switched off” by setting the coupling rates to zero; competing changes can be modeled by introducing multiple changed forms; and so on. Of course, as we extend the expressive power of the
model, so we render more complex the task of discovering the history of a lexical diffusion process. As we consider in the next section, however, computational models that have been used to study other aspects of language evolution may perhaps be adapted to the task of modeling lexical diffusion.

Figure 8: The progress of a syntactic change (periphrastic *do* in English), after Ogura (1993). Lines represent the progress of the change in sentences of distinct types.

3. Modeling language evolution—dynamical systems

The area of lexical diffusion discussed in the previous section deals with language evolution at an intermediate window size, with changes
taking place over decades or centuries. Language evolution can be studied at many time scales, however, from the short interchanges between mother and child in early language acquisition to the many millennia over which language has evolved since its first emergence. We now turn to the question of language emergence.

Computational studies of language emergence provide a very valuable antithesis to the currently popular innatist position that there is literally an autonomous organ for language, or that language requires a special bioprogram, or that language is based on any instinct exclusive to it. If language had emerged only very recently as recent studies in population genetics indicate, then the likelihood is small indeed that biological evolution could have put such an organ in place.

On the other hand, it is obvious that a very wide array of abilities must have been in place before our ancestors were ready for language, ranging over sensory, motor, memory and cognitive dimensions, as well as social skills in courtship, forming alliances, collaborating in group activities, and strategizing against enemies. Some years back, Wang (1978:116) characterized this point of view with words like ‘mosaic’ and ‘interface’. It was in this spirit that Tzeng & Wang (1983) carried out a set of experiments to argue for a common neuro-cognitive mechanism for both language and movements.

A basic assumption that many computational studies make is that numerous interactions among members of a community, as well as among members across communities, over a long span of time can result in behaviors and structures that are quite complex. When Murray Gell-Mann (1994) wrote of the evolution of “highly complex forms,” he could have easily included languages among his examples.

The field of modeling language evolution by computer essentially began with Hurford’s (1989) discussion of the emergence of a consistent lexicon. Hurford considered the relative merits of three highly idealized learning strategies. Individuals adopting the Imaginator strategy produce a particular utterance to indicate a certain object when they observe that nearby individuals typically produce that utterance to indicate the object. Similarly, they attend to a particular object when perceiving a certain utterance when they observe that nearby individuals typically attend to that
object in response to the utterance. Calculators, however, use a particular utterance to indicate a certain object when they observe that others attend to the object when perceiving that utterance. Similarly, they attend to a particular object in response to an utterance when nearby individuals use that utterance to indicate the object. Hurford shows, however, that a better strategy is to follow the approach he refers to as Saussurean. Individuals adopting the Saussurean strategy copy the speech production patterns of nearby individuals, like Imitators. But unlike Imitators, they make their perception consistent with their production. Thus, Saussureans base both their production and perception on the speech production of other individuals.

Surprisingly, perhaps, Hurford’s paper did not stimulate an immediate interest to take up the challenge of modeling language by computer. The field had to wait for further stimulation from the fields of game theory, cellular automata, and artificial neural networks before contributions in the mid-'90s, such as by Clark & Roberts (1993), Batali (1994), Hutchins & Hazlehurst (1995), Steels (1995), and Noble & Cliff (1995), began to appear.

While many models for language evolution have adopted the agent-based simulation paradigm (which we discuss in Section 4), comparatively few models have been based on dynamical systems. An appropriately designed dynamical system is most useful for describing the qualitative behavior of a system rather than predicting the exact behavior for a particular instance of the system. Thus the behavior of a system can be described in terms of its stable and unstable equilibria, oscillations, the values of certain parameters at which the system bifurcates, and so on. Much of the recent work on modeling language change from a dynamical systems perspective has come from Martin Nowak and his associates. They have focused particularly on modeling the evolution of universal grammar, which we now discuss.

3.1 Evolution of universal grammar

Nowak, Komarova & Niyogi (2001) have proposed a dynamical system model of the evolution of Universal Grammar (UG) among a population of heterogeneous language agents. Universal grammar is an
abstract representation of one currently popular view of language acquisition. UG consists of a “mechanism to generate a search space for all mental candidate grammars” and “a learning procedure that specifies how to evaluate sample sentences” (Nowak et al. 2001). The UG model fits very well with the innatist viewpoint of language emergence: that our language faculties derive from a language-specific organ, the Language Acquisition Device (LAD). However, the model can equally well be applied without adopting an innatist position, provided we assume that children are able to develop a learning algorithm for acquiring language without recourse to exclusively innate capabilities of some LAD. Nowak and colleagues also assume that the UG consists of a finite set of grammars.

In the model, each agent uses a single grammar. In some cases, a sentence that can be parsed in one grammar may also be parsed in another grammar. Thus users of different grammars may often be able to communicate with some degree of success. For example, while the sentence “I might could do” is not grammatical in many English dialects (although many people would understand the intended meaning), it is grammatical in parts of Scotland and the USA (Trask 1996); nevertheless, many sentences that are grammatical in those particular dialects are also grammatical in many other English dialects. The result is that such dialects are mutually intelligible to such a degree that communication between them is essentially unimpaired.

For each pair of grammars, the probability that a user of grammar $G_i$ produces a sentence that can be parsed by a user of grammar $G_j$ is denoted by $a_{ij}$, with $0 \leq a_{ij} \leq 1$ and $a_{ii} = 1$. The probability that two agents who use grammars $G_i$ and $G_j$ respectively can communicate successfully is therefore given by

$$F(G_i, G_j) = \frac{1}{2} (a_{ij} + a_{ji})$$

The payoff of grammar $G_i$ is then defined as the average probability that a speaker of $G_i$ produces a sentence that can be parsed by an arbitrary individual, i.e.

$$f_i = \sum_{j=1}^{n} x_j F(G_i, G_j),$$
where $x_j$ is the frequency of individuals using grammar $G_j$. The payoff associated with a particular grammar is assumed to be linked to the reproductive success of individuals who use that grammar. Nevertheless, children do not always learn the grammar of their parents.

In the model, the distribution of the grammars that a child may learn depends on a matrix whose values indicate the probability $Q_{ji}$, that a child with a parent who uses $G_i$ learns to use $G_j$. Nowak et al. propose the following language dynamical equation to explain the evolution of the system:

$$ \frac{dx_i}{dt} = \sum_{j=1}^{n} x_j f_j Q_{ji} - \phi x_i, $$

where $\phi = \sum_j x_j f_j$ is termed the grammatical coherence, which also measures the average probability of mutual understanding among members of the population.

Equation (18), much like equations (8) and (10) in the section on lexical diffusion, describes a kind of diffusion process, although, as Nowak and colleagues show, the diffusion is not necessarily complete. They display a graph, repeated in Figure 9, to indicate under what (simplified) conditions a population converges to a single dominant grammar—distinct grammars are assumed to be equidistant in the sense that $a_{ij} = a$ for some constant $0 \leq a \leq 1$ and $i \neq j$. They find that a single dominant grammar only diffuses across the population when the probability that children learn the grammar of their parents $Q_{ii}$ is sufficiently high; otherwise a state in which each grammar has equal frequency emerges. Even when a single dominant grammar does emerge, it does not diffuse across the entire population unless children learn the grammar of their parents perfectly—other grammars continue to be used by some members of the population, although with relatively small frequency.

The UG model, (18), appears to have a formulation very similar to that of the lexical diffusion (LD) model, (8), which we discussed earlier. Since, as we have just observed, the UG model sometimes produces diffusional behavior, the question arises whether this model can be used to model some features of lexical diffusion. Equation (18) can be rearranged as
The lexical diffusion equation (8) can be written as

\[
\frac{dx}{dt} = \sum_{j=1}^{n} (Q_{ji} - x_{i}) f_{j} x_{j}.
\]

The two sets of variables, \(x_{i}\) in (19) and \(c_{i}\) in (20), both represent the frequencies of a set of time-varying, measurable linguistic quantities. In the UG system the frequencies must sum to 1—the various grammars must therefore “compete” for users. However, in the LD system there is no such competition—the frequency of changed forms of any particular word may reach 1 without forcing the frequencies for other words to fall. This substantially different behavior is introduced by the factor \((Q_{ji} - x_{i})\); the UG frequencies increase whenever \(Q_{ji} > x_{i}\) but fall whenever \(Q_{ji} < x_{i}\).

In his initial formulation of lexical diffusion (1969), Wang had in...
mind that sound changes sometimes compete for words, causing apparent irregularities to appear in the changing phonological system. For example, consider the two hypothetical sound changes $R_1 : A \rightarrow B$ and $R_2 : A \rightarrow C$ that describe the change of a segment $A$ either to $B$ or to $C$. Suppose further that $R_1$ commences first but is interrupted by $R_2$ before it is complete. Those words that have already acquired $B$ due to $R_1$ will not be affected by $R_2$. Only those words that still maintain the unchanged segment $A$ will be affected by $R_2$. Some words, however, may be in the process of change, with some members of the population still using the unchanged form $A$, while others use the changed form $B$—in such cases the two sound changes will tend to compete for both words and speakers. The result then will be that some words will have replaced $A$ with $B$, while others will have replaced $A$ with $C$ (some free variation between $B$ and $C$—and perhaps $A$ also—might persist among some speakers). A similar, apparent irregularity would emerge if the second sound change were instead $R_2' : B \rightarrow C$.

What then can Nowak’s model of the evolution of universal grammar tell us about lexical diffusion and, more generally, about linguistic diffusion processes at large? Equation (18) tells us that the adoption of a single grammar by an entire population occurs only rarely, if ever. Instead of using the $x_i$ in (18) to represent the frequencies of the numerous possible grammars, we could consider using the $x_i$ to represent the frequencies of the various sounds that can arise due to a set of competing sound changes. If we take the parameters $Q_{ji}$ to indicate the probability that sound $i$ is replaced by sound $j$, and substitute for $f_i$ a coefficient that measures the affect of coupling between both pairs of words and between sounds, we obtain a dynamical system describing the evolution of a set of competing sound changes. Developing such a model might allow us to gain more insight into the circumstances under which qualitatively different types of behavior emerge when a set of sound changes compete.

4. Modeling language evolution—multi-agent systems

The model proposed by Nowak et al. in studying language evolution is concerned with the evolution of universal grammar from the perspective
that language as a whole evolves under the mechanism of natural selection. While this analytic approach is promising in providing a framework to study the dynamics of language evolution as a whole system, it gives us little hint on how a language system came into being. When speculating on the process of language emergence, there can be few who believe that a complex language system with elaborate lexicon, morphology, and syntax could have sprung up all of a sudden from scratch. Language must have emerged and evolved gradually and incrementally to reach its modern form. And this process must have developed due to the communication interactions among individuals. It is through these communication interactions that language emerged and evolved to meet the increasing communication needs.

Our proposition is that, given the cognitive and physiological prerequisites being available, language emergence and evolution is basically a continuous conventionalization process, from the individual innovation of a new linguistic item, a word, a phrase, or a syntactic construction, to the diffusion of the innovation through interactions among individual language users by imitation and learning during language acquisition. A compelling scenario for the emergence of language is that first a set of early words or holistic signals emerged, and that later different word orders or relationships between words came to be used to signify different aspects and moods etc. in order to cope with the increasing need to express more complex meanings. Our view, therefore, is that one of the first steps in exploring the emergence of language in the macro scale should be to study how the first words emerged.

4.1 The emergence of vocabulary

Words are the fundamental functional unit in language, characterized by pairing meanings with sounds in arbitrary ways. A modern individual typically has many thousands of words in his lexicon through which he sees his universe, and by means of which he communicates his needs and desires. At the outset, however, such symbols were much fewer. Zoologists tell us that no animal in its natural state has more than several dozen symbols, be they vocal calls, facial expressions or body gestures (Hauser 1996). While animals’ communication systems are mostly imprinted innately...
(though some of them are affected by learning, e.g. bird song, reported by Marler (1987)), the words used in a linguistic community are mostly established by conventionalization. Two great philosophers have pronounced similar ideas regarding the conventionalization of naming. Xunzi in China taught that “names have no intrinsic appropriateness” and “names have no intrinsic reality”, while at about the same time in Greece, Plato wrote that “any name which you give is the right one, and if you change that and give another, the new name is as correct as the old” (Wang 1989).

We have designed several models to simulate the process of conventionalization leading to emergence of a shared lexicon (Ke et al. 2002b). Assuming the existence of a set of meanings and a set of utterances, a lexicon is a set of mappings each of which associates a meaning with an utterance. The association can be either probabilistic or binary, with a score measuring the frequency of usage. Such formulation has been used in early simulation studies of the emergence of a lexicon or a language as a set of holistic signals (Hurford 1989, Steels 1995, Nowak et al. 1999). The principal questions regarding the emergence of lexicon therefore include how these associations are formed and how members of a population reach the same set of associations. The answers to these questions lie in the modes of interaction among agents during communication.

Imitation is one of the most likely interaction mechanisms. The strong ability of imitation in humans, even from early infancy, has been extensively documented in the studies reported by many investigators, e.g. Meltzoff (1996). While other social animals, particularly the primates, also imitate (Dugatkin 2000), it appears that the tendency is by far the strongest and most general in our species. We assume that imitation may serve as the most explanatory mechanism for the formation of a common vocabulary. Before establishing a consistent way of naming things, early humans very likely made use of their propensity for imitation; the younger ones imitating their elders, the followers imitating the leaders or, just by chance, their neighbors. In the simulation model, we assume there to be a number of agents, each of which initially has its own set of mappings between meanings and utterances. When two agents interact, one imitates the other according to some strategy, either randomly, by following the majority in the population, or by avoiding homophones. We demonstrate that the
agents in the population always converge to a single identical vocabulary. Some mathematical models using Markov chain have been used to prove the convergence (Ke et al. 2002b).

While in the imitation model we adopt only one set of mappings for the associations between meanings and utterances, in another simulation model, we distinguish the speaking and listening vocabulary, each represented by a set of mappings between meanings and utterances. This representation is considered to be more realistic, and is more effective in dealing with the fact that active and passive vocabularies are often not identical. An example of the two matrices with three meanings and three utterances is given in Table 3.

We hypothesize another type of interaction, in which probabilistic changes are applied to the mappings, rather than imitation in the discrete manner used in the above model. When two agents interact, a successful interaction occurs when the listener interprets the received utterance as the meaning intended by the speaker—this will result in a reinforcement of the mapping used in the speaker’s speaking matrix and the listener’s listening matrix. On the contrary, a failed interaction will lead to the mappings used by the two agents to be weakened. At the beginning of the simulation, the speaking and listening matrices of each agent are both randomly initialized.

After a number of interactions, the system converges: a common lexicon emerges in the population (inter-agent convergence), and the speaking and listening matrices of each agent reach a compatible state (intra-agent convergence). When the number of meanings equals the number of utterances, the speaking and listening matrices of each agent are identical. However, when the number of meanings is larger than the number of utterances, the speaking matrix of each agent is a subset of the listening matrices. Nevertheless, in both cases, the speaking and listening matrices of each agent become compatible.

The intra-agent convergence is an emergent property of the system as there is no explicit and obligatory mechanism forcing the speaking and listening matrices to be compatible. This leads us to speculate that it might not be necessary to presume a Saussurean strategy for the formation of vocabulary as proposed by Hurford (1989)—Hurford shows that a Saussurean strategy, in which the speaking and listening matrices are
assumed to be identical, has the advantage of high communication effectiveness over other strategies and therefore might have been selected by biological evolution.

Figure 10 shows the trends of three measures of convergence: similarity (SI), population consistency (PC) and individual consistency (IC), from a typical run of simulation. In the run, the population consists of 10 agents, each starting with a vocabulary in which each meaning is randomly associated with each utterance with a different probability. A consistent vocabulary is formed and shared by all agents after a long period of fluctuation.

It can be seen that the convergence is not gradual but rather is quite abrupt after about 3,000 interactions, exhibiting a “phase transition” characteristic. For a long period of time, the interactions among agents only result in fluctuation, and there is little consistency in the population’s vocabularies. However, at some instant, there is an abrupt rise of the consistency, and the population converges quickly after that period. The conditions of the model have not changed at all in the process toward
convergence. The abrupt emergence of order in the population is the result of a sequence of interactions which bring a converging momentum.

Table 3: An example of the speaking and listening matrices in the interaction model in Ke et al. (2002b)

<table>
<thead>
<tr>
<th></th>
<th>u₁</th>
<th>u₂</th>
<th>u₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>p₁</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>m₁</td>
<td>0.1</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>p₂</td>
<td>0.4</td>
<td>0.55</td>
<td>0.05</td>
</tr>
<tr>
<td>m₂</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>p₃</td>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>m₃</td>
<td>0.4</td>
<td>0.4</td>
<td>0.1</td>
</tr>
</tbody>
</table>

For this interaction model, we have also investigated the effects of various parameters. When the agents are prohibited from interacting with themselves, we observe an interesting window in the optimal population size, between 5 and 15, in which the population converges the most quickly. When "self-talk", which means the agents can talk to themselves, is allowed, however, the convergence becomes much faster, and there is no such window effect: the smaller the population size, the easier it is for the population to achieve a consistent vocabulary. However, the convergence time does not increase linearly with the increase of the population size, which suggests that there may be a threshold of population size for the convergence to be realistically possible in time. This may be linked with speculations that our ancestors tended to gather into populations of 50 to 100 members. It has been proposed that the social structure and size of groups of *homo sapiens* is one of the most important factors in the emergence of language (Dunbar 1993).

The observation of an optimal population size under the condition of prohibiting "self-talk" is unexpected. A preliminary explanation is that in a very small population consensus accumulates with great difficulty—agents’ mappings are very sensitive to small adjustments after each interaction and thus keep oscillating. When "self-talk" is allowed, however, the indirect imitation helps to suppress such oscillation to a large extent because the speaking behavior of the group of agents influences not only each individual agent’s listening vocabulary but also, indirectly, that agent’s speaking behavior through interaction with him/herself.
4.2 Homophony and ambiguity

The existence and abundance of ambiguity in languages has intrigued linguists for a long time. It would seem that in an ideal code, one signal should correspond to exactly one message, and that ambiguities due to one-to-many correspondences would cause miscommunication. Yet all languages are rife with such ambiguities at various levels, from polysemy to homophony to syntactic ambiguities. Indeed ambiguity has been the most formidable barrier to computational linguistics since its start—from automatic summary, to machine translation, to speech recognition—and remains so today, even as methods of disambiguation are becoming increasingly sophisticated and powerful.

Recently, we have started to investigate a typical phenomenon, homophony in the lexicon, which is usually considered to be a major source of ambiguity, undertaking both empirical analysis and modeling approaches (Ke et al. 2002a). Why and how do homophones arise, and how do they survive? Nowak et al. (1999) use a mathematical model to demonstrate that within a given limited signal space, homophony is unavoidable to achieve the least error limit. We propose a simulation model of naming games (Steels 1997) to elaborate this view.

Similar to the interaction model reported above, agents in the model are assumed to be able to produce a number of distinctive utterances and make use of such utterances to communicate a set of meanings. At the beginning, the agents do not have any words, a word here referring to an association between a meaning and an utterance. Agents can create new words at random, and learn the words created by other agents as well.

The simulation model is designed within the “name games” framework proposed by Steels (1995). The flow is as follows. At each time step, two agents are chosen to communicate, one as the speaker and the other as the listener. The speaker decides a meaning he/she wants to communicate, looks for or creates an utterance which is associated with the

---

5 Ambiguities sometimes serve various purposes in linguistic play—in puns, jokes, etc.—but these are surely developments which arose much later after ambiguities had taken root in languages.
meaning, and transmits the utterance to the listener. The listener perceives the utterance and tries to interpret the meaning by searching his existing lexicon. If he interprets the same meaning for the utterance, then this is considered to be a successful communication. Each word has a score; after each successful communication, the score of the word is increased. Otherwise, the score is decreased. When the score of the word becomes too small, the word is removed from the lexicon. Upon failure, the listener learns the word from the speaker by adding an association between the perceived utterance and the intended meaning of the speaker. At the beginning, all agents have an empty lexicon. However, after a long period of interaction, we observe that a set of associations between objects and utterances are shared by all agents.

