[*Translating and the Computer 21*. Proceedings… 10-11 November 1999 (London: Aslib)]

Integrating Translation Technologies Using SALT

Gerhard Budin, University of Vienna <gerhard.budin@univie.ac.at>, **Alan K. Melby,** Brigham Young University at Provo <akm@compuserve.com>, **Sue Ellen Wright,** Kent State University, <swright@kent.edu>, **Deryle Lonsdale,** Brigham Young University at Provo <lonz@byu.edu> , and **Arle Lommel,** Brigham Young University at Provo <fenavad@ttt.org>

Summary

The acronym *SALT* stands for *S*tandards-based *A*ccess to multilingual *L*exicons and *T*erminologies. The objective of the SALT project is to develop and promote a range of tools that will be made available on the World Wide Web to various user groups, in particular translators, terminology managers, localizers, technical communicators, but also tools developers, database managers, and language engineers. The resulting toolkit will facilitate access and re-use of heterogeneous multilingual resources derived from both NLP lexicons and human-oriented terminology databases.

Motivation: Availability of and Accessibility to Language Resources

The increasingly global nature of business, trade, and industrial research and development has led to a higher visibility for language technologies, including such multilingual activities as translation and localization, all of which require terminology management. At the core of all language technologies is an infrastructure based on words: lexicons used by computers and terminologies used by humans. The availability of these resources in the form of dictionaries, terminologies, ontologies such as thesauri, text corpora, and the like is essential for the development and use of all other language technologies. Both today's human translation tools and machine translation systems require sophisticated, machine-processable information about terms and the concepts they designate. In addition, it remains apparent that human and machine translation will continue to work hand in hand for the foreseeable future. Consequently, terminologies for human translators and lexicons for machine translation systems must be coordinated in order to obtain consistency, which is a primary requirement for the re-useability of documentation. However, each of these types of data has its own historical tradition, methods, and representations, which means that these data can occur in mutually incompatible forms. Thus terminology exchange and coordination pose a serious challenge to our efforts to increase efficiency throughout the multilingual document production chain through the integration of a full range of tools and methodologies.

An integrative Format (XLT) for Lexical and Terminological Resources

One of the consequences of this incompatibility among resources is that there is an urgent need for a universal electronic format for the interchange and dissemination of information from lexicons, terminologies, and ontological systems. This need has been felt for some time and substantial progress has been made toward filling it in EU-funded projects such as TRANSTERM, INTERVAL, INESTERM, and OTELO, along with ISO standards 12200:1999 and 12620:1999. SALT will integrate the best features

and most successful products of these predecessor projects by developing and then implementing an integrative format (XLT) for the free exchange of lexicons and terminologies (lex/term data). Yet another parallel development, broader than the language industries, has been the definition of XML by the World Wide Web Consortium. XML is a widely accepted new method of transporting any kind of structured data. In XLT, the SALT project brings together the results of these three threads of development (EU-funded work, ISO standards, and XML) to form a baseline for future work in this area by (1) integrating the best features of existing formats, (2) creating toolkits for use by developers in the language industries, (3) disseminating these resources as freeware via a permanent website, and (4) publishing research findings.

As a result of the project, sharing lex/term-data will become easier, more convenient, and more standards-based, which in turn will assure more consistency across documents, a tighter integration of human and machine translation technologies, and the freer interchange of language data organized according to its associated concepts.

Who is SALT?

The SALT project brings together a transatlantic team of experienced researchers who have worked on various aspects of word-related data representation and exchange for a number of years, frequently in collaboration with one another. Experts from academic, governmental, and commercial organizations form the core group of main partners who will oversee and manage the project. In addition, the project includes a wide array of industrial support partners who work in software tools development, localization services, information technology, and language industry oversight. Advisory groups that have joined the SALT effort include international organizations overseeing terminology, translation, computational linguistics research, and localization. Overall, the project's composition reflects the wide range of interests represented worldwide with respect to issues involving the exchange of data from lexicons and terminologies. It also unites researchers, developers, and consumers in the task of solidifying, leveraging, and popularizing an infrastructure that will be crucial to multilingual technologies in the twenty-first century.

Gaining Critical Mass and the Impact of the SALT Project

SALT will refine and promote XLT, using input from its partners in various sectors. By involving all these key players, SALT will be able help XLT reach "critical mass" and become the universally adopted and implemented linguistic exchange format that is needed. Furthermore, careful monitoring by industry leaders concerned with actual implementation of the XLT format will ensure that solutions proposed within SALT will actually be used in industry after the completion of the funded project instead of gathering electronic dust in some computer archive.

