[*Translating and the Computer 20.* Proceedings from Aslib conference, 12 & 13 November 1998]

Towards a Multi-language Multi Script Web Based Reference & Terminology System

Olaf-Michael Stefanov, Chief
Linguistic Support Unit
Translation and Editorial Service
Division of Administrative and Common Services
United Nations - Vienna

Vienna International Centre
Wagramerstrasse 5
P.O. Box 500
A-1400 Vienna, Austria

Olaf-Michael.Stefanov@unvienna un or at

Introduction

The United Nations Office at Vienna (UNOV) has what may be one of the first databases containing terminology and bibliography in more than two scripts and accessible via the World Wide Web. In this paper I would like to give an overview of why we took the route we did, what we have achieved, and where we expect to go.

Rather than reading this paper, I will present a set of slides summarising key points and show examples from the system. I also wish to reach out to those of you who are also assembling and developing reference and terminology data, and ask you to consider sharing with us and looking at mutual exchange and / or access to our respective materials.

Multi-language Multi Script Texts

Publishing in multiple languages increases in complexity as the number of languages used grows, and in proportion to how the written languages diverge. While most issues were solved quite a few decades ago, as traditional typesetting goes, a look at e-mails from other countries or across language boundaries suffices to remind us that when computers are involved problems still abound.

Multilingual publications often have different languages in different sections or in parallel columns. This permits parts written in languages based on the Latin alphabet, such as English, French, and Spanish to be typeset together, while parts in Arabic, Chinese, and Russian are set separately. The masters of these four distinct language sets, each based on its own form of writing, i.e., script, are then assembled into a single publication, either on a composition light table, or, in the bindery of the print shop, each part having been printed separately.

Some publications, such as multi-lingual terminology bulletins and glossaries, pose a higher level of challenge. They require different languages, in multiple scripts, to coexist in the same text lines or text blocks. Such publications require either a very large amount of manual cut and paste, or typesetting systems able to handle all required scripts. Some such systems became available between the late 1950s and early 1970s. Due to their expense, however, few organizations ever published and printed works in multiple scripts in house.

As traditional typesetting was replaced by systems using computer-generated fonts the lines between languages began to blur. First, it became generally expected that texts in one Latin-based language correctly show all accents and diacriticals for all other Latin-based languages in quotations, terminology, or official titles in other languages. Later, it became increasingly common to include parts written in Greek, Russian, or other Cyrillic languages, into the base corpora along with Latin-based languages as texts were composed and published.

Overall, though, the path from source to printed word, for texts in multiple scripts, continued to be a long and costly process, and therefore updates were comparatively infrequent, which was a major drawback for terminology bulletins and glossaries.

Early United Nations terminology databases

By the early 1970s, as mainframe computers were introduced in international organizations, these were also put to use to store terminology, in the hope of reducing the time and cost of keeping up-to-date. Early systems, such as United Nations headquarters' UNTERM, restricted themselves to the character-set of the mainframe, EBCDIC[1]. Since this character-set was designed only to accommodate English, no possibility existed to include French or Spanish accents. These were therefore ignored (on the understanding that good translators would know where to put them). For Russian, a transliteration using the Latin alphabet was employed.

In Vienna and Paris a system was developed that made use of the possibility of a so-called Programmed Symbol Set (PSS) of characters to extend and modify EBCDIC. This system, based on CDS/ISIS[2], required special terminals to display and accept input in such modified character-sets, and made similar requirements of printers. With the right terminals and printers, ISIS permitted French and Spanish accented characters to be used by the late 1970s. To support requirements to translate local material, including legal texts, German (albeit without special German characters) was also included, even though German was not an official language. An extension to cover Cyrillic characters allowed the inclusion of Russian by the early 1980s.

The language service in Vienna, then operated by UNIDO, had, with ISIS, a good basic tool with which to build up a solid reference and terminology database.

Initially work was limited largely to the industrial development sectors of UNIDO. After 1981, the establishment of the Centre for Social Development and

[1] Extended Binary Coded Decimal Interchange Code, the de facto mainframe standard encoding for IBM and compatible mainframes.

[2] Computerized Documentation System/Integrated Set of Information System, started by Giampolo Del Bigio at IAEA in the early 1970s, and continued by him at UNESCO until recently. The version used for UNOV's database was enhanced by UNIDO in the 1980's.