Figure 11: Simulation results: the homophone evolution under four conditions. (Upper left: $M = U$, one-word communication; upper right: $M = 3U$, one-word communication; lower left: $M = U$, two-word communication; lower right: $M = 3U$, two-word communication)
With the above construction, we compare the convergence for two different ratios between the number of meanings (M) and the number of utterances (U). Figure 11 shows simulation results for the two different ratios. When M = U, we can see that agents are able to acquire the same lexicon, and their communications are successful 90% of the time, 20% of the words having homophones. When we increase the number of meanings that are to be communicated by agents, for example, setting the meaning-utterance ratio to 3, the lexicons of the agents no longer converge—every word has homophones—resulting in a low rate of communicative success (only about 30%).

This situation M > U simulates a realistic situation in which our lexical needs far exceed the number of words we can utilize. Cheng (1998) has shown that there exists a similar limit to the size of the active vocabulary, which suggests a cognitive constraint in the size of lexicon. As a result of this limit, lexicons are abundant with homophones, not to mention the polysemes which are ubiquitous across languages. In spite of the considerable ambiguity implied by homophones, our daily communication does not seem to be much hampered by it. We are seeking explanations for effective communication in real situations which the current model seems unable to demonstrate.

In the above simulation, only one meaning is transmitted each communication event. In a real situation, most of the time, we communicate with a phrase or a sentence. Words in the phrase or sentence are always semantically related. To simulate this situation, we have designed a two-word communication model. In a communication event, the speaker chooses two meanings (m1 and m2) which are close to each other in the semantic space, and produces two utterances (u1 and u2) to communicate with another agent. The listener receives the two utterances. If u1 has only one meaning m1, and u2 has two or more meanings, say m3 and m4, then the listener will choose between m3 and m4 according to which is closer to m1 in the semantic space. If neither u1 nor u2 has a unique meaning, the same principle of disambiguation can be applied: the listener will choose that pair of meanings from the u1-meanings and u2-meanings that are the closest in the semantic space.

In this formulation, while the semantic proximity helps to disambiguate
homophonous utterances, random learning could cause trouble when the listener learns the wrong order of association. However, we observe a gradual increase in the rate of communicative success through successive interactions. When $M = U$, we see that the communicative success reaches 100%, much better than the early case of one-word communication, even with a degree of homophony as high as 70%. When we increase the semantic demand, we can see a much clearer improvement of the system owing to the two-word communication. Homophony can be tolerated up to about 100%, with the rate of communicative success still rising to more than 80%. This simulation demonstrates clearly that, with the help of context, the lexicon can tolerate a high degree of homophony, even when the number of meanings greatly exceeds that of utterances.

5. Discussion

We would like to offer some concluding remarks on this exciting new area of CSLE regarding the assumptions and limitations of current methodologies, as well as regarding the road that lies ahead.

While computational models can demonstrate how certain linguistic structures emerge and/or change, most of them, such as the first and third cases we reported above, simulate the interactions between individuals under certain assumptions and constraints, with large degrees of idealization and simplification. It is an advantage of computational models that various assumptions must be made explicit and implementable, and thus can be examined, verified, and compared. For example, in simulating communication interactions among agents, the models have to clearly specify various details as to how meanings are represented by the agent, how meanings are transmitted by the speaker, and how the listener interprets the signals received. And in simulating language acquisition, the models must be explicit regarding the properties that the learners are assumed to be endowed with, such as the learning algorithm (if any) which determines how the learners construct their own language by memorizing and extracting regularities from the linguistic input.

Currently most models make rather strong assumptions or great
simplifications of the real situations. For example, in our model of lexicon formation, we assume that meanings are transmitted explicitly and listeners have no problem at all in knowing the meaning intended by the speaker. Many other computational models simulating the interactions between individuals adopt a similar assumption, especially those in which agents are represented by neural networks and learn the meaning-signal mappings by some training process (e.g. Batali 1998). However, the transparency of meaning in communication may not be true in many real situations as ambiguous interpretations are almost always possible.

Moreover, in the case of the acquisition of the first language, it is not well known how children can identify the meanings intended by adults. There have in fact been some studies addressing this problem by making the meaning interpretation more realistic. For example, in Steels (2002) the meanings being communicated are embodied by object detection and feature extraction; Cangelosi et al. (2002) incorporate environmental information together with the signal during communication, and the listener interprets the meaning from the environmental information. With these embodiments implemented in the models, consistent communication still emerges without the prerequisites of meaning transference. It is through the process of identifying and then relaxing assumptions in the simulation models that more realistic frameworks are obtained and utilized.

A hypothesis supported by one model might not be supported by another model that is implemented based on different assumptions. For example, Kirby (2001) demonstrates that a compositional language can emerge from a set of random meaning-signal mappings by an iterative-learning model. However, in his model the mappings are represented by a version of Definite Clause Grammar and learners are assumed to have an induction algorithm that can look for common substrings and infer generalized rules generating them, which are highly biased toward language-like systems. He hypothesizes that a bottleneck effect, by which the learner is only exposed to a small subset of the possible language, is necessary for the emergence of compositional language. However, in a critique of this model by Tonkes & Wiles (2002), a neural network model is implemented for which no explicit rules or generalizations are required. They show that compositionality still emerges without an explicit bottleneck
hypothesis. From these models, we can see that computational models allow different hypotheses to be evaluated and compared objectively as long as the assumptions and representations have been explicitly stated.

It is clear that we should be encouraged by what the new area has achieved so far. The knowledge base for research on language evolution must rest on what linguistics has to offer, regarding how the several thousand languages available to us are organized, what is the common core of this organization that is shared by all languages, extending to the most idiosyncratic features observed for just a few languages, which marks the outer periphery of what a language can be like. This knowledge base grew tremendously in the 20th century, when linguists described a broad range of languages in many parts of the world which had not been studied scientifically before. This linguistic knowledge has been joined by genetic knowledge since the 1980s in research on language evolution, and by computational studies since the 1990s.

As we look back on this decade or so of CSLE, it is clear that the achievements have been impressive and encouraging. At the same time, we see that there are many central topics on language evolution awaiting careful formulation and investigation.

While there have been exciting simulations on the emergence of the lexicon, and on the formation of phonological systems, not much is known about how hierarchical syntax emerged. Hierarchical structures are a hallmark of complexity, as Herbert Simon noted decades ago. When a chimpanzee takes off the top of a box to get at the banana inside, it presumably recognizes that the two parts of the box are discontinuous constituents of a single hierarchical unit that holds the banana. Cognitively it is comparable to separating constituents of language, such as taking apart ‘call up’ in ‘call him up’, or embedding large constructions within expressions like ‘what for’, such as in ‘what did you call him up for?’

---

6 It is sad to note concurrently that the 20th century also marks the accelerated extinction of indigenous languages as these are replaced by a few international languages, empowered by economic and technological success. This development has a homogenizing effect which simultaneously expands the common core and shrinks the outer periphery of the space within which language locates.
Linguists have studied the dependency relations of constituents in great depth in a variety of languages—what can be moved, what can be deleted, what can cross over, etc.—and we can hope that computer simulations will soon be able to model such dependency relations within hierarchical structures.

Hierarchical structures are the bases of recursiveness, and recursion is the central mechanism that makes language infinite via repeated conjoining and embedding. While it is undeniable that there is no longest sentence, the fact remains that most utterances in everyday language are quite short, and statistical approximations to these utterances can be very useful in helping us understand the structure and function of such language.

Another question that has intrigued us a lot in recent years is that of ambiguity. From a CSLE vantage point, an interesting research topic would be to see at which points various types of ambiguities emerge as the most rudimentary languages with the simplest lexicons gradually grow toward the level of complexity of modern languages. Embedded in this topic are several questions concerning a typology of ambiguities in the languages of the world: are there universal ambiguities, how do we typologize them, and how do we predict them from the structures in which they reside? Since ambiguities are at once a robust phenomenon and probably unique to human communication, simulating their emergence can tell us much about the nature of language.

As our last point here, we would like to emphasize the tremendous heterogeneity of language. To get our computer simulations started, it is natural to have small and simple models, with a limited community of members who speak a homogeneous language. However, as the simulations continue, as the members and generations multiply, and as the number of interactions grows very large, we should expect the languages to become greatly diversified and the linguistic behaviors of the speakers increasingly heterogeneous.

The fact that two people are talking with each other by no means leads to the conclusion that they really understand each other, or that they share the same grammar and linguistic representations. As communities become larger and more complex, their speakers become more diverse as well. One school of linguistics once claimed that the central focus of its research was
on an ideal speaker-listener situated in a homogeneous community, an attitude that someone has criticized as ‘monastic.’ As the empirical foundations for linguistics have grown, however, there has come to be a fuller realization of how much speakers differ from each other, even within the same family. It is such variability, of course, when amplified manifold across time and space, which produces dialects, and eventually distinct languages. It would be a worthy goal for CSLE to eventually be able to simulate such evolutionary processes with realism. Given that the area has been progressing at such an exciting pace, such a goal may not be too far away.

References


Komarova, Natalia L., Partha Niyogi, and Martin A. Nowak. 2001. The


Biographic Sketch and Research Outline of Authors

William S-Y. Wang

William Shiyuan Wang was born in Shanghai, where he received his early schooling. In 1951, he won a four-year scholarship to Columbia College in New York City. While still a freshman, he co-authored a musical production of the Chinese legend, *The Cowherd and the Weaving Maid*, which was performed at the International House of New York. He went on to pursue graduate studies in linguistics at the University of Michigan, and wrote his Ph.D. dissertation in the laboratories of Prof. Gordon E. Peterson. The dissertation, completed in 1959, was one of the first studies to apply the combined knowledge of linguistics and acoustics to the problem of machine recognition of speech.

After his graduate studies, Wang undertook research on machine translation from Russian to English at the I.B.M. Research Center at Yorktown Heights. He also held a post-doctoral appointment at the Research Laboratory of Electronics of the Massachusetts Institute of Technology, working on problems of speech analysis. He returned to teach at the University of Michigan for a year before accepting a position at the Ohio State University. There he established a Department of Linguistics and a Department of East Asian Languages, and served as the first Chairman of both.

After a visiting professorship sponsored by the Departments of Chinese and of Anthropology of the National Taiwan University, Wang accepted a position as Professor of Linguistics at the University of California in Berkeley in 1966, and served in that capacity until his retirement in 1994. In 1973, he founded the *Journal of Chinese Linguistics*, the first international publication in the field, and continues as its editor till this day, with its office at the Project on Linguistic Analysis at Berkeley.

The honors he has received include a Guggenheim Fellowship [New York], two fellowships from the Center for Advanced Studies in the Behavioral Sciences [Stanford], a National Professorship from Sweden [Stockholm, Umea, Upsala, and Lund], and a resident fellowship from the Center for Advanced Studies at Bellagio [Italy]. He was elected President
of the International Association of Chinese Linguistics when it was founded in 1992 [Singapore]. He is an Academician of the Academia Sinica [Taiwan], and serves on the Advisory Committees of the Institute of Linguistics [chairman] and of the Institute of Information Sciences [member].

At the City University of Hong Kong, Wang has served as Chair Professor of Language Engineering in the Department of Electronic Engineering since 1995. He directs the Language Engineering Laboratory, and was Chair of the Quality Assurance Committee and of the University Press Committee. He is an Adjunct Professor of the Hong Kong University of Science and Technology, Lanzhou University, Nankai University, and Yunnan University.

His publications include numerous articles in technical journals, and several encyclopedias, as well as in general science magazines, including *American Scientist*, *Nature*, *Proceedings of the National Academy of Sciences* [USA], *Scientific American*, and *Scientific Chinese*. He has contributed to several encyclopedias, including “Speech” in the *McGraw-Hill Encyclopedia of Science and Technology*, “Sino-Tibetan” in *Encyclopedia of Language and Linguistics*, and “Origins of Language” in the *Oxford International Encyclopedia of Linguistics*.

In recent years, Wang has been collaborating with biologists and computer scientists in a common search for the origin of language, and for patterns in language differentiation and language endangerment. His research on language is from an interdisciplinary perspective, involving engineering, linguistics, and biological sciences.

List of Publications:

http://www.ee.cityu.edu.hk/~wsyw/

**Jinyun Ke**

Ms. Jinyun Ke is currently a Ph.D. candidate in the Language Engineering Laboratory, Department of Electronic Engineering of City University of Hong Kong. Her main research interest is to apply a self-organization theory to study language evolution, considering linguistic structures to be emergent phenomena at two levels, that is, in the individual
language user and in the language community. Language changes are the inevitable consequence of heterogeneity in a self-organizing language community. She is studying the evolution of homophones in different languages from this perspective, investigating the interwoven relationship between homophony with phonological inventory and syllabicity in the lexicon. Also, she is interested in applying complex network theory to study the semantic structure of lexica, as well as the effect of social networks on language evolution. In addition to empirical studies on language change, computational models, particularly multi-agent models, are used to simulate language change and language emergence.

List of Publications:
http://www.ee.cityu.edu.hk/~lelab/publication.htm

James W. Minett

James Minett is a research fellow at the Language Engineering Laboratory, City University of Hong Kong, having completed the B.Sc. and M.Sc. degrees in Mathematics and the Ph.D. degree in Electronic Engineering.

He has brought his training in mathematical and computational techniques to bear on various aspects of both historical and evolutionary linguistics. In historical linguistics, his main aim has been to develop algorithms that can be used to detect language contact and borrowing, and has proposed both cladistic and lexicostatistical methods for this purpose. In evolutionary linguistics, he has been investigating linguistic diffusion processes and their impact, his current focus being on the computational modeling of the dynamics of language endangerment and maintenance, taking particular account of the effects of both bilingualism and social structure. Along with other members of the Language Engineering Laboratory, he is also working on developing models for the emergence of language, with the aim of describing how the lexicon and syntax, as well as social structure, might plausibly have co-evolved.
Selected Research Output:

List of Publications:
http://www.ee.cityu.edu.hk/~lelab/
Introduction to
New Chances for Deep Linguistic Processing

Winfried Lenders

During the last decade, large-scale, statistically based methods for natural language processing has become more and more prominent in practical application systems like machine translation (MT) and information extraction (IE). They are favored from a technological point of view because they are better equipped to overcome the traditional bottlenecks of rule based systems: lack of efficiency, robustness, and coverage. One of the most prominent language processing projects where statistically based methods were used and investigated was VERBMOBIL, a collaboration of many German university and industrial labs, with partners in the US (Stanford) and Japan (Tokyo). In this project, a basic decision was made for a hybrid architecture of deep and shallow processing. Its success means that deep linguistic methods offers new opportunities for further developments, in particular systems working with HPSG.

The following paper by Hans Uszkoreit demonstrates from several points of view how under these circumstances deep linguistic processing could be successfully integrated and further developed. In a certain sense, this article represents a position opposite to recent trends in language technology, which have tended more towards statistically based systems and large-scale training methods. Uszkoreit’s approach is to combine deep processing methods with shallow methods, not in such a way that the latter assists the former, but to “let deep processing assist a shallow processing IE system”. This application field has been chosen, since, as Uszkoreit points out, the particular demands of these systems often require a deep grammatical analysis in order to assign the “detected entities to the appropriate argument slots”.

Given this general background, the first part of the paper outlines the reality of an international collaboration between natural language labs in
different countries, in the field of grammar optimization and parser evaluation using HPSG-based parsing.

The second part goes more into the details of the preconditions and the architecture of applicable systems. The author first discusses what robustness and coverage mean for practical natural language applications, and then he demonstrates—using the WHITEBOARD approach, which has been elaborated at his own lab in Saarbrücken—how such a hybrid architecture of shallow and deep linguistic processing can be realized in the practical environment of IE.

Uszkoreit’s paper as a whole shows not only where the chances of deep linguistic processing in the context of the modern language industry are, but it is also a perfect demonstration how international coöperation may function and bear good results. This is also the author’s message concerning the future of Computational linguistics: IE “will become the leading research paradigm in Computational Linguistics” and “massively collaborative research” must be the new paradigm for research and development. For this a suitable infrastructure must be available by which common usable resources like annotated (multilingual) corpora, ontologies, tools, etc. can be provided.
New Chances for Deep Linguistic Processing*

Hans Uszkoreit

Recent developments in deep linguistic processing give rise to renewed optimism concerning the practical applicability of advanced grammatical analysis. For HPSG parsing a breakthrough in efficiency was achieved through a new form of international collaboration that led to improved combinations of methods in a rather systematic way. Yet efficiency is just one of the obstacles to the utilization of deep processing in real-life applications. A novel approach to the exploitation of deep parsing offers a strategy to compensate for deficiency in coverage and robustness. Through a combination of

* The project Verbmobil was funded by the Bundesministerium für Bildung und Forschung under grant 01 IV 701. The project WHITEBOARD is funded by the Bundesministerium für Bildung und Forschung under grant 01 IW 002. The project Perform is funded by the Deutsche Forschungsgemeinschaft as part of the SFB 387 “Resource-Adaptive Cognitive Processes”. My special thanks go (in alphabetical order) to all participants of our international collaboration, mainly to Tania Avgustinova (Saarland U.), Liviu Ciortuz (U. of York), Ann Copestake (Cambridge U.), Dan Flickinger (Stanford U.), Emily M. Bender (Stanford U.), Ulrich Callmeier (Saarland U.), John Carroll (U. of Essex), Ann Copestake (Cambridge U.), Andreas Eisele (DFKI), Frederik Fouvy (Saarland U.), Anette Frank (DFKI), Bernd Kiefer (DFKI), Valia Kordoni (Saarland U.), Hans-Ulrich Krieger (DFKI), Chris Manning (Stanford U.), Takaki Makino (Tokyo U.), Rob Malouf (San Diego State U.), Mitoishi Minoru (Tokyo U.), Takashi Ninomiya (Tokyo U.), Stephan Oepen (Stanford U.), Detlef Prescher (DFKI), Ivan A. Sag (Stanford U.), Melanie Siegel (DFKI), Kenjiro Taura (Tokyo U.), Kentaro Torisawa (Tokyo U.), and Jun-Ichi Tsujii (Tokyo U.).

I am especially grateful to Stephan Oepen, who created the engineering platform tsdb and served as the main driver of the collaboration as well as to the members of the DFKI project WHITEBOARD Berthold Crysmann, Anette Frank, Bernd Kiefer, Stephan Müller, Günter Neumann, Jakub Piskorski, Ulrich Schäfer, Melanie Siegel, Feiyu Xu, Markus Becker and Hans-Ulrich Krieger for their realization of the hybrid architecture.
deep and shallow processing, the robustness of shallow processing is preserved for information extraction applications that exploit parsing methods of different depth. The underlying strategy and the realized architecture also provide a basis for distributed collective research on novel combinations of processing methods including applications of hyper learning.

Introduction

The grammatical analysis with linguistically designed grammars has always been a central topic of investigation in theoretical computational linguistics. However, deep grammatical analysis has so far played a negligible role in the development of language technology applications. Deep parsers have lacked the efficiency and robustness required for real life applications. Building up coverage has been slow and costly. The few linguistic grammars that truly exhibit large coverage have resulted in a constant battle with extensive ambiguity. In my talk I shall present recent developments that provide new challenges and opportunities for linguistic methods. Increased efficiency of deep parsing and the embedding of a selective deep analysis into a robust shallow regime for information extraction offer ways to employ deep parsing in an environment where it can improve results without ruining robustness. Such a selective utilization was realized by our integration of HPSG parsing into a hybrid IE system, comprised of statistical and FST components. I shall argue that such an approach provides a promising direction for gradual, measurably controlled progress in the development of deep grammars and lexica.

The most serious shortcoming of today’s language technology is the lack of methods that get at the real contents of text and speech—systems approximating language understanding. Therefore, the central problem in computational linguistics has always been the realization of deep linguistic processing involving an accurate mapping between written and spoken utterances on the one hand and useful semantic representations on the other.

