



# MT News International

Newsletter of the International Association for Machine Translation

## In this issue . . .

Spotlight on the News . . . .	1
Conference Reports . . . . .	4
Speaking of MT . . . . .	10
Association News: AAMT. . .	12
Conference Announcements	20
Resources . . . . .	22
Calendar. . . . .	23

## Of special note . . .

- **Maghi King reviews Huitchins' "Early Years in MT"** (page 16)
- **MT Users' Desiderata** (page 17)
- **AMTA 2002 Call for Papers** (page 14)
- **Profile: Junichi Tsujii** (page 13)
- **Chats with Michael Quinlan and Jaap van der Meer** (page 10-11)

## Spotlight on the News

### 5th AMTA Conference — From Research to Real Users Tiburon, California, USA - October 8-12, 2002

#### Call for Papers, Tutorials, Workshops, and Exhibitors

The Association for Machine Translation in the Americas (AMTA) is pleased to announce its fifth biennial conference. The main conference program (October 10-12) will be preceded by a day of focused workshops (October 8), and a day of tutorials (October 9). Submission of papers to be presented in the main conference program, or proposals by individuals who would like to conduct workshops or tutorials are solicited now. Please note that all IAMT members will be entitled to the AMTA member registration rate.

#### Call for Papers

The program committee seeks submissions of three types: theoretical papers, user studies, and system descriptions/demonstrations. Submissions focusing on key questions posed in the call for papers are particularly encouraged: Why aren't any current commercially available MT systems primarily data-driven? Do any commercially available systems integrate (or plan to integrate) data-driven components? Do data-driven systems have significant performance or quality issues? Can such systems really provide better quality to users, or is their main advantage one of fast, facilitated customization?

If any new MT technology could provide such benefits (somewhat higher quality or facilitated customization), would that be the key to more widespread use of MT, or are there yet other more relevant unresolved issues, such as system integration? If better quality, customization, or system integration aren't the answer, then what is it that users really need from MT in order for it to be more useful to them?

The detailed requirements and submission procedure for each of the three categories

*Continued on page 20 ►*



# MT News International

Issue No. 30 (vol. 11, no. 3)

April 2002

## EDITOR-IN-CHIEF:

**Laurie Gerber**

Tel: +1 (619) 200-8344

Fax: +1 (619) 226-6472

E-mail: [mtni@eamt.org](mailto:mtni@eamt.org)

## CONSULTING EDITOR:

**John Hutchins**

E-mail: [info@eamt.org](mailto:info@eamt.org)

## CONTRIBUTING EDITOR:

**Colin Brace**

Fax: +31 (20) 685-4300

E-mail: [webmaster@eamt.org](mailto:webmaster@eamt.org)

## REGIONAL EDITOR, AAMT

**Hitoshi Isahara**

Fax: +81-774-95-2429

E-mail: [isahara@crl.go.jp](mailto:isahara@crl.go.jp)

## REGIONAL EDITOR, AMTA

**David Clements**

Fax: +1 (619) 490-3684

E-mail: [David.Clements@lhsl.com](mailto:David.Clements@lhsl.com)

## REGIONAL EDITOR, EAMT

**Jörg Schütz**

Fax: +49 (681) 389-5140

E-mail: [joerg@iai.uni-sb.de](mailto:joerg@iai.uni-sb.de)

Copyright 2002 by IAMI. Permission is hereby granted to reproduce articles herein, provided that they appear in full, and are accompanied by the following notice:

"Copyright 2002 International Association for Machine Translation (IAMT). Reprinted, with permission, from the IAMT newsletter, Machine Translation News International, issue #30, March, 2002."

Electronic copies available upon request: [mtni@eamt.org](mailto:mtni@eamt.org)

## SDL Acquires ALPNET and Language Partners

SDL International has again been making the news for its recent acquisitions. On December 13, 2001 it released a joint announcement with ALPNET of a merger between the two localization giants. Both are publicly traded corporations, ALPNET on the American OTC exchange, and SDL plc on the London stock exchange. The merger agreement, now completed, enabled SDL to acquire all of the outstanding shares of ALPNET at \$0.21 per share for all of the outstanding common shares of ALPNET. The merger agreement provided that ALPNET shareholders would receive the same price paid in the tender offer in cash. The merger will take SDL from approximately US\$50 million in annual revenues, to US\$100 million. More recently, SDL also announced that it has bought Language Partners, a U.S. value-added reseller of MT, computer-aided translation and localization software in the U.S.

ALPNET'S precursor company, ALPS, was an early commercializer of both machine translation and translation memory technology developed at Brigham Young University in the United States.

The March 2002 issue of Multilingual Computing includes an 8-page ad for SDL, introduced by SDL's Chairman and CEO, Mark Lancaster, describing the merger, additional service capabilities, and presenting ALPNET technology products NETXchange (globalization management tools) and GLOBLIX (translation memory), alongside SDL's home-grown globalization tool SDLWebFlow, and translation memory tool SDLX. The section on enterprise products concludes with an intriguing description of an integrated TM+MT system: "SDL is currently working on the integration of the two major translation technologies... Translation memories contain 'human quality' translations that allow the intelligent updating of language rules and dictionaries for MT. By using MT techniques, TM can intelligently pick translations from the database, based on "non-

inflected forms." In addition to this complementary functionality, SDL's new system will also track regularly translated words, phrases and sentences that are not currently in the TM database or MT dictionary. Users can then download these words, phrases and sentences and 'teach' the system how they should be translated. Using this 'learning' approach allows the quality of the translations from the new system to steadily improve over time - effectively becoming a knowledge-based translation system."

Such a description sounds great, with real potential to increase both quality and usability. But is it really part of the development plan? Or the fruit of a good imagination?

MTNI spoke to SDL's Hedley Rees-Evans in the last issue (#29) about SDL's recent purchase of the Transcend MT system from Transparent Language. This issue, we speak with Jaap van der Meer, who was president of ALPNET from 1995 until January of this year, and to Michael Quinlan, president of Transparent Language.

For more information, contact Mark Lancaster, Chairman and CEO of SDL International: [mlancaster@sdlintl.com](mailto:mlancaster@sdlintl.com); Tel +44-1628-410100; or see the website: [www.sdlintl.com](http://www.sdlintl.com) □

## EUROMAP: Promoting Language Technology

The EUROMAP LT project aims to provide awareness, bridge-building and market-enabling services to boost opportunities for market take-up of the results of national and European HLT RTD projects.

The project is currently implemented by a team of eleven National Focal Points in Austria, Belgium/Netherlands, Bulgaria, Denmark, Finland, France, Germany, Greece, Italy, Spain and United Kingdom. Each NFP will draw on the skills and knowledge gained in previous HLT awareness-raising actions to achieve the following objectives: Increase the number of projects that deliver ready-for-market results. Accelerate awareness of the benefits of HLT

## International Journal of Translation MT Issue

The 2001 issue of the IJT (Vol. 13, No. 1-2) is a special issue on Machine Translation, guest-edited by Michael Blekhman of *Linguistica 98*. The journal contains a collection of 17 short articles on MT from developers, users, and researchers around the world. Topics are wide-ranging, and include an account of Language Engineering in the USSR from the 1950s through the 1970s, editorial perspectives "Is it worth learning translation technology?", and practical guidance like "Machine Translation Methods: Text Structure and Translator Work". Authors include John Hutchins, Michael Blekhman, Sergei Nirenburg, Joseba Abaitua, and many others. Copies of the journal are available for US\$ 90 from *Bahri Publications*, 997 A/9, Gobindpuri, P.O. Box 4453, Kalkaji, New Delhi 110019, India. Email: [bahrius@vsnl.com](mailto:bahrius@vsnl.com); Website: [bahripublishations.com](http://bahripublishations.com) □

enabled systems, services and applications within user sectors, policy makers and national administrations. Boost the number of best-of-class technology developers participating in research projects. Improve the relevance of project targets and technology supplier/user needs. Improve the match between HLT design and supplier/end user expectations. Facilitate user partnerships and communities for beta testing, demonstration, real-time utilisation monitoring and other close-to-market application activities. In January, ELDA (the European Language resources Distribution Agency, Paris, France) announced that the European Commission has selected ELDA to promote language engineering in France. Since October 2001, ELDA has been the French National Focal Point.

For success stories, news, events, and more see: [www.elda.fr/proj/euromap.html](http://www.elda.fr/proj/euromap.html) and [www.hltcentral.org/euromap](http://www.hltcentral.org/euromap). □

## PAL—Professional Association for Localization

Since 1995, LISA (Localization Industry Standards Association) has been virtually the only professional forum for the localization community. As such it attracted vendors and clients alike to conferences in which best practices are developed and discussed, projects reviewed and critiqued, war stories shared. Now LISA is joined by PAL which has a somewhat different focus. Unlike LISA, which emphasizes corporate memberships and institutional participation, PAL has only a US\$75 individual membership category.

The mission of PAL is to provide its members with "Structured education and training; standardization and best practices; information about translation trends, tools, and career opportunities; discussion forums; and representation in the fields of Localization, Internationalization, and Globalization. PAL intends to serve and represent those who prepare software and documentation for the world. For more information, including online membership application, links to dictionaries, and other information on localization, see: [www.pal10n.org](http://www.pal10n.org) □

## AMTA and ACL Memberships - Discounts and Online Application Forms

The ACL and AMTA have agreed to offer reciprocal discounts of 10% to members of the other organization. Both discount membership agreements are for first-time members and only apply for one (the first) year of membership. The ACL membership application is available online at: [www.cs.columbia.edu/~radev/newacl/membership.html](http://www.cs.columbia.edu/~radev/newacl/membership.html).

Note that AMTA's membership application is also available online (or use the one on page 22 of this newsletter) at [www.amtaweb.org/membership.html](http://www.amtaweb.org/membership.html) □

## New MT Compendium!

Ever since its first release in April 1999, the "MT Compendium", compiled by John Hutchins, has been an invaluable reference for people in the MT world. A completely revised edition is now available for free download to IAMT members, and for a small fee to non-members.

The full title is "Compendium of Translation Software: commercial machine translation systems and computer-aided translation support tools." This comprehensive reference guide to current commercial products and vendors is published by the EAMT on behalf of the IAMT.

The latest version, released in February 2002, has a significant amount of new information: many new systems, many old systems no longer available, changes to existing systems (new language pairs, new platforms, new facilities, etc.)

Prepared with the assistance of Walter Hartman, Hutchins estimates that the rapid changes in the market over the last year resulted in changes to 70% of the entries between its second release in January 2001, and its third in February 2002.

The Compendium is truly an amazing reference. The January 2001 edition included 60 pages of product listings, with full contact information for all vendors. Among the many helpful features of the Compendium, products are classified into one of 8 categories based on the level and type of functionality they offer, for example "electronic dictionary", "MT services", and "translator workstations", each of which is given a clear description at the back of the book. Finally, a language-pair index allows readers to quickly target products and vendors that offer the languages of interest.

For more information, and to buy or download the Compendium, see [www.eamt.org/compendium.html](http://www.eamt.org/compendium.html) □

---

## Conference Reports

---

### ASLIB—23rd Translating and the Computer Conference

*Federico Gaspari*

The 23rd “Translating and the Computer” Conference was held at the end of November 2001 in London. It was an enjoyable, interesting, and very informative event bringing together a wide international audience of professional translators, business managers, software developers, researchers, institution officials, language experts and consultants.

For the past twenty years the focus of the “Translating and the Computer” series organised by Aslib/IMI has been the role played by computer technology in the translation business and related areas. Attracting speakers and delegates working in the industry (e.g. translation agencies, localisation vendors, media companies, multinational corporations, etc.) as well as in the public sector (i.e. international institutions, government agencies, research centres, universities, etc.), this annual Conference has established itself as a leading forum for discussing the impact of technology on language services from a variety of perspectives.

The Conference provided a valuable opportunity for delegates to receive a thorough update on the latest issues and developments relevant to their interests, promoting networking and cooperation with colleagues in a friendly environment. After each talk the audience was given the opportunity to address queries to speakers or voice comments, thus encouraging lively debate and a fruitful exchange of views and experiences.

The keynote speech on day one of the Conference was given by Gregor Thurmair (Sail Labs, Germany), who presented a comprehensive overview of the MEMPHIS (EU-funded) project for the distribution of multilingual knowledge, including e.g. information extraction, summarisation, cross-lingual and multilingual search strategies. His discussion

encompassed applications designed to manage Internet content, such as the search for and identification of useful information and its subsequent delivery in natural language to serve a number of purposes for profiled users.

Within this framework he illustrated the components of a system providing this kind of service in three steps, namely acquisition, transformation into useable format and final distribution of relevant content via a variety of means such as PC, mobile telephone (SMS) or handheld PDA devices. Particular emphasis was laid on the delicate initial stage of identifying informative sources and documents that may be relevant for this service, which relies heavily on automatic language detection and topic identification (e.g. through keywords and other more advanced extraction technology techniques).