Humanitarian Affairs in Vienna brought the need for social issue terminology, while the transfer to Vienna of three drug control units brought their special terminology requirements with them, while the arrival of UNCITRAL brought with it the need for international trade law terminology, and the Office of Outer Space Affairs needed space science vocabulary as well as meteorological and various scientific subjects areas. The recent establishment of the Preparatory Commission for the Comprehensive Test-Ban Treaty Organization (CTBTO) has added hydrology, geology, seismology, as well as disarmament issues, and the needed to expand meteorology and space sciences terminology and reference materials. For some of these subject areas United Nations headquarters and the Office at Geneva provided much of the terminology material.

Limits of conventional character-sets

It is important to have a basic understanding of computer character-sets, to understand their limitations, to understand why "powerful computers" have so much difficulty handling accents and multiple scripts.

Virtually all computers, from early mainframes to computers of all sizes developed over the past 20 years, use an internal code wherein 8 bits (individual on/off switches) are used to represent each character. Since $2^8 = 256$, 8-bit character-sets can accommodate only a maximum of 256 characters. In practice, 30, 50, or even more of these possibilities must be reserved for internal purposes, reducing the actual number of possible viewable characters to around 200.

The Latin alphabet, without accented characters, i.e., as used in English, has 26 letters, each in two formats, upper and lower case. Since these are defined separately, the base alphabet for English requires 52 characters. To these are added: the 10 digits 0-9, 10 common punctuation marks: full-stop, comma, colon, semi-colon question-mark, and exclamation mark, open and close parentheses, as well as a basic pair of so-called "straight" single (') and double (") quotes. The base set includes 14 additional signs: ampersand (&), asterisk (*), at sign (@), equal sign (=), percent sign (%), plus sign (+), hash sign (#), hyphen (-) or minus sign, underscore (_), forward slash (/), oblique slash (\ ), vertical bar (|) less than sign (<), and greater than sign (>). It also includes a "space" character, and a character called "carriage-return-line-feed" now often referred to as a hard return or hard-line-break. It also includes, as stand-alone characters, a grave-accent-mark (`) and a tilde (~).

Within each major family of character-sets, such as EBCDIC, ASCII[3], and ANSI[4] all the above 90 characters have the same code values. These are therefore the only characters that will always look the same.[5]

[3] American Standard Code for Information Interchange

[4] The Acronym stands for the American National Standards Institute. In fact, "ANSI" is for a code originally defined under the aegis of this Institute, not for the Standards Institute itself.

[5] Character-sets must not be confused with Keyboard layouts. A given character set may be used with a number of different keyboard layouts. On the other hand, similar keyboard layouts can be used with different character-sets. A classic example is the standard US-English keyboard which will result in a different internal character-set encoding when used with a mainframe (EBCDIC/US), a PC running

Of course, all character-sets have more characters, beyond this base, but unlike those listed above, their encoding varies from character-set to character-set and even from implementation to implementation[6]. Even the dollar-sign ($) designed into the "base" set by Americans is, in some implementations, replaced by local currency symbols. A file into which a dollar-sign is keyed in one computer, may, when viewed using a different code page, or printed on a printer with a different character-set, suddenly appear as a pound-sterling sign (£) or other character.

The fate of the dollar sign is shared, to a greater or lesser extent by the pairs of square brackets ([and]), curly brackets ({ and }), and the caret (^), all of which are also defined as part of the "base" character-set. They are coded differently in at least some implementations. For example, someone typing what they believe to be a German "scharfes-s" (ß) will get a square bracket in some implementations, or a lower case a-acute (á) in others. A "§" (symbol to denote a section of a legal document, as used in several European countries) converts to (õ) (lower case letter o with a tilde).

The full set of accents and accented characters for the Latin alphabet is so large that most modern character-sets differentiate between "Western Latin", covering the needs of the countries of the European Union and EFTA, and "Central European Latin", which covers the Slavic languages using the Latin alphabet and Hungarian.

What applies to Latin applies equally to Greek, Cyrillic and Arabic, as well as to other families of alphabets.

No matter how hard one tries, it is not possible to squeeze the complete set of Latin characters plus more than one other character-set into 256 characters.