Modern linguistics has been able to provide theories and formalisms for the specification of grammars that express this mapping in a declarative and transparent way. Computational linguistics has contributed elaborate
platforms and tools for grammar development. A few large-scale grammars have been designed exhibiting sufficient accuracy and coverage for real application tasks. However, these encouraging developments were seriously hampered by a lack of methods for language analysis that fulfill the minimal requirements in efficiency, robustness, and specificity. This simply means that all systems working with these grammars have been too slow and too brittle for real applications.

Furthermore, they have not been able to manage the vast ambiguity in natural language; i.e., they could not select among large numbers of linguistically correct analyses.

Still the most critical problem has been time and space efficiency. If an NLP system cannot process everyday sentences in a reasonable amount of time on a normal PC, it is not suited for most applications. Moreover, there was no chance to improve coverage and solve the issues of robustness and specificity if researchers had to wait for hours for a sentence to process. The performance problem was so severe that many promising lines of research ended without yielding any applicable results. The lack of efficiency became a major obstacle for several large endeavors involving extensive grammar development such as the IBM LILOG project with HPSG parsing on the LILOG STUF Parser in Prolog (Herzog and Rollinger 1991), the Pargram conducted by XEROX with LFG parsing on the old Interlisp XLE platform (Butt et al. 1999), and the EU project LS-GRAM with HPSG parsing on the Prolog ALEP platform (Schmidt et al. 1996).

The situation seemed hopeless since all laboriously achieved gains in efficiency were almost immediately offset by efficiency losses due to increases in coverage or sophistication of the grammars.

**1. The Battle for Efficiency**

When Verbmobil (Wahlster 2000), the largest research project ever conducted in speech technology, adopted deep linguistic processing on the basis of HPSG as one of the central methods for speech analysis in real-time translation of spoken face-to-face dialogues, this decision faced
considerable criticism from both inside and outside the consortium. Why should the slowest and most complex processing method be employed in a system that strives for real-time processing? The decision could only be maintained because in the hybrid Verbmobil architecture, deep processing was just one of several processing methods and could therefore always be preempted by an analysis from a faster processing module. We shall return to this point.

During the first phase of the project from 1993-1996, a team at IBM Germany in Heidelberg had been responsible for deep parsing. They tried hard to overcome the efficiency problem by combining statistical language modeling with their HPSG parsing scheme. In 1996 at the end of Phase One of Verbmobil, we were still far from getting even close to the performance requirements for the final Verbmobil prototype. When my lab was entrusted with the responsibility for deep linguistic analysis in the second halftime, it was not clear whether we would be able to deliver a component that would not always time-out against the faster shallow processing modules. In the beginning we used the existing parser of our HPSG development platform PAGE that had been implemented in the Project DISCO (Uszkoreit et al. 1994). The HPSG grammar writers at Stanford University and DFKI had already worked with the PAGE development system for several years.

Interestingly, it was the immense pressure for efficiency in this speech application project that caused two members of the consortium, DFKI LT-Lab and CSLI at Stanford University, to join forces in developing completely new strategies for performance research in deep linguistic processing. Also included in the collaboration was the project PERFORM that I am conducting at Saarland University.

PERFORM contributed most of the methodology and evaluation technology. The methodological basis of the effort was the systematic, sophisticated and very detailed measurement of all relevant performance data for each version of a parser. For each parser and each parsed sentence a database record was created containing all data on numbers and sizes of successful and unsuccessful, complete and partial results, and on overall time and space consumption. The controlled utilization of the same grammars and the same test corpora was a precondition for obtaining comparable results.
The test corpora had to be annotated by the correct results and linked to previous performance data.

The novel engineering platform tsdb, developed by Stephan Oepen (Oepen 2002), produces detailed diagnostic reports and complex multidimensional comparisons between alternative systems.

Later the Natural Language Processing Lab at Tokyo University joined the collaboration. The groups agreed on shared test suites and a reference grammar for comparative evaluation. The Stanford LinGO grammar for English developed at CSLI by Flickinger (Flickinger 2000) and others was selected, although extensive tests were also run with large HPSG grammars for German (Müller & Kasper 2000) and Japanese (Siegel 2000).

At this time the LinGO grammar consisted of about 8,000 types specified in 100,000 lines of source code. The lexicon contained some 6,000 lexemes. An average feature structure in processing consists of more than 300 nodes.

The test suites, systematically composed of diagnostically relevant
generic examples and representative examples for the Verbmobil domain, had a combined size of more than 3,000 test items, in most cases sentences. They also included negative examples.

Through the tsdb platform, five systems were evaluated in this phase of the joint effort. The insights gained by the evaluation soon led to successful efforts in combining the most promising methods.

As coördinator of the collaboration, I organized: (1) a workshop in Berlin in March 1999 where initial results and new goals were discussed; and (2) together with Jun-ichi Tsujii, a symposium at Schloß Dagstuhl in October of 1999 where the achievements were compared with progress in other research camps, e.g., parsing with LFG, TAG, and categorial grammar.

A number of new methods were developed by the three sites and tested in many combinations. In the end, it was a synthesis of methods reached by a true scientific and technological cross-fertilization process that brought about the breakthrough in the battle for efficiency. A first joint publication on such combination was Kiefer et al. (1999).

All participating parsers improved in efficiency. Contrary to the predictions of our partners from the speech community, the Verbmobil HPSG parser did not end up at the tail end of the performance scale. The average distribution of total run-time among the Verbmobil processing stages demonstrates that deep linguistic processing had successfully overcome the initial efficiency problems: speech recognition 38%, prosody processing 17%, **syntactic/semantic analysis by HPSG 25%**, semantic interpretation and dialogue processing 14%, transfer 3% und generation 3% (Wahlster 1997). It should be pointed out here that the short processing time for the transfer stage became possible because the results of semantic processing permit a reduction of transfer to mere lookup in most cases. The role of HPSG processing in Verbmobil and the reasons behind the choice of the grammatical framework are discussed in Uszkoreit, Flickinger, Kasper, & Sag (2000). The Verbmobil parser is described in Kiefer et al. (2000).

The fastest parser that came out of the joint effort was the system PET developed by Ulrich Callmeier (2000) in the project PERFORM. PET could not be integrated anymore in the final Verbmobil demonstrator. It is freely available under an open source for academic and commercial use.
The preferred grammar engineering platform is Anne Copestake’s completely reworked LKB system, which is also freely available under an open source license (Copestake 2002).

The methods that were combined in the most successful systems involve all components of parsing starting with optimizations of the grammar and changes to the formalism, but ranging all the way to improved algorithms for feature unification and a better representation of parse forests. The list of the techniques that contributed most profoundly to the efficiency gains include:

- the elimination of general disjunction in feature structures;
- methods for optimized structure sharing;
- improved rule filtering techniques;
- quick check computation of relevant parts of feature structures before full unification;
- efficient subsumption checking for ambiguity packing.

The overall run-time efficiency gain accomplished after three years was a factor of 2,000 (Oepen et al. 2000). Space consumption was also reduced by more than an order of magnitude. Time measurements were performed on comparable hardware. This means that the gains achieved on the software side was also complemented by well-known progress in hardware efficiency. Sentences can now be analyzed in fractions of the time needed for real-time speech applications. Larger texts can be analyzed in a few seconds. The fastest parser can now be run on a standard PC. Thus HPSG parsing now meets the speed and working memory requirements for a wide range of applications. Papers on major results of the collaboration are compiled in Flickinger et al. (2000) and Oepen et al. (2002).

This breakthrough led to increased interest in HPSG processing in several areas of research and in industry. Many theoreticians and practitioners of grammar have expressed their interest in using the software for grammar development. The first industrial applications have been developed.
2. Robustness and Coverage

2.1 The Verbmobil Approach

Now we turn to the problem of robustness and coverage. Even if efficiency had been the main stumbling block on the road to real-life natural language processing applications that work with highly accurate deep linguistic grammars, the real and potentially much more challenging problem is the design of systems able to properly deal with the rich variation of human language input in realistic application situations. We want systems to master the proper subset of a natural language that seems to be needed for a specific application. Moreover, we want a system that does not fall over if faced with some input outside this imagined subset.

It has always been tricky to distinguish issues of missing coverage from those connected to a lack of robustness. If we consider the definition of robustness from the IEEE standard glossary of software engineering terminology (IEEE 1990) then robustness is “the degree to which a system or component can function correctly in the presence of invalid input or stressful environmental conditions.” At first glance this definition seems to apply quite straightforwardly to language processing applications. Invalid input may be ungrammatical utterances or grammatical utterances outside the appropriate sub-language for the application. Stressful environmental conditions may result from background noise or distortions in the speech signal or from hundreds of queries simultaneously sent to a question-answering system.

However, the problem is how to define the appropriate sub-language. If we want to account in a spoken language input system for the wide variety of idiolects, dialects, accents, and sociolects of speakers, and if a written input system is supposed to properly deal with very large volumes of unseen texts produced by a large number of authors, then the sub-language should be defined rather large. Such a demand poses a serious problem to most contemporary linguistically sophisticated competence grammars. It would require the grammar to analyze the union of many widely overlapping languages instead of relying on an additional mechanism mapping any sentence within the sub-language but outside
some standard variant of the language into the most closely corresponding sentence within the standard language. The more permissive a large-scale grammar becomes by increasing coverage within one variant of the language, by going beyond this variant, or by even accounting for performance errors and distorted input, the more the system will be confronted with the problem of excessive local and global ambiguity. We shall return to this problem later.

At this point we can state that the differentiation between problems of missing robustness and those of missing coverage with respect to some specific system depends on the specifications underlying its system design. If the implemented grammar itself is seen as the specification of the relevant sub-language, the difference becomes useless. All inputs that cannot be analyzed by this grammar automatically fall under the category of deficiencies in robustness.

In the project VERBMOBIL, coverage was restricted along two dimensions. One dimension was the size of the lexicon. Since the performance of the speaker-independent recognition of continuous speech is strongly dependent on a restricted vocabulary, the size of the lexicon was first limited to 2,500, later to 10,000 words. The second dimension follows from the goal to realize a plausible application scenario for such a restricted vocabulary and to deeply model a domain for the integration of dialogue and knowledge processing with the analysis and generation of spoken utterances. Therefore the domain was restricted to dialogues on the scheduling of appointments. This restriction limited lexical ambiguity and excluded certain types of constructions.

In VERBMOBIL two main approaches were employed for realizing robustness. One method was based on deep grammatical processing with HPSG and utilized minimally recursive under-specified semantics. Whenever the HPSG parser could not produce an analysis covering the whole sentence, the resulting chart was sent to a processing component for robust semantic interpretation that attempted to exploit the intermediate results in the chart and knowledge about the domain and the sub-language to hypothesize at least a partial semantic representation of the input utterance (Pinkal et al. 2000).

The second approach is reflected in the fundamental design of
VERBMOBIL. The architecture of the system is based on the concurrent application of several different processing methods and the final selection among alternative analyses and even translations. In case deep grammatical analysis and translation based on a semantic representation could not provide a solution, i.e. a translation in a target language (or failure to deliver a result before a preset timeout) a translation could be selected from one of the other three processing strands; translation based on dialogue analysis and dialogue acts, a statistical translation or a case-based translation. In this way translations could be produced for virtually all inputs including even extreme cases of ill-formed utterances. In about 74% of all cases the translations were correct or approximately correct. A scientifically challenging issue is the optimal selection among several results proposed by different methods or components. Within the time frame of the VERBMOBIL project very little research could be dedicated to this exciting theme. The derivation of stable confidence measures and the search for statistically reliable quality indicators could have become the objective of a follow-up project. In the realized and demonstrated VERBMOBIL system the result of deep analysis and translation was always preferred over the results of shallow processing if it could be provided.

To sum up the VERBMOBIL experience with respect to progress in deep grammatical processing, we can report breakthroughs in the areas of efficiency and robustness. Progress in run-time efficiency was achieved by the careful and systematic combination of several methods ranging from the optimization of the grammar via generic program optimization to a number of novel algorithmic solutions in parsing with unification grammars. The required robustness was achieved by combining deep processing methods with several shallow analysis and translation techniques that worked in parallel and almost independently of each other. In addition, promising first results could be achieved in making deep grammatical processing more robust by exploiting the incomplete chart and in under-specified semantic representation for the construction of partial interpretations with the help of domain-dependent heuristics.
2.2 The WHITEBOARD Approach

In this section a new approach for utilizing deep grammatical processing in real-life applications is described. The approach is based on an integrated architecture for a variety of shallow and deep processing components.

The combination of deep and shallow methods for improving the performance of analysis is not a new idea. Several approaches have been proposed and demonstrated for augmenting a deep parser with shallow methods for preprocessing such as POS tagging and the analysis of complex names or other fixed multiword expressions. Another way of combining the virtues of deep and shallow methods is the employment of shallow analysis tools such as statistical phrase or chunk parsers for selecting among alternative readings.

The strategy I proposed as the basis of our WHITEBOARD project, differs from these solutions. Instead of letting shallow processing assist the deep parser, we let deep processing assist a shallow processing IE system.¹ This decision is in parts based on our experience with information extraction.

Let me start with the demands of the application. In IE recall and precision are constantly improved through the design of more sophisticated rule systems and through the application of statistical or symbolic machine learning techniques. However, in the detection of relations, IE systems are often confronted with texts in which the exact assignment of detected entities to the appropriate argument slots of hypothesized relations seems to require a deep grammatical analysis—frequently even with some semantic filtering. The combinatorics behind such constructions makes each type rather rare in data collections, too rare at least for contemporary learning approaches.

Assume IE in medical and pharmaceutical texts by which we try to find relevant relationships between medicines and medical conditions. A medicine may be indicated to treat a condition, however a condition may also be a side effect or a counter indication. Since both conditions and side

¹ Actually this master-slave relation is complemented by several ways exploiting shallow processing for the deep parser.
effects can occur very close to the reference of the medicine.

A similar problem arises in personnel news such as management succession reports where references to several persons can occur in close proximity to the reference of a management function. Consider the following pair of examples from German:

(i) Peter Mischke zufolge wurde Dietmar Hopp gebeten den Vertrieb zu übernehmen  
   According to Peter Mischke, Dietmar Hopp was asked to take over the sales department.

(ii) Peter Mischke wurde von Dietmar Hopp gebeten den Vertrieb zu übernehmen  
   Peter Mischke was asked by Dietmar Hopp to take over the sales department.

In order to correctly fill the slots, the system needs to recognize which of the two persons mentioned in these sentences is the successor in the position. To this end, the subjects of gebeten and übernehmen need to be determined. This cannot be done without an analysis taking into account the interaction of word order, control and passivization. I would claim that a pseudogrammar which can assign the correct slot fillers in these cases will come so close to the functionality of a deep grammar that we might be better off if we employ an already existing deep grammar. Here we seem to back at the performance arguments because in typical IE applications thousands of documents need to be analyzed within seconds or minutes.

However, if a reasonably efficient deep parser is consulted by the IE system in just the cases where the IE system cannot decide among several slot assignments and if the IE system can accept or disregard the results of
the deep parser, neither efficiency nor coverage or robustness will be severed by the deep parser. This is the avenue we proceed.

The next logical step is the consequent specialization of deep grammars with respect to certain IE domains and tasks. Our Verbmobil grammar has already been a mix of generic and specialized components, so we are now specializing the HPSG grammar for IE tasks.

The WHITEBOARD\textsuperscript{2} architecture that has been implemented is very general. It supports the combinations of processing methods currently under investigation in the WHITEBOARD project but also countless others. The underlying idea is as simple as it is powerful: several shallow and deep processing components annotate an input text with XML markup. In theory, the components could perform their annotation in any order and even be called several times. They can look at and exploit previously annotated markup. Many types of linguistic information cannot be merged into a single XML markup structure. Among them are identifiers for multi-part objects such as discontinuous constituents or co-references and annotations involving conflicting bracketings, such as the markup of ambiguous readings or alternative structuring hypotheses. Therefore we employ a multitude of annotation layers. Hyperlinks and information on covered spans in the text connect the different layers. Some layers may introduce (additional) auxiliary layers that may contain representations that are not coded as XML markup. In this way the layer encoding the labeled bracketing produced by the HPSG parser can connect to an auxiliary layer containing feature structures that do not have to be merged into the annotated text. These auxiliary layers may also be used for storing analog data such as sounds or pictures. The core part of the architecture is the WHITEBOARD annotation machine (WHAM). The WHAM can be called by an application that needs to know about the WHAM’s generic OOP interface and about the components that may be requested. The architecture of the WHAM is presented in Figure 2.

\textsuperscript{2} We discovered that the term “Whiteboard architecture” has been used before. In computational linguistics it was proposed at the 1994 Coling by Christian Boitet and Mark Seligman (Boitet & Seligmann 1994) to refer to a powerful extension of blackboard architectures.
The application passes an input text to the WHAM together with specifications about the requested components, the sequence of their activation, and the nature of the requested result. The WHAM interface can now call the specified components in the requested order. It consists of iterators that walk through the different annotation levels and reference and seek operators that jump to corresponding annotations on different levels. These can return, for instance, all part-of-speech label tokens of the current sentence or the next-named entity of a certain type starting from the current token position or the next temporal prepositional phrase. Other components of the OOP are accessor methods that return the information encoded in the chart as well as general methods that enable navigation through the type system and the feature structures of the deep processing components. The WHAM works with two mechanisms for the representation of results: (1) an external offline representation in which the input text is enriched with XML markup in an additive monotonic fashion; and (2) an internal
online annotation chart with multiple levels, in which the more abstract structural information (bracketing hypotheses even involving discontinuous constituents) and multi-dimensional characterization of complex objects (such as syntactic and semantic feature structures) are encoded in a uniform linguistically motivated form.

Each interface to a specific processing component is realized as a subclass of the generic WHAM interface. If a new component has to be integrated, a new subclass has to be defined and its component-specific representations have to be specified by a new DTD for the XML markup or a set of transformation rules for the chart or both. Shallow processing components that have been integrated include our statistical part-of-speech tagger and phrase parser TNT by Thorsten Brants (1999) that is based on cascaded hidden Markov models, the chunk parser CHUNKIE by Wojciech Skut (1999) that extends the functionality of TNT, and our shallow processing platform SPPC by Jakub Piskorski and Günter Neumann (Piskorski & Neumann 2000), a rule-based system that employs cascades of weighted finite-state machines for tokenization, morphological analysis, part-of-speech filtering, named entity recognition, and the detection of chunk, sub-clause, and sentence boundaries. In the current setup, deep processing is represented by the HPSG parser PET. Ulrich Callmeier and Ulrich Schäfer extended the PET system to allow—instead of single words—multiple-word input items that may even be overlapping and ambiguous forming word hypothesis graphs. Another extension permits the dynamic creation of atomic type symbols so that arbitrary symbols can be added to feature structures for flexible interfaces to external components such as morphology tokenization or named-entity recognition.

In addition to the central shallow and deep processing components, a further knowledge source was added exploiting the same interface mechanisms to the WHAM. In order to obtain more fine-grained semantic information on the lexical level, especially sortal information, the German version of WordNet, GermaNet graphs was utilized. Compared to GermaNet the current HPSG lexicon is quite small. Its semantic classification is rather sophisticated in the Verbmobil domain, but lacking fine-grained sortal information in most other subject areas. In order to draw benefits from the integration of the rich information source GermaNet, a mapping from the
sortal categories of GermaNet synsets to the HPSG sort system had to be obtained. To this end a supervised learning algorithm was employed that was trained on those words annotated with semantic sorts in the HPSG lexicon and with synsets in GermaNet. The learning algorithm then computes such mappings for words that do not yet have corresponding sort information assigned in the HPSG lexicon. It does this by calculating a relevance measure for each possible association. The hypothesized mapping was evaluated in the domain of business news. From a corpus of business news, 4,664 nouns were extracted that were not represented in the HPSG lexicon; 2,312 of these unknown words were contained in GermaNet. GermaNet assigned 2,811 senses to these words. These word senses were then automatically mapped to HPSG sorts. An evaluation of these sort assignments by human judges yielded a rather promising result. In 76.52% of all assignments the mapping suggested by the algorithm (i.e., with the highest relevance measure) turned out to be correct. In 27% of all cases, the correct sort had received the second or third highest relevance assignment.