Chris Callison-Burch (Amikai Inc., USA) presented an interesting perspective on machine translation (MT) provision. Amikai has developed the concept of “best of breed” MT to provide a value-added service that goes beyond the reselling of output from individual systems judged to offer overall best performance for specific language pairs. The developers at Amikai have written a program that automatically selects the best translation of a particular sentence from target-text versions generated by several different commercial systems. The approach is based on a statistical language model of English derived from a two million word corpus from Internet articles and texts. The strategy aims to maximise the quality and reliability of MT. It rests on the underlying crucial assumption that the most fluent output corresponds to the best translation for a passage of the source text. The speaker reported experiments and evaluations which support this guiding assumption.

Stephanie Schachtl and Audrey Fraser-Tillmann (Siemens, Germany) reported on the effective use of the in-house interactive machine translation system TopTrans, which is currently employed to translate the user guides for Siemens mobile phones and for Gigaset telephones from German into English. The talk focused in particular on the advantages of-

ferred by the implementation of additional features to enhance the quality of the system's performance, thus reducing the need for post-editing to a minimum. Some of these crucial improvements were determined by the adoption of a controlled language environment, by the presence of an interactive syntactic and lexical disambiguation facility and by terminology support.

Carmen Heine (Fachhochschule Flensburg, Germany) described the QUATRE research project that is being carried out by the Department of Technical Translation of Flensburg University of Applied Sciences in collaboration with a number of commercial partners located in northern Germany. The aim of the project, funded by the German Federal Ministry of Education and Research, is to develop an online quality assurance manual for technical documentation and translation processes. Taking into account existing standards, the QUATRE approach provides a modularised framework for companies to develop a standardised workflow which is also tailor-made for their specific needs.

The afternoon session of day one of the Conference, chaired by Daniel Grasmick and Reidar Fischer (SAP AG, Germany), comprised four presentations and a panel discussion devoted to the current hot topics of internationalisation and globalisation. Representatives of four leading companies presented their experience and views: Peter Gottlieb (Uniscape), Liam Garstang (Globalsight), Yann Meerseman (Idiom) and Bjorn Austraart (eTranslate). Each had adopted pragmatic and user-oriented approaches to globalisation and translation of Internet and Web documentation. They had the opportunity to introduce their own products and services to the audience, reporting on projects their companies were involved in. It was the first time in Europe when four key players in the exciting arena of global content management had got together to share their expertise and insights with an audience of professionals. Although their companies have special areas of interest and have different approaches to tackling certain issues, a considerable degree of overlap emerged in their main concerns, in the general principles guiding their strategies and in the philosophy of the services offered. The panel discussion at the end of the session enabled the audience to take advantage of the presence of

the speakers to expand on some of the most interesting issues that had been raised in their presentations.

The second day of the Conference was opened by a talk given by Dieter Rummel (Translation Centre for Bodies of the EU, Luxembourg) and Sylvia Ball (European Parliament Translation Service, Luxembourg). They presented an overview of the work in progress within the IATE project, launched at the beginning of the year 2000. The aim of this initiative is to create a single central terminology database for all the institutions, agencies and bodies of the European Union. The speakers drew attention to the thorny issues entailed in uniting the massive amount of terminology that has been created in different EU institutions over some decades of work. Different approaches to similar problems have resulted in duplication and overlap of effort. The existence of parallel and independent methods for the creation and maintenance of terminology poses daunting challenges in a cooperative undertaking such as the IATE project. The merging of legacy data calls for wide consensus on the adoption of new standardised procedures and on their introduction into the terminology and translation workflow of the participating bodies.

Jon Wells (SAP AG, Germany) described a practical example of combining machine translation and translation memory technology to facilitate the initial and subsequent translation of large volumes of documents on a daily basis in a corporate environment. Thanks to the implementation of this solution and to the flexibility and scalability of both the software and its users, it is claimed that the company has dramatically reduced translation costs.

Kirsty Macdonald (SAP AG, Germany) reported on a series of experiments on the WinAlign program (part of the Trados package), testing it on a collection of German-English parallel texts. Her investigation focused specifically on the accuracy of proposed alignments and on the causes of misaligned segments. Since the quality and accuracy of alignment is crucial to the effective use of translation memory software, she put forward guidelines for the design of a proposed misalignment checking tool.

Steve McLaughlin (SAIL Labs, Germany) described DTS (Distributed Tasks and Services), a client-server system for delivering multilingual and translation services to users within a distributed environment. He concentrated in particular on

some of the scenarios in which the high-level services provided by DTS might be applied, as well as on its modular and scalable architecture and its role in the MALT framework. After addressing some security issues that crop up for any application within a distributed environment, the talk discussed the possible services and future enhancements that could be developed.

Celia Rico (Universidad Europea de Madrid - CEES, Spain) proposed a user-oriented reproducible model for the evaluation of computer-assisted translation tools. The growing importance of using CAT tools in present-day professional translation means that the evaluation of these tools has emerged as a key element in the overall strategy for project management in the translation business. Her model takes into account the main interacting components in the translation workflow, the client, the product, the process and the CAT tool under consideration. For defining system performance ISO 9126 is taken as a starting point. The model is designed for a wide variety of translation needs in different scenarios, e.g. industry, public administration, agencies and freelance translators, but as yet not tested in practice.

Christian Sestier (Linguistique & Technologies, France) presented the eTermino Q&A Multilingual Internet Terminology Assistant. This is a tool which is not intended to replace existing terminology management software, but has been designed as a companion tool for teams of translators and project managers to deal successfully with terminology problems during the course of a multilingual translation project. It consists of a shared terminology list developed and updated interactively by a question and answer process over the Internet. As a case study, the speaker described the typical situation in which a team of freelance translators works on the same project. He illustrated how in this case eTermino Q&A could be integrated successfully with the use of a translation memory tool such as WordFast (from Champollion & Partners).

In the last talk of the conference Antonio Sánchez Valderrábanos (Sema Group sae, Spain) reported on the LIQUID project, which is funded by the European Commission and aims at developing a cost-effective solution to cross-language access to databases containing technical and scientific information. LIQUID's purpose is to design a system that returns documents

relevant to a search topic in all the languages that are found in the document base, irrespective of the language in which the initial query was originally phrased. The project started at the beginning of 2001, and the languages covered are French, Spanish, English and German.

The main challenge lies in organising unstructured textual information according to its information contents and irrespective of language. The solution proposed combines a terminology extraction tool and a domain-specific language-independent ontology. The former identifies the keywords that describe the contents of the documents for indexing, and the latter provides a model of the domain-specific knowledge to which the keywords can be linked. The main goal is to solve problems of term variability by the enrichment of initial term sets, using existing linguistic resources.

Papers of the conference will be available from the Aslib website: [www.aslib.com](http://www.aslib.com), where PDF copies of the papers by Thurmair and Valderrábanos can be downloaded.

*Federico Gaspari (fedegasp@tin.it), Advanced School of Modern Languages for Interpreters and Translators, University of Bologna, Italy.* □

---

## MT2010—Roadmap workshop at MT Summit VIII

Steven Krauwer

The aim of the workshop, held September 18, 2001 in Santiago de Compostela, Spain just before the MT Summit, was to contribute to ELSNET's ongoing action to establish a roadmap for MT for the next decade. A roadmap comprises an analysis of the present situation, a vision of where we want to be in ten years from now, and a number of intermediate milestones that would help in setting intermediate goals and in measuring our progress towards our goals.

The function of the roadmap is not to

*Continued on page 19 ►*

---

# CICLing-2002: Pyramids, Butterflies, and Computational Linguistics

Alexander Gelbukh

*CICLing was held February 17-23,  
2002 in Mexico City. —Ed*

If it were not called Conference on Intelligent Text Processing and Computational Linguistics, CICLing might have been called Everything You Need for Development of Quality MT Systems—at least its computational linguistics part. Just look at the table of contents: semantics, disambiguation, anaphora resolution and generation, syntax and parsing, part-of-speech tagging, lexicon and corpora, text generation, morphology, speech technology.

CICLing is a yearly conference series, of which this one was the third. CICLing by design is a small professional meeting bringing together scientists working on cutting-edge problems of computational linguistics in both their theoretical and applied aspects. This year it was attended by approximately 40 specialists from Bulgaria, The Czech Republic, France, Germany, Ireland, Italy, Japan, Mexico, Poland, Romania, Russia, Spain, Switzerland, Taiwan, UK, and USA. The proceedings of this conference are available as Lecture Notes in Computer Science volume N 2276 from Springer-Verlag.

Invited speakers **Ruslan Mitkov**, **Ivan Sag**, and **Yorick Wilks**, each organized an informal event—a combination of a tutorial and a discussion, in addition to presentation of their papers published in the proceedings. Unfortunately, two other invited speakers—**Nicoletta Calzolari** and **Antonio Zampolli**—could not present their paper in person due to a health problem, though it is available in the proceedings.

The conference opened with a talk by Mexican linguist **Teresa Carbo**, member of Mexican Academy of Sciences. In her speech she emphasized the huge gap that exists between the traditional, pencil-and-

paper linguistics and modern computational technology. Unfortunately, computational specialists often reinvent the wheel, ignoring the facts long ago known in classical linguistics. On the other hand in many cases computational techniques can instantly solve problems that take months of manual work for a linguist. Surprisingly, however, it seems that she and her colleagues had over-estimated the current state of linguistic software, probably expecting from computational linguistics something like what one can see in science-fiction movies.

## Translation systems

Two fully implemented translation systems were presented, both for somewhat “unusual” languages.

**Éva Sáfár** and **Ian Marshall** demonstrated a system that translates written English into British sign language. Sign language translation is a challenging task since the structure of such languages is very different from that of spoken languages: the roots and morphemes it uses are finger, arm, and body movements and facial expressions, as pronouns, spatial locations and directions, etc. The morphology and syntax of a sign language usually have nothing to do with those of the spoken language used in the same region—so language translation is in no way simple transliteration.

The authors use a knowledge-rich interlingual approach. Using existing tools, written English is transformed into a discourse representation structure (developed in the framework of Discourse Representation Theory). From this semantic representation, the sign shapes are generated using an HPSG grammar. Perhaps the generation component is the most interesting part of the system. It was very impressive to watch a nice-looking artificial feminine character signing the sentences as the authors type them in the input window.

**Arantza Casillas** presented a Spanish-Basque multilingual document generation environment. Unlike most of the European languages, Basque is an agglutinative, ergative, subject-object-verb word order language, i.e., both its morphology and syntax are very different from Spanish.

A feature of this system is structured input. A traditional MT system translates text sentence by sentence. However, the

global structure of legal documents can differ significantly between languages: e.g., in one language the date is indicated at the beginning of the document using figures while in the other at the end using words, etc. The authors use the Document Type Definition structure to mark up the logical parts of the source document and a translation memory technique to translate the corresponding pieces of text. The translation memory is trained on a bilingual corpus.

## Knowledge-rich versus knowledge-poor approaches

As usual, the papers presented at the conference could be classified into two practically non-intersecting classes: knowledge-rich and knowledge-poor approaches. The latter approaches are mainly statistically based (data driven): the necessary data are automatically acquired by means of statistical analysis of very large corpora. Also, simple heuristics that work “most of the time” are used. In contrast, the former approaches try to precisely describe everything that is known about the behavior of language units. This usually results in huge manually crafted dictionaries and grammars.

Most of the authors from both sides claimed their approach is better if not the only one possible. In an informal discussion that he organized, **Yorick Wilks** characterized this difference as following: while most academic researchers construct knowledge-rich systems (usually very limited in their domain and/or lexicon), it seems that the only really working (industrial) systems are knowledge-poor. However, pointed out Yorick Wilks, there is a tendency for convergence: “practical” data-driven systems begin using richer data structures and linguistic knowledge, while “academic” approaches try to incorporate statistical information as an important kind of knowledge. In another discussion, **Ivan Sag** assured us that in the near future statistical information would be incorporated into HPSG—one of the most successful knowledge-rich approaches nowadays. On the other hand, **Ruslan Mitkov**, the author of the famous Mitkov’s knowledge-poor approach to anaphora resolution, emphasized the necessity to incorporate more knowledge into anaphora resolution systems.

## Semantics, lexicon, and knowledge representation

Several talks were devoted to the knowledge-rich formalism known as Head-driven Phrase Structure Grammar (HPSG; see [hpsg.stanford.edu](http://hpsg.stanford.edu)) and other constraint and unification based formalisms.

In his invited talk, **Ivan Sag** showed how a wide range of multiword expressions (non-compositional phrases) can be described using various mechanisms of the HPSG formalism. In the informal discussion he presented more examples describing so-called core constructions in HPSG. Most importantly, he mentioned that the efficiency of existing parsers for computational interpretation of HPSG grammars has passed the threshold of feasibility for practical applications in industrial systems. Personally, I recommend that developers have a closer look at this very promising formalism combining the best ideas of many different paradigms. Probably machine translation, where the key issue is accurate transfer of semantics, is the area that can benefit from it most.

The sign language translation system mentioned above, was based on the HPSG description of the sign language. **Ivan Meza** and **Luis Pineda** presented a description of Spanish auxiliary verbs in HPSG. It seems that Spanish has a greater number of auxiliary verbs than English, and HPSG formalism allows to accurately describe the subtle differences in their meaning and usage. A similar representation of Tatar morphology was presented by **Djavidet Suleimanov**.

