La Traducción en el mundo del Software Libre

Análisis del estado de las herramientas lingüísticas, proyectos actuales y necesidades de la comunidad del software libre.

Juan Rafael Fernández García
Coordinador responsable de herramientas lingüísticas
TLDP-ES

La Traducción en el mundo del Software Libre: Análisis del estado de las herramientas lingüísticas, proyectos actuales y necesidades de la comunidad del software libre.

por Juan Rafael Fernández García

0.9 Edición
Publicado $Id: ponencia9.xml,v 1.1 2003/11/09 18:32:37 norax Exp $

Análisis de las herramientas lingüísticas de que dispone la comunidad de software libre, de los proyectos de traducción y de los desafíos que presenta el tratamiento informático de los lenguajes naturales.

- Última versión: http://es.tldp.org/Articulos/0000otras/doc-traduccion-libre/ (http://es.tldp.org/Articulos/0000otras/doc-traduccion-libre/)
- Fuente: http://cvs.hispalinux.es/cgi-bin/cvsweb/doc-traduccion-libre/ (http://cvs.hispalinux.es/cgi-bin/cvsweb/doc-traduccion-libre/)

Este documento está íntimamente ligado al de especificaciones de requisitos de las herramientas lingüísticas de TLDP: http://es.tldp.org/especificaciones/herramientas-linguisticas/herramientas-linguisticas/ (http://es.tldp.org/especificaciones/herramientas-linguisticas/herramientas-linguisticas/)

Historial de revisiones

Revisión 0.9 2003.09.28 Revisado por: jr
Actualizado tras VI Congreso de Hispalinux
Revisión 0.8 2003.02.26 Revisado por: jr
Convertido a DocBook XML. Comienzo a integrar información de un-paso-adelante.xml
Revisión 0.7 2002.07.22 Revisado por: jr
Capítulo ‘Lo que un informático quizás no sepa’
Revisión 0.6 2002.07.04 Revisado por: jr
Corregido spanish-team.sgml
Revisión 0.5 2002.06.23 Revisado por: jid
Correcciones menores de marcado
Revisión 0.4 2002.06.20 Revisado por: jr
Convertido a DocBook SGML
Revisión 0.3 2002.06.08 Revisado por: jr
Puesto en el CVS de LuCAS
Revisión 0.2 2002.05.24 Revisado por: jr
Corregidos errores de concepto señalados por Santiago Vila y Jaime Villate
Revisión 0.1 2002.05.09 Revisado por: jr
Creación en LaTeX como trabajo final para un curso

Tabla de contenidos

Lista de tablas

Introducción

Este documento tiene su origen en el trabajo final de un curso de la UNED sobre traducción que debía presentarse por escrito. Gran parte de su estructura y algunas de sus limitaciones tienen ese origen: inicialmente estaba dirigido a traductores profesionales o estudiantes de traducción e iba a presentarse exclusivamente en papel. No obstante, está inspirado por el espíritu contributivo de la comunidad del software libre y mi compromiso es seguir desarrollándolo como una pequeña aportación al conocimiento de este mundo y como introducción al tema para aquellas personas que quieran colaborar y no sepan por dónde empezar. Modestamente, no pretende ser más una instantánea del estado de las herramientas y proyectos en un determinado momento.

Todo el software y los recursos utilizados para la realización de este documento son libres.

1. ¿Por qué Debian GNU/Linux?

Una precisión: «Linux» es el nombre del núcleo del sistema operativo (el *kernel*); el sistema operativo que uso se llama «GNU Linux», porque gran parte de los programas que utiliza (y de las herramientas utilizadas para programarlo y compilarlo) proceden del proyecto «GNU».

El software libre no se limita a GNU Linux (existen FreeBSD u OpenBSD, existen GNU Hurd, Cygnus…). Sin embargo sí es uno de sus desarrollos principales. Y el entorno en el que trabajo y sobre el que puedo escribir.

El Proyecto Debian fue fundado oficialmente por Ian Murdock [1] el 16 de agosto [2] de 1993. La creación de Debian fue patrocinada por la FSF durante un año (entre noviembre 1994 y noviembre 1995) y actualmente tiene el respaldo de *Software in the Public Interest, Inc.*, una organización sin ánimo de lucro con base en Nueva York.