3. Conclusion and Future Research

Although the number of centers conducting projects in deep grammatical processing has decreased and therefore also the visibility of this line of research, considerable progress has been achieved.

For time and space limitations and for other obvious reasons, I have concentrated here on developments in processing with HPSG. Other breakthroughs have recently been accomplished in parsing with Lexical Functional Grammar, Tree-Adjoining-Grammar, Categorial Grammar, and Dependency Grammar. Many of those improvements are based on a combination of statistical methods and grammatical processing, on better and larger grammars, as well as on clever and careful engineering. The current XLE system for LFG, for instance, exhibits an impressive efficiency and integrates a powerful machinery for dealing with massive ambiguity (King et al. 2000).

One of the goals should be the systematic comparison of deep parsers
across linguistic, formal, and technological frameworks. For this purpose the evaluation against the Penn Treebank (Marcus et al. 1993) is the only game in town. Such evaluation in the spirit of Parseval that has revolutionized the field of parsing research is much more suited for shallow- or medium-depth parsers (such as PCFG system) than for parsers producing a sophisticated semantic representation. Also for evaluating the value of partial results with respect to applicability, the current evaluation setup is of limited value.

Nevertheless it is a great achievement of our colleagues from the LFG and TAG communities that they made their parsers ready for the evaluation against the WSJ corpus and were able to achieve impressive and encouraging results (e.g., Riezler et al. 2002).

In a workshop at LREC 2002 called “Beyond Parseval” (Carroll et al. 2002), the participants representing several distinct schools in deep linguistic parsing almost unanimously acknowledged the shortcomings of the Penn Treebank annotation. In the final discussion of the workshop they agreed with a proposal by Briscoe et al. (2002) showing an annotation scheme based on grammatical relations that is better suited for comparing results of different grammar frameworks as well as for robust partial parsing and multilingual annotation.

Although we can point at considerable progress in deep linguistic progressing, hard problems remain to be solved. Our international collaboration has entered into a new phase. The partnership has been broadened by including additional groups, among them Cambridge University, University of Sussex, and University of Edinburgh. In the current phase of collaboration, methods for improving robustness, coverage, and disambiguation constitute the main objectives of theoretical research. A focus lies on statistical methods for extending grammar and lexicon and for learning disambiguation preferences. The group at Saarland University has achieved progress in exploiting the NEGRA/TIGER treebank of German, the Stanford group is building the Redwoods Treebank (Oepen et al. 2002), a dedicated HPSG treebank.

In addition, the combination of deep and shallow processing for IE and other applications will be further pursued by the consortium. New results on robust parsing and partial interpretation with robust minimal recursion semantics open new application domains.
Some partners (DFKI, Cambridge University, University of Sussex, University of Trondheim, and Stanford University) have obtained funding through a recently approved EU-sponsored project named Deep Thought, in which they will work together with three innovative industrial companies (CELI in Torino; Edify in Edinburgh, and XtraMind in Saarbrücken) on the practical exploitation of combined deep and shallow processing starting from the WHITEBOARD architecture.

The third area of collaboration is multilingual grammar development. Continuing a theme of our Verbmobil research, linguistic, formal, and practical issues in sharing information among different grammars are investigated (Müller & Kasper 2000, Siegel 2000). Among the practical goals of this research are shorter development times for new grammars, re-usability of semantic resources and uniform multilingual applications. Emily Bender et al. (2002) and others have already developed a so-called matrix grammar containing the shared components of the English, German, and Japanese grammars that is used to get a warm-start on grammars for additional languages.

The work plan for the current phase of the collaboration can be viewed at:

http://www.coli.uni-sb.de/~hansu/Collaboration.html

4. Outlook

Only very few of us still expect to see one day the automatic understanding of natural language texts as a result of a sudden breakthrough in research. Certainly I do not foresee it.

Instead, I would predict that information extraction will become the leading research paradigm in computational linguistics. In our discipline, this paradigm subsumes all applications that can recognize relevant types of information in human language texts. These can be topics, the most important sentences of a text, named entities, binary relations, event templates, or complex relational concepts. On one side of the scale we have text filtering, as the extraction of a category in a binary classification, on the other extreme, we have the extraction of complex relational objects,
representing the meaning of an entire discourse.

In this way, information extraction spans the continuum between the most modest language technology applications and true language understanding.

If we want to convince the research community funding agencies and industrial clients of the value of sophisticated grammatical analysis, it will not suffice anymore to insist on the necessity of deep processing for reaching the ambitious long-term goal. We shall have to demonstrate that deep processing can contribute to progress on the long way of small and controlled steps that lie ahead of us. Thus the advancement of linguistic processing needs to become subject to similar measurement and evaluation as progress in shallow processing. This requires improved methods for combining and comparing alternative approaches and techniques. Common tasks in today’s technology evaluations and shared data collections mark the beginning of a new era of massively collaborative research.

From all we learned during the past decades of research, human language is not less complex than the subjects of chemistry, genetics, or geology. Yet the organization of contemporary research in language and language processing stands in stark contradiction to this insight. Each individual center builds its own software systems for the processing of one or more languages. Sometimes systems and tools from other centers are adopted or re-implemented, but there is no infrastructure or common tasks that permits the comparison, exchange, and combination of methods and systems. An international multi-site collaboration within one school of research is already very difficult to organize and maintain. In the future we need ways of comparing and combining results across theoretical frameworks and research communities.

Heterogeneous architectures such as the WHITEBOARD can provide means for combining methods and components. A new R&D paradigm of collective research can only be realized, however, if the necessary infrastructure is available. Without common corpora, annotation schemes, and tasks, the envisaged combination of methods cannot be achieved. We need large multilingual corpora automatically annotated by a variety of processing tools. These corpora also constitute common tasks. For each participating language, several shallow and deep processing systems
should enrich the text by layers of annotation:

- categorizers
- segmentizers/tokenizers
- statistical and rule-based POS taggers
- morphological analysers
- full text indexing
- shallow chunk and phrase parsers
- wordnets and other thesauri
- information extraction systems NEE, relation extraction, template filling
- statistical parsers
- deep linguistic parsers such as LFG, HPSG, DG and CG parsers
- systems that determine temporal or discourse relations
- summarization systems

Some portions of the texts should be hand-corrected for obtaining measures of reliability and gold standards for evaluation. The multi-layered annotations can be encoded in XML. For maintaining a uniform correspondence between annotation and text spans and for allowing the extension to speech documents, an approach of annotation graphs (Broeder et al. 2000) will have to be exploited.

A number of the components utilized for annotation will be based on machine-learning techniques. The envisaged corpora could thus serve as an extremely valuable resource for higher-order learning, so-called hyper-learning.

The annotation of complete and partial semantic analyses poses a special challenge since compatibility between the results of deep processing and information extraction needs to be established. Robust minimal recursion semantics or other semantic formalisms accommodating underspecified and partial representations will be needed for encoding semantic information. For annotating the corpora by meta-information the OIL/DAML (Connolly et al. 2001) format may be utilized in order to exploit existing connections with the Semantic Web community and to improve the value of the corpus.
In my opinion, the question is not whether such resources will be created, but rather when the work is going to start. A consortium of research centers from several European countries has already decided to initiate actions into the outlined direction.

References


Biographic Sketch and Research Outline of Author

Hans Uszkoreit

Hans Uszkoreit is Professor of Computational Linguistics at Saarland University. At the same time he serves as Scientific Director at the German Research Center for Artificial Intelligence (DFKI) where he heads the DFKI Language Technology Lab. By courtesy appointment he is also Professor of the Computer Science Department.

Uszkoreit studied Linguistics and Computer Science at the Technical University of Berlin and the University of Texas at Austin. During his time in Austin he also worked as a research associate in a large machine translation project at the Linguistics Research Center. In 1984 Uszkoreit received his Ph.D. in linguistics from the University of Texas. From 1982 until 1986, he worked as a computer scientist at the Artificial Intelligence Center of SRI International in Menlo Park, Ca. During this time he was also affiliated with the Center for the Study of Language and Information at Stanford University as a senior researcher and later as a project leader. In 1986 he spent six months in Stuttgart on an IBM Research Fellowship at the Science Division of IBM Germany. In December 1986 he returned to Stuttgart to work for IBM Germany as a project leader in the project LILOG (Linguistic and Logical Methods for the Understanding of German Texts). During this time, he also taught at the University of Stuttgart.

In 1988 Uszkoreit was appointed to a newly created chair of Computational Linguistics at Saarland University and started the Department of Computational Linguistics and Phonetics. In 1990 he became the head of the newly founded Language Technology Lab at DFKI. He is co-founder and principal investigator of the Special Collaborative Research Division (SFB 378) “Resource-Adaptive Cognitive Processes” of the DFG (German Science Foundation). He is also co-founder and professor of the “European Postgraduate Program Language Technology and Cognitive Systems”, a joint Ph.D. program with the University of Edinburgh.

Uszkoreit is President of the European Association for Logic, Language and Information, Permanent Member of the International Committee of Computational Linguistics (ICCL), Member of the European Academy of
Sciences, of the Executive Board of the European Network of Language and Speech, the Board of the European Chapter of the Association for Computational Linguistics and serves on several international editorial and advisory boards. He is co-founder and Board Member of XtraMind Technologies GmbH, Saarbrücken and acrolinx gmbh, Berlin, and he serves on the Board of Directors of GlobalWare AG, Eisenach.

His current research interests are computer models of natural language understanding and production, advanced applications of language and knowledge technologies, cognitive foundations of language and knowledge, grammar formalisms and their implementation, syntax and semantics of natural language and the grammar of German.

List of Publications:
http://www.coli.uni-sb.de/%7Ehansu/publ.html
Introduction to
The Roles of Natural Language and XML in the
Semantic Web

Winfried Lenders

Tim Berners-Lee described the main goal of a future Semantic Web as enabling intelligent questions instead of conventional string matching and obtaining intelligent answers that satisfy the user with respect to both precision and recall. From a Computational Linguist’s point of view, this idea includes three aspects where linguistic methods can help make the web more intelligent.

The first point concerns the simple fact that most of the knowledge distributed all over the world is recorded in a natural language. In order to process this knowledge in natural language form, the content must be described by a knowledge markup. Computational Linguistics has contributed significant research to support semantic processing and knowledge markup, e.g., concept description, semantics networks, ontologies, representation and description languages, and general language understanding systems. Furthermore, there is a long tradition in Computational Linguistics in the field of information retrieval and information extraction (e.g., Martin Kay & Karen Sparck Jones 1973, Linguistics and Information Science) and there is much ongoing work in the annotation of texts on different levels and the extraction of the relevant knowledge from documents. COLING conferences like the last COLING-2002 in Taipei are a good indicator of the efficiency of Computational Linguistics in this area.

The second aspect concerns the user’s site, and the human-computer interface more specifically. A future ‘intelligent’ semantic web has to operate with user questions that are not just key words computable as simple character strings. The user’s question must be dealt with as a bona fide natural language expression with its full range of meanings and nuances. Since both intelligent questions and intelligent answers normally
are formulated in natural language, they have to be manipulated in an adequate manner by language “understanding” programs. The processing of these expressions has to be done by future “intelligent” semantic web browsers that therefore have to provide not only a morphological and syntactical analysis, but in particular a conceptual description which is able to reduce the multitude of possible expressions to the underlying domain depending conceptual structure.

The third area where natural language processing is crucial for the Semantic Web involves the existing language barriers. As recent studies have shown, Marshal McLuhan’s vision in the 1960s of the future world as a Global Village is still utopian. On the one hand, the World Wide Web already is established as a worldwide communication medium, but on the other, there are still hundreds of different languages spoken in this ‘global village’. As has been shown by these studies, people surfing in the web normally stay in their particular language domain and only a limited number of individuals all over the world are able to communicate in English. In this situation a more intelligent web should also reflect these language problems. Computational Linguistics can help to break the existing language barriers, in particular between languages which are typologically different and/or which are not common to the majority of users as their lingua franca. Again there is a long tradition of Computational Linguistics in machine translation and there are already proposals to combine semantic or conceptual descriptions with interlingual facilities.

These three aspects pose important challenges for Computational Linguistics, and these challenges are the subject of the following contribution. It consists of a series of articles or extended statements dealing mainly with the first of the above-mentioned aspects. As a whole, they represent something like the minutes of the panel discussion at *The 2nd Workshop on NLP and XML (NLPXML-2002)* chaired by Nancy Ide, Laurent Romary, and Graham Wilcock, conducted as one of the 2002 post-COLING-workshops.
The Roles of Natural Language and XML in the Semantic Web

Graham Wilcock, Paul Buitelaar, Antonio Pareja-Lora, Barrett Bryant, Jimmy Lin, and Nancy Ide

1. Introduction (Graham Wilcock)

This chapter is based on the panel session “The Roles of Natural Language and XML in the Semantic Web” at the 2nd Workshop on NLP and XML (NLPXML-2002), held at COLING-2002 in Taipei. The workshop¹ covered a wide range of topics in which both NLP and XML were central: XML-based NLP tools, corpus annotation standards, XML in document generation, XML in spoken dialogue systems. Whereas most of the workshop papers presented tools and systems that had already been implemented, the main aim of the panel session was to look ahead to future development of the Semantic Web. The panel members, nevertheless, were researchers who already had practical experience in using Semantic Web technologies.

The chapter contains six sections, each written independently. I chaired the panel session and wrote this introduction. The next four sections were contributed by the four panel members: Paul Buitelaar presents an overall vision of the Semantic Web and its implementation technologies; Antonio Pareja-Lora describes experiences with XML and RDF; Barrett Bryant describes experiences with DAML+OIL; Jimmy Lin argues for a wider role for natural language. The final section by Nancy Ide forms both a conclusion and a link to the 3rd workshop² in the NLPXML series.

In this introduction I outline why we had a panel on “The Roles of Natural Language and XML in the Semantic Web”. The Semantic Web has become an important topic, but what do we mean by “the role of XML”

¹ http://www.ling.helsinki.fi/~gwilcock/NLPXML
² http://www.cs.vassar.edu/~ide/events/NLPXML3
and “the role of natural language”? What kinds of roles are required in the Semantic Web? The sections by the panel members will give more detailed descriptions of some of these roles.

One of the questions raised at the panel session, leading to extensive discussion, concerned whether it was possible or not to have a single all-embracing ontology as the basis for the Semantic Web. I shall summarize some of the points from this discussion.

1.1 The Roles of XML

XML[^3] (eXtensible Markup Language) can be used in two distinct ways, which are both important for the Semantic Web. First, XML is a language which can be used directly and by itself to represent information. Second, XML can be used to define more specialized languages. In fact, XML has been used as the basis for many different specialized languages. Here are just a few examples:

- MathML[^4] (Mathematics Markup Language) for mathematical formulæ
- WML[^5] (WAP Markup Language) for WAP mobile phones
- JSML[^6] (Java Speech Markup Language) for speech synthesizers
- SVG[^7] (Scalable Vector Graphics), a language for XML-based graphics
- XHTML[^8], a form of HTML that conforms to XML syntax.

In addition, XML Schema[^9] is a language for defining the permitted structures and data types of an XML document type. The schema definition, itself an XML document, is used for validating the contents of other XML...

[^3]: http://www.w3.org/XML/
[^4]: http://www.w3.org/Math/
[^5]: http://www.oasis-open.org/cover/wap-wml.html
[^7]: http://www.w3.org/TR/SVG/
[^8]: http://www.w3.org/TR/xhtml1/
[^9]: http://www.w3.org/XML/Schema
documents. An XML Schema can therefore be used to define a specialized language like XHTML. Unlike the earlier DTD form of document-type definition, an XML Schema itself conforms to XML syntax. A language that conforms to XML syntax (like XHTML, but unlike ordinary HTML) can be efficiently parsed, validated, and transformed by standard XML processors.

Among this multitude of XML-based languages, there are only a few languages (XML itself, XML Schema, RDF, and RDF Schema, DAML+OIL, OWL) that we are concerned with here because they are used particularly in the Semantic Web. In fact, when we talk about “the role of XML in the Semantic Web”, we are really using “XML” to refer to this small group of XML-based languages which plays a particular role in the Semantic Web. This group of Semantic Web languages and their relationships are described by Paul Buitelaar in §2.

1.2 The Roles of Natural Language

Natural language is, of course, ordinary language like English or Chinese. We use the term “natural language” in order to explicitly exclude artificial languages like Java or XML. When we talk about “the role of Natural Language in the Semantic Web” we are referring to the use of natural languages to play some particular role in the Semantic Web, as opposed to the use of the group of XML-based languages (XML, RDF, DAML+OIL) mentioned above.

Of course, the existing World Wide Web already contains enormous amounts of natural language in the texts of many millions of web pages. The problem is that it is difficult to find relevant information and extract it from this huge mass of texts. Most of the texts are marked up in HTML, but the markup mainly specifies the presentation format of the text, not its contents. By contrast, the vision of the Semantic Web is to mark up the semantic content of the information on the Web. The information whose semantic content needs to be marked up may be in many different forms. In addition to natural language texts, the information may be in table format, or in graphical images, audio, video, or other forms.
This leads to an interesting question. What form should the markup itself take? Should the markup language for the Semantic Web be XML? Or should it be one of the XML-based languages such as RDF or DAML+OIL, or some combination of these? Or would it be better to use natural language as the markup language? When we talk about “the Roles of XML and Natural Language in the Semantic Web”, we are referring to this question about what form the markup language should take, not merely to the existence of natural language texts in the Web.

1.3 Layers in the Semantic Web

Sometimes the Semantic Web is described in terms of a layer model. There are different versions of this model, such as the one by Tim Berners-Lee in Figure 1-1.

---

10 http://www.w3c.org/2000/Talks/1206-xml2k-tbl/slide10-0.html
The layers we are most interested in are the XML layer, the RDF layer, and the ontology layer. We are not concerned here with the underlying Unicode/URI layer, and we do not have much to say about the upper layers dealing with logic, proof, and trust.

To some extent, the different sections in this chapter focus on different layers. Paul Buitelaar introduces the overall architecture and describes the relationship between the different representation languages. Antonio Pareja-Lora deals mainly with the XML and RDF layers, while Barrett Bryant discusses DAML+OIL used in the ontology layer. As an alternative to these XML-based languages, Jimmy Lin puts forward an alternative proposal for a greater use of natural language.

1.4 Ontologies

The question at the panel session which generated the most discussion, for which we are indebted to Eduard Hovy, concerned the possibility or impossibility of having a single all-embracing ontology as the basis for the Semantic Web. Where does this question come from? I believe it arises partly from the overall vision of the Semantic Web, and partly from the arguments put forward for the benefits that the Semantic Web will bring. The Semantic Web is based on the World-Wide Web, which is a single all-embracing web, so the vision of the Semantic Web is that it will be a single all-embracing Semantic Web. Among the arguments put forward for the expected benefits of the Semantic Web is the vision that, once all the information in it is semantically annotated clearly and unambiguously in some agreed way, the Semantic Web will make smart searching and inferencing possible, and will eliminate all the confusion and contradictions in the present chaotic World-Wide Web. This annotation will be based on ontological classification, and it will only be fully effective once the Semantic Web reaches a certain “critical mass”.

This argument can clearly be challenged from a practical point of view, as the problems involved in building such a semantically annotated web on a global scale are enormous. However, the question raised at the panel session was not whether a single all-embracing ontology is difficult in practice, but whether it is even theoretically feasible. As an example,
Eduard Hovy brought up this question: “How many colors are there?” It is well known that different languages and cultures divide the spectrum in various different ways, and there is no possible way to arrive at one single globally agreed-upon list of colors. Similarly, different languages and cultures divide up everything else in the world in different ways, and there is no possibility of a global, fundamental agreement about classification of time and space, or entities and events—in short, it seems that there is no possibility of a single all-embracing ontology.