**Philippe Blache** suggested a constraint-based formalism (similar to soft constraints) that tolerates ungrammatical input: When you parse a text, you are normally not interested in whether it is grammatical or not but instead in what was its intended meaning. In the suggested formalism, there is no issue of grammaticality: what is constructed is the semantic representation, no matter whether the input was correct.

**Karin Harbusch** described an integrated text generation system based on Schema-Tree Adjoining Grammars with unification. The system uses large existing knowledge bases, which makes feasible its practical application. **Hermann Helbig** presented a sophisticated semantic

network formalism for meaning representation, called Multilayered Extended Semantic Network (MultiNet). He emphasized that this formalism is especially suitable as an interlingua for machine translation. Unlike some other semantic network based formalisms, MultiNet is supplied with comprehensive, systematic, publicly available documentation.

Multilingual parallel corpora are of great importance for MT. The method for measuring cross-lingual document similarity presented by **Ralf Stainberger** can be used to search for translation equivalents of documents for compilation of parallel corpora out of existing large document collections. **Alexander Gelbukh** suggested using the Internet to compile a special type of corpora rich in the occurrences of specific words, which is useful for learning collocation or sub-categorization patterns, as well as translation equivalents in some cases. **Arantza Casillas** presented a hybrid approach to aligning multiword terms in a bilingual corpus.

The paper by **Nicoletta Calzolari** and **Antonio Zampolli** discussed in detail the standards for multilingual dictionaries (EAGLES/SLE project). **Igor Bolshakov** criticized Spanish EuroWordNet dictionary for poorly structured and poorly ordered word senses, as compared with three traditional dictionaries.

## Word sense disambiguation and anaphora resolution

These are one of the most important issues in machine translation, and perhaps most understandable for non-specialists. For example, to translate "John took a cake from a table and washed it." into Spanish you have to choose between Spanish mesa 'table: furniture' and tabla 'table: matrix' as well as between lo 'it: masculine' interpreted as referring to pastel 'cake' (washed cake) and la 'it: feminine' interpreted as referring to mesa 'table' (washed table).

Ideally, the best methods of such disambiguation are knowledge-rich ones (such as HPSG). However, again, as was mentioned above, current knowledge-rich methods still don't seem very feasible in practice since the development of the necessary dictionaries is too labor consuming. On the other hand, knowledge-poor statistical and heuristic methods prove to be surprisingly efficient for these

tasks.

Traditionally, word sense disambiguation (WSD) is a hot topic at CILCling. **Ted Pedersen** presented several variants of a simple (he called it "baseline") knowledge-poor approach to WSD. It relies on elementary, easy to identify features; yet it scores within seven to ten percentage points of accuracy of the best existing systems. Though the author discussed it in the context of comparing WSD systems, in my personal opinion this methodology can be an excellent practical option for MT systems, taking into account that the approach was successfully tested across a variety of languages.

In another talk, **Ted Pedersen** presented a different, knowledge-rich approach: a generalized Lesk algorithm, which relies on a WordNet-like thesaurus (any explanatory dictionary can also be used). Unlike the standard Lesk algorithm, which considers each word independently, this modification resolves the ambiguity globally in a context window, which leads to better overall results. **Susana Soler** and **Andrés Montoyo** suggested formulas for measuring semantic similarity in a Lesk-like WSD method. **Armando Suárez** discussed feature selection for WSD.

In many languages, diacritics are used to distinguish meaning, like in Spanish te 'you' and té 'tea'; toco 'touch' and tocó 'touched'. However, people tend not to use diacritics in electronic texts such as email because of software compatibility issues; this presents problems to practical translation systems. **Rada Mihalcea** presented a "zero-knowledge" approach to diacritic restoration: to apply it to a new language one only needs to train it on a relatively small sample of correct text, without any need for dictionaries or grammars. For Romanian, she claimed 98.75% to 99.69% accuracy (depending on the diacritic).

Classical word sense disambiguation relies on the assumption that a word has several clearly defined senses listed in the dictionary. However, the notion of a word sense is quite fuzzy: the intended meaning in a context may—and probably almost always does—significantly differ from any "standard" one, which may seriously affect translation. In his very detailed talk,

*Continued on next page ►*

---

## CICLing-2002

...continued from previous page

Yorick Wilks discussed a method (which he called lexical tuning) for such on-the-fly contextual adjustment of lexical sense and compared this method with some other approaches such as lexical closeness and relaxation, underspecification, and lexical rules.

Part of speech (POS) tagging is the most basic kind of disambiguation. Lourdes Araujo presented an evolutionary (genetic) algorithm that globally optimizes the choice of the POS marks. Dariusz Kogut described a fuzzy set approach. Jesús Vilares discussed tokenization problems in POS tagging. Finally, Héctor Jiménez presented a system for POS tagging in Spanish.

Anaphora resolution (AR) was the topic of the invited talk of Ruslan Mitkov. He presented a fully automatic AR system called MARS. As he pointed out, most existing AR prototypes which are claimed to be automatic in fact crucially rely on manual post-editing of the outcome of the pre-processing modules such as POS tagging or parsing. In his unpublished second talk, devoted to evaluation issues in AR, he again emphasized this difference: one can evaluate either an AR algorithm presented with accurate, manually marked input, or a complete AR system presented with raw, unprepared, randomly selected texts; the results may be very different. As a rich source of training and evaluation data, he proposed to use parallel multilingual corpora: resolving anaphora in the sentence "John took a cake from a table and washed it," is trivial if the parallel Spanish sentence ends in "y la limpió" (not "lo limpió") since *la* 'it' clearly corresponds to the feminine *mesa* 'table' and not to the masculine *pastel* 'cake'.

Jesús Peral discussed the opposite task: generation of anaphoric pronouns (he / she / it / they / them...) in an interlingua-based English-Spanish MT system. Most of the errors in generation of such pronouns, he said, are due to the problems of anaphora resolution or syntactic-semantic parsing of the source text, which thus (no surprise) must be improved in future MT systems.

### Pyramids and butterflies

Why do we go to conferences instead of

reading books? To make friends. And CICLing gave us a unique opportunity to make friends with the best experts in the field. To speak with them in an informal atmosphere. To share memories. Really exciting memories. In fact, half of the time of the conference was devoted to the sightseeing program—which, in common opinion, made the other half much more productive.

The conference began with a full day excursion to the 2000-year-old pyramids of legendary Teotihuacan. In the Aztec language, Teotihuacan means "the place of the Gods". It's really difficult to describe the feeling of illimitable space that one feels in front of these solemn stone giants, among the regular structures of the dead city! When the talks began on Monday, all participants were acquainted with each other and the friendly atmosphere allowed for free discussion; neither speakers nor the listeners felt constrained as is usual with other conferences. The conference deliberately lacked any banquet: instead, the welcome party was combined with the presentation of posters and demos, so that the participants could freely move between the stands and computers sipping good Mexican wine.

Another excursion (in the middle of the conference) was to a unique place in the world—a Monarch butterfly wintering site. Normally, it is a pine wood which looks like a leafy forest due to the millions of butterflies clustering on the branches and covering the trunks and the sky; you can see photos from the past conferences at [www.CICLing.org](http://www.CICLing.org). However, the reports about massive death of the butterflies due to the frosts proved to be true: the forest was covered with dead butterflies, as though with snow. I hope their population will recover by the next CICLing. Even with this, the few trees still covered with live butterflies looked very impressive.

During a half-day excursion to the largest anthropological museum in the world the participants could appreciate the amazing diversity of Mexican pre-Hispanic cultures. Finally, already after the last day of the talks, participants visited yet another archeological site and a very beautiful two-kilometer-long cave with an underground river flowing right out of the mountain. And best of all, in the bus hey had a lot of time to speak,

discuss their work, and make friends!

Alexander Gelbukh is at the Natural Language Laboratory, Center for Computing Research (CIC), National Polytechnic Institute (IPN), Mexico. [gelbukh@cic.ipn.mx](mailto:gelbukh@cic.ipn.mx) □

---

## PACLIC 16

Yong-Beom Kim

The Sixteenth Pacific Asia Conference on Language, Information and Computation (PACLIC 16) was held on January 31-February 2, 2002 at Hotel Green Villa, Jeju-do, Korea. There were about 70 participants from various countries including Canada, Germany, Hong Kong, Japan, Korea, Singapore, Taiwan, and United States. The papers presented at the conference covered wide range of topics in theoretical and computational linguistics. The proceedings from the PACLIC 16 contains 43 high quality papers, which were virtually reviewed by three anonymous referees.

The PACLIC Conference, which is annually held in Pacific Asia, has a tradition of aiming at the mutual advancement of theoretical and computational linguistics. In PACLIC 16 this tradition has been well exemplified by two keynote speeches by Erhard Hinrichs and by Kiyong Lee. Hinrichs talked about syntactic annotation of corpora and memory-based parsing. He combines symbolic parsing with finite-state methods with memory-based parsing in order to perform morpho-syntactic and syntactic annotation as well as annotation of function argument structure for the Tübingen corpus of spoken German and German Reference Corpus. Kiyong Lee's lecture focused on establishing database semantics as a model of computational semantics combining it with Hausser's time-linear syntax. Lee demonstrates his theoretical constructs by analyzing some constructions such as semi-free word order, conjunction, temporal anchoring, adnominal modifica-



tion and antecedent binding.

Roughly speaking, during the Conference, 22 papers dealt with computationally oriented natural language applications; many of them dealt with machine translation, parsing, and computational semantics, and they used language corpora of various kinds. In the area of theoretical linguistics about 11 syntactic papers, 10 semantic papers and one phonology paper were presented.

Subsequent to the PACLIC 16, two other academically related events were held in the same place. One is KORTERM, an Annual International Roundtable on Terminology and Language Resource Management and the other is ISO/TC37/SC4 Preliminary Meeting. These meetings were attended by many internationally renowned scholars including Christian Galinski (Austria), Allan Melby (U.S.A.), Harry Bunt (Netherlands) and Laurent Romary (France). In sum, the whole series of academic gatherings have certainly helped to bring about the mutual understanding of theoretical and computational linguistics in this region of Asia.

*Yong-Beom Kim, Kwangwoon University, Korea. Chair, Organizing Committee of PACLIC 16. ybkim@daisy.gwu.ac.kr*



## European SAE J2450 Meeting – Translation Quality Metric

Jörg Schütz

*A taskforce to establish an industry-wide translation quality metric was established in 1997 in the U.S., consisting of representatives of major car makers and translation suppliers. An SAE Standard documenting the taskforce's findings was published in December 2001, and is available from the SAE website, mentioned at the end of this article. Although the standard is aimed narrowly at automotive service manuals, it is an interesting example of a trend in the "human" translation community toward easy-to-implement,*

*objective metrics for translation quality. Such metrics enable client and provider to communicate about expectations for translation jobs, facilitate decision-making about cost vs. quality, and help resolve complaints. –Editor*

The second meeting of the European SAE J2450 Committee was hosted by Opel AG at their German headquarter in Rüsselsheim on January 22, 2002. The participants are representatives of European car and truck manufacturers, translation companies and service providers which all act globally, and so they have to maintain and quality assure all technical information in multiple languages.

Don Sirena of GM presented how they have successfully implemented a J2450-based process with their translation provider, and the impressive figures they gained in terms of translation quality increase and cost reductions. It is important to note that currently only a set of samples consisting of 300 to 500 words are manually checked against the J2450 translation quality metric but these seem sufficient to achieve the improvements. Most errors were in the area of terminology, a problem that is shared by all participants: terminology errors are the most crucial and severe mistakes in source and target language products.

In the follow-up discussion, the participants identified a certain gap between US and European translation processes. Almost every European translation company has set up their own quality assurance process, and the J2450 metric seems to be only a subset of the existing European quality metrics and criteria. One very good example of such a set of measurable quality criteria is the "Black Jack" metric of ITR, a London-based translation company. "Black Jack" was previously presented at the SAE TOPTECH Symposium in Paris last autumn, and at the first European J2450 meeting at Bowne in Wuppertal in November 2001. This time, Helen Eckersley of ITR demonstrated a Microsoft Word based support utility for translators and editors for the assignment of J2450 errors to a translation product. The utility is based on their implementation of "Black Jack", and it is freely distributed to J2450 committee members for test purposes.

Although various companies have implemented their own quality criteria and processes, they acknowledged the existence of an independent quality metric for reasons of "cross platform" compatibility. Meanwhile, the J2450 proposal has reached the "best practice" level of the SAE which is the second level of becoming an SAE standard. All participants agreed to set up test implementations of J2450, and to report their findings at the next meeting which is scheduled for April 2002.

The last topic that was discussed focused on the application of J2450 to the source language, and its impact on the translation quality. This discussion led directly to the tool presentation which was scheduled for after the official meeting. In this presentation, Jörg Schütz of IAI Saarbrücken introduced their flagship product CLAT (Controlled Language Authoring Technology) which is an application and tool that supports technical writers and editors in producing high-quality documentation in terms of general language correctness and of corporate language correctness. General language correctness concerns rules for spelling and grammar, and corporate language correctness relates to consistency in terminology use and in existing common and in-house writing guidelines, so-called style rules. CLAT checks the natural language documents against these sets of rules, reports all rule clashes, and allows the technical writer to correct her mistakes and to instantly recheck corrections made. In addition, error reports can be generated to allow for the setup of an efficient and effective quality assurance process in an existing workflow. CLAT's flexible and scalable design provides a perfect computational platform for the implementation of a J2450 application for source and target languages. Jörg has been involved in the J2450 initiative since its start.