The impact of the SALT project on the typical end-user will be significant. Commercial translation technology products will begin to include a new filter that converts between the native format of the product and XLT. Terminological data will then flow much more easily between colleagues. Instead of being obliged to have a

software engineer write a custom conversion program for each new lex/term resource, end-users will simply request data in XLT format. Within an organization, SALT will help companies coordinate their machine translation lexicons and their human-oriented termbases, thus achieving increased consistency in the use of technical terms. This integration of lex/term resources has already been implemented in a few companies using expensive proprietary development. SALT will facilitate the spreading of this type of integration to many other companies that will be able to build on the SALT open standards and toolkit. A universal lex/term exchange format is important to all translation technologies, not just to integrate machine translation and human translation but also to facilitate the flow of termbase information from one translation tool to another and the dissemination of terminological data in machine processable form over the Internet.

In order to ensure maximum impact in the language industries and to respond directly to real needs in different end-user groups, the work will be performed by the project consortium in close cooperation and coordination with leading industrial partners in the language, translation, publishing, and localization industries, as well as with academic and commercial partners, and with an advisory group of governmental and non-governmental organizations. The project will start with a description and analysis of existing and emerging standardized or proprietary formats, in particular the ISO standard MARTIF, related new work items such as GENETER, industrial standards such as TBX of the OSCAR group within LISA, the Interval interchange format, OLIF from the OTELO project, as well as other relevant formats for data encoding and interchange. This work will be complemented by gathering and analysing sample data from real-life databases and data repositories (human-use databases as well as machine-translation lexicon entries) in different sectors of the economy in Europe and beyond. A data analysis and mapping procedure will provide the required knowledge and specifications for refining the integrative format XLT. Careful attention will be paid to multilingual ontologies, as they are usually a special problem in data interchange. Utilities and tools will be developed on the basis of this work, allowing the user to perform a diversity of tasks including data conversion, interchange, validation and presentation. The technical encoding level will be in XML, with processing mainly in XSL and Perl. A website will be developed that will be tailored toward different user groups, responding to their specific needs and offering a diverse range of services. Based on the active integration of numerous industrial partners, user feedback and user evaluation of prototypes and demonstrators are built into the project. A dissemination, exploitation and promotion strategy will be developed right from the start and will be constantly fine-tuned as the development work progresses.

The SALT Approach: Building on Previous Projects and Integrating Results

The approach used in SALT is to build on the work that has already been completed in several projects that have dealt with sharing lex/term-data (INCLUDING OTELO, INTERVAL, TRANSTERM, and INESTERM) and to integrate existing standards, including ISO standards from Technical Committee (TC) 37 (ISO 12200 and ISO 12620:1999) and industry standards from the Localization Industry Standards Association. The multilingual aspects of the project, coupled with the long-standing tradition of collaboration that exists among partners, including industrial partners, will be important to the long-term success of SALT.

Mapping Ontologies and Data Structures

The interfacing of ontologies by mapping positions in one system onto positions in another is recognized as a challenging problem that arises when attempting to minimize information loss going from one termbase to another if the two termbases do not use the same ontology. The incongruity between database-related classification systems has long impeded the acceptance of foreign data into any kind of translation-oriented lexbase or termbase because the slightest variation in the position of a concept within a system can change the foreign equivalent required in context. Nevertheless, it is widely agreed that no single ontological system will ever meet the wide variety of needs present in science and industry. Consequently the facilitation of exchange between and among different ontological approaches promises to remove a major stumbling block on the road to universal sharing of lex/term data.

SALT Results: A website with Tools to Download

The overall goal of the entire project is practical: to build a useful website that will allow end-users (1) to submit presumed XLT files for formal validation of syntactic correctness and permissible element values (not semantic validation such as checking the internal consistency of definitions), (2) to merge two XLT files, and (3) to convert data files from one format to another via XLT as a pivot format, so long as they are in one of the specific formats supported by the website. The inclusion of many partners in Salt is intended to ensure that XLT will reach "critical mass" so that major tools developers will incorporate some level of XLT support in their commercial products and so that developers in charge of proprietary lex/term-bases will use XLT to communicate with other lex/term-bases. In other words, the project is intended to create a universal format for exchanging lex/term-data. The demonstration website will of course use the XLT toolkit and will provide a venue that will foster rapid uptake of the terminology exchange. However, the website cannot include conversion routines between XLT and every possible format. As soon as the XLT toolkit is ready, all user groups are encouraged to use the XLT toolkit to facilitate