One way to approach this issue is to remember the origins of the World-Wide Web and the Internet. The World-Wide Web is a single all-embracing web, based on the Internet, which is a single all-embracing network. But the Internet grew from many small local networks based on local communications protocols, which were gradually joined together into regional and national networks, and in some cases into industry-specific or sector-specific networks by adopting agreed-upon protocol standards. Eventually these large networks joined together globally by internetworking based on the Internet Protocol. Similarly, the Semantic Web is growing from many applications based initially on small ontologies stored in local databases. These local ontologies will be combined with others to produce national or industry-specific or sector-specific ontologies, by adopting ontology standards and making the agreed ontologies more widely available. These wider ontologies will become part of the Semantic Web, which will have a sufficient critical mass to produce the benefits predicted. This growth process will be accelerated because the vital importance of standards has been understood, and local applications based on local ontologies are already using W3C standard ontology languages from the outset. The ideas and experiences of some who have pioneered the use of these technologies are presented in the following sections.

2. The Semantic Web: Vision and Implementation
   (Paul Buitelaar)

2.1 The Semantic Web Vision

The Semantic Web is a vision of a future version of the World-Wide Web, in which all web-based knowledge is encoded in an explicit, formal
way to allow for increasingly intelligent and therefore autonomous agents (Berners-Lee et al. 2001).

As illustrated in Figure 2-1, this entails the definition of formal, web-based ontologies to express the knowledge that is understood by humans as well as agents, and knowledge markup of (textual, multimedia) documents and databases using these ontologies. Knowledge markup is an elaboration of so-called metadata as currently defined and in use for a restricted set of applications, e.g. the Dublin Core\footnote{http://dublincore.org/} set of bibliographical metadata such as ‘title,’ ‘author’, etc. It is to be expected that over the next decade the knowledge structures of many more such applications will be formally encoded in web-based ontologies. Specifically in the context of e-business this will become apparent, as companies will need a common and explicit understanding of their products and services in order to allow for an automatic commercial exchange by artificial agents.
2.2 The Implementation of the Semantic Web

The Semantic Web is not a new technology in itself, but rather a vision of how existing technologies could be combined in establishing a more intelligent interaction with web-based information. These technologies consist of ‘core’ technologies in knowledge markup (i.e., markup languages, knowledge representation) and knowledge processing (i.e., intelligent agents, web services) and “enabling” technologies in knowledge organization (i.e., information science, machine learning) and knowledge access (i.e., database systems, language technology). In the next sections we shall discuss the relationship between these technologies and the semantic web in more detail.

2.2.1 Knowledge Markup and Processing

The definition of web-based knowledge representation languages is currently an active field of study, which has led to a number of proposals and emerging standards. Foremost among these are RDF Schema\(^\text{12}\) and DAML+OIL\(^\text{13}\) (recently redefined as OWL\(^\text{14}\)), the latter of which is defined on top of the other. Besides these, also XML Schema\(^\text{15}\) and Topic Maps\(^\text{16}\) are sometimes seen as knowledge representation languages.

In Figure 2-2, an overview is given of some important aspects of the XML/RDF family of knowledge markup languages—an overview based on Gil & Ratnaker (2001). From a syntactic point of view, RDF is written in XML, whereas DAML+OIL is written in RDF. On the semantic side, ontologies written in XML Schema, RDF Schema, or DAML+OIL are all based on the notion of a namespace, which defines the interpretation context of any XML, RDF, or DAML+OIL expression.

\(^\text{12}\) http://www.w3.org/TR/rdf-schema/
\(^\text{13}\) http://www.daml.org/2001/03/daml+oil-index
\(^\text{14}\) http://www.w3.org/TR/owl-guide/
\(^\text{15}\) http://www.w3.org/XML/Schema
\(^\text{16}\) http://www.topicmaps.org/xtm/1.0/
For instance, defining the following XML statement to be in the jobs namespace ensures that the job of John Smith as a systems analyst is interpreted exactly as defined in this particular ontology.

```xml
<xmls:jobs="http://www.jobs.org/daml+oil-jobs#">
<jobs:systems-analyst>John Smith</jobs:systems-analyst>, a senior systems analyst with IBM, concluded that...
</xmls:jobs>
```

In this way, a semantic web agent will be able to identify John Smith as a systems analyst and look up additional knowledge on this concept in the daml+oil-jobs ontology, which it can access in a distributed fashion at the indicated namespace address.

2.2.2 Knowledge Organization and Access

2.2.2.1 Information Science

Within information science there is a long tradition of defining classification schemas (thesauri) for the organization and retrieval of
available information in libraries and other archives. The already mentioned Dublin Core set of metadata results from this tradition. Also in future semantic web developments it is to be expected that tools and best practice techniques developed in information science will play a central role.

2.2.2.2 Machine Learning

Although classification and organization of knowledge is a highly intellectual and therefore human task, there is definitely a need for automatic support as the amount and complexity of the knowledge to be organized is rapidly growing. Machine learning approaches and tools are therefore needed to support the development, adaptation, and use of ontologies on the semantic web.

2.2.2.3 Database Systems

The efficient organization of and access to knowledge largely depends on the availability of powerful database systems that can handle the storage and retrieval of large amounts of semantic objects on the semantic web, represented in RDF or other markup languages. Semantic objects may range from simple facts like “John Smith:systems-analyst” to complex objects such as instantiations of multi-agent negotiation protocols in e-commerce.

2.2.2.4 Language Technology

As for humans, since the use of language is still the most natural way of expressing knowledge, there will remain a need to transform this ambiguous medium into structured knowledge to be accessed by agents and other web services on the semantic web. Therefore, language technology tools will be central in semantic web development in the following three areas: Knowledge Markup, Ontology Development, Intelligent Interfaces.
• **Knowledge Markup:** Turning the web into a semantic web implies widespread annotation of documents with ontology-based knowledge markup. Many of these documents consist of free text in different languages, which can only be marked up in an efficient way by use of automatic language technology tools.

• **Ontology Development:** Ontologies evolve rapidly over time and between different applications. Therefore, semi-automatic ontology learning that combines natural language processing (text mining, information extraction) with machine learning is essential for their efficient use.

• **Intelligent Interfaces:** Communication between humans and agents on the semantic web will be driven by natural language input, i.e., speech dialog. Obviously, language technology will be essential here in analyzing user responses and in generation of appropriate synthetic responses by artificial agents.

3. **Hybrid Web Page Annotation: RDF(S) Experiences**

(***Antonio Pareja-Lora*)

Following the guidelines of the Semantic Web initiative, as expounded throughout this paper, much research has already been carried out by AI researchers on the semantic annotation of web pages. However, these researchers have somehow neglected the decades of work and the results obtained in the field of *Corpus Linguistics* on corpus annotation, not only on the semantic level, but also on other linguistic levels. These other linguistic levels, whilst not being intrinsically semantic, can also add some semantic information and help a computer understand a text.

This section presents some preliminary results from the *ContentWeb* project\(^{17}\) (Aguado 2002) on how linguistic annotation can help computers

---

\(^{17}\) Supported by MCyT (Spanish Ministry of Science and Technology): “ContentWeb: Semantic Web Technologic Platform: Ontologies, Natural Language Analysis and E-Business”—TIC2001-2745. We would like to thank Guadalupe Aguado, Inmaculada Álvarez de Mon, Rosario Plaza and the LIA-Ontologies group for their collaboration in the research presented in this section.
understand the text contained in a document—a Semantic Web document, for example. Special efforts are being devoted in the aforementioned project to identifying complementarities between the semantic annotation models from AI and the annotations proposed by Corpus Linguistics, and combining both of them together; far from being irreconcilable, they may be considered complementary. Thus, an introduction to corpus annotation is offered in §3.3.1. An example of the integration of both paradigms (AI’s and Corpus Linguistics’) in ContentWeb–OntoTag—is presented afterwards, in §3.3.2. Finally, the main advantages and drawbacks of such a model are discussed in §3.0.

3.1 Text Annotation in Corpus Linguistics

The idea of text annotation was originally developed in Corpus Linguistics. An annotated corpus “may be considered to be a repository of linguistic information [...] made explicit through concrete annotation” (McEnery and Wilson 2001). The benefit of such an annotation is clear: it makes retrieving and analyzing information about what is contained in the corpus quicker and easier.

In EAGLES (1996a), a list of the main different levels of linguistic annotation can be found, namely: lemma, morphosyntactic, syntactic, semantic, and discourse annotation. They are shown in Figure 3-1 (Annotation Level Pyramid), together with their corresponding tools (Linguistic Tool Stack) and applicable criteria, recommendations and guidelines (Linguistic Annotation Criteria Heap). A deeper analysis of these concepts than the one included below can be found in EsperOnto (2003).
3.1.1 Lemma Annotation

Lemma annotation (lemmatization) accompanies every word-token in a text with its lemma, that is, the form of the head word found when looking up that word in a dictionary. In English, lemma annotation may be considered redundant, but in more highly-inflected languages, such as Spanish, the ratio of word-forms per lemma makes lemma annotation a very valuable contribution to information extraction (Leech 1997). There are no specific guidelines for lemmatization, so Leech’s (1997) and CES (1999) criteria should be applied to this annotation level, which is usually carried out as a subtask by POS-taggers.

Once annotated, the lemma of a word can be linked to the concepts of an ontology and to the entries in a lexicon, functioning as a lexical semantic key for processing that word.
3.1.2 Morphosyntactic Annotation

Together with the syntactic annotation (next subsection), this is one of the most extended types in Corpus Linguistics. **Morphosyntactic annotation, part of speech annotation, POS tagging or grammatical tagging** is the annotation of the grammatical class (e.g., noun, verb, etc.) of each word-token in a text,\(^\text{18}\) together with (possibly) the annotation of its morphological analysis. As claimed in McEnery & Wilson (2001) POS information establishes an essential foundation for further forms of analysis such as syntactic parsing and semantic field annotation and can be carried out automatically with an acceptable error rate by **POS-taggers**. EAGLES (1996a) offers a very valuable set of guidelines for this annotation level.

Disambiguation of homographs, identification of idiomatic word sequences and compounds, and separation of contracted forms are some of the different irregularities an annotator must face at this level and the main contribution of this annotation level to the semantic analysis and processing of a document.

3.1.3 Syntactic Annotation

Once morphosyntactic categories in a text have been identified, **syntactic annotation** adds information about the higher-level syntactic relationships between these categories, which are determined, e.g., by means of a phrase-structure or dependency parse. Different parsing schemes are employed by different annotators; according to McEnery & Wilson (2001) these schemes differ in:

- The number of constituent types they employ (typically, the number of tags in the POS tagset).
- The way in which constituents are permitted to combine with one another.

\(^{18}\) In other words, a POS tagging system holds the answers to these problems: (1) how to divide the text into individual word tokens (words); (2) how to choose a tagset (= a set of word categories to be applied to the word tokens); and (3) how to choose which tag is to be applied to which word (token).
The grammar followed to parse and annotate the text.

Two kinds of tools can be applied to syntactic annotation: **chunkers** (shallow parsing) and **parsers** (full, deeper parsing). The most interesting guidelines for this level can be found in EAGLES (1996b).

Some syntactic phenomena and properties can determine some—not necessarily minor—changes in the semantics of an expression, such as the ordering of the words in a compound, the change of meaning of certain adjectives, in Spanish for example, when pre- or post-modifying a substantive or the determination of PP-attachments.

### 3.1.4 Semantic Annotation

As asserted in McEnery and Wilson (2001), two broad types of semantic annotation may be identified, related to:

- **A. Semantic relationships between items** in the text (i.e., the agents or patients of particular actions). This type of annotation has scarcely begun to be applied.
- **B. Semantic features of words in a text**, essentially the annotation of word senses in one form or another. There is no universal agreement in semantics about which features of words should be annotated.

As shown in Figure 3-1, the only tools available for automatic semantic annotation are sense taggers. Regarding the annotation criteria for this level, no EAGLES semantic corpus annotation standard has yet been published, although some preliminary recommendations on lexical semantic

---

19 Compare, for example, “a Semantic Web page annotation model” and “a Semantic Web annotation model page”.

20 For instance, cf. the difference between *un pobre hombre* ‘a poor (i.e., unlucky) man’ vs. *un hombre pobre* ‘a poor (i.e., having little money) man’.

21 The participants involved in an event, process, or state described, the different phrases and clauses in the syntactic level, etc.

22 See, for example, the controversies within the SENSEVAL initiative meetings (Kilgarriff 1998; Kilgarriff & Rosenzweig 2000).
encoding have already been posted (EAGLES 1999); nevertheless, for the second type of semantic annotation enunciated, B, a set of reference criteria has been proposed by Schmidt (1988) for choosing or devising a corpus semantic field\(^{23}\) annotation system. These criteria can be summarized as follows:

1. *It should make sense in linguistic or psycholinguistic terms.*
2. *It should be able to account exhaustively for the vocabulary in the corpus, not just for a part of it.*
3. *It should be sufficiently flexible* to allow for those emendations that are necessary for treating a different period, language, register, or textbase.
4. *It should operate at an appropriate level of granularity (or delicacy of detail).*
5. *It should, where appropriate, possess a hierarchical structure.*
6. *It should conform to a standard, if one exists.*

The use of ontologies as a basis for a semantic annotation scheme fits perfectly and accomplishes the criteria posited by Schmidt. Clearly, the mostly hierarchical structure of ontologies fulfils by itself criterion (5) and, as a side effect, criteria (2) and (4), since the ontology can grow horizontally (scope extension) and vertically (specialization). Criterion (3) is also satisfied by an ontology-based semantic annotation scheme, since we can always specialize the concepts in the ontology according to specific periods, languages, registers, and textbases. Ontologies are, by definition, consensual and thus are closer to becoming a standard than many other knowledge models, as criteria (6) requires. Concerning criterion (1), quite a lot of groups developing ontologies are characterized by a strong interdisciplinary approach that combines Computer Science, Linguistics and (sometimes) Philosophy; then, an ontology-based approach should also

\(^{23}\) A semantic field (sometimes also called a conceptual field, a semantic domain, or a lexical domain) is a theoretical construct which groups together words that are related by virtue of their being connected—at some level of generality—with the same mental concept (Wilson & Thomas 1997).
make sense in linguistic terms. Hence, the ontologic and the linguistic points of view of the Semantic Web can be considered complementary and mutually enlightening.

3.1.5 Discourse Annotation

This is the least frequently developed kind of annotation, at least in corpora. Still, two main different kinds of approaches to annotation at this level can be found. Stenström’s approach (McEnery & Wilson 2001) is based on what she called discourse tags, derived empirically from an initial analysis of a corpus subsample. These tags included categories such as “apologies” (e.g., sorry, excuse me) or “greetings” (e.g., hello, good evening) and were used to mark items whose role in the discourse dealt primarily with discourse management rather than with the propositional content. This first approach has never become widely used in corpus linguistics. Conversely, the pronoun reference or anaphoric annotation approach considers cohesion a crucial factor in our understanding of the processes involved in reading, producing, and comprehending discourse, which can be considered the main contribution of this annotation level to the semantic interpretation of a document. A clear exponent of this approach is the UCREL discourse annotation scheme, together with many other anaphoric annotation schemes, such as those by De Rocha, Gaizauskas & Humphries, and Botley (Garside et al. 1997).

As shown in Figure 3-1, no automatic annotation tool or specific criteria has been developed for this level yet. Hence, as in the case of lemmatization, criteria according to Leech (1997) and CES (1999) should be applied when annotating at this level.

3.2 Hybrid Annotation: OntoTag

One of the four subtasks of ContentWeb (Aguado 2002) is the development of OntoTag, a model and environment for the hybrid—

24 Cohesion (Halliday & Hasan 1976) is the vehicle by which elements in texts are interconnected through the use of pronouns, repetition, etc.
linguistic and ontological—annotation of web documents.

![Morpho-syntactic annotation excerpt](image)

**Figure 3-2: Morpho-syntactic annotation of the Spanish article “la”**

Within the elaboration of OntoTag, a first exploration phase has been performed. A short example of this first phase, implemented in RDF(S), is presented below; the annotation of the Spanish sentence *{Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos.}* at the first three levels is shown in Figures 3-2, 3-3, and 3-4.

At the morpho-syntactic level (Fig. 3-2), every word or lexical token

---

25 “After five years of expectation and gossiping, here comes the most expected film for the time being.”
is given a different Uniform Resource Identifier (URI henceforth) and three possible categorizations are included, according to three different tagsets and systems we want to evaluate. Each tagset has been assigned a different class in the morphAnnot namespace: TradAnnot–CRATER tagset–, MBTAnnot–MBT (2002) tagset– and ConstrAnnot–Constraint Grammar, FDG (Tapanainen and Järvinen 1997) tagset–. For the sake of space, just the annotation of the article *la* has been included in the figure.

At the syntactic level (Fig. 3-3) every syntactic relationship between morpho-syntactic items is given a new URI, so that it can be referenced in higher-level relationships or by other levels of the annotation model (i.e., `<synAnnot:Chunk rdf:ID= “1_510”>`). Also for the sake of saving space, just the annotation of the phrase *la película más esperada de los últimos tiempos* has been included in the figure.

Figure 3-3: Syntactic annotation of the Spanish chunk “*la película más esperada de los últimos tiempos*.”
At the semantic level (see Fig. 3-4) some components of lower level annotations are tagged with semantic references to the concepts, attributes and relationships determined by our (domain) ontology, which is implemented in DAML+OIL. Further elements susceptible of semantic annotation are being sought and research is being done towards their determination by the linguist team in our project. The pragmatic counterpart of OntoTag has not yet been tackled and thus this level is not included in the example.

![Semantic annotation excerpt](image)

Figure 3-4: Semantic annotation of the Spanish sentence “Tras cinco años de espera y después de muchas habladurías, llega a nuestras pantallas la película más esperada de los últimos tiempos”.

### 3.3 Hybrid Annotation: Discussion

The integration of these two approaches (the linguistic and the ontological) into a hybrid annotation scheme entails many advantages for language engineering and AI applications. First of all, from the point of view of language resources:

- **Linguistic tools are re-used**: the tools developed so far for lemmatization, POS tagging and chunking/parsing can be exploited
for the generation of the linguistic counterpart of the annotation.

- **Annotated documents are multi-purpose and can be reused:** once a document—web page—has been annotated, there is no need to POS-tag it, to parse it, etc., anymore, no matter what kind of processing it must go through afterwards (i.e., machine translation, information retrieval, information extraction, text mining, and so forth). Since parsing, for example, is a high time-consuming task, we can have an additional advantage, that is, reducing our overall Semantic Web page processing time.

The second main advantage is that the meaning of a page with explicit semantic annotation can be reinforced by the meaning contribution provided by all of the linguistic levels; semantic analysis can also benefit from the invaluable work done so far on the development of ontologies as conceptual and consensual hierarchical models, specially (UNSPSC 2002, RosettaNet 2002) in specific domains (i.e., e-commerce).

However, the main disadvantage lies in the limitations imposed by current technologies: the process of obtaining automatically compact, readable and verifiable pages is quite a hard task to be fully specified and delimited, but the work being done in our laboratory tries to shed some light upon it.

4. Using DAML for Representing Domain Specific Knowledge* (Barrett Bryant)

The Semantic Web envisioned by Berners-Lee, Hendler, & Lassila (2001) requires software “agents” which operate across the web giving “intelligence” to web pages in the form of “services.” Our view is that the Semantic Web may be thought of as a software system composed of many interoperating distributed heterogeneous software components. The software

* This research is based upon work supported wholly or in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number DAAD19-00-1-0350 and by the U.S. Office of Naval Research under award number N00014-01-1-0746.
realizations of these components may span various component models but
their interoperability is facilitated by their “knowledge” of a particular
domain. Our research involves the formalization of these domains, or more
precisely, development of the technology required to realize this
formalization. We have applied this work to two problem areas: (1)
formalization of domain knowledge to assist in understanding natural
language requirements documents; and (2) formalization of domain models
under which heterogeneous components may interoperate within a
common understanding. Both types of formalizations have required a
suitable knowledge representation language and toward this end we have
experimented with XML (Decker et al. 2000), the eXtensible Markup
Language, and DAML, the DARPA26 Agent Markup Language (Hendler &
McGuinness 2000). This section describes experiences with using DAML.
We first briefly describe DAML and related semantic representation
languages and then describe how we used DAML to represent domain-
specific knowledge in our two research projects. Finally we conclude with
some summary observations.