*Further information on the SAE J2450 can be obtained at: [www.sae.org](http://www.sae.org). Search on j2450 to reach pages describing the task force, and the separate order page where the report is available for US\$59. Information on CLAT can be obtained from Dr. Schütz through email: [joerg@iai.uni-sb.de](mailto:joerg@iai.uni-sb.de).*



---

## Special Feature: Speaking of MT

---

### A Chat with Michael Quinlan

*Michael Quinlan is the founder and president of Transparent Language, a software company headquartered in New Hampshire, U.S.A., specializing in language teaching/learning software. In 1996, Transparent Language bought an MT system called Transcend from Intergraph. Transcend had already changed hands and names before. It started out as the Weidner MicroCAT system, developed at Brigham Young University. Just a year ago (February 2001) Transcend was again sold, this time to British localization giant SDL International. MTNI spoke to Mr. Quinlan in December, 2001 to find out why Transparent Language got out of the MT business.—Editor*

**MTNI:** What is the pattern to recent sales of MT systems?

**MQ:** In 1958 when MT was first demonstrated it was a compelling concept. Now it is starting to be a commercial market, but there is still more money being put into it than is being earned from it. This defines the MT market. You can't have an industry that loses money on average.

It is obvious that in the future, everything that is published online will change language to suit its audience in a fully connected world. In such a world 99% of the content that changes languages will do so via MT. In that situation, there should be big money in MT, but it hasn't happened yet.

Transcend is particularly suited to that application. It is a good, fast MT system, and is a good candidate to function as a real-time language transfer object in a network.

What changed for us was the April 2000 collapse of the technical market, and the impossibility of getting funding for language technology with long-term promise, but no short-term profitability. At that point we started to look for a buyer who could use it as a component in

an already profitable application. SDL was ideal for that.

As people have the vision of ubiquitous MT, people will put big money into MT. We had the vision, but we would have needed 3-4 more years and capital investment, which we didn't have and couldn't get. It would have taken 3-4 years to penetrate a real commercial market and bring in more money than it was costing.

There are already applications where MT is being used commercially, but it is not that widespread.

(Given the level of public awareness and level of the development of the technology), MT is still an evangelistic sale. Selling it takes a long time and is expensive. Our goal was to get to the point of profitability with the product and then put the money back into development where we could make significant improvements to the underlying technology. However, at the same time that we couldn't get funding, we couldn't compete against companies like IBM who could afford to put tremendous amounts of money into both marketing and development. We weren't interested in playing and losing. In a small company, you have to find a place where you have a product or service that people will give you money for right now.

**MTNI:** How much did you pay for Transcend in 1996, when you bought it from Intergraph?

**MQ:** We don't talk about that, but we got a good deal. We saw the coming of real-time MT and the seller didn't.

**MTNI:** Did any of the staff come with it?

**MQ:** Just a couple of people, the chief developer and a marketing person.

**MTNI:** How difficult was it to put together a development team?

**MQ:** We already had language technology staff who worked on development of our language learning software, so we had people who could do dictionary development. However, we felt that the system was basically ok, but that it was not in the right form to be a solution for online real time translation. The main development we did on the system was to

create the translation server. If we had had a success with that, we would have begun putting money into the MT engine itself.

**MTNI:** Is the development team still housed at your facility?

**MQ:** The same development team is still working in our facility, though in the next few months they will move to their own facility nearby.

**MTNI:** You did an interview with Karen Lake of Strategy Week in February, right near the time the sale was announced – I think it was even on the day that the sale was announced. In the interview you were so positive about the future of MT in enterprise applications, it seemed odd that you were getting out of that market.

**MQ:** I'm still bullish on MT. It is obvious that there will be ubiquitous translation over networks. It's exciting as a human being who lives in a multilingual world. The MT that will be available won't be 100% accurate, but it will be instantaneous and *incrementally free*. That is, you pay a licensing fee, as you do, for example, with word processing software, and the cost doesn't change no matter how many translations you do. I don't think that the per-word model will take off for automatic online translation.

The strange thing is that it has taken so long to get to where we are. MT has had a *negative slip differential*. 1 year after MT was conceived, people thought that human quality translation was 5 years away. 5 years after it had been conceived, people thought it was 10 years away. At 10 years, years it was 20 years away, and now it's leveled off at 30 years away.

**MTNI:** You had Transcend for 4 or 5 years. On balance, what is your feeling about the MT market?

**MQ:** It is like electric cars. For three decades, people have wanted to build great electric cars, but people have just lost money on them. Now we're starting to see hybrids that people really buy, and so finally electric cars are going to be a reality.

**MTNI:** I noticed that Transparent Language rarely participated in MT community events. For a group like yours, what would have made such events more rele-

vant/valuable?

**MQ:** Transparent Language is a different kind of company from most MT vendors, our primary business is language teaching software. We were looking for big market penetration, and felt that no investment in MT quality over a two year period was going to make much of a difference in our market. Improvements to MT quality are so slow and so expensive. We decided to tailor the product configuration to a market, win the market, and create the cash flow that would allow us to improve the MT. The first few years we just worked on the package. The peer groups (MT conferences) were not talking about our interests.

**MTNI:** Any closing thoughts?

**MQ:** For all services you have three variables: quality, cost, and speed. The conventional wisdom is that you can get two but not all three. MT has cost and speed propositions that HT can't approach. That's why real time translation is going to be the place where MT is successful. □

---

## A Chat with Jaap van der Meer

*In December 2001 and February 2002, MTNI spoke to Jaap van der Meer, President of ALPNET from 1995 through January 2002. ALPNET was one of the top 5 localization companies in the world before its recent sale to SDL International. It is an interesting accident that in acquiring both the Transcend MT system and the company, ALPNET, SDL International acquired whatever remains of the two main MT systems that came out of Brigham Young University in the early 1980s: ALPS, and Weidner MicroCAT. Van der Meer, who now works as an independent consultant, will address AMTA 2002 in a keynote presentation on the use of translation technology in production environments.—Editor*

**MTNI:** (Our original question, based on incorrect information, is repeated

here) ALPS went off the market as a commercial MT system in the late 80s, but was incorporated into the services offered by what then became a translation company. ALPNET really pioneered this approach, which was later adopted by Logos as well. How did it work out?

**JvdM:** ALPS was originally a research MT system developed at Brigham Young University in Salt Lake City, Utah. In 1982 or 83, the ALPS company was started to commercialize the system. The developers had also, almost accidentally, produced an interactive translation tool, which was really a the first translation memory system. They found that customers didn't want the fully automatic machine translation, they were more interested in the translation memory tool. So ALPS marketed their translation tool, Autoterm, more than the MT system. Autoterm (which later became the Joust translation memory system used and sold by ALPNET) was ported to many different operating systems including IBM, Controlled Data Systems, and NCR. However, after selling this and ALPS MT from 1983-1987, management decided that it wasn't a very good business to be in – customers really wanted translation services more than tools. So in 1987 or 1988, the translation tools were taken off the market and the company ALPNET was started as a translation services company. It went public on NASDAQ, and raised enough money to acquire several top level translation services in Europe and Canada. The ALPS MT system was never used as part of the ALPNET services, but the translation memory tools did create a competitive advantage by allowing faster turn around and more cost-effective completion of translation jobs.

Now, however, we are using MT – SYSTRAN and Websphere. Both of them are integrated into our workflow system ALPNETXchange and can be invoked from the translation memory system for sentences with no matches in the translation memory database.

The translation memory system we use now is called Globelix, a second genera-

tion tool based on the Euramis system developed for the European Commission. We also did some co-development on it with Sun, so they are using essentially the same system, but they call it SunTrans. An important feature of Globelix is that it is sever-centric. All of the other translation memory systems are designed for use by individual translators and reside on a single desktop system. Globelix keeps all translations in a central repository. The translator logs in via a Java Web interface, submits a translation, and then downloads all relevant translation memory segments to work on. The fact that translation memories are centralized means that translators from all parts of an organization can take advantage of them. Most importantly, the centralization has virtually eliminated the file-management burden that used to accompany every project.

**MTNI:** The selling price for ALPNET seemed rather low.

**JvdM:** The company had run out of assets. We had banked very heavily on the development of the ALPNETXchange workflow system, and the combination of development costs and low licensing was a heavy burden on the company. In the translation/localization industry, it is very hard to make a profit on the translation part of the business, so ALPNET, like many other companies, pursued additional consulting services, and software solutions to make money. (In the last couple of years, workflow solutions have seemed like a very attractive products as translation projects become more complex. Many companies had the same idea and the market is now full of such products. –Ed.) However, the IT managers, CIOs and CTOs who make such purchases are not convinced that translation workflow systems are necessary, given their tight budgets. Even in the case of companies who have bought such systems, it's not clear that they are actually using them. For the companies that specialize in this type of software, I don't expect most of them to survive. □

## Association News

### IAMT and the Regional Divisions: AAMT

The Asian-Pacific Association for Machine Translation was founded in 1991 by Professor Makoto Nagao, at the same time as its sister organizations, AMTA (profiled in MTNI #27) EAMT (#29), and IAMT, the International Association for Machine Translation, which is an umbrella organization linking the three regional associations. The IAMT will be profiled in the next issue, #31, with special coverage on the history of the IAMT which had its 10<sup>th</sup> anniversary last year.

The AAMT has a special challenge: How to adequately represent the entire Asia-Pacific region, when a) It started out as a Japan-specific organization; b) it has an extremely strong and active membership in Japan; and c) it is affiliated with the Japanese government (the Ministry of Trade and Industry - MITI, via the Japanese Electronics and Information Technology Industries Association)?

The AAMT's efforts to be more inclusive in Asia were discussed in an AAMT meeting held at the MT Summit in Spain. Highlights from that meeting include an effort by the AAMT board to add four or five directors from other Asian countries such as Taiwan, China, and Thailand. This expansion of representation has already begun with the appointment of Professor Key-sun Choi from Korea last year. The AAMT will begin a series of workshops or conferences on MT and Multilingual NLP that will include the whole region. The AAMT website will include space for information the broader Asia-Pacific region. The AAMT will enlist correspondents from each of the active areas within the region to channel information about their activities to the AAMT and MTNI. And finally, the AAMT will cooperate with AFNLP (The Asian Federation of NLP) in areas such as linguistic resource collection and organization of conferences.

The AAMT also has some special advantages. Corporate members allocate some manpower to the AAMT for three

working groups on technical, publication, and market issues. These committees have published some valuable reports about MT. The most recent is Machine Translation -- Vision for 21st Century. Nov.2000, (340 Pages) Member price : ¥2,500, Non-Member price ¥3,500. The book was a joint effort of the three working groups, and was two years in the making. It introduces "the present condition and the vision for the 21st century of the machine translation" via explanatory text, and many tables and figures. (Available only in Japanese from the AAMT website: [it.jeita.or.jp/aamt/pub-e.html](http://it.jeita.or.jp/aamt/pub-e.html))

**Regular publications:** The AAMT Journal is published twice a year in Japanese and includes conference re-

ports and announcements, technical articles, and periodic listings and reviews of the MT systems available in Asia, and for Asian languages. A recent issue included two tables listing many Asian MT systems, as well as links to online MT engines. (This list is also available on the AAMT website, see below.)

**Public information seminars:** Once or twice a year, the AAMT offers seminars to the general public introducing machine translation technology. (Held in Japan, in Japanese.)

**Membership Meetings:** Held in June of each year. A special seminar meeting is held following the membership meeting, in which the working groups report on their activities and accomplishments.

**AAMT Forum:** This email bulletin alerts members to mentions of MT in the popular press and media. It also includes information and reminders about MT-

### AAMT Board

Jun-ichi Tsujii	President	Prof. Info. Science, Univ. of Tokyo
Taizo Kotani	Vice President	President, Intergroup Corporation
Eiichi Yoshikawa	Vice President	Executive V.P., NEC Corporation
Yuzo Murata	Secretary General	AAMT
Hozumi Tanaka	Director	Prof., Tokyo Institute of Technology
Key-Sun Choi	Director	Prof., KAIST
Makoto Nagao	Director	President, Kyoto University
Hozumi Tanaka	Director	Prof., Tokyo Institute of Technology
Hirosato Nomura	Director	Prof., Kyushu Institute of Technology
Shun Ishizaki	Director	Prof., Keio University
Shoichi Yokoyama	Director	Prof., Yamagata University
Hitoshi Isahara	Director	Group Leader, Communications Research Lab.
Junji Maeyama	Director	Senior V.P., Fujitsu Ltd.
Isamu Washizuka	Director	Senior Executive V.P., Sharp Corp.
Hiroo Okuhara	Director	Corporate Senior V.P., Toshiba Corp.
Masaaki Hayashi	Director	Senior V.P. & Director, Hitachi Ltd.
Yoshiaki Kushiki	Director	Member of the Board, Matsushita Electric Co.
Shintaro Ushio	Director	Adviser, Oki Electric Industry Co.
Tatsuo Tanaka	Director	President, JEITA
Kiyotoshi Sato	Auditor	Executive V.P., JEITA
Mihoko Katsuta	Auditor	CEO, Toin Corporation

related events, and product releases in Asia.