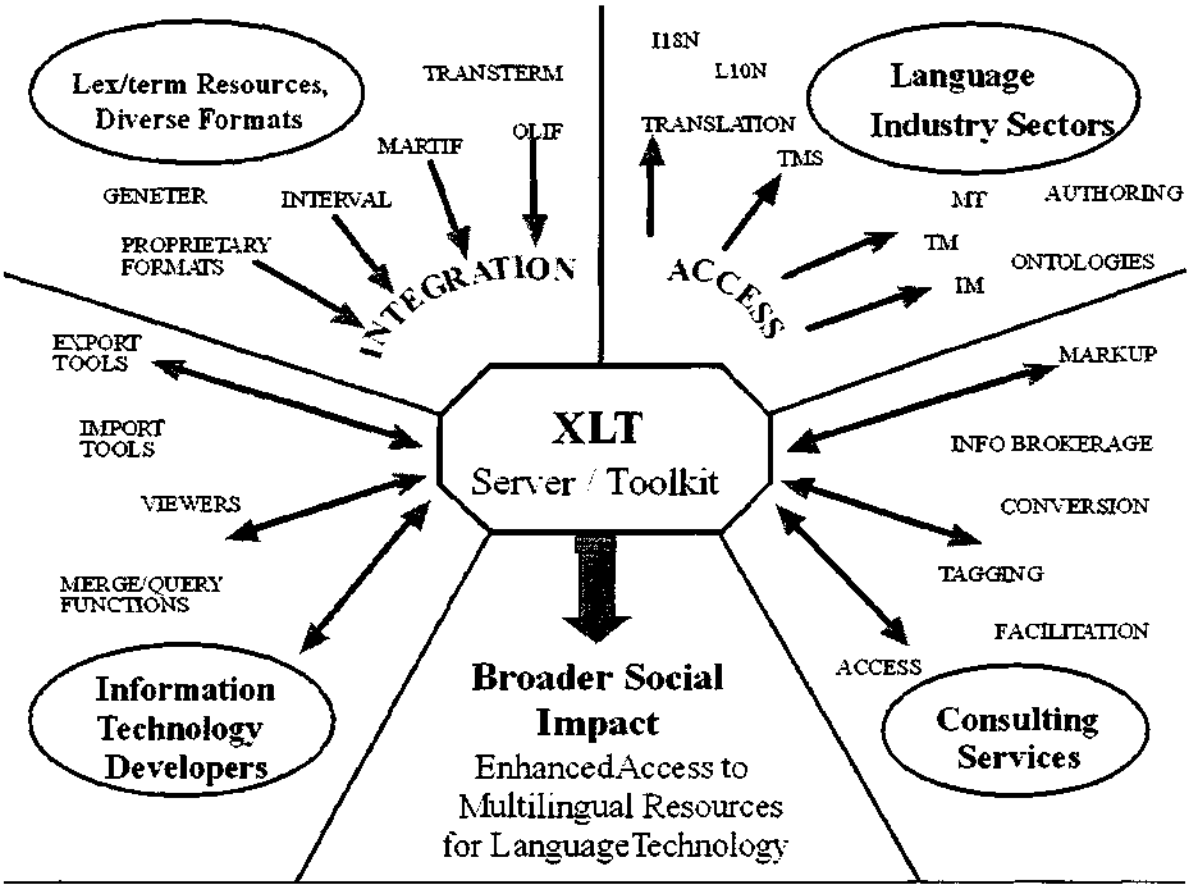

Figure 1: *The Impact of SALT*

the task of converting their proprietary lex/term-data into XLT format or vice versa. A tutorial will also be provided to assist end users in developing their own custom conversion routines to and from XLT. Detailed specifications, sample files, and XLT tools will be made universally available as freeware to commercial developers, consultants, and end users, with the only restriction being strict adherence to the specifications in order to claim XLT compliance. Figure 1 shows the objectives in visualized form and illustrates in their interrelationships: (1) Integration of formats, (2) convertibility and the resulting access to resources in various industry sectors, (3) applicability in different user environments through the development of utilities by tools developers, and (4) consulting services. The interaction of these four dimensions leads to the broader social impact of the overall project.

This figure shows the integrated XLT format at the center of a series of potential interactions, beginning with lex/term data undergoing the integration process from numerous original sources and formats and then flowing into XLT and back out on the other side of the diagram to provide access from a wide variety of applications. The bottom left section of the graphic represents the developers of information technology tools who will contribute conversion utilities for use with their tools in order to enable users to employ XLT without undue difficulty and without having to know a great deal about the theoretical aspects of data interchange. The lower right-hand section outlines some of the services that can be performed either using the server or the XLT toolkit for the processing and dissemination of lex/term information in a wide range of environments. The broader social impact is reflected in enhanced multilingual

resources for language technology.