4.1 Semantic Representation Languages

XML has revolutionized the way in which the syntax of data may be
represented for portability across a wide variety of platforms, languages,
and applications. Whereas XML has achieved syntactic interoperability,
DAML strives for semantic interoperability. The development of DAML is
due to a DARPA sponsored research project to extend XML with semantic
relationships and the ability to express ontologies required for
“understanding” such semantics. The emphasis is on describing semantic
information in such a way that the aforementioned software agents will be
able to seamlessly integrate Web pages, the software systems which are
embedded in such pages, and the databases that such systems interact with.
The original definition of DAML shared many characteristics with OIL
(Fensel et al. 2001), the Ontology Inference Layer, especially an object-

mil)
oriented type system. DAML+OIL represents a merger of these two approaches which has been proposed as a W3C\textsuperscript{27} standard for representing ontologies. Bechhofer, Goble, & Horrocks (2001) point out that the DAML+OIL representations sometimes do not preserve all important information present in semantic models. A related effort to DAML+OIL is OWL (Smith et al. 2002), the Web Ontology Language, which provides semantics in the form of logical assertions constituting a knowledge base about classes and relationships between them that are omnipresent in Web applications.

Baclawski et al. (2001) use UML, the Unified Modeling Language, to represent ontologies. They found that there were many compatibilities between UML and DAML. Since UML is an OMG\textsuperscript{28} standard for object-oriented modeling of software systems and hence has many associated tools, this work allows such tools to be used in representing semantic knowledge. This is also highly relevant to another OMG (2001) initiative, Model Driven Architecture (MDA), which proposes that software architecture be constructed around models developed within the context of standardized application domains.

We have used Two-Level Grammar (TLG) (Bryant & Lee 2002) to represent semantic information. The two levels of the grammar are a “meta-level” for describing the domains to be used at the “hyper-level,” the logical rules over the domains. These grammars are encapsulated in classes and structured in an object-oriented manner for compatibility with objects and components, which they are used to describe the semantics of.

4.2 Representing Domain Specific Knowledge in DAML+OIL

Our research has used DAML+OIL for representing knowledge found in software requirements documents and domain knowledge about a particular application domain for which a software system is to be assembled from a collection of components. Each of these projects will be described in terms of a simple example, namely a bank account.
management system. Such a system typically consists of a database with account information, and a server exposing the database to clients such as customers using Web browsers, ATMs, or telephone interfaces. This system (and all other systems) may be viewed at two levels: (1) the software requirements for constructing the system; and (2) a collection of components which may be assembled to construct the system.

4.2.1 Software Requirements

A software requirements document typically describes the functionality of the system to be constructed, usually in natural language. The document may contain specific information about how the system should work, including non-functional aspects such as timing constraints, security, etc. A human reader may understand this information due to his/her knowledge of the domain; e.g., most software engineers would have a basic understanding of the functionality of a bank account management system. However, as software engineers are not domain experts, it is often required that domain knowledge be explicitly described. Recent research in “generative programming” (Czarnecki & Eisenecker 2000) has shown that much of the software development process may be automated (i.e., through generation of program components) when the domain is well specified. The domain knowledge required describes the relationship between components and other constraints which are usually presumed in requirements documents or too implicit to be extracted easily from the original documents. For example, the requirements for an ATM component in our banking system might say “the user inputs the 4-digit PIN number by pressing the buttons.” The fact that the set of the buttons is a component of the ATM machine is implicitly assumed and therefore not explicitly mentioned in the requirements documents. This kind of information is domain specific knowledge. The units of measurements, who passes what to whom, which synonyms of a word are used, what each acronym stands for, etc., are some of the examples of domain specific knowledge that can supplement software requirements documents.

29 Automatic Teller Machines
We have used DAML to specify the domain knowledge required for a software system (Lee & Bryant 2002). The following examples show the use of DAML as domain knowledge for the ATM example. The disjointUnionOf notation in DAML can be used to list the subcomponents of a component, as in three data fields of an account:

```xml
<daml:Class rdf:ID="Account">
  <daml:disjointUnionOf
    rdf:parseType="daml:collection">
    <daml:Class rdf:ID="ID"/>
    <daml:Class rdf:ID="PIN"/>
    <daml:Class rdf:ID="Balance"/>
  </daml:disjointUnionOf>
</daml:Class>
```

where rdf stands for Resource Description Framework on which DAML and XML are built.

The sameClassAs definition in DAML may be used to indicate that the word Machine used in the ATM requirements is a synonym of the word ATM and that the word ATM stands for Automatic Teller Machine.

```xml
<daml:Class rdf:ID="Automatic_Teller_Machine">
  <daml:sameClassAs rdf:ID="Machine"/>
  <daml:sameClassAs rdf:ID="ATM"/>
</daml:Class>
```

Using ObjectProperty notation in DAML, the fact that a Balance is passed from a Bank to an ATM may be expressed as follows:

```xml
<daml:ObjectProperty rdf:ID="passBalance">
  <rdfs:domain rdf:ID="Bank"/>
  <rdfs:range rdf:ID="ATM"/>
</daml:ObjectProperty>
```
The data type or the measurement unit of a component may be expressed using DatatypeProperty notation in DAML as shown below for the type of Amount.

```xml
<daml:DatatypeProperty rdf:ID="Amount">
  <rdfs:range rdf:resource="http://www.w3.org/2000/10/XMLSchema#float"/>
</daml:DatatypeProperty>
```

In summary, the precise formal semantics of DAML provides a very useful way to specify the domain-specific knowledge explicitly. This knowledge is used as supplementary information for the conversion from a natural language requirements document into a formal specification in Two-Level Grammar, further details of which are described by Lee & Bryant (2003).

### 4.2.2 Semantic Web of Software Components

The UniFrame 30 (Raje et al. 2002) project aims to facilitate the construction of software systems via the integration of heterogeneous and distributed components deployed on the Web. Each software component will be accompanied by a formal specification indicating the functional and non-functional (also known as QoS—Quality of Service) properties of the component. The system assembler submits a query to a specialized search engine called a “head-hunter” which locates a possible set of components which satisfy the query using the formal specifications of the component. This process is facilitated by a formalized domain model of the application domain (e.g., bank account management systems) which each deployed component adheres to, as advocated in Model Driven Architecture (OMG 2001). The formalized domain model is called a Generative Domain Model (GDM) since it contains rules to generate programs (in the sense of Generative Programming (Czarnecki & Eisenecker 2000)) which connect

---

30 Unified Framework for Seamless Integration of Heterogeneous Distributed Software Components (http://www.cs.iupui.edu/uniFrame)
heterogeneous components, called “wrapper/glue code” to make a unified whole. The formalization aspects of this process are described further by Cao et al. (2002). For our bank account management system example, assume that on the Web are deployed various database components for storing bank account information, server-side components for accessing the database, and client-side components for interacting with the server. Each of these is accompanied by a formal specification of how the component adheres to the domain model, what technology it uses (e.g., CORBA, J2EE, or .NET), what types of component interactions are expected, and the Quality of Service expected for the component. The system assembler indicates the domain of the system and what functionality is expected based upon the domain, as well as system QoS requirements. The UniFrame process assembles the various sets of compatible components into a collection of candidate software systems. Each software system is then tested with a set of test cases to validate that QoS requirements are met for the assembled system.

The GDM in our system is expressed in Two-Level Grammar with UML used to provide a visual representation. These may further be expressed in XML and related dialects, including DAML+OIL. However, we found that the expressiveness of TLG could not be conveniently captured in the current form of these notations. It is not difficult to represent the object-oriented type structure of TLG or the GDM ontological information in XML or DAML+OIL, but the generation rules of TLG do not have such straightforward representations.

4.3 Summary

In conclusion, we found DAML+OIL to be a convenient representation for domain specific knowledge representations associated with software requirements documents, but not so suited for expressing generation rules in generative domain models. We did not investigate the use of DAML+OIL.
OIL for expressing non-functional requirements such as Quality of Service but believe this would be an interesting exercise. We shall continue to monitor the developments in languages for the Semantic Web for opportunities to integrate generation rule technologies and also to explore Two-Level Grammar as another possible language for the Semantic Web. Certainly TLG is able to capture semantic information pertaining to the integration of software components that is not currently available in Semantic Web languages.

5. Bringing Natural Language to the Semantic Web
(Jimmy Lin)

Because the ultimate purpose of the Semantic Web is to help users better locate, organize, and process content, we believe that it should be grounded in the information access method humans are most comfortable with—natural language. However, the Resource Description Framework (RDF), the foundation of the Semantic Web, was designed to be easily processed by computers, not humans. To render RDF friendlier to humans, we augment it with natural language annotations, or metadata written in everyday language. We argue that natural language annotations, parsed into computer-readable representations, are not only intuitive and effective, but can also accelerate the pace with which the Semantic Web is being adopted. We believe that our technology can facilitate a happy marriage between natural language technology and the Semantic Web vision.

5.1 The Semantic Web Vision

The vision of the Semantic Web is to convert existing Web information into a more machine-readable form, with the goal of making the Web more effective for users. This goal grew out of the recognition that although a wealth of information readily exists today in electronic form, this information lacks any machine-understandable semantics, and hence cannot be easily processed by computer systems. By adding semantics to existing Web data, e.g., \(This\ \text{particular} \ \text{number} \ \text{is} \ \text{a} \ \text{price} \ \text{in}\)
Euros., we can create an environment that allows intelligent software agents to interoperate easily. However, we argue that such metadata alone is not enough to bring out the full potential of the Semantic Web.

Fundamentally, Semantic Web research is attempting to address the problem of information access: building systems that help users locate, collate, compare, and cross-reference content. As such, we believe that users should be able to access information using everyday language, and that the Semantic Web should be grounded in linguistically motivated constructs. Natural language is intuitive, easy to use, rapidly deployable, and requires no specialized training. In our vision, the Semantic Web will be equally accessible by computers using specialized languages and interchange formats, and humans using natural language. The scenario of being able to ask a computer \{When was the president of Taiwan born?\} or \{Find me the cheapest vacation package in the Bahamas this month.\} and getting back “just the right information” is very appealing.

Because the first step to building the Semantic Web is to transform existing sources (stored as HTML pages, in legacy databases, etc.) into a machine-understandable form (i.e., XML/RDF), it is sometimes at odds with a human-based natural language view of the world because formally and precisely defined ontologies necessary for computer comprehension may seem very unnatural to humans. Although the general framework of the Semantic Web includes provisions for natural language technology, the actual deployment of such technology remains largely unexplored. We fear that if greater consideration is not given to the integration of language technologies at the outset, future solutions might be little more than ad-hoc patches.

We believe that the fundamental disjoint between the current Semantic Web vision and actual end-users is that the Resource Description Framework (RDF), the standardized Semantic Web language for describing metadata, was meant for consumption by computers, not humans. Given this philosophy, how can we be sure that we are creating useful metadata? How can we be sure that our ontologies mirror the way users organize and think about content? Since the final beneficiary of the Semantic Web
should be the user, we advocate a human-centered organization of metadata grounded in natural language. We accomplish this by weaving natural language annotations directly into the Resource Description Framework.

5.2 Natural Language Annotations

Use of metadata is a common technique for rendering information fragments more tenable to processing by computer systems. Using natural language itself as metadata presents several additional advantages: it preserves human readability, allows for easy querying, and encourages non-expert users to engage in metadata creation. To this end, we have developed natural language annotations (Katz 1997), which are machine-parsable sentences and phrases that describe the content of various information segments. These annotations serve as metadata that describe the kinds of questions that a particular piece of knowledge is capable of answering.

To illustrate how this technology works, consider the following paragraph about Mars:

Mars has two small moons: Phobos and Deimos. Phobos (fear) and Deimos (panic) were named after the horses that pulled the chariot of the Greek war god Ares, the counterpart to the Roman god Mars…

This paragraph may be annotated with the following:

Mars’ two moons
Phobos and Deimos are two moons orbiting Mars.

34 It is true that many parts of the Semantic Web will never have any contact with humans, and may be created only for the benefit of software agents, e.g., inventory management systems communicating with warehouses. For these applications, natural language may not be necessary. Nevertheless, a large fraction of the Semantic Web involves end-users, where we believe natural language forms the best information access medium.
A question answering system would parse these two annotations and store the parsed structures (e.g., ternary expressions (Katz 1988)) with pointers back to the original information segment. To answer a question, the user query, parsed into the same type of structures, would be compared against the annotations stored in the knowledge base. Because this match would occur at the level of parsed representations, linguistically sophisticated machinery such as synonymy/hyponymy relations, ontologies, and structural transformation rules (e.g., “S-Rules” (Katz 1988; Katz & Levin 1988)) could be brought to bear on the matching process. If a match were found, the segment corresponding to the annotation would be returned to the user as the answer. Because sophisticated natural language processing could be invoked in matching questions with annotations, precision far beyond that of standard keyword-based information retrieval techniques could be achieved. In addition, a linguistically-based system allows for variations in user queries, e.g., alternate formulations, active/passive voice, nominalizations, etc.

In the above example, the natural language annotations would allow a question answering system to answer the following questions:

What satellites orbit Mars?
How many satellites orbit Mars?
What are the names of the Martian moons?

We have implemented the above technology in START35 (Katz 1997), the first question answering system available on the World Wide Web. Since it came online in December 1993, START has engaged in exchanges with hundreds of thousands of users all over the world, supplying them with useful knowledge.

5.2.1 Natural Language Annotations for the Semantic Web

An important feature of our annotation concept is that any information segment can be annotated: not only text, but also images, multimedia,
database queries, and even procedures. Attaching natural language annotations to RDF serves as the basis of our framework for bringing natural language access capabilities to the Semantic Web. Here we provide an illustrative example of this technology; more detailed descriptions of the underlying mechanisms can be found elsewhere (Katz et al. 2002).

Suppose we want to answer the following “family” of questions about various attributes (e.g., state bird, state flower, state motto, population, area, etc.) of U.S. states:

What is the state bird of California?
Tell me what the state motto of Massachusetts is.
Do you know Colorado’s population?
What is the capital of Kentucky?

Fortunately, the data necessary to answer such questions can be easily found on the Web. Assuming that this information has been structured into RDF, our method of bridging the Semantic Web and question answering can be conceptualized in the following ontology fragment:

![Ontology Fragment](image)

---

36 http://www.50states.com/
37 Currently, this must be accomplished manually, but in the future, information sources will be directly published in RDF.
In this simple ontology, a state “class” has a number of properties, e.g., has-bird and has-population. Attached to these properties are natural language annotations that describe, in stylized English, a linguistic realization of the information contained within the ontology. That is, the annotations are parameterized so that a single lexical item, e.g., any-state, can serve as a stand-in for a larger class of lexical items, e.g., all the 50 U.S. states. In effect, the annotation attached to the has-bird property is a shorthand way of specifying a large class of annotations: the state bird of Alabama is Yellowhammer, the state bird of Alaska is Ptarmigan, etc. One single parameterized annotation describes all individual instances of the has-bird property and relates all states to their respective state birds.

The question answering process proceeds as follows: A natural language question like {What is the population of Massachusetts?} is parsed and matched against the annotations. Because this match occurs on the level of syntactic representations, the matching mechanism can handle various formulations of the same question, e.g., dealing with lexical variations, semantic alternations, etc. The question matches the annotation of the has-population property, which triggers generation of the answer: the object of the has-population property of Massachusetts.

We have built a prototype implementing the technology described above (Katz et al. 2002) in the first Semantic Web question answering system that we are aware of. Although the system is currently limited in the types of questions that it can answer and the domain, we believe that the system is a proof of concept that demonstrates a viable method of integrating natural language techniques with the Semantic Web. By separating the knowledge, stored in RDF, from the natural language processing components, we can isolate ontology building from language engineering and vice versa, but still maintaining the connection between the two through natural language annotations.

Natural language annotations are part of our larger effort to build a uniform metadata framework for the Semantic Web (Karger et al. 2003). We wish to create the digital equivalent of physical “sticky notes”, with the ability to attach annotations to anything and anywhere. Annotations written in natural language would complement other forms of more traditional, structured metadata such as author, creation date, subject, etc. Such a
uniform framework would afford users the flexibility to query on multiple levels, using different sources of information. For example, consider a query such as \{*Show me mystery novels published within the last two years that other people have liked.*\}. This request can only be fulfilled by searching over structured metadata (genre and publication date) and unstructured free text (user opinions). Our generalized annotation framework is built to accommodate exactly this type of user queries.

In our vision of the Semantic Web, natural language annotations would co-exist alongside other metadata under our generalized annotation framework. Metadata could be distributed (e.g., embedded directly into web pages) or centralized (e.g., stored in a common server); either way, a software agent would compile these digital “sticky notes” into a question answering system capable of providing natural language information access to users.

5.2.2 From Here to the Semantic Web

We believe that natural language annotations are not only an intuitive and helpful extension to the Semantic Web, but will assist in the deployment and adoption of the Semantic Web itself. The primary barrier to its creation is a classic chicken-and-egg problem: people will not spend extra time marking up their data unless they perceive a value for their efforts, and metadata will not be useful until a “critical mass” has been achieved. Although researchers have been focusing on technology to reduce barriers to entry (via authoring tools, for example), such initiative may not be sufficient to overcome the hurdles. As Hendler (2001) remarks, lowering markup cost is not enough; for many users, the benefits of the Semantic Web must come for free. Semantic markup should be a by-product of normal computer use and there is no process of metadata creation that is easier and more intuitive than the use of natural language—users simply describe the contents of objects in everyday language. By divorcing the majority of users from the need to understand formal ontologies and a precisely defined vocabulary, we can dramatically lower barrier-of-entry, easing the transition into the Semantic Web vision.
Ultimately, let us not forget that the purpose of the Semantic Web is to benefit humans, not computers. The original idea was that instead of waiting for computers to become smart enough to understand human language, we should focus on the slightly less difficult problem of making human data more understandable to computers. To this end, the foundations of the Semantic Web are grounded in language and information access. However, to achieve interoperability and to facilitate interaction between software agents, we have had to sacrifice a lot of human understandability—precise ontologies and formally defined semantics are foreign concepts to the average user. By reintroducing natural language annotations and rendering the connection to human language explicit, we can achieve a satisfying middle ground between computer and human needs.

6. Conclusion: Language Technology and the Semantic Web (Nancy Ide)

In its broadest definition, the Semantic Web is intended to associate knowledge to web objects, whether they be documents, images, audio signals, or other media or processes. This knowledge will be represented in ontologies that define the relations among and properties of various bits of knowledge. While the bits of knowledge in the ontology may represent abstract concepts with no reference to a realization in any human language, it is unavoidable that the web objects with which they are associated will in the vast majority of cases consist of language data. Thus the role of language technology—which is the means by which we automatically discover meaning in language data or generate language data intended to represent specific meanings—in the Semantic Web is at the very least the means by which many of web objects will be identified, appropriately associated with ontological knowledge, and presented to the human user.

In terms of its relevance for language technology, it is useful to identify three different ways of thinking about the Semantic Web: (1) the Semantic Web as a long-term goal, realizing the vision of a vast web of objects linked to a common ontology representing all knowledge, which in turn enables us and our intelligent agents to discover and manipulate these
objects in sophisticated ("meaningful") ways; (2) the Semantic Web as it could be realized in the relatively near future using the technologies so far developed and under development intended to implement it; and (3) the means to actually construct the Semantic Web in both its near- and long-term realizations. Language processing technology will certainly play a seminal role in the construction of the Semantic Web by providing the tools required to automatically identify relevant objects in language data. Identification demands the usual range of language processing capabilities, depending on the kinds of objects under consideration: broad topical information can be identified using standard (often statistics-based) document retrieval techniques; information extraction can provide more specific information (e.g., more precise topical information, names, dates, etc.); detailed information about, for example, an event, its participants, causes, and outcome may require deeper linguistic analysis. Similarly, machine learning techniques can be harnessed in the service of ontology-building for Semantic Web applications, and language/speech understanding and generation will be critical to the implementation of user-friendly interfaces. So, in the near future, we can assume that many of the activities central to language processing work will continue on in much the same way as they have done—except in one fundamental way.