**List of MT systems and links:** Note that AAMT is NOT responsible for any damaged caused by use of the sites accessible from the following URL: [it.jeita.or.jp/aamt/list-e.html](http://it.jeita.or.jp/aamt/list-e.html) (or [list-j.html](http://list-j.html) in Japanese)

**Articles of Association:** Available upon request to AAMT headquarters, the articles are slated for future inclusion in the association website. Unlike AMTA and EAMT, whose role is mainly educational, the AAMT includes specific technical goals for the association, including establishment of standards and specifications for MT systems, and development of evaluation methodologies, as well as guidelines for the introduction and use of MT systems. These functions have been performed largely in conjunction with JEITA, the Japan Electronics and Information Technology Industries Association.

**Election of officers:** The executive board within the board of directors consists of the president and up to three vice presidents. Among these, two may be from academia, and two from industry. One of the industry executive board members should be from a translation company, and the other an MT vendor. The presidency runs from one Asian MT Summit to the next, for a 6-year term. Board members include representatives of the largest corporate members, and academic board members. Unlike the European and American associations which have only a few corporate or institutional members each (and which pay a set fee) AAMT has 32 corporate members which contribute at varying levels.



## AAMT People

The AAMT has had three presidents so far, Professor Makoto Nagao, the founder (we'll speak to him next issue, along with the other founders of the IAMT), Hozumi Tanaka, who was profiled in MTNI #20, and now Junichi Tsujii, professor of Information Science at the University of Tokyo.

### Introducing

### Junichi Tsujii

Professor Jun'ichi Tsujii has been president of the AAMT since 1999, when he succeeded Hozumi Tanaka at the conclusion of MT Summit VII in Singapore. Professor Tsujii is the third president of the AAMT, following Professor Tanaka, and AAMT founder, Professor Makoto Nagao. Like many of the enduring figures in the MT community, he has spent time at various research labs in Europe and Asia. MTNI spoke with Professor Tsujii in Spain during Summit VIII, and found him to be a practical optimist, a philosophical scientist, and a hopeful community builder for the growing pan-Asian MT R&D community.

Jun'ichi Tsujii earned bachelors, masters, and doctoral degrees in electrical engineering at Kyoto University. At the time, Kyoto University did not have a computer science department, and the electrical engineering department included computers and artificial intelligence. He studied under Professor Makoto Nagao, but did not tackle machine translation as a student. Tsujii's graduate

research included such topics as natural language understanding, question answering, and knowledge representation. Following completion of his PhD, Tsujii remained for a while teaching at Kyoto University. Then in 1981, he had a chance to spend a year at Grenoble as a visiting researcher. During this time he worked under Vauquois, which he counts as an important and influential time. The next year was spent as a visiting researcher in China. Upon his return, he joined the newly formed Mu MT project with Professor Nagao. Not long after the Mu project finished in 1986, he went to England, teaching at UMIST from 1988 to 1995, and getting involved in Eurotra towards the end of that project. Although his career and research have covered many topics in AI and natural language processing, he has continued to stay involved in machine translation because "It is the most exciting. It incorporates the challenges from all parts of NLP. And many of the components developed for machine translation can be broken out and reused in other NLP tasks."

#### 30 years of MT

The Mu project was a seminal event for machine translation in Japan. Six or seven private companies, including electronics companies like Oki, Hitachi, Fujitsu and Toshiba, as well as two translation agencies, sent researchers to join the Mu team, which developed a full-scale transfer-based MT system. The project began in 1982 and ended in 1986. While the Eurotra in Europe and the Mu in Japan started at almost the same time, they were very different in nature. The research and development in Eurotra was distributed across many research institutions, while the MU project was highly centralized, i.e. all research and development staff

stayed in Kyoto University to work together. The project successfully delivered an English-to-Japanese and a Japanese-to-English MT system in four years. Considering that they started from scratch with no substantial

*Continued on next page ►*

## AAMT Committee and Workshop chairs

Steering Committee	Akio Taneda	NEC Corporation
Market-Research Committee	Prof. Yoshiyuki Sakamoto	Tokyo Kasei Gakuin, Tsukuba Jr. College
Technical Research Committee	Hitoshi Isahara	Communications Research Laboratory
Editorial Committee	Prof. Hirosato Nomura	Kyushu Institute of Technology
Network Translation Workshop	Shoichi Yokoyama	Yamagata University

---

## Junichi Tsujii

...continued from previous page

dictionaries or grammar writing software, it was an impressive accomplishment. When the project at Kyoto University ended, each project member got a copy of the source code. All of the MT systems that came out of Japan in the late 1980s were influenced in varying degrees by the Mu system. Some were based on that original code, while others modified it independently. Later, 3 or 4 of the groups



built their own systems which weren't copies of the Mu system, but stemmed from that intellectual heritage.

Professor Tsujii has some useful perspective on the history of endeavors to build MT, "Each generation of researchers has tried to build systems that are more modular and maintainable than their predecessors, but still, when building MT systems, the complexity of the system tends to get out of hand. In spite of efforts to separate procedures and data, there is inevitably some compromise. It is only after building one's first system, that one knows how to build an MT system. But that effort tends to exhaust people."

In the early years, much of the discussion was theoretical, focusing on formalisms to improve translation quality. Although the primary goal of improved quality did not see much progress for a long time, the fact that there were active research communities discussing MT formalisms led to some equally important practical advances. In the 1980s, linguistic resources such as dictionaries and rule bases were tied to particular MT systems and were unshareable. A change in approach - towards reusable resources - meant that rules and lexicons could be reused by different projects, or subse-

quent generations of the same project. For example, development practices moved toward a declarative style of encoding knowledge (separation of linguistic rules from the procedures that make use of them). This way, even if developers were exhausted by their development efforts, they began to leave a legacy that lightened the burdens of subsequent developers. Later projects could focus on the problem of improving output rather than the problem of building up enough rules and lexical entries to have a system that could be tested operationally. "Unfortunately", he observes, "the effort to make knowledge purely declarative is never completely successful. The Mu project tried to avoid mixing data and programs and did not succeed. Eurotra also tried and failed." In spite of this, Tsujii believes that the progress over 30 years has been quite impressive. The extensive study of how to do machine translation has resulted in the combination of many areas relating to natural language, and this has brought much more knowledge of text and language to bear on the problem.

### Looking for quality

"There is no need to be pessimistic about MT because it does better in specialized domains," Tsujii observes, "human translation is the same way". The best human translators operate within limited domains of expertise. The best way to improve MT quality may be to emulate this model with multiple specialized MT systems. "It should be possible to make many systems with different specializations available via an internet portal, for example. A pre-analysis of the text could determine the topic or genre of the text and select the best MT system for the translation, the way translation agencies select the best translator for a job."

One way of developing such specialized systems may be to pool the expertise of a user community via the web. Oki has recently formed an experimental community to do this. The user community is divided into interest groups based on the type of text they translate. Users give feedback to a moderator and submit dictionary entries to improve the quality of output. The moderator reviews and incorporates submissions. All users benefit from the improvements.

Another current direction in Japan is context-dependant translation. Translation

systems are being trained on huge linguistic resources so that when a text is submitted, the appropriate resources can be searched and retrieved for that particular text, rather than just a lexical/syntactic match of translation rules, which may not yield the best translation for the source language context.

Although quality tends to be discussed as a single uniform problem, each MT system design has its own characteristic quality and performance problems. For EBMT, for example, the problem is how to store and categorize a huge database of examples. The problem is closer to an IR problem than a linguistic problem.

### MT use in Japan

A few years ago there was a low point for MT in Japan, and MT companies were losing money. But now MT developers are becoming profitable. The assimilation market (translation into Japanese) is the biggest. Tsujii asserts that the quality of currently available English->Japanese systems is quite good. Lately, the typical user profile has been changing.



While MT was formerly used primarily by big companies, now casual individual users dominate the market.

In the long run however, the dissemination market is potentially much larger. Japan exports software, cars, clothing, electronics, and many other products all over the world. The need for translation is not just for product documentation, but for communication with offshore manufacturing facilities. However more specialized translation systems will be necessary to produce good enough quality for this application to take off.

Junichi Tsujii may be reached at [tsujii@is.s.u-tokyo.ac.jp](mailto:tsujii@is.s.u-tokyo.ac.jp) □

---

## MT Users Corner: A User's Desiderata, Part I

Jackie Murgida

### MT is for *who*?

Several years ago I attended a presentation on one of the big-name MT products at an AMTA conference. The company representative gave the usual run-down. Although the system was good by the standards of the time, it looked cumbersome, from a user's perspective, and the quality of the output would make it useful mostly for indicative, information-only purposes. In the question period I asked what kind of user the product was aimed at, and the presenter's unhesitating answer was, "translators."

Gulp! I was really amazed that anyone thought a translator would want to jump through the electronic hoops required by the average MT system of the day. What translator would want to update the dictionary with words unknown to the system but very well known to translators, by means of a very unfriendly interface? Or go through the arduous procedure of preparing a whole glossary to import into the MT user-defined dictionary. Who would want to wrestle with formats, system requirements and so on? And above all, why would a translator willingly post edit raw output that would take three times as long as translating the document from scratch?

### What do they want?

Since that time I've been thinking about what users really want from MT - what they would buy and actually use, if it existed for their language pair/direction. This is the first in a planned series of articles on different types of users and what they would find useful in an MT system. The first two installments address the needs of translators, traditionally the group expected to use MT. Subsequent articles will look at MT desiderata for such groups as researchers and analysts, multinational businesses, and individuals using the Internet and Web.

So, what do translators want? Well, which translators? There are highly competent experts who know their languages and domains inside out, and they enjoy the *entire process* of translation. They love taking a text in one language and crafting a target-language text that conveys the meaning of the original but doesn't read like a translation. They love writing in the target language, honing the text until it's a joy to read. They don't want a computer to do their favorite task. They don't want to post edit MT any more than they want to edit a translation done by another human being.

Related to these creative translation artists are the people who just don't like editing someone else's translation. Some of them barely edit their own work. Forget about them. If someone gives them an MT program, they'll only use the terminology module, or they'll use the CD as a coaster for their coffee mug.

Another category of users we'll skip over is the recreational MT user. An in-house translation supervisor who was in charge of keeping an eye on software tools for his unit told me that there wasn't anything available that met their requirements. But he had a copy of a low-end, off-the-shelf MT product. He also had a French father-in-law, so he wrote him newsy letters in English, filled with colloquialisms and idioms. He then translated the letter with the fifty-dollar software and sent it, unedited, along with his own fully-manual high-quality human translation. Papa reportedly really enjoyed the MT's fractured French.

The other type of translators we'll keep out of this discussion are the ones who are techno geeks and love any new software or electronic gadget, no matter how user-hostile it is. They'll take hours installing it because it's new and does something swell. Then they'll take days or weeks to climb the learning curve, no matter how steep and long. They'll convert this to that and back again. They'll create macros and find workarounds to the most galling gli--, I mean features. They'll spend all night trying to solve one problem. Forget them, too. They install sounds of bodily functions on their own and other people's computers because they're way cool.

### Everyone else

That leaves everyone else. People who translate to make money, because they enjoy working with words or they happen to know a couple of languages and answered a want ad, or any number of reasons. They use the available technology to make life more pleasant and convenient. Translators want to be more productive and serve their clients better. They want to be able to accept a very large project that couldn't be done quickly enough by a lone free lance. They want to make more money or have more time to spend in a better climate or bond with their kids. Whatever.

Now, finally, the list of desiderata. Based on my discussions with colleagues over the years, these are the features that I believe would make MT attractive to a large number of translators, whether in-house, staff translators or free-lancers, part- or full-timers, highly experienced or novice. Not everyone would use every feature, as with any software tool, because everyone has differing comfort levels with the technology, levels of expertise in their languages and domains, work styles, personal preferences, and needs of the end-user of their translations.

### MT Desiderata for Translators

#### Quality of raw output must be very good

This is the *sine qua non* of MT for translators. Out of the box, with only the addition of unknown words, the linguistic quality (syntax, morphology, lexicon) of the raw output has to be good enough that it can be post edited in significantly less time and with less aggravation than translating and editing it oneself. This means that word order and word sense disambiguation have to be state-of-the-art.

My rule of thumb on this is that the raw output has to be as good as a so-so human translation. It can be quite literal and a little awkward, with some mistakes, but correctable without needing to take Valium. The mistakes will be *different*, but *they should be no more difficult to correct*.

#### The dictionary should be really big and good

Continued on page 18 ►

---

## Book Review:

# Early Years in Machine Translation

W. John Hutchins

Amsterdam Studies in the Theory and History of Linguistic Science, Volume 97.

John Benjamins, 2000

Reviewed by Maghi King, TIM/ISSCO, School of Translation and Interpretation, University of Geneva.