De entre las diferentes distribuciones de GNU Linux, Debian es en mi opinión la más coherente. No está creada según el modelo de una empresa, sino como un conjunto de voluntarios que se obligan mediante un Manifiesto y un «Contrato Social»:

1. Debian seguirá siendo 100% Software Libre.
2. Devolveremos nuestras contribuciones a la Comunidad de Software Libre.
3. No ocultaremos los problemas.
4. Nuestras prioridades son nuestros usuarios y el Software Libre.

En el espíritu de ese «contrato» escribo este documento.

Notas

1. El nombre procede de la contracción de su nombre y del de su mujer Debra.
2. ¡Mi cumpleaños!

I. Conceptos previos

Cuando me planteé escribir este documento pensaba en qué cosas debía saber un traductor que se asomara al mundo del software libre. Progresivamente he descubierto que es igualmente interesante plantearse, desde el punto de vista del usuario comprometido o del desarrollador, qué puede aportar la traducción «profesional». Por un lado está escrito para quien no sabe lo que es GNU o la FSF; por otro lado intenta explicar lo que es una Memoria de Traducción o la utilidad de un *corpus*. No soy un experto en ninguno de los dos campos, espero encontrar ayudas mientras este documento crece, y aprender a medida que se desarrolla.

Capítulo 1. Lo que un traductor quizás no sepa

Hung Chao-Kuei ha creado un diagrama que explica las diferentes categorías de software[1]: La figura 1.

Clases de software

Clases de software

Vemos que los términos extremos, *Software Libre* y *Software «Privativo»*[2], están bien delimitados pero que hay una maraña de variantes intermedias. Entre ellas está el concepto del *Software Abierto* o *de Código Abierto*.

1.1. El concepto de software libre

A principios de los ’80 Richard Stallman [3] decidió negarse a usar software propietario en su trabajo en el Laboratorio de Inteligencia Artificial del MIT y emprendió el desarrollo de un sistema completo de software libre [4] llamado «*GNU*». [5]

> El término «free software» [N. del T.: en inglés *free* = libre o gratis] se malinterpreta a veces — no tiene nada que ver con el precio. El tema es la libertad. Aquí, por lo tanto, está la definición de software libre: un programa es software libre, para usted, un usuario en particular, si:
>
> - Usted tiene libertad para ejecutar el programa, con cualquier propósito.
> - Usted tiene la libertad para modificar el programa, para adaptarlo a sus necesidades (para que esta libertad sea efectiva en la práctica, usted debe tener acceso al código fuente, porque modificar un programa sin disponer del código fuente es extraordinariamente dificultoso).
> - Usted tiene la libertad para redistribuir copias, tanto gratis como por un canon.
> - Usted tiene la libertad para distribuir versiones modificadas del programa, de tal manera que la comunidad pueda beneficiarse con sus mejoras.
>
> Como «*free*» [libre] se refiere a libertad y no a precio, no existe contradicción entre la venta de copias y el software libre. De hecho, la libertad para vender copias es crucial: las colecciones de software libre que se venden en CD-ROM son importantes para la comunidad, y la venta de las mismas es una manera importante de obtener fondos para el desarrollo de software libre. Por lo tanto, si la gente no puede incluir un programa en dichas colecciones, dicho programa no es software libre.

En http://www.gnu.org/philosophy/free-sw.es.html (http://www.gnu.org/philosophy/free-sw.es.html) precisa

> «Software Libre» se refiere a la libertad de los usuarios de ejecutar, copiar, distribuir, estudiar, cambiar y mejorar el software. Más precisamente, se refiere a las cuatro libertades de los usuarios de software:
>
> - La libertad de correr el programa, con cualquier propósito (libertad 0).
> - La libertad de estudiar cómo funciona el programa, y adaptarlo a sus necesidades (libertad 1). El acceso al código fuente es una precondición para esto.
> - La libertad de distribuir copias de manera que se puede ayudar al vecino (libertad 2).
> - La libertad de mejorar el programa, y liberar las mejoras al público de tal manera que toda la comunidad se beneficie (libertad 3). El acceso al código fuente es una precondición para esto.

> Un programa es software libre si los usuarios tienen todas estas libertades.

¿Cuáles son las razones de Stallman? Él habla de razones éticas: de compartir conocimientos en la comunidad de software.