Once identified by language processing applications, the relevant objects must be annotated to record the discovered information. Before the advent of the Semantic Web idea, annotation of language data typically involved identification of relevant segments (tokens, utterances, sentences, discourse units, etc.) and "labeling" them with some linguistic information for morpho-syntax, syntax, co-reference, parallel alignment, and the like. Sometimes the labels are included in-line in the data; more recently, "stand-off" markup has been used to both delimit segments and link them to the relevant linguistic information in another document. Often, the labeling system is idiosyncratic to the data, although recent standardization efforts have enabled some homogeneity in linguistic labeling among annotation projects.38 Annotations at various linguistic levels are occasionally

38 However, note that in general the language processing community has resisted a common labeling scheme, for the good reason that different theoretical
linked to specify relations among them (usually, constituent relations); but rarely, if ever, is an ontology of linguistic categories used in the background.

In the Semantic Web, annotation of objects is accomplished by associating an object with a category in an ontology, which in turn specifies its properties and relations to other categories. The big advantage of this idea, in addition to avoiding the duplication of information and ensuring a standard annotation scheme, is the ability to perform inferencing over the annotated data that enables the extraction of information that is not explicitly given. The notion that annotation will be accomplished via linkage to a common ontology of information brings up a number of questions about the ways in which language processing work can and should be accomplished in the immediate future.

The answer depends in part on what kinds of annotations we expect to be a part of the Semantic Web. It is possible to imagine that someday, as ontologies become increasingly rich, language processing capabilities become more robust, and computers become orders of magnitude faster, the only annotations that will be retained will be those at the highest levels, such as the representation of an event or state, together with its participants, their roles, etc., and lower-level linguistic information, such as syntactic structure, co-reference information, etc., may be computed on the fly and discarded once the really useful information is obtained. Eventually, it might be possible to generate much—possibly all—required information even at higher levels on the fly, and annotation of any kind will become obsolete.

However, for the foreseeable future, we can expect that intermediate annotations will be retained, and this brings up yet another question for language processing in the near term: should lower-level annotation types themselves be integrated into the Semantic Web technology? That is, should we be creating ontologies of linguistic categories together with their properties and relations, to be used in and by language processing applications? It may seem circular to utilize Semantic Web technologies to create ontologies to support the development of Semantic Web ontologies, but in fact the process is a bootstrap rather than a self-feeding loop. And it approaches cannot be represented with a common set.
is a critical bootstrap, because we cannot expect semantic homogeneity to
any degree at the higher levels if it is not achieved at the lower ones first.

The role of language technology in the Semantic Web is, then,
twofold: first, established and evolving language processing techniques
will play a crucial part in identifying objects to be integrated into the
Semantic Web, developing the ontologies to support it, and enabling
effective human-computer interaction that exploits the results. Second, it is
up to the language processing community to employ the same technologies
that will support the Semantic Web by encoding the requisite linguistic
information in ontologies and exploiting inferencing capabilities in order to
feed this effort. This second activity is in fact far more difficult than the
first because it will demand, above all, an international collaborative effort
to achieve it. This activity has barely begun, and it is not entirely clear how
it can be accomplished. Some language processing researchers are
developing ontological information to support Semantic Web applications
(for example, the DAML effort sponsored by the U.S. Department of
Defense) without full involvement of the international community, that are
almost certainly bound to be domain-specific and ultimately unacceptable
as off-the-shelf solutions. Other groups, such as the International Standards
Organization Committee on Language Resources (ISO TC37 SC4), are
attempting to work with the international community to achieve common
standards by allowing for variation via formalized definitions of categories
deviating from the stock of established norms; but even here, it is not clear
how such deviations will be handled or tolerated by inferencing engines
and other processing software. It will indeed be a very long road to achieve
what is needed, but it is a road we must take with full awareness of not
only the nature, but also the magnitude and complexity of the task.

Of course, at this point the Semantic Web is only a vision. Although it
has been energetically embraced by much of the research community, its
full realization is a very long way off. We are, in fact, in the stage where
only the most fundamental groundwork for a Semantic Web is being laid,
and the vision itself is so enormous and, to some extent, vague, that we
cannot be sure exactly how the final product will turn out. Nonetheless, the
Semantic Web seems to be a good idea (or at least the best idea we have at
the moment), and we need to work towards achieving it even if along the
way we find that the architecture has changed or the foundation needs major renovation. I cannot help but think of current work in language processing as “brain-building”, where we are attempting to cobble together a few hundred neurons here, a few hundred there, without much idea of how it all fits together in an interdependent network involving billions of such neurons that can accomplish language understanding at anything like the human level. The idea behind the Semantic Web, I believe, is one of those “intuitive leaps” that enabled us to have a suddenly clearer idea of how at least some of the pieces could be integrated, and this is likely the reason why so many have embraced and begun to pursue it. It is a step, however modest, toward the eventual goal.

References


Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. May,


Biographic Sketch and Research Outline of Authors

Graham Wilcock

Graham Wilcock is Docent of Language Technology at University of Helsinki. He has a Ph.D. in Computational Linguistics from University of Manchester.

He was co-founder of the NLPXML series of workshops, serving on the program committee of the 1st workshop (Tokyo, 2001), co-chair of the 2nd workshop (Taipei, 2002), co-organizer of the 3rd workshop Language Technology and the Semantic Web (Budapest, 2003), and program chair of the 4th workshop RDF/RDFS and OWL in Language Technology (Barcelona, 2004).

His early career included twelve years in the European computer industry with ICL, debugging mainframe operating systems and supporting EU database systems in Luxembourg. His language technology experience dates from 1985 in a UK national project on Japanese-English machine translation at University of Sheffield.

Through the 1990s he worked in Japanese industry with Sharp Corporation, developing the DUET Japanese-English machine translation system. During this period he was a visiting researcher at the Centre for Computational Linguistics, UMIST (University of Manchester), doing research on machine translation and natural language generation with Systemic Functional Grammar and HPSG.

Since 1998 he has been a university lecturer in linguistics and computer science, working in HPSG linguistic theory, natural language generation, spoken dialogue systems, Java- and XML-based NLP tools, and the Semantic Web. His current focus is an ontology verbalizer (http://www.gate.ac.uk/conferences/iswc2003/). He has given courses and seminars on XML-based natural language generation in Finland, Estonia, Japan, and UK, and is responsible for the generation component of an ontology-based dialogue system in a Finnish research project. He will be co-organizer of the 10th European Workshop on Natural Language Generation (Aberdeen, 2005).
List of Publications:
http://www.ling.helsinki.fi/~gwilcock/publications.html

Research Accomplishments:
1. Natural Language Generation with HPSG (Ph.D. thesis)
   http://www.ling.helsinki.fi/~gwilcock/Pubs/PhDThesis.ps
2. Papers in the on-line HPSG Bibliography
   http://www.cl.uni-bremen.de/HPSG-Bib/Bib/W.html
3. XML Finland tutorial: XML-based Natural Language Generation
   http://www.ling.helsinki.fi/~gwilcock/Pubs/XMLFinland-02.pdf
4. Finnish USIX “Interact” spoken dialogue project (2001-3)
5. Finnish FENIX “4M” ontology-based dialogue project (2003-7)

Paul Buitelaar

Paul Buitelaar is a senior researcher in the Language Technology Lab at DFKI GmbH, Saarbrücken, Germany. He received his Ph.D. in Computer Science from Brandeis University, Waltham MA, USA, in February 1998 and then joined DFKI. He taught classes at Brandeis University, at Saarland University, Saarbrücken, Germany, and in the context of several summer schools (e.g., Eurolan-1999, Eurolan-2001 and the OntoWeb summer school on “Ontological Engineering and Semantic Web”) on the topics of computational lexical semantics, concept-based information retrieval, and semantic web technologies.

He organized several international workshops on topics in lexical semantics, semantic annotation, and ontology development (e.g. in the context of ESSLLI-1998, COLING-2000, and ECAI-2004) and has been an invited speaker at many panels and workshops in these research areas. His main research interest is in Concept-Based Information Access in relation with research and development in Semantic Annotation, Ontology Development, Information Extraction, and Corpus-Based Lexical Semantics.

At DFKI he has been working as a senior researcher on several EC funded projects in the area of concept-based information access (e.g.,
MIETTA, OLIVE, and MEMPHIS) and as coördinator of the EC/NSF funded project MUCHMORE on concept-based cross-lingual information retrieval in the medical domain. In recent years he has been involved in the world-wide effort on development of the semantic web as coördinator and originator of the Special Interest Group on “Language Technology in Ontology Development and Use” within the EC funded thematic network OntoWeb. Finally, he is also co-chair of the DFKI Competence Center Semantic Web.

List of Publications:
http://www.dfki.de/~paulb/pub.html

Research Accomplishments:
1. CoreLex: An Ontology and Lexical Semantic Knowledgebase of Systematic Polysemous Classes (Based on WordNet)
   http://www.dfki.de/~paulb/corelex.html
2. MuchMore Bilingual Corpus: a German/English Parallel Medical Document Collection with Corresponding Queries and Relevance Assessments and Evaluation Sets of Disambiguated Terms
   http://muchmore.dfki.de/resources_index.htm

Antonio Pareja-Lora

Antonio Pareja-Lora is a professor at the Facultad de Informática of the Universidad Complutense de Madrid (UCM) and a research fellow at the Artificial Intelligence Laboratory (LIA) of the School of Computer Science at the Technical University of Madrid (UPM).

He received his six-year degree with honours in Computer Science and Engineering from the UPM in March 1998, and he is expecting to receive his Ph.D. in Computer Science and Artificial Intelligence by September 2004. He has held research grants or positions in other academic and research institutions, such as the Antonio de Nebrija and Pontificia Comillas universities and the Austrian Research Institute for Artificial Intelligence (ÖFAI).
From the time of graduation, he has been actively involved in diverse areas of Natural Language Processing as, for instance, Machine Translation, stochastic parsing, linguistic resources standardisation, and corpora and Semantic Web annotation from a hybrid—linguistic and ontological—point of view. He elaborated a mini-prototype for high-quality machine translation called HAT, based on Dik’s Functional Grammar Theory and, after acquiring some supplementary background on stochastic parsing, collocation extraction, and ontology development, he is currently co-directing the construction of the framework for standardised annotation OntoTag, and the development and annotation of the ODE-Corpus on Entertainment. His research vision is to develop multilingual and ontological resources and infrastructure, focusing on the Spanish components, and linking them to OntoTag framework and to the development environment WebODE, from the LIA-UPM, in order to bring a better human language understanding to machines, making them more helpful in information processing tasks to human beings.

List of Publications:
http://www.ucm.es/info/dsip/directorio/APL.html

Barrett R. Bryant

Barrett R. Bryant is a Professor and the Associate Chair of the Department of Computer and Information Sciences at the University of Alabama at Birmingham (UAB), which he joined after receiving his Ph.D. in computer science from Northwestern University in January. He has held visiting research positions at Ibaraki University in Japan, the IBM Database Technology Institute, Rome Laboratory of the U.S. Air Force, the U.S. Army Research Laboratory Software Technology Branch, Oak Ridge National Laboratory, the U.S. National Air and Space Administration (NASA) Marshall Space Flight Center, and the U.S. Naval Postgraduate School. He is currently the Chair of the Association for Computing Machinery (ACM) Special Interest Group on Applied Computing (SIGAPP).
His primary research interests are the formal specification of software systems, the development of syntactic and semantic formalisms, implementation technologies for formal specifications, and the integration of natural language processing with formal methods. The main development of this work has been the use of the Two-Level Grammar (TLG) formalism as a successful approach to unifying these interrelated themes. TLG has been extended from van Wijngaarden’s original concept to have object-orientation, functional forms, and a declarative style of expressing logical rules. Recent results have been the establishment of TLG as an intermediate representation between informal natural language specifications of software and formal specifications, and as a language for specifying software components. His future direction is to investigate TLG for programming the Semantic Web to further extend the technologies discussed in this paper.

List of Publications:
http://www.cis.uab.edu/info/faculty/bryant/papers

Jimmy Lin

Jimmy Lin is a Ph.D. candidate at the MIT Computer Science and Artificial Intelligence Laboratory. His research focuses on question answering, which lies at the intersection between information retrieval and natural language processing. The goal is to develop intuitive methods for accessing knowledge and to build machines capable of satisfying user information requests phrased in natural language. He has worked extensively on question answering techniques that take advantage of the World Wide Web, structured and semistructured databases, corpora of newspaper articles, and even videos. Jimmy’s work attempts to integrate syntax- and semantics-based strategies with robust keyword-based and statistical techniques to achieve both high-precision and broad coverage in question answering. Previously, he has worked at the IBM T.J. Watson Research Laboratory and Microsoft Research.
Nancy Ide

Nancy Ide is Professor and Chair of Computer Science at Vassar College in Poughkeepsie, New York. She has been involved in humanities computing and computational linguistics for twenty years. She was president of the Association for Computers and the Humanities from 1985-1995, and is currently co-editor of the journal *Computers and the Humanities*. She is also co-editor of the Kluwer book series Text, Speech, and Language Technology, and has co-directed the past four EUROLAN summer schools on various topics in computational linguistics. In 1987 she spearheaded the Text Encoding Initiative and served on its steering committee until 1997. She has published numerous papers on the application of statistical methods to language analysis, including computational lexicography, word sense disambiguation, and discourse analysis. Most recently she has been involved in developing standards for the representation of linguistically annotated resources, creation of the American National Corpus, development of an “intelligently searchable” corpus for historical research comprising materials from the Franklin D. Roosevelt Library, and adapting language processing practices to the Semantic Web.

List of Publications:
http://www.cs.vassar.edu/faculty/ide/pubs.html
Introduction to
Chinese Language Processing
at the Dawn of the 21st Century

Chu-Ren Huang

Chinese Language Processing (CLP) offers an ideal case-study for
natural language processing (NLP) because the advantages gained by using
NLP confer a tremendous advantage over traditional ways of text preparation.
At the same time, however, the non-alphabetic writing system of Chinese
poses great challenges because there is no obvious solution to basic issues
involving input and output. This hurdle, as described in the current panel
report by Benjamin T’sou, remained almost a deterrent until the last decade
or so.

Hence it is of great interest to observe how research in Chinese
computational linguistics has developed. It would be no exaggeration to
say that issues in Chinese Language Processing forced computational
linguists to think beyond the man-made traditions of keyboard positions
and spaces as wordbreaks. In other words, CLP moved the focus of NLP to
the content of language, as opposed to focusing solely on how written text
is conventionalized.

Although the next paper restricts its coverage to the last decade and to
issues relating to the next decade, we should keep an eye on the wider
perspective as well. The potential for applying the computer to Chinese
processing issues did not escape the attention of pioneering scholars. In the
1960s, both the Phonology Laboratory and the Project for Machine
Translation and General Automated Linguistic Systems at UC Berkeley
(later merged to form the Project on Linguistic Analysis, POLA) worked
on various aspects of computer processing of Chinese, including machine
translation. Results were published in a COLING proceedings (e.g.,
Dougherty 1969) and in Scientific American (Wang 1973). Professor T’sou
brings his privileged perspective as a member of that pioneering group as
well as from his forty years of continuous research on Chinese language processing.

The rejuvenation of interest in Chinese language processing in the 1980s has a lot to do with the availability of personal computers in a Chinese environment. The accessibility and representation problems of Chinese characters that dogged the early Berkeley group were given a scalable solution by Bangfoo Chu’s group working out of “garages” in the Tāibēi suburbs. The introduction of the PC brought affordable computers to the general public, while Chu’s CANJIE made the PC accessible to a Chinese market. PCs with Chinese capability led to a boom in the information industry among the “Chinese dragons” (China, Tāiwān, Singapore, Hong Kong). Chu has brought to his panel the perspective of one who has worked with Chinese language processing both as a tool for modernization and for profit.

The next breakthrough came when Huang, Sproat, and Su led Chinese language processing into the mainstream of computational linguistic studies in the late ‘80s. In the ‘90s, CLP research became industrialized as Microsoft established MSRA in Bēijīng as its second global research center, with Ming Zhou among the first group of researchers involved.

The members of this panel represent these stages in the development of CLP, and their views integrated in this paper herald a new age for Chinese language processing.

References


Chinese Language Processing
at the Dawn of the 21st Century

Benjamin K. Tsou

1. Overview

European languages have not had to confront the three major hurdles that Chinese language processing (CLP) has to deal with: (1) input problems because of the use of logographic or morpho-syllabic Chinese characters in a non-alphabetic writing system; (2) word segmentation problems because of the prevalence of long continuous character strings (rather than tokenized words) and general absence of overt word morphology; and (3) syntax, again because of severe paucity in morphology.

In the 1960s, great efforts went into basic CLP in the PRC and the USA. Both countries were engaged in the development of Machine Translation: between Chinese and Russian in the PRC, and between English and Russian in the USA. The rise to prominence of generative-transformational grammar was to provide another boost to these efforts. While the ALPAC report put a damper on this line of research in the West for more than two decades, CLP developments in Chinese-speaking communities continued, although no large-scale operational systems were available for production runs because of (1) the universal limitation on computing power from the 60s to the 80s, and (2) the specific linguistic problems of Chinese.

The first hurdle was overcome when the CPU went from 8 to 16 bits, and more recently from 16 to 32 bits. This development has meant more wholesome and esthetically acceptable characters could be produced when research results in computational linguistics could be fruitfully harnessed to facilitate the final selection of characters. It has allowed the replacement of the use of cumbersome four-digit telegraphic codes by human operators for input, or the large and unwieldy multi-layer keyboards, by a small number of keys available on the QWERTY keyboard. The second hurdle is
not so daunting, but is still in the way. This is because the tokenization of words for continuous text in Chinese encounters more obstacles than English, for example, because written Chinese uses no capital letters for proper names, or, German, for another example, where nominal compounds are spelt together. However, extensive developments utilizing massive language databases and other linguistic resources such as labor-intensive lexical and semantic analyses, together with rapid expansion of computing power, have led us to the brink of new breakthroughs. This is especially true when we look, for example, at the proliferation of applications ranging from speech products to the development of the Internet and PC-based search engines and semi-automatic systems involving queries and translation. The third hurdle remains as a major challenge primarily because POS tagging in Chinese can rely on little or no case markings, which therefore cannot be used to facilitate the necessary syntactic analysis. It is likely that spectacular results would become much more difficult to obtain, and investment in economic terms could be seen to have reached a point of diminishing returns, just as user requirements become much more critical than before. However, this is not the case with basic research as new and specific efforts have enhanced our understanding of the Chinese language with increasingly more efforts on empirically based fundamental studies. When these same developments become increasingly successful in attracting the interests of leaders in the IT industry while the remaining obstacles are confronted, they will accelerate developments to close the gap in the maturing technology. The fact that China has not only recently entered WTO but is quickly taking its rightful place in an IT-oriented world will provide additional impetus, especially when strategic significance will be boosted by commercial importance. This development will be important not just within and across the vast range of Chinese speech communities, the largest in the world, but for linguistic mediation among the Chinese speech communities as well as between Chinese and other language communities.

The SigHAN panel consisted of established figures in CLP ranging from applied R&D to basic research, and consumers of CLP products. (See Appendix). The phenomenal developments of the last ten to twenty years have allowed the panelists to each review his work and to look into his crystal ball for the next decade and reveal what he sees as the new
breakthroughs in basic and applied research in Chinese IT at the threshold of the Twenty-first Century. Following a long-standing, concerted government policy to promote successful efforts to become a global player in the IT world, Taiwan already has a comprehensive e-Taiwan vision, and Hong Kong has introduced an e-Government initiative. All this will certainly accelerate development. But undoubtedly the 2008 Beijing Olympics, being touted as a “digital Olympics”, will provide a showcase for CLP achievements, as comparisons will be made with advances in applied natural language processing shown off in Japan and Korea during the World Cup Soccer Finals held in 2002, and at other international sporting events like the Winter Olympics.