This is a lovely book. The editor has succeeded in persuading more than twenty five of the early pioneers of work in machine translation to record their memories in their own words. To quote from the preface

*"contributors were encouraged not just to 'tell it how it was' and to recount the origin and development of their own research programmes, but also to include personal details and anecdotes and to give their impressions of the impact of political and social events on the development of the field." (p. vi).*

He is to be congratulated on having produced a book which fascinates and grips the reader from start to finish.

One reason, is that, as might be expected from contributors who have worked with language for most of the lives, all the contributors write extremely well. It may be possible, with some effort, to find the occasional boring passage, but I would be prepared to challenge anyone to find a badly written one. Some samples will be found in the quotations included in this review.

Secondly, the contributors have very distinctive and very different voices, which in turn reflect their individual characters. Gentle, unwilling to say harm of anyone, calm and rational, passionate or intolerant, vindictive even –

all manner of men are to be found here. The result is a constantly shifting variety.

It is important, though, to realise that establishing a definitive and coherent history of early work in MT is not one of the book's aims. It is true, as one of the contributors says, that MT tends to live with a short memory:

*"In other branches of science, people try to capitalize on previous results. Mathematicians study and redemonstrate theorems, physicists reproduce experiments, computer scientists read 'The Art ...' and dissect compilers. In MT and some other areas in computational linguistics, however, research seems to function 'as a finite-state device' (as J. Rouault put it ...), that is, with bounded memory." (p.331).*

But the memories recorded here are too disparate to make up a coherent story. Indeed, there are times when contributors disagree in their recollections, to the point occasionally of flagrant contradiction. The editor has taken care to correct where he could purely factual matters, such as in the extended series of footnotes to p. 255, where he suggests that at least the chronology of his contributor's memories must be inaccurate, but has otherwise resisted the temptation to try and impose coherence.

Even with this caveat, the book does serve to point out that some of the received truth about the early days of machine translation is, in fact, rather less clear cut than is usually believed. For example, a very common claim is that the early workers were extremely naive in their thinking about language, and as a consequence rather arrogant in what they thought could be achieved. A couple of quotations suggest that some, even at the time, thought that this was so:

*"What impressed me more than anything else when I got into MT research and looked around at what other researchers were doing was that most of them had little or no training or experience in linguistics, and they didn't seem to think that such training and experience was necessary for their work." (p. 177)*

*"Pretty soon it came to me that I knew what we should be doing, and that eve-*

*rybody else was trying not to do it." (p.129)*

But it was not universally the case. Here, for example, is a quotation from Bar Hillel's statement at the public opening session of the first conference on MT in 1952, which shows that at least some were aware of the complexity of the task they had set themselves and were prepared to be modest in their aims:

*"... completely automatic and autonomous mechanical translation with unique correlates to the original text is, in general, practically excluded, even with respect to scientific texts... This being so, mechanical translation means no more than mechanical aids to translation. Only some kind of brain-machine partnership is envisaged." (p.304).*

This in its turn destroys another of the myths that I at least was brought up on: that of the devil Bar Hillel, who started out enthusiastic (and naive), grew disenchanted and in revenge destroyed the field. In actual fact, he shows a remarkable consistency from the beginning of his involvement with MT through to a study done in 1971 where he is still saying that there are three main options:

*Machine aided human translation  
Man-aided machine translation  
Low-quality machine translation*  
holding them all to be "eminently practical." (p309).

I have picked out Bar Hillel's remarks also because they foreshadow later developments, in this case the realization that translation aids, such as translation memory systems or computerized terminology banks might prove extremely valuable. In other ways too both he and others come up with ideas which have a very modern flavour. Here is Bar Hillel again, this time suggesting that the notion of translation quality may be variable, an idea which might be extended to suggest that even bad translation may have its uses.

*"A translation which is of good quality for a certain user in a certain situation might be of lesser quality for the same user in a different situation or for a different user, whether in the same or a different situation. What is satisfactory for one need not be satisfactory for*



another." p309.

Bar Hillel is far from being the only pioneer to surprise us with the modernity of his ideas. For example, although Yngve in MIT accepted that one should take a word-for-word translation as a first approximation, he also suggested that syntactic analysis was a crucial tool to resolving what he calls the "multiple meaning problem" – and this was before the publication of Chomsky's Syntactic Structures revolutionized the linguistics world in 1957 (p42). Similarly, current work on ontologies and their use is foreshadowed by, for example, work at the Cambridge Language Research Unit, with an awareness of the philosophical and ontological issues involved that is not always found today:

*"To that extent whether such categorial primitives are out there in the world, or here in the head, is immaterial: the issue was whether some set of primitives could be found that was functionally sufficiently effective for MT. Thus the analogy is with the practical lexicographer who, though he lists three senses for a word, does not maintain the number three is absolute, only appropriate for the purpose to hand."(p272)*

And the use of controlled language to achieve a higher quality of translation – a topic which has in recent years received much attention – was, once again, suggested by Bar Hillel as far back as 1951:

*"restricting, by voluntary convention, the richness of expression ... to such a degree that sentence-pattern translation might easily and quickly be applied." p303.*

Perhaps less surprising is that almost every possible architecture subsequently used as a basis for MT systems is already present in early work: there are proposals for interlingual systems (Reifler, Montgomery, Mel'cuk, Andreev and others), for transfer based systems (Yngve, Kulagina, Marcuk, Kirschner and others), for statistics based systems (Yngve, Ljapunov and especially the Speech Statistics Group, p236). Even those who were insisting that the key to successful MT lies in the lexicon, and who were subsequently derided as proposing something close to

word-for-word translation, are more sophisticated than history holds them to be, and might well be re-interpreted as pre-cursors of modern work according a central role to the lexicon.

The book also brought home to me how much influence early work in MT had on pure computer science. In 1967, I was a straight computer scientist, working on the development of compilers. Among the hot topics of the day were parsing algorithms for formal languages and the development of high level languages for specific application areas – both technologies developed originally within pioneering work on MT, at RAND, at MIT and elsewhere.

So, if the early pioneers were not as naive as popular history would have them be, it is legitimate to ask what went wrong. It is after all salutary to realise that the period which is the main focus of this book lasted little more than the duration of a standard European Union Framework Programme: although some research pre-dates 1955, the flowering only really started then, and by 1960 there were already congressional hearings challenging the advisability of investing in MT research. Indeed, although popular history once again would have us believe that it was the ALPAC report which sounded the death knell of early MT work, it is clear from reading this book that a number of important projects had closed even before the committee started its hearings.

The easy answer is to blame it all on the lack of computing resources: such computers as existed were, it is true, primitive by modern standards, and there were many groups, especially in the old Eastern block countries, who did not even have access to such machines as did exist. It was commonplace to do research on MT by playing at being a computer – trying to simulate manually what a computer would have done.

But I suspect that that is too easy. One of the contributors remarks

*"At least as interesting as the achievements of the time were the colourful personalities involved and the dynamics that existed between research groups."(p97)*

This is the nice way of putting it. Listen to some of the contributors talking

about their contemporaries nearly fifty years later (I have deliberately suppressed names in the hope of whetting the reader's own appetite for finding out who is talking about whom):

*"The unscientific conduct of discourse in linguistics was most disturbing. Instead of reliance on scientific criteria, too often one found ex-cathedra pronouncements, polemics, intellectual bullying, bowing to authority or tradition, trading on charisma and personal reputation, adherence to the orthodox views of a school or –ism, and other sorts of appeals and forms of pressure irrelevant to deciding scientific questions."(p68)*

*"... but as a researcher I was a bit unsure of him, whether he was just a figure head or whether he was a bit of a fraud."(p79)*

*"... was of a different colour altogether, a smooth slick operator, always looking to promote himself."(p79)*

*"My contact with the other ... people came only many years later; I did not find their work particularly interesting or profitable."(p218)*

*"Of particular damage was the approach of ..., who put forward the contention that machine translation at that time should not be made practical until theoretical questions were solved. In his opinion, machine translation should be 'translation without translation, without machines, without algorithms'. ... Any practical experiment in MT was declared 'creeping empiricism'."(p249)*

*"... one of the less esteemed members of the American MT community."(p255)*

These people do not like one another. Stronger still, they do not hesitate to accuse one another of unethical conduct:

*"The ... MT demonstrations seemed always to be contrived; they made impressive publicity for the sponsors, but they soured the atmosphere by raising expectations nobody could fulfil." (p79)*

*"Among amusing memories was the demonstration that turned out excellent translations for Russian ... the translations had been input earlier with the source sentences."(p159)*

Continued on next page ►

## King Review of Hutchins

...continued from previous page

*"And at the 1965 World Fair in New York, the ... exhibit included a computer that had an amazing control of language. Such demonstrations suggested that the output of operational systems ... were also based on dubious procedures." (p161)*

And lest anyone should think that only in the USA were practices suspect, this quote concerns the USSR:

*"After formulating some rules for translation of a single short text, they proclaimed that they had achieved the goal of MT in general. There were protests from a lot of people working in MT. ... and ... were publicly exposed, and their results recognised as false. Since both are now dead, it may be unkind to write about this, but the truth must be told." (p228)*

But of course, dislike and accusations of unethical behaviour also have to be explained, and I do not know where to seek for that explanation. Perhaps an earlier generation was more passionate than our own, perhaps all can be explained through a sordid competition for funding, perhaps politics of one sort or another plays a role: after all, we are explicitly told that one project obtained CIA funding by arguing that the opposition (i.e. the USSR) was distributing free technical literature in South East Asian countries in "local dialects" as a way to spread communism. The only way the USA could compete was by emulating them, using MT to produce the translations (p113). On the other side of the fence, there are also direct claims that politics played a major role in the USSR:

*"Unfortunately, after the brief period of general enthusiasm, MT became a coin of barter in the intrigues and rivalries between the Committee for Science and Technology and the Academy of Sciences, between the KGB and the Soviet Army, and between the Communist Party bureaucracy and the military-industrial structures. Anti-Semitism and Russophobia, chauvinistic Stalinism and turmoil over dissidents, as well as unending denunciations and complaints*

*to Communist party committees, to the KGB or to the Presidium of the Academy of Sciences, served as permanent ingredients for tragic conflicts where everyone fought everyone." (p236)*

The role of politics and the motives behind so much dissent remain the major mysteries: the book does much to increase our awareness of them, but little to clear them up. But I suspect that it might be almost impossible to reconstruct an accurate and unbiased history, even if one set oneself the task of doing so.

There are other mysteries too, in fact so many that at one point I seriously thought of writing this review in the form of series of questions to the reader, asking him to solve the puzzles by reading the book in much the same way that one tries to out-think a detective story. Here is a small selection for those who might be intrigued.

What public demonstrations, if any, really were faked? The accusations are many and various, as are the denials.

Why did Pankowicz's collection of documents about MT disappear? We are told both that he was forced to destroy them, and that it was somehow the result of his dying intestate.

Why did both Anthony Oettinger and David Hays agree to serve on the ALPAC committee? Oettinger's contribution suggests that it was because they both, through their place of work, had connections to the military, but he also says that he no longer has any clear recollection of how he came to be a member. He furthermore suggests that another mystery, that of how the ALPAC Committee actually functioned and how it reached its conclusions is not yet adequately elucidated:

*"It did seem to me that the decision about withdrawing support for MT was reached very quickly, and that evidence to back it up was collected afterwards." (p83)*

This is only a small selection of topics on which the contributors do not agree: But, as I said earlier, their various opinions and the strength of them is part of what makes the book so fascinating. I learnt an enormous amount from it, and, more importantly, enjoyed it immensely. Do read it. □

## MT Users Desiderata

...continued from page 15

This is almost as important as linguistic quality. Unless the user just loves playing with words and has a lot of time to enter unknown words, the MT lexicon has to be large, with lots of senses for different domains and the ability to choose the right one most of the time.

In addition to the usual domain stacking capability, for the dictionary to be most useful it should also have proper nouns (personal names, entities, organizations, place-names). Furthermore, for source languages such as Chinese and Arabic that don't consistently use capitalization or some other way of indicating proper nouns, the MT system should be able to flag any noun that can be both proper and common for the post-editor's attention. This would be especially useful for novice translators who don't have a vast knowledge of names in the source culture.

For instance, the MT system should be able to identify nouns and noun phrases like "carpenter," "victor," and "servant of God," in Arabic, determine that they are likely to be personal names, and transliterate them as Najjar, Nasir, and Abdallah, rather than translate them. These could be flagged when there is ambiguity, in case the literal translation is actually appropriate.

The same thing would be handy for idioms and collocations that have a non-literal meaning most of the time but at times should be translated literally. This could be a great help to people who don't translate into their native language and to novice translators who may not be aware of all the pitfalls of faux amis and idiomatic usages in the source or the target language.

### Diagnostic tools

There are several types of facilities that would be useful as one is post editing:

for a questionable target-language term, highlight it and get the corresponding term in the source text to see quickly where it came from so that the user can correct the translation or analyze and research the term, and correct the dictionary, if necessary

- highlight a source term and find out what the other senses of the word are in the source language and what the alternate translations are

- select a source sentence and get the parse tree so that the user can see if that seems correct

- get alternative parses (especially useful when translating from a language that has long sentences with complex embedding)

- get the corresponding translations for the other senses and alternative parses

The user should be able to select the alternate translations and drop them into the translation during post editing. This type of facility could be an aid to novice translators, helping them do more of their translation work on their own and reduce the amount of review needed by a senior translator.