Las consecuencias son el movimiento por el Software Libre.

En el otro extremo está el concepto de *propiedad intelectual*.

Sobre este concepto, ver http://oasis-open.org/who/intellectualproperty.shtml (http://oasis-open.org/who/intellectualproperty.shtml)

* *[ToDo. Desarrollarlo brevemente]*

1.2. Localización e internacionalización del software

Se entiende[6] por «internacionalización» (abreviado «*i18n*») la operación por medio de la cual se modifica un programa, o conjunto de programas en un paquete, para que pueda adecuarse a múltiples idiomas y convenciones culturales. [7]

Por «localización» («*l10n*»), nos referimos a la operación por la que, sobre un conjunto de programas que ya han sido internacionalizados, se le proporciona al programa toda la información necesaria para que pueda manejar su entrada y su salida de un modo que sea correcto respecto a determinados hábitos lingüísticos y culturales (por ejemplo el signo de la moneda de un país, el orden en que se expresan mes, día y año en una fecha…).

Utilizaremos la expresión «adaptación a la pluralidad lingüística» (traducción improvisada de *Native Language Support*, NLS), para hablar de las actividades o rasgos genéricos que engloban tanto la internacionalización como la localización, de manera que sean posibles interacciones plurilingüísticas en un programa.

1.2.1. Juegan los locales

> *LOCALE* is a basic concept introduced into ISO C (ISO/IEC 9899:1990). The standard is expanded in 1995 (ISO 9899:1990 Amendment 1:1995). In LOCALE model, the behaviour of some C functions are dependent on LOCALE environment. LOCALE environment is divided into a few categories and each of these categories can be set independently using setlocale().
>
> POSIX also determines some standards around i18n. Almost all of POSIX and ISO C standards are included in XPG4 (X/Open Portability Guide) standard and all of them are included in XPG5 standard. Note that XPG5 is included in UNIX specifications version 2. Thus support of XPG5 is mandatory to obtain Unix brand. In other words, all versions of Unix operating systems support XPG5. [8]

Un ejemplo vale más que mil explicaciones. date es un programa internacionalizado, que devuelve la fecha y la hora del sistema. Para un usuario que no ha configurado sus *locales* o que ha elegido el *locale* POSIX la salida sería igual a la de un programa no internacionalizado, con mensajes probablemente en inglés:

```
[Mi_maquina]$
LC_ALL=C date
Wed May  8 20:46:09 CEST 2002
[Mi_maquina]$
```

Un usuario español habrá configurado su cuenta de manera que reciba los mensajes en español (para el ejemplo usamos variables de entorno para cambiar de *locale*).

```
[Mi_maquina]$
LC_ALL=es_ES@euro date
mié may  8 20:46:22 CEST 2002
[Mi_maquina]$
```

Podemos ver que la salida es en español. Para ver la salida en francés basta con usar

```
[Mi_maquina]$
LC_ALL=fr_FR date
mer mai  8 20:46:31 CEST 2002
[Mi_maquina]$
```

¿Cómo se logra esto?

* *[ToDo. Explicar brevemente funcionamiento del sistema]*

Un ejemplo de Santiago Vila [9] servirá, porque además no se refiere a un programa en c: sino a un *script* de *shell*

```
#!/bin/sh
if [ -x /usr/bin/gettext ]; then
    echo=/usr/bin/gettext
else
    echo="echo -n"
fi
export TEXTDOMAIN=libc
$echo "cheese"
echo ""
```

```
[Mi_maquina]$
LANG=es_ES; ./test-script
queso
```

1.3. La necesidad de los estándares

Es curioso leer cómo *Sun* defiende el uso del software libre y de los estándares abiertos en las páginas de OpenOffice (http://xml.openoffice.org/xml_advocacy.html), donde presenta su nuevo formato basado en XML

> open and free licensing guarantees that you are not at the mercy of a single company for improvements and fixes of the format or its supporting software, thus providing very strong protection for all investments and efforts you put into this format.

Continúa haciendo una relación de las ventajas de su nuevo formato

1. Separation of Content, Layout, and Meta Information
2. Standards Based
3. Uniform Representation of Formatting and Layout Information
4. Structured Format
5. Idealized Format
6. Common Format Across All