The reuse of data in integrated environments involves accessing, managing, and exchanging information in very practical multilingual contexts.

SALT Benefits

SALT promises a comprehensive set of advantages within the information management and document production chain:

1. Efficiency: faster insertion of new terms into databases and between tools
2. Consistency across documents
3. Integration of various standards and data formats
4. Innovation: The XLT format is based on XML and uses XML schemas and XSL.
5. Economy: The freeware toolkit will facilitate development.

Because of these features, the XLT format is a candidate to become a universal exchange format.

Integration with Existing Standards

Rather than starting with a totally new formalism, XLT essentially integrates MARTIF (ISO 12200) and OLIF (from the OTELO project). Neither MARTIF nor OLIF is XML compliant, but XLT adapts both of them to XML and integrates them into a single format. The GENETER format from the INESTERM project does not allow for the representation of MT lexicon data, but it is highly compatible with MARTIF, since both MARTIF and GENETER are based on the data categories specified in ISO 12620 and have a similar high-level conceptual data model. Therefore, it is anticipated that one of the conversions developed in the SALT project will be between XLT and GENETER. Lossless bi-directional conversion between the terminology data aspect of XLT and GENETER may require proposing some modifications to GENETER and MARTIF. Modifications to MARTIF have already been proposed to ISO Technical Committee 37, some of which are intended to facilitate MARTIF-GENETER conversions. The TRANSTERM conceptual data model from the TRANSTERM final report will be used in the SALT project as input to the XLT conceptual data model, with adaptations as needed. Although XLT is not intended to exchange translation memory databases, it takes into account certain features of TMX (Translation Memory eXchange format) from LISA'S Oscar in the areas of Unicode handling and meta-markup. (The *O*pen *S*tandards for *C*ontainer/content *A*llowing *R*e-use (OSCAR) special interest group (SIG) of the Localisation Industry Standards Association (LISA) is responsible for the development of data interchange formats between and among various translation tools.) By integrating MARTIF and OLIF, relating formally to GENETER and TRANSTERM, and by coordinating with OSCAR, existing standards for lex/term-data exchange are brought under the scope of one format, XLT. Incorporating features of TMX makes XLT more easily acceptable to translation memory tools developers who are already implementing TMX.

XLT Modularity

MARTIF and OLIF are both modular, which facilitates their integration. To begin with, OLIF is used primarily for machine translation lexicons, MARTIF for complex

termbases, and TMX for translation memories. Integrated in XLT, one module provides a simple structural framework for exchange files, and other modules specify particular data categories available within the structural framework described by the simple core structure. Particular data categories on the human-oriented terminology side are introduced by what is called a "data category" module. This module consists of a file that lists the data categories and associated properties used in any given exchange environment. There is also an "NLP description" module on the "lex" side consisting of an element that lists specific NLP feature value pairs used by the OLIF aspects of XLT. This modularity makes it easier for a developer to understand XLT than is the case with a monolithic format that includes both structure and data category specifics in a single description. Another benefit of modularity is that the format is more easily adapted to the needs of a specific user group. The flavor of XLT developed for the SALT project will include a data category module called the master set, which lists the total inventory of permissible data categories that can be used in the SALT environment. Various user-group-specific subsets will also be defined. For example, the subset for the localization industry will be called TBX (TermBase eXchange) and will be defined by the OSCAR group. Although the SALT project will specify a particular master data-category module, the XLT approach can be adapted to other contexts for other projects by specifying a different master data-category module while using the same structural framework. This modular approach allows the configuration of specific formats and is more attractive to tool developers than an approach that begins with a completely specified format and then defines subsets of it without a master plan that relates those subsets to each other.

Conceptual Data Model

The basis of the XLT format is a conceptual data model that defines the data categories and the relationships among them. In the SALT project this data model has to be as generic, modular, and configurable as possible. Meta data categories possess one or several instances, i.e. data categories that are collected in an inventory. For instance, the meta data categories *Term* and *Note* consist solely of those elements, but other categories include a set of sub-categories.

In the process of defining the conceptual data model, the SALT project team will coordinate with ISO Technical Committee 37, subcommittee 3, working group 4. The data model (a high-level meta model, as it is currently called in the ISO working group) will likely include global information, concept entries, and shared resources such as bibliographic entries. The concept entries will include four levels: the concept level that applies to multiple languages [which is slightly different than being language independent] such as definitions, images, and subject field, the language level that separates the various languages, the term level, and the term component level.]