*Jakie Murgida is an MT consultant, translator, and language trainer. She can be reached at: [jmurg@star.net](mailto:jmurg@star.net)* □

## MT 2010

...continued from page 5

impose anything on anyone, but rather to provide a broadly supported definition of a context in which to position the MT community's efforts, which would allow us to identify common priorities for joint activities in e.g. research, resources and training.

For this workshop we invited papers that:

gave critical analyses of the present state of the art in machine translation of written and spoken language, presented visions of the future of machine translation, both from a technological and from an application point of view, or identified major milestones and challenges on our way towards the future, and/or ways to measure our progress along the road.

As the workshop was planned as a half-day event, we could only accommodate a limited number of papers.

The nine papers that were selected for this workshop covered broadly three different aspects of MT. The first group of papers (by Gerber, Schütz, Boitet, Farwell and Helmreich, Och and Ney) all addressed the question: where to go? Each of them sketched its own perspective of a potentially successful future for MT.

The second group of papers

## From the Editor

*Dear Readers,*

*When I took over editing MTNI almost a year ago, I wanted to accomplish two things. The first was simply to put out the newsletter on a regular schedule at the established high level of quality! The second was to make some minor changes that would increase the value of the information provided in MTNI. The first goal has begun to be accomplished. The second can now be addressed.*

### MTNI—Paper or Digital?

*The MTNI editorial board and IAMT board of directors have been discussing electronic distribution of this newsletter to overcome both the delay and expense of mailing for those members who would like it. The consensus is that readers should be given the option of receiving MTNI in digital form (as a zipped pdf file) in addition to, or instead of, the paper version. An email survey will be circulated by your regional association. Please don't forget to respond!*

### Timeliness

*In a publication that appears infrequently (see the publication schedule below), it is difficult to report all the news in a timely fashion. This is particularly true in the case of the time-sensitive information that appears in the Calendar section.*

*I would like to divide the channels by which the content is delivered, so that time-sensitive information can be accessed/delivered outside of the newsletter publication schedule. This may be done via email, or simply made available on the website, leaving it up to members to access at their convenience. It may take some time to settle into a pattern that is both doable for the editorial staff, and ideal for readers, but there will be some experiments along these lines in the coming months. We would greatly appreciate any suggestions or feedback.*

### Frequency

*Subscribers have trouble understanding the publication schedule of MTNI, and events in the last year have compounded the problem. By the IAMT charter, MTNI is to be published three times per year. However, over ten years we got two issues behind. To correct this, four issues were published in 2001, (#26-29) and four will be published in 2002. The schedule for this year is March, June, September, and December. In 2003 we will revert to three issues per year and adopt a regular schedule.*

*Laurie Gerber, Editor in Chief, [mtni@eamt.org](mailto:mtni@eamt.org)*

(Macklovitch and Valderrábanos, Flournoy and Callison-Burch) dealt with the role of interaction in MT, and made it clear that the topic may be old, but far from obsolete.

The last two papers were different in that they did not specifically address the future development of MT, but rather reminded us of two issues that we should not ignore when discussing the Big Problems and the Grand Challenges that lie ahead of us. The paper by Tsou and Kwong clearly illustrated that the notion of problem is relative: what counts as a non-problem when translating between Western European languages (translation of personal names) turns out to be a really hard problem when translating between e.g. Chinese and English.

Probst *et al.* drew our attention to the existence of low-density languages,

where the lack of commercial interest may make it hard for speakers of those languages to get access to translation facilities and hence to the multilingual information society.

All participants were asked to complete a questionnaire identifying the main challenges reflected in each paper. The evaluation of the questionnaires is ongoing, and the results will be presented as a discussion paper at the next MT Roadmap workshop, which will take place in conjunction with the TMI 2002 conference in Keihanna, Japan, in March 2002.

The papers of the MT Summit workshop can be found on the ELSNET website at <http://www.elsnet.org/mt2010.html>, and the input documents and results of the TMI workshop can be found at <http://www.elsnet.org/roadmap-tmi2002.html>. □

## AMTA 2002—From Research to Real Users

...continued from page 1

of submission are available in the call for papers at [www.amtaweb.org/AMTA2002](http://www.amtaweb.org/AMTA2002). The deadline for submission of papers to the conference is April 15, 2002. Accepted papers will be included in the proceedings for AMTA 2002, to be published by Springer in the Lecture Notes in AI series.

### Call for Tutorials and Workshops

Proposals for tutorials and workshops are also being solicited on these and other topics of direct interest and impact for MT researchers, developers, vendors or users of MT technologies. We welcome and encourage participation by members of AMTA's sister organizations, AAMT in Asia and EAMT in Europe, as well.

Workshops will be held on Tuesday October 8th. Approximately 7 hours may be allocated per workshop.

Tutorials will be held on Wednesday October 9th. Tutorials would typically last 3 hours, although other arrangements might be possible.

Proposals should state the topic(s) to be addressed, the rationale for addressing it, and the structure of the activities. Proposals should be in English and not longer than 4 pages.

Please submit proposals as soon as possible to Bob Frederking at [ref@cs.cmu.edu](mailto:ref@cs.cmu.edu). Proposals must be submitted on or before Friday, April 12, 2002.

### Call for Exhibitors

Every two years, the MT community gathers to hear what's new in the research labs, share success stories, and see what's happening in commercial MT. Attendees at AMTA conferences include MT researchers, developers, users, MT shoppers and consultants—all of whom look forward to seeing the concrete progress that has made its way into commercial products and research prototypes. Exhibits open

Wednesday evening, October 9th, during the welcome reception, and remain open for the rest of the conference. A booth is \$600, including a table, chairs, sign, and pipe-and-drape booth enclosure. We will only be able to accommodate a few exhibits, so please contact Laurie Gerber <[lgerber@gerbersite.com](mailto:lgerber@gerbersite.com)> to reserve your place.

All of the information in this announcement, and more, is available on the conference website:

[www.amtanet.org/AMTA2002](http://www.amtanet.org/AMTA2002) □

## Conference Program

### Keynote Speakers

- Ken Church, AT&T Labs-Research
- Jaap van der Meer, Consultant, former President of ALPNET
- Yorick Wilks, University of Sheffield

### Paper presentations

- Technical papers on current research
- User studies of MT deployments
- System descriptions

### Product Exhibition

Companies will exhibit and demonstrate their translation technology and services in the exhibition hall.

### Tutorials

Pre-conference tutorials will give attendees an accelerated introduction to practical and technical topics about machine translation and its use.

### Workshops

Special interest groups meet for day-long focused sessions of theoretical or practical interest.

## 6th EAMT Workshop

Manchester, England

November 14-15, 2002

### Call for Papers

The sixth EAMT Workshop will be hosted by the Centre for Computational Linguistics, UMIST, Manchester, England. Organised by the European Association for Machine Translation, in association with the Natural Language Translation Specialist Group of the British Computer Society. The Workshop focus is *Teaching Machine Translation*.

See the call for papers on the website for a full list of recommended topics including: Why and to whom should MT be taught? Teaching: theoretical background of MT; Pre- and post-editing skills; MT evaluation. Building 'toy' MT systems in the laboratory. Translation studies and MT. Etc.

We invite submissions of extended abstracts, up to two pages, summarizing the main points of the full paper. The abstract should be sent by email in (doc, html, pdf, ps) format, or as plain text, to [Harold.Somers@umist.ac.uk](mailto:Harold.Somers@umist.ac.uk).

Programme Committee: Harold Somers, UMIST, Manchester; Derek Lewis, University of Exeter; Ruslan Mitkov, University of Wolverhampton; Mikel Forcada, Universitat d'Alacant.

For details, including final paper format and length requirements, see: [www.ccl.umist.ac.uk/events/eamt-bcs/cfp.html](http://www.ccl.umist.ac.uk/events/eamt-bcs/cfp.html) □

### EAMT Important Dates

Abstract deadline	July 31, 2002
Notification to authors	September 6, 2002
Full papers due	October 14, 2002

### AMTA 2002—Important Dates

Tutorial and Workshop Proposals due	April 12, 2002
Paper submission deadline	April 15, 2002
Notification to authors	May 31, 2002
Final papers due	July 15, 2002
Pre-conference workshops	October 8, 2002
Pre-conference tutorials	October 9, 2002
Main AMTA conference	October 10-12, 2002

### AMTA-2002 CONFERENCE ORGANIZERS

Elliott Macklovitch, General Chair, [macklovi@iro.umontreal.ca](mailto:macklovi@iro.umontreal.ca)  
Stephen D. Richardson, Program Chair, [steveri@microsoft.com](mailto:steveri@microsoft.com)  
Violetta Cavalli-Sforza, Local Arrangements Chair [vcs@sfsu.edu](mailto:vcs@sfsu.edu)  
Bob Frederking, Workshops and Tutorials, [ref@cs.cmu.edu](mailto:ref@cs.cmu.edu)  
Laurie Gerber, Exhibits Coordinator, [lgerber@gerbersite.com](mailto:lgerber@gerbersite.com)

---

## Translation Technology: Present and Future

Barcelona, Spain  
April 4-6, 2002

This international congress is being organized by the Association of Centers Specialized in Translation, with the backing of the EUATC (European Union of Associations of Translation Companies). It will be a meeting point for all representatives of the sector, and the first step towards an event of greater importance to be held in the Barcelona Forum 2004.

The congress is aimed at everyone involved in translation and new technologies, such as companies, translators, students and language and communication enthusiasts. Representatives of technology, universities, literature and politics will be taking part.

Full information on the conference program, speakers and registration, is available at: [www.act.es/congreso/uk](http://www.act.es/congreso/uk). Special rates for ACT members, EUATC members, associated free-lance translators and students. Contact: [congreso@act.es](mailto:congreso@act.es) □

---

## EMNLP-2002

Philadelphia, Pennsylvania, USA  
July 6-7, 2002

SIGDAT, the Association for Computational Linguistics' special interest group on linguistic data and corpus-based approaches to NLP, will hold its Empirical Methods in Natural Language Processing (EMNLP), conference immediately preceding ACL.

Submissions have been solicited around the theme "The Next Big Thing in Data-driven NLP", that describe attempts to substantially and radically deviate from the current practice of simple adaptations of existing and usually well-studied methods. Directions into previously unknown territory for NLP will be included, such as: using Really Large Corpora (cf. last year's Brill's

---

## Three Workshops following ACL

Philadelphia, Pennsylvania, USA

---

### 1: Speech-to-Speech Translation: Algorithms and Systems

Construction of speech-to-speech MT systems involves research in Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Machine Translation (MT), Natural Language Understanding (NLU) and Generation (NLG). Although substantial progress in each of these components individually has been made over the last two decades, simply integrating individual ASR, NLU, MT, NLG, and TTS components to produce S2S systems is not sufficient to produce acceptable results.

This workshop aims to bring researchers together to present and discuss the current state-of-the-art in speech-to-speech translation. Organizers: Yuqing Gao, IBM T. J. Watson Research Center, [yuqing@us.ibm.com](mailto:yuqing@us.ibm.com); and Alex Waibel, Carnegie Mellon University. See [www.acl02.org](http://www.acl02.org), Conference Workshops. □

talk); using previously neglected methods, including those from non-NLP fields, such as biology, nuclear physics, or finance, with promising results and/or reasonable potential for the future; employing known methods in a radically different way or new problems with significant improvement; combining intuition-based and data-based methods (finally!) with substantially improved results on known problems.

Authors may provide extended final section in their submissions, discussing perhaps even slightly speculatively what the future might look like.

Jan Hajic, Charles University, Prague, Czech Republic (chair) [hajic@ufal.mff.cuni.cz](mailto:hajic@ufal.mff.cuni.cz); Yuji Matsumoto, Nara Institute of Science and Technology (co-chair) [matsu@is.aist-nara.ac.jp](mailto:matsu@is.aist-nara.ac.jp). See [ufal.mff.cuni.cz/~hajic/emnlp02](http://ufal.mff.cuni.cz/~hajic/emnlp02) □

---

### 2: Unsupervised Lexical Acquisition

July 12, 2002

This workshop will be a forum for the presentation of new work in the field of unsupervised or minimally supervised lexical learning, as well as an opportunity to survey the state of the art. In addition to novel approaches to the acquisition of subcategorization or translation information, we will welcome and encourage papers that address the coverage of all aspects of the lexicon, such as morphological, semantic, or collocational information, as well as the identification of out-of-vocabulary words and their lexical properties. Workshop Chairs: Joseph Pentheroudakis, Microsoft Research, [josephp@microsoft.com](mailto:josephp@microsoft.com); Nicoletta Calzolari, Istituto di Linguistica Computazionale del CNR, [glottolo@ilc.pi.cnr.it](mailto:glottolo@ilc.pi.cnr.it); and Andi Wu, Microsoft Research, [andiwu@microsoft.com](mailto:andiwu@microsoft.com). See [www.acl02.org](http://www.acl02.org), Conference Workshops. □

---

### 3: Morphological and Phonological Learning

July 12, 2002

Two groups of researchers are converging on the need to construct morphologies and phonologies of low density languages. Natural language engineers hope to develop machine translation, speech recognition, and other NLP technologies for these languages. Meanwhile, linguists and native speakers want to document the languages for scientific or humanitarian reasons. Mike Maxwell, Linguistic Data Consortium (chair); Invited Speaker: David Yarowsky, Johns Hopkins University. See [www.acl02.org](http://www.acl02.org), Conference Workshops. □

## Lexicom—Lexicography Master Class

July 14-19, 2002  
Brighton, England

The Information Technology Research Institute at the University of Brighton will host a lexicography master class with trainers Sue Atkins, Adam Kilgarriff, and Michael Rundell. The class will bring together lexicographers, linguists, computer scientists, project computer officers, translators and terminologists for hands-on training in creating, managing and exploiting lexical data.