Virtually all implementations today therefore apply the following compromise. Each national or linguistic character-set includes a full set of characters, diacriticals, and symbols needed for a particular language and/or country, plus US-English[7]. Depending on how many character possibilities are taken up by local requirements, some additional symbols from Western European Latin are usually included in other character-sets. French accents or accented characters are often included. Scandinavian or German extensions, on the other hand, are less likely in other characters-sets, such as Greek, Russian, Arabic, or any of the Asian alphabets.

Some Asian languages pose a different challenge, each one needing considerably more than 256 characters by itself. 2-byte[8] codes were developed in response.

DOS (e.g., ASCII DOS code page 437), a PC running Windows (ANSI 1252 US standard), or an Apple Macintosh computer (e.g., Eur. West Macintosh 10000)

[6] Encoding implementations are often known as code-pages. These may differ along national or language lines, or between different types of computers.

[7] Actually, the basic character-set as defined above, i.e., Latin without accents, plus basic punctuation.

[8] A "byte" is defined as consisting of 8 bits (on/off values) to define a character. Since each bit is a single on/off switch in the computer, 8 bits permit 2^8 or 256 possibilities, while 2^{16} allows 65,536 possibilities. A "byte" is often equated simply as being a "character" in most computer documentation.

One mode of implementing a mixture of Asian languages such as Chinese or Japanese is to add one or two internal control characters[9] signalling the computer to switch between characters in 8 or 16 bit format. English text is thus represented in 8-bit format. Chinese or Japanese text is preceded by a "switch" indicating that the characters which follow are composed from 16 bits each (or the equivalent of 2 English characters in size, hence the term "2-byte encoding"). Another switch toggles back to 8-bit format, as needed. A similar need, to switch direction, exists for texts in languages such as Arabic or Hebrew, which flow from right to left, rather than from left to right.

UNOV opts for UNICODE in its broadest implementation - UTF8

Recently the concept of a universal encoding scheme has gained ground. The most prevalent versions of such a scheme are referred to as Unicode. Unicode that stores the basic English characters noted above in 8 bits and other language scripts in 16 bits is defined as UTF-7. Unicode that stores all languages in 16 bit character encoding is defined as UTF-8.

In order to handle all the Latin alphabet requirements for French and Spanish, as well as German, along with Arabic, Chinese, and Russian, we, at the United Nations Office at Vienna chose UTF-8 as the encoding for VINTARS[10], our new reference and terminology database.

Ease of Access

As indicated above, UNOV's traditional reference and terminology system required special terminals, able to work with the mainframe Programmed Symbol Set (PSS).

These special terminals are, of course, long gone, having been replaced by personal computers with mainframe terminal-emulation software. But PSS proved difficult to emulate. UNOV found itself limited to a particular emulation programme, with fewer features than other emulations. Recently, as PCs were upgraded from 16-bit to 32-bit operating systems (i.e., Windows-95 and/or Windows/NT in place of Windows 3.1x) it was discovered that support for PSS was not being migrated to a 32-bit version. In Windows-95 it continues to be possible to run 16-bit software; under Windows-NT this is not possible. Additionally, this software conflicts with certain other Windows options; e.g., Arabic Windows-95.

[9] If only one "switch" character is used it works as a toggle. Depending on whether the current state is 8-bit or 16-bit, it switches to the "other" mode. Two distinct characters permit a defined "switching". This assures that if a section of text in 8-bit mode (e.g., English) between two sections that require 16-bit mode (e.g., Chinese) is removed, the now adjoining Chinese texts both continue to be interpreted as such. A single switch character could result in the 2nd Chinese part being suddenly viewed as a series of 8-bit units, resulting in "typical computerese gibberish".

[10] VINTARS - Vienna Internet Terminology And Reference System, a multi-lingual multi script web-based system.

Similar problems have plagued UNOV in terms of printing from the custom symbol set on the mainframe.

It was therefore a paramount requirement of a new system that access, input, and output should be as universal and easy as possible.

Any authorized user should have access from her or his PC.

In fact, the need is now for access by translators and others, not only within the local environment at the United Nations in Vienna, but also by translators at other duty stations as well as contractors working on behalf of the United Nations. The latter may be anywhere in the world.