And so the panel has focused on possible developments within the next ten years or so; not an easy task considering the pace of developments in the cyber world in the preceding decade.

This report is based on the wide-ranging discussions at the Panel meeting arising from (1) views expressed by two opening speakers of the Panel (Chu and Huang), (2) views expressed by discussants (Sproat, Su & Zhou), and (3) general views expressed by the other participants, including the Panel Chair. Part 2 of the report is a recapitulation of some recent milestone developments, and Part 3 looks forward to the next ten years.


The marketing of affordable, desktop personal computers in the ’70s and ’80s resulted in an urgent need for a simple, non-technical way to input Chinese characters. A broad range of methods was made available, but the most popular turned out to be the CANG-JIE system introduced by Chu BongFoo (朱邦復) in 1976. (Cāng Jié (倉頡), legendary inventor of Chinese writing, was said to have been inspired by the use of knots on strings to convey messages.) What perhaps made the CANG-JIE system, which requires regular practice to maintain the acquired skills, most successful in terms of gaining wide acceptance immediately was that Chu made the noble and admirable decision to make his system available to the
public without making monetary claims for the associated Intellectual Property Rights. This system drew on traditional etymology and folk knowledge on the composition of Chinese characters common among Chinese and enables the well-practiced user to gain speed in character input comparable to English word processing.

This success has emboldened Chu, ever the intellectual maverick, to seriously begin working in the 1990s on a computer Operating System based on the logographic Chinese writing system rather than the established Western alphabet-based system.

The availability of the CANG-JIE and similar input systems laid the foundation for meeting the pent-up demand for massive text processing in the ’70s and ’80s, beginning with word processing, which briefly allowed WANG Computer to become a major global player, and thus allowed CLP to gain popular exposure and to really launch the IT age for Chinese. The increasing demand also expanded to cover other developments such as production of increasingly versatile electronic dictionaries and more central natural language processing, such as Machine Translation systems. These systems drew mostly on the linguistic intuition of their creators and were mostly rule-based. Their performance was spectacular at first blush, but by no means acceptable.

The most significant trend at that time was probably the shift towards recognition of the need for massive language databases. Such databases, instead of the abstracted generalizations based on a small number of individuals, would provide the most reliable and practically accessible basis for language use and realistic norms on which the machine could model.

This development brought to a head the problem of the segmentation of Chinese texts from continuous strings of monosyllabic characters into words which generally average about between 1.5 to 1.7 characters per word depending on the segmentation standard and database. The availability of massive raw texts gave rise to the use of statistical methods, which also became increasingly sophisticated on the basis of experience gained from English and other languages. But this massive textual material is no substitute for processed texts which minimally would have to be properly and consistently segmented. The requirement on consistency has
led to the need for agreed standards on what constituted words in Chinese and for the recognition of sub-norms for the three recognizable varieties of the Chinese language used in the PRC, Taiwan, and Hong Kong. Some clear and standardized norms have begun to emerge since the 1990s, as well as authoritative segmented corpora.

The following is a simplified chronology initially based on the perception of one of the panelists, Chu-Ren Huang in Taiwan, with supplementation by the author.

1992 –Corpora
Completion of the first Chinese corpus for linguistic research (Huang and Chen, COLING’92:1214-1217). The corpus contained untagged and non-segmented material, but searchable.¹

1992 –Segmentation Standard
Announcement of the first national standard for word segmentation by the Government in Běijīng. [GB 13715-信息處理用現代漢語分詞規範].

1993 –Lexicon
Completion and Release of the first version of CKIP lexicon (with the category set and ICG thematic roles; first version of K. Chen’s parser for Chinese).

1994 –Corpus
a. 10th anniversary for the Automation of Chinese historical textual databases; completion of the pre-Qin Classical Chinese corpus at Academia Sinica.
   b. Establishment of the Language Information Sciences Research Centre at the City University of Hong Kong.

¹ This achievement, though perhaps pale by current standards, represented a great leap forward compared to an earlier effort in computerized concordance for Chinese in 1976 (see T’sou 1976).
1995—Corpus
Completion of Sinica Corpus (v. 1.0 1 million words), the first balanced and tagged Chinese corpus.

1996
a. 10th Anniversary of the Institute of Computational Linguistics at Peking University.
b. 10th Anniversary of the CKIP (Chinese Knowledge Information Processing) Group at Academia Sinica.
c. First Anthology of Papers:
Readings in Chinese Natural Language Processing (Journal of Chinese Linguistics Monograph No.9) (Editors: C R Huang, K J Chen, and B K T'sou).
d. Sinica Corpus on Web, the first fully searchable Chinese language corpus on the WWW (the original website: http://www.sinica.edu.tw/ftms-bin/kiwi.sh; current website: http://www.sinica.edu.tw/SinicaCorpus/).

1997
a. May—The LIVAC corpus from Hong Kong on Web—a dynamic Chinese corpus (Linguistic Variations in Chinese Communities) consistently and synchronously drawing parallel Chinese media material from Bēijīng, Hong Kong, Macau, Shanghai, Singapore, and Taiwan initially with 7 million characters of segmented text becoming accessible through the web. (current website: www.LIVAC.org).
b. Publication of the first Chinese dictionary compiled directly from a corpus (The Dictionary of Nominal Classifiers in Chinese).
c. The Tenth Annual ROCLING conference.

1998
a. KnowledgeNet
b. Segmentation Standard  
Official announcement of CNS14366 for Taiwan.

1999  
Second Anthology of Papers:  
*Quantitative and Computational Studies on the Chinese Language*  
(Editors: B K T’sou, T B Y Lai, S W K Chan and W S-Y Wang),  
(LISRC Monograph on Computational Linguistics, City University of Hong Kong).

2000 –Treebanks  
a. July – Peking University releases its word lists.  
b. Simultaneous completion and announcement of two Chinese Treebanks:  
   *Penn Chinese Treebank*  
   *Sinica Treebank*  
c. The 38th Association for Computational Linguistics (ACL) Annual Conference, held in Hong Kong for the first time, included a workshop on Chinese Language Processing.

2001 –Association formed and more corpus activities:  
a. Formal approval for the formation of ACL SigHAN, the first international organization on Chinese Language Processing.  
b. May – Peking University has made available 1 year of text of the *People’s Daily* through its website.  

2002  
a. The 19th International (Biennial) Conference on Computational Linguistics, (COLING2002) held in Taipei for the first time, included a Steering Committee meeting for the establishment of the Asian Federation of Natural Language Processing (AFNLP)
and the first international conference to launch it.
b. Formal launch of Hsieh’s Intelligent Character Encoding System
   (a solution to the missing character problem).
c. The 2nd SigHAN Workshop on Chinese Language Processing to be
   held in conjunction with the 41st ACL in Sapporo in 2003 would
   report on the results of the 1st International Competition on
   Chinese Word Segmentation Bakeoff. The competition drew on
   four corpora for benchmarking: (i) Academia Sinica (Taipei), (ii)
   LIVAC (Hong Kong), (iii) Peking University (Beijing), and (iv)
   University of Pennsylvania Tree Bank (Philadelphia).
d. The LIVAC Corpus reached 100 million characters.

3. The Next Decade (2002 to 2012)

The views expressed at the Panel focused on: (1) user demands and
expectations relating to CLP in five years and then ten years, and (2) the
specific priorities set for overcoming the remaining hurdles.

3.1 Computer technology after 10 year

According to Moore’s law, the speed, memory, hard disk, and
bandwidth of the computer after ten years would increase thirty times, while
the computer would shrink in size and weight. Moreover, the computer
would be the center of computation, communication, control, entertainment,
involving, for example:

a. Telephone, game machine, learning, home control, Internet
b. Any time, anywhere, any device
   c. Multi-modal input: speech, pen input, keyboard, gestural and face
      recognition
   d. E-business, internet transactions in daily life
   e. Much easier information access such as personal newspaper and
      customized news clipping, notification, information retrieval…
   f. STT/TTS email will be popular
g. Popular cross-language functions which will include Translation, cross-language inquiry, retrieval
h. Automatic phone/e-mail helpdesks
i. More consumer-friendly natural language user interface

3.2 Chinese Computer Applications in 5 Years

They will have the biggest market in the world because:

a. A majority of adult Chinese will have a mobile phone. There would be strong demands for improvements in Chinese mobile computing (which became available five years ago) with respect to ease of input and maximal short message display
b. Chinese internet users could reach 500 million, or one fourth of web-users will use the Chinese Language

c. Nearly all Chinese households will have a computer
d. Chinese webpage number will be only second to English; Chinese and English will be most significant Internet language, with about 15% to 20% of web content originating in the Chinese Language
e. Need for multi-modal input: Chinese pen input, speech, keyboard
f. Demand for Chinese-English, Chinese-Japanese Machine Translation products will reach unprecedented new heights
g. Chinese language learning and teaching software will be in great demand
h. STT, TTS needed as standard features of computers
i. Increasing demand for mobile phone translator, information access – depending on quality
j. Enhanced content screening and cyber monitoring for strategic and commercial purposes

---

2 In an urban setting such as Hong Kong, there are 1.1 mobile phones per person for a population of nearly seven million.

3 It may be noteworthy that even though major ISP’s such as Google (and Yahoo) have had less than five years of history in allowing access to search engines via Chinese, it has become a phenomenal growth area.
3.3 Infrastructural Support in 5 Years

a. Chinese WordNet (CWN): At least two versions will be completed. CWN will utilize various levels of semantic information encoded in the Chinese language, including, but not limited to: components (部件), radicals, roots (in compounds), and classifier constructions such as those proposed by Hsieh and Chuang’s intelligent character encoding scheme will make missing character no longer a problem, and will create an open (i.e. efficient resource sharing) environment for character fonts,

b. Segmentation: New and successful hybrid models where segmentation and POS tagging are integrated and not modularized, with add-on modules to cater for domain and local needs in better recognition and tolerance of linguistic diversity within enhanced social harmonization,

c. Tagging: more sizable corpora available for benchmarking NLP tasks such as segmentation, POS-tagging, and syntactic analysis

d. Successful implementation of various digital archive projects, e.g.:
   (i) Massive and comprehensive diachronic Chinese corpora covering nearly 3000 years will enhance research on the diachronic developments of the language
   (ii) Serious attempts at archiving and documentation of minority languages in different parts of China and including Austronesian language in Taiwan (before more endangered languages become extinct because of modernization and globalization).
   (iii) Additional successful attempts to include historical and geographical dialects (such as a digitized language atlas of China) will gain increasing significance in an expanding milieu of social consciousness and, like the other databases, will become an essential source for the study of social and evolutionary development because of the uniquely long and continuous written tradition of China.

e. With the Semantic Web slowly becoming a reality, most Chinese
web services and pages will use semantic features of the language as a skeletal part of their ontology. The semantic information is realized from components, radicals, root meaning, classifier collocation etc. Other language sites will also adapt the skeletal ontology of Chinese.

f. There will be authoritative bilingual resources relating to at least the domains of finance, law and technology.

g. There will be comprehensive as well as domain specific, chronologically as well as geographically localized, named-entity resources, maintained dynamically.

h. More and updated global properties of the Chinese language such as character and word entropy will be available on the basis of massive corpus analysis

3.4 In 5 to 10 Years

a. The population of computational linguists whose work involve Chinese could exceed computational linguists working on any other languages.

b. Code-switching (between Chinese and English) will become even more popular. English words (either transcribed or in alphabet form) could take up more than 10% of a random corpus in some domains. This would naturally facilitate segmentation and POS-tagging, and thus could be fully exploited to overcome some of the remaining CLP difficulties, as has been the case with Japanese language processing where the combined use of Kanji (Chinese characters), Katakana, and Hiragana in written texts could be fruitfully exploited.

c. Super large annotated corpus will be available.

d. High performance integrated system of Chinese word segmentation + Named Entity Identification (NEI) + POS tagging: Word segmentation ambiguity resolution would be improved by 50% when compared with the early 2000’s, human and geographical NEI should achieve better than 90% success rate.

e. Open source Chinese parser will be available
f. Chinese NLP could surpass other Asian Languages as the hot topic in the world

g. Parsing technology
   (i) Hybrid of statistical shallow (chunk) parser and rule-based dependency parser
   (ii) Improved statistical approach, including unsupervised learning, for Word segmentation, POS tagging, Named Entity Identification, Base phrase identification
   (iii) Rule based approach will be enhanced: long distance analysis, semantic role determination, sentence structure determination

h. Information Retrieval Technology
   (i) Search engines for site search/domain search
   (ii) Information service platform will be ready: covering named entity identification, information extraction, automatic indexing, summarization, classification
   (iii) Personal information service
       Cross-language information access, collection, reading, abstraction, and evaluative summary

i. New Theory on CLP, and enhanced portability and cost-effectiveness in CLP

The panel discussions also identified areas requiring considerable attention in the second decade: a. Linguistic resources (Corpus/lexicon cultivation), b. Evaluation platforms (e.g. Chinese-DARPA), c. Chinese Named Entity Identification, well articulated with non-Chinese sources, d. SLM (Structured SLM – long distance dependency), e. Information extraction (Unstructured to structured), f. Machine learning (Finer knowledge learning, Large scale knowledge learning [system level], Learn FSTs and their probabilities, CFG, translation models, word classes), g. Information retrieval and QA (Specific domain/site Structured information Any time, any software), h. Machine translation (Adaptive, customized EBMT, SMT).
4. Conclusions

Shortly after the panel was held in 2002 and while the report was being compiled, already some of the predicted developments have come true. They included the first SigHAN sponsored international competition on the segmentation of Chinese texts, the popularization of convenient and economical short messaging facilities in Chinese via the mobile phone, as well as the availability of the low cost Chinese Tablet PC, which contains sizable texts, and would allow Internet access to such archives and other information.

When Chu BongFoo suggested in his opening remarks that the day would not be far for wrist-worn computers to be operating in Chinese and for the peasants in the vast rural areas of China to gain easy and convenient access to the Internet via Chinese, there were mixed reactions among skeptics and enthusiasts. It remains to be seen in ten years not whether, but how these and other glimpses of further CLP might become reality, and how the consequential developments will impact on Chinese society in a global context.

4 The author wishes to thank the panel members as well as the participants for their input and lively exchange of ideas. Because of various constraints, it has not been possible to refer to the panelists for detailed feedback on this report, and credit for ideas has been attributed to the panel which discussed them as a whole rather than to individuals. He is especially grateful to Chu-Ren Huang for further useful contributions in the subsequent write-up, and to O Y Kwong and B Y Lai for assistance. Some of the research reported here has received support from a Competitive Earmarked Research Grant (CERG) (Project No. CityU 1241/02H) from the Research Grants Council of the Hong Kong SAR.
References


Copenhagen, Denmark.

**Websites**

HowNet
http://www.keenage.com

SigHAN
Http://www.sighan.org/

Sinica Corpus
http://www.sinica.edu.tw/ftms-bin/kiwi.sh (original website in 1996)
http://SinicaCorpus.sinica.edu.tw
http://BOW.sinica.edu.tw
LIVAC Corpus
http://www.livac.org
http://www.livac.org/celebrity

Appendix

2002 The 1st International Competition on Chinese Word Segmentation Bakeoff

Panelists: BongFoo CHU 朱邦復 (Culturecom, Macau)
Chu-Ren HUANG 黃居仁 (Academia Sinica, Taipei)

Discussants: Richard SPROAT 史伯樂 (AT&T Labs)
Keh-Yih SU 蘇克毅 (Behavior Design Corporation, Taipei)
Ming ZHOU 周明 (Microsoft Research Asia)

Moderator: Benjamin K T'SOU 鄒嘉彥 (City University of Hong Kong)
Rapporteur: Olivia KWONG 鄺藹兒 (City University of Hong Kong)
Tom LAI 黎邦洋 (City University of Hong Kong)

* * * * *

Biographic Notes

CHU, BONGFOO became interested in the origin and special aspects of
the Chinese writing system while living in Brazil after having studied
and worked in Taiwan. He has devoted much of the last two decades
and his earnings in the pursuit of an ideal world for Chinese IT, first
by his popular CANG-JIE input method for Chinese and more recently
by working in Macau as an intermediary among IT communities on
both sides of the Formosan Straits.

HUANG, CHU-REN obtained his B.A. from National Taiwan University
and his Ph.D. from Cornell University. He has specialized in lexical
semantics and is one of two key scholars who have played a pivotal role in the development and construction of the Sinica Corpus. He was also the Chair for the local organizing committee of COLING 02 in Taipei.

**SPROAT, RICHARD** obtained his B.Sc. and Ph.D. from MIT and has been working in computational linguistics with Chinese as a target language. He recently joined the Faculty at the University of Illinois, Champagne-Urbana after having worked for many years at Bell-lab and AT&T.

**SU, KEH-YIH** obtained his B.Sc. in National Taiwan University and his Ph.D. in Computer Sciences at the University of Washington, Seattle. He was on the Faculty at National Tsinghua University, Hsinchu, before joining BDC where he heads its efforts in Machine Translation.

**Zhou, Ming** studied Computer Science at the Harbin Institute of Technology where he obtained his PhD. He did research on NLP at the Department of Computer Science of Tsinghua University before joining Microsoft Research Asia in Beijing and subsequently heading its NLP section.
Biographic Sketch and Research Outline of Author

Benjamin K T'sou, (M.A. Harvard University, Ph.D. University of California, Berkeley), member of the Royal Academy of Overseas Sciences of Belgium (elected 2001), is Chair Professor of Linguistics and Asian Languages, and Director of the Language Information Sciences Research Centre of the City University of Hong Kong, where he was previously Pro-Vice-Chancellor (Research and Development).

He also taught at the University of California, San Diego and University of Hong Kong, and held visiting professorship or other research positions at Australian National University, Běijīng University, University of British Columbia, Vancouver, University of California, Berkeley, Chulalongkorn University, École des Hautes Études en Sciences Sociales (Paris), University of Melbourne, and Nanyang Technological University.

Since 1995 he has developed and cultivated the largest synchronous corpus of Chinese (130 million Chinese character, http://www.livac.org) and has continued with research in Natural Language Processing involving summarization and more recently categorization and other aspects of text processing which he began in the 1960s under Victor Yngve at MIT. His recent research interests have focused on the extent and nature of linguistic differentiation on the basis of very large corpus, which has led him to look at the innovation and diffusion of lexical structure and a re-examination of entropy in Chinese and other languages. He has authored several books and monographs, and over 100 articles covering the above areas from the perspectives of qualitative and quantitative studies. Among these is the Language Atlas of China, of which he is a general editor and contributor. It was awarded a Class I prize for social science research in the PRC in 1998.

(Anthology on Linguistics) 北京大學 (Peking University) and he was the Founding Editor-in-Chief of the International Review of Chinese Linguistics (IRCL) [國際中國語言學評論].

He also serves on the Executive Board of the Chinese Information Processing Society of China 中國信息學會, and is President of the Chinese Language Computer Society, and of the Asian Federation of Natural Language Processing.
Computational Linguistics and Beyond

主 编：Chu-Ren Huang (黃居仁)
Winfried Lenders
出版者：中央研究院語言學研究所
發行者：中央研究院語言學研究所
台北市南港區研究院路二段 128 號
電話：(02) 2786-3300
網址：http://www.ling.sinica.edu.tw

印刷：長紅印刷事業有限公司
初版：中華民國九十三年七月
定價：新台幣 300 元/US$12 (精裝)
        新台幣 250 元/US$10 (平裝)
        國外郵寄及手續費每本另加美金 10 元
        Please add US$10 handling charge for each volume.

ISBN  957-01-7609-1 (精裝)
       957-01-7610-5 (平裝)
GPN   1009301959