Applicants should have experience in at least one of the following disciplines: lexicography, linguistics, computational linguistics, computer support for dictionary projects, translation, or terminology.

The principal topics to be covered are: writing dictionary entries; building and working with text corpora; automatic corpus analysis for lexicography.

All sessions will include practical work at a computer terminal.

See [www.itri.bton.ac.uk/lexicom](http://www.itri.bton.ac.uk/lexicom) □

## SAAKM 2002

Lyon, France  
July 23, 2002

Semantic Authoring, Annotation, and Knowledge Markup. This workshop aims at bringing together members of communities that share an interest in semantic authoring and annotation for developing methods and tools:

- Semantic Web researchers who use semantic authoring and annotation to enrich the web with distributed relational meta-data in order to enable a machine-readable web
- members of the computational linguistics community, developing information extraction systems for the generation of meta-data
- people from the multimedia content domain, indexing and searching of multimedia (and multilingual) data.

The workshop is held in conjunction with the European Conference on AI. **Submission Deadline: April 28, 2002.** See [saakm2002.uib.uni-karlsruhe.de](http://saakm2002.uib.uni-karlsruhe.de) □

## Resources

### Laurie's Links

So far I've emphasized free resources. This time, I look at some of the "strategic market research" reports on MT and language technology, and a forward-thinking (and free) whitepaper on the use of translation technology.

### Strategic Market Research

There are an amazing number of companies who do nothing but write and sell reports about other companies. These reports are frequently cited as support in newspaper articles that report on market size, growth and trends. A relatively well-known example in the MT community is the Ovum report from 1995. Recently, there have been a number of others. The team of Steve McClure and Mary Flanagan at IDC lead the pack in sheer numbers, with some reports comparing MT systems, overviewing and introducing MT, and in some cases profiling individual companies or implementations. SYSTRAN's deployment for tech support translation at AutoDesk, and an overview of recent developments at SYSTRAN S.A. are recent examples. IBM WebSphere, and LogoMedia have also been profiled. The reports are pricey—from US\$500 to US\$2500. See [www.idc.com](http://www.idc.com) and search for analyst Mary Flanagan. The SYSTRAN report is also available on the SYSTRAN website at: [www.systransoft.com/IDC/26459.html](http://www.systransoft.com/IDC/26459.html)

### Industry Whitepapers

Localization giant Lionbridge has been putting some thought into the pros and cons of MT in localization projects. This has resulted in a very useful white paper: "When to Automate Translation Processes". Although they have not quite gotten to the point of providing a foolproof mathematical formula for determining the potential return on investment from MT, they provide a conceptual framework that gets close. Go to [www.lionbridge.com](http://www.lionbridge.com), ROI Challenges, Content Globalization. It's among the whitepapers listed along the right side of the page. □

## Calendar

### 2002

**March 1-3:** ICON 2002 (Indian Conference on Natural Language Processing). Chennai, India. See [www.iit.net/icon\\_conf.htm](http://www.iit.net/icon_conf.htm)

**March 13-17:** TMI 2002. The 9th Conference on Theoretical and Methodological Issues in Machine Translation. Keihanna, Japan. See: [www.kecl.ntt.co.jp/events/tmi/](http://www.kecl.ntt.co.jp/events/tmi/)

**March 29-30:** HKTerm. Hong Kong workshop on terminology in education, research, and applications. See: [cpct92.cityu.edu.hk/TAHK/index.htm](http://cpct92.cityu.edu.hk/TAHK/index.htm)

**March 24-27:** HLT 2002. San Diego, California. Sponsored by DARPA, NSF, ARDA et. al. Organizers: Mitch Marcus, University of Pennsylvania (general chair), David Yarowsky, Johns Hopkins University (co-chair). See: [hlt2002.org](http://hlt2002.org).

**April 4-6:** Translation Technology: Present and Future. Barcelona, Spain, at the World Trade Center. Organized by the Association of Centres Specialized in Translation. See [www.act.es/congreso/uk](http://www.act.es/congreso/uk)

**April 24-25:** Language Technology for Business Information Systems. Poznan, Poland. Organizers: Jakub Piskorski, DFKI GmbH Germany, [piskorsk@dfki.de](mailto:piskorsk@dfki.de), Feiyu Xu, DFKI GmbH Germany, [feiyu@dfki.de](mailto:feiyu@dfki.de). See: [bis.kie.ae.poznan.pl](http://bis.kie.ae.poznan.pl)

**May 9-11:** The Fifth Symposium on Natural Language Processing 2002 + Oriental COCOSA Workshop 2002. Hua Hin, Prachuapkirikhan, Thailand. See: [kind.siit.tu.ac.th/snlp-cocosa2002](http://kind.siit.tu.ac.th/snlp-cocosa2002)

**May 20-23:** TAG+6: The Sixth International Workshop on Tree Adjoining Grammars and Related Frameworks. Venice, Italy. [www.ircs.upenn.edu/tag](http://www.ircs.upenn.edu/tag)

**May 26-28:** LREC Pre-Conference Workshops: Machine Translation Evaluation, OntoLex2: Ontologies and Lexical Knowledge Bases; Acquiring, Organizing, and Evaluating Terminology; International Standards of Terminology and Language Resource Man-

agement; And more! Las Palmas, Canary Islands, Spain, immediately preceding LREC. See: [www.lrec-conf.org/lrec2002/index.html](http://www.lrec-conf.org/lrec2002/index.html)

**May 29-31:** LREC (Language Resources and Evaluation Conference). Las Palmas, Canary Islands, Spain. Organized by ELRA (European Language Resources Association). Antonio Zampolli, Instituto di Linguistica Computazionale, Pisa, Italy (Conference Chair). See: [www.lrec-conf.org/lrec2002/index.html](http://www.lrec-conf.org/lrec2002/index.html)

**June 1, 2:** LREC Post-Conference Workshops: Arabic Language Resources and Evaluation; First International Workshop on UNL, other Interlinguas and their Applications; Towards a Roadmap for Language Resources; Beyond PARSEVAL: Toward Improved Evaluation Measures for Parsing Systems; And more! Las Palmas, Canary Islands, Spain, immediately following LREC. See: [www.lrec-conf.org/lrec2002/index.html](http://www.lrec-conf.org/lrec2002/index.html)

**June 24-27:** TALN 2002. Organized by l'ATALA (Association pour le Traitement Automatique des Langues). Nancy, France. See: [www.loria.fr/projets/TALN](http://www.loria.fr/projets/TALN)

**June 10-12:** ISWC—The International Semantic Web Conference. Sardinia, Italy. See: [iswc.semanticweb.org](http://iswc.semanticweb.org).

**June 24-26:** Recital 2002—Student Sessoin at TALN. Nancy, France. See: [www.loria.fr/projets/TALN](http://www.loria.fr/projets/TALN)

**July 1-3:** INLG 2002 (Second International Natural Language Conference). See: [www.research.att.com/~rambow/inlg/inlg.html](http://www.research.att.com/~rambow/inlg/inlg.html)

**July 6-7:** EMNLP—Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania, Philadelphia, PA, USA. Organizers: Jan Hajic (chair); Yuji Matsumoto (co-chair). **Submission deadline: April 4, 2002.** See: [ufal.ms.mff.cuni.cz/~hajic/emnlp02/](mailto:ufal.ms.mff.cuni.cz/~hajic/emnlp02/)

**July 7-10:** ACL-2002. University of Pennsylvania, Philadelphia, PA, USA. Organizers: Pierre Isabelle (general

chair), Eugene Charniak and Dekang Lin (program co-chairs) Martha Palmer (Local Arrangements). See: [www.acl02.org](http://www.acl02.org)

**July 11-12:** ACL Post-Conference Workshops. University of Pennsylvania, Philadelphia, PA, USA. Computational Approaches to Semitic Languages; EMNLP; Morphological and Phonological Learning; Speech-to-Speech Translation: Algorithms and Systems; Unsupervised Lexical Acquisition; Word Sense Disambiguation. **Submission deadlines range from February 22 to April 5.** See: [www.acl02.org](http://www.acl02.org) for exact dates.

**July 14-19:** Lexicom: Lexicography Master Class. At the Information Technology Research Institute (ITRI), University of Brighton, England. Trainers Sue Atkins, Adam Kilgarriff, and Michael Rundell. **Early registration deadline March 31, 2002.** See [www.itri.bton.ac.uk/Lexicom](http://www.itri.bton.ac.uk/Lexicom)

**July 15-26:** ELSNET Summer School. Odense, Denmark. **Pre-registration deadline: April 15, 2002.** See [www.elsnet.org](http://www.elsnet.org), or contact Holmer Hensen [hensen@nis.sdu.dk](mailto:hensen@nis.sdu.dk) for more information.

**July 23:** SAAKM: Workshop on Semantic Authoring, Annotation, and Knowledge Markup. **Submission deadline: April 28, 2002** See [saakm2002.aifb.uni-karlsruhe.de](http://saakm2002.aifb.uni-karlsruhe.de).

**August 7-10:** 16th World Congress of the International Federation of Translators, "New Ideas for a New Century". See: [www.stibc.org/fit2ki.htm](http://www.stibc.org/fit2ki.htm).

**August 12-16:** Recent Advances in Speech Translation Systems. Trento, Italy. Workshop held in conjunction with 14th European Summer School in Logic, Language and Information (ESSLI). **Submission deadline: April 15, 2002.** See: [www.essli2002.it](http://www.essli2002.it) or contact [alavie@cs.cmu.edu](mailto:alavie@cs.cmu.edu).

**August 24-September 1:** COLING 2002. Academia Sinica, Taipei, Taiwan. Organizers: Chu-Ren Huang, Academia Sinica (Conference Chair); Winfried Lenders (Program Chair), Univ. of Bonn, Germany [Lenders@uni-bonn.de](mailto:Lenders@uni-bonn.de); Antonio Zampolli (Workshop Chair) [pisa@itc.pi.cnr.it](mailto:pisa@itc.pi.cnr.it). See: [www.coling2002.sinica.edu.tw](http://www.coling2002.sinica.edu.tw)

**September 2-3:** Third International Workshop on Natural Language and Information Systems. Aix-en-Provence, France. [Www.ifs.univie.ac.at/~wnwnlis2002.htm](http://www.ifs.univie.ac.at/~wnwnlis2002.htm).

**September 9-12:** TSD 2002—Fifth International Conference on Text, Speech and Dialog. Brno, Czech Republic. Organized by the faculty of Informatics, Masaryk University, Brno, and the Faculty of Applied Sciences, University of West Bohemia, Pilsen. **Submission deadline for regular papers: March 15, 2002; for demonstration papers: July 31, 2002.** See: [www.ft.muni.cz/tsd2002/](http://www.ft.muni.cz/tsd2002/)

**September 18-20:** DAARC2002, Discourse and Anaphora Resolution Conference. Lisbon, Portugal. Organisers: Antonio Branco, Tony McEnery and Ruslan Mitkov *Contact: daarc2002@di.fc.ul.pt*

**September 16-20:** ICSLP2002, 7th International Conference on Spoken Language Processing. Denver, Colorado, USA. **Submission deadline: March 30, 2002** Organised by the Center for Spoken Language Research, Univ. of Boulder. See: [www.icslp2002.org](http://www.icslp2002.org)

**October 8-12:** AMTA 2002. Tiburon, California, "From Research to Real Users". Organizers: Elliott Macklovich, University of Montreal (General Chair) [macklovi@iro.umontreal.ca](mailto:macklovi@iro.umontreal.ca); Stephen Richardson, Microsoft, (Program Chair) [steveri@microsoft.com](mailto:steveri@microsoft.com); Violetta Cavalli-Sforza, University of California at San Francisco (Local Arrangements) [vcs@sfsu.edu](mailto:vcs@sfsu.edu); Robert Frederking, Carnegie Mellon University (Workshops, Tutorials, and Website) [ref@cs.cmu.edu](mailto:ref@cs.cmu.edu); Laurie Gerber (Exhibits) [lgerber@gerbersite.com](mailto:lgerber@gerbersite.com). **Submission deadline: April 15, 2002.** See [www.amtaweb.org/AMTA2002](http://www.amtaweb.org/AMTA2002) for the latest developments.

## 2003

**July 7-12:** ACL-2003. Sapporo Convention Center, Sapporo, Japan. Junichi Tsujii, University of Tokyo, (general chair); Kenji Araki, Hokkaido University (local organization chair) See: [www.ec-inc.co.jp/ACL2003/](http://www.ec-inc.co.jp/ACL2003/)