A proposal for a solution

The need to overcome the obstacles and difficulties noted above, coupled with the general introduction of PCs for translators, the outsourcing of some translation work, as well as the general demise of mainframes resulted in our looking for a replacement system. After determining in 1995-1996 that no commercial system existed that could handle the needed range of language scripts, UNOV chose a customized solution.

A design contract was issued in spring 1997. The recommendation was for a system based structurally on the ISO draft standard 12 620 for Terminology - Computer Applications - Data Categories.

Implementation was proposed by the use of a relational database running on a Windows-NT server platform. Data was to be stored in Unicode. Retrieval was proposed via an Internet web browser.

The need to be able to handle Unicode and the special requirements for data entry in the required scripts posed a bit of a problem, since neither was fully supported by all products on the market.

Implementation of the proposal

The original proposal was for an Adabas-D database. However, Software AG announced late last summer that implementation of Unicode was being postponed. Since a development contract had already been made, the contractor was asked to propose an alternative. The alternative used is therefore Borland's InterBase.

While Netscape and Explorer were indicating presentation support for Unicode by mid-1997, neither had yet implemented input for Unicode or dynamic support for switching to and from Right-to-Left (for Arabic) and Left-to-Right (for other languages). Support for multiple keyboards is likewise limited.

One web browser, however, did support all requirements in early 1997, Alis Technologies Inc.'s, Tango. It was chosen for input or queries covering all languages.

Tango has some special features. One is the ability to choose in a simple manner from a wide range of keyboard layouts (and, if desired to see the keyboard layout on the screen). Tango also allows different input methods: switching to Right-to-Left for Arabic input is supported, as are a number of different input modes for Chinese. Tango also permits choosing from a very broad range of character-sets and code pages, including Unicode UTF-8, but also most widely used code pages for DOS, Windows, and Macintosh platforms for all United Nations languages as well as a broad range of other languages.

Searching, in English or in the language(s) supported by the particular version of MS-Windows running on a PC can also be carried out using Netscape Version 4, as can updates limited to this / these language(s). No tests with MS Internet Explorer 4.xx have been carried out so far.

A standardized approach to Terminology and Official Titles

Term-related information designed into the database includes fields for common names, short form, abbreviations and acronyms, as well as synonyms, initialisms, and variants (spelling, different organizational or geographic usage).

Usage-related information includes context, definition, source, and temporal qualifiers (including whether a term or title is outdated, obsolete, superseded, or subject to a time-restriction).

Term status related information includes normative authorization, normative status, approved access (public, internal, UN only, etc.), and whether it is a proposal or a preferred term or title, and if so, according to whom.

Improved Quality

The project wishes to enlist the active support of all its users. Translators should not only have access from their PCs; they should also be able to post suggestions and proposals for improvements. These could then be checked online by responsible terminology focal points, before becoming, on approval, available to the designated user community. The approval parts of the system are still under construction.

The recent implementation of hyperlinks speeds checking and greatly enhances the user-friendliness of the system. Three types of hyperlinks have been implemented: those linking records to each other (e.g., "formerly:" and "superseded by:"), links to authoritative sites (anywhere on the World Wide Web), as well as links to context and source documents (in WordPerfect, MS-Word, or HTML formats). Hyperlinks are possible in any text field. Further refinements to the feature are planned.

Sources

Each term-related data item in each language can have a corresponding source field.

This may appear as overkill; and in many cases would be. However, given the nature of texts translated by the Translation Service of the United Nations Office at Vienna, official titles especially may have a number of different authoritative sources, depending on language. A title may be from an organization with only three official languages, rather than the United Nations' six[11], or a term may stem from a treaty published in 2 or more original languages. These organizations or treaties are the source for their official languages. Equivalents in the other official languages would be from other sources such as the chiefs of the respective translation sections, they being authorized to define such translations. Acronyms may also have sources distinct from full titles, and different acronyms may be used by different organizations for one and the same term or title.

Such references can be substantiated and made verifiable in VINTARS by providing contextual information, or references to official correspondence or official organizational web sites.