The SALT data categories are a subset of ISO 12620, in particular:

1. Term
2. Term-related information
3. Equivalence
4. Subject Field
5. Concept description

6. Concept relations
7. Concept system information
8. Note
9. Documentary languages
10. Administrative information

How to participate?

There are many ways to participate in the SALT project:

- For tools developers: to implement an XLT interface functionality in their tool(s)
- For companies and organizations with active terminology management teams: to supply test data; to give feedback to XLT conversion tools or their specifications provided
- For professional organizations: to help promote the SALT ideas and objectives to increase its critical mass and critical impact in the translation and localization industries

Bibliography

[Note: all references given represent the complex and dynamic knowledge spaces that serve as the basis for the SALT project.]

Ahmad, K., Ogonowski, A., Dauphin, E., Sta, J.-D., Arppe, A.. 1996. Engineering Terminology —A Case for a Linguistically-informed Terminology Database. [Transterm reference] In: *TKE '96: Terminology and Knowledge Engineering.* Indeks Verlag, Frankfurt am Main, pp. 166-178

Bosak, J., Bray T. 1999. "XML and the Second-Generation Web: The combination of hypertext and a global Internet started a revolution. A new ingredient, XML, is poised to finish the job." *Scientific American* 5/1999.

Carroll, S. 1999. The Technical content of TMX 1.1: A format for the exchange of data between competing translation database systems. *Multilingual Computing & Technology.* No. 22 Vol. 9/6, pp. 48-49.

Connolly, D. 1998. "Nature's Web Matters: The XML Revolution." *Nature,* October 1998.

Galinski, C. 1988. Advanced Terminology Banks Supporting Knowledge-Based MT: Some reflections on the costs for setting up and operating a terminological data bank (TDB). In: *Conference Proceedings of New Directions in Machine Translation.* Foris Publishers, Dordrecht/Providence, pp. 1-15.

ISO 12 200:1999, *Computer Applications in Terminology — Machine-readable Terminology Interchange Format (MARTIF) — Negotiated Interchange.* International Organization for Standardization, Geneva.

ISO 12620:1999, *Computer Applications in Terminology — Data Categories.* International Organization for Standardization, Geneva.

Melby, A.K. 1982. Multi-level Translation Aids in a Distributed System In: *Proceedings of COLING '82.* Jan Horecky, ed. North Holland Publishing Company, Amsterdam, pp. 215-220.

Melby, A. K. 1988. Electronic Dictionaries and the Translator's Workbench. In: *Manuscripts and Program of the International Symposium on Electronic Dictionaries.* ISED, Tokyo, pp. 62-64.

Melby, A.K., Wright, S.E. 1999. Leveraging Terminological Data for Use in Conjunction with Lexicographical Resources. In: Sandrini, P. [ed.]. Proceedings of TKE' 99. Terminology and Knowledge Engineering, 23-27 August 1999, Innsbruck, TermNet, Vienna, pp. 544-569

OSCAR (Open Standards for Container/Content Allowing Re-use). 1999. http://www.lisa.unige.ch/tmx/index.html.

Sager, J.C. 1990. *A Practical Course in Terminology Processing.* John Benjamins Publishing Company, Amsterdam.

Schmidt, G., Hejl, M, Jornitz, G., Luedde, A., Painke, E., Schoeck, E. 1999. Overview of Current IBM Translation Technology. In: *TAMA '98-Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium: Tools for Multilingual Communication.* TermNet, Vienna, pp. 25-39.

Thurmair, G., Ritzke, J., McCormick, S. 1999. The Open Lexicon Interchange Format — OLIF. In: TAMA '98 — Terminology in Advanced Microcomputer Applications. Proceedings of the 4th TermNet Symposium: Tools for Multilingual Communication. TermNet, Vienna, pp. 237-262.

Walmer, D.I999. "One Company's Efforts to Improve Translation and Localization." *Technical Communication: Journal of the Society for Technical Communication.* 46/2, pp.230-237.

Wright, S. E.., Budin, G. 1997. "Lexicography vs. Terminology," In: *The Handbook of Terminology Management,* Vol. I. John Benjamins Publishing Company, Amsterdam/Philadelphia, p. 328.

XML Schema Part 1: Structures. W3C Working Draft 6-May-1999. http://www.w3.org/1999/05/06-xmlschema-l/.

XML Schema Part 2: Datatypes. World Wide Web Consortium Working Draft 06-May-1999 (with accompanying schema and DTD). http://www.w3.org/1999/05/06-xmlschema-2/