Copy - Paste

It is possible to copy any part of a term/title entry using the simple Windows copy functionality. The results can be copied into any Windows application supporting the respective Windows character-sets. In simple terms this means that you can copy a block containing all 7 languages into, e.g., MS-Word 97. If you choose a Unicode font such as Bitstream's Cyberbit font, all characters in all languages will be correct. If your copy of Word supports directional flow (for Arabic), the Arabic text will flow correctly. Otherwise, while each individual Arabic character is correct, the order will be reversed.

For users who wish to have the results converted into other character-sets the VINTARS development contractor has designed a companion program (Convertlt) to convert the contents of the clipboard to the standard Windows code pages for the relevant languages, 1250 (Latin-2), 1256 (Arabic), and 1251 (Cyrillic). This program, including all required parts, as well as a third party routine for conversion of Chinese from/to Unicode can be downloaded and installed directly from the application.

As Convertlt is bi-directional, it can also be used to convert text from the above-mentioned Windows code pages into Unicode.

For users of Word Perfect for Windows Cyrillic, UNOV has developed a macro that converts the Word Perfect Cyrillic encoding to true Windows Cyrillic for copy/paste into VINTARS. This permits re-use and inclusion of data from pre-existing glossaries or documents in Cyrillic.

[11] Arabic, Chinese, English, French, Russian, and Spanish

Implementation Plan

System development work started in August 1997. A first six-language prototype, in all languages, using a single field, was ready by mid-September 1997.

Difficulties, especially relating to the underlying web-enabling software, delayed the project into early 1998. Testing with a limited set of data from the old mainframe system began in March 1998. Data conversion problems (practically each character within the PSS code page of the mainframe had to be individually identified and converted), stability problems and programming shortcomings were identified. It was also determined that UNOV required a more powerful server, in fact more powerful than was available when the project was started. Bidding was initiated; acquisition is underway. Provisions are also being undertaken to mirror the application outside the local firewall[12] for access by authorized users from the Internet[13].

Following a series of corrective measures testing resumed in mid-July and a reasonably stable version was achieved by early September. Productive use started shortly thereafter. Another update is expected before the end of October 1998.

While the problems indicated did set the project back, they also had their positive side. All in-house users' computers have been upgraded to Windows-95 and have developed reasonable familiarity with many related concepts and software (e.g., WinWord 97).

The testing period has also permitted us to better evaluate our existing data. Substantive assessment, checking, and corrective data gathering is being undertaken.

The immediate future

The number of users will be increased from the current 15 to 30 and following the server upgrade to cover all translators, editors, interpreters, and reference and terminology staff in house (some 70).

The reference staff and terminological focal points will work together and with contact persons in substantive areas to review our complete stock of terminology, official titles, and references.

Just a ways beyond

The integration of this database should proceed along a number of routes. First and foremost is the integration of users working away from the office and of contractual translators working under contract for the United Nations Office at Vienna. Then comes the integration with other computer systems in support of translation. This

[12] Due to the possibilities of HTTP traffic through the firewall leading to security problems, no such traffic is allowed.

[13] These will include contractual translators, staff working away from the office, e.g., on mission, staff at meetings held away from the Vienna International Centre, as well as organizations with whom we may set up terminology exchange programmes which include access to our system.

includes translators' workbenches, voice input (and possibly also output), and further automation of macros and hyperlinks to provide the right information in correct doses. Then, the integration of projects in other areas towards automated identification and pre-selection of candidates for a terminology and reference database. A fourth goal would be possibilities for advanced hyperlinking, i.e., the ability to dip into repositories to which access is possible only via complex login and search procedures. Foremost among these would be the full-text repository being built up within the United Nations itself, on Optical Disk, called the United Nations Optical Disk System or ODS for short. Other possible candidates would be databases such as the World Health Organization's WHOTerm, the European Union's EURODICAUTOM or its successor, or such national terminology databases as Canada's TERMIUM.

Wrap Up

As the title of this paper and my presentation at ASLIB-20 states, what I am describing is work in progress.

The presentation will highlight the key points of this paper and demonstrate the advantages and possibilities inherent in the system's features. It will show the status of VINTARS at its latest stage, including recently added functions such as hyperlinks. I will show concurrent and equal handling of Arabic, Chinese, English, French, Russian, and Spanish, as well as German, accessible using Internet web browsing tools, for the maintenance of up-to-date terminology and bibliographic data, from authorized desktops anywhere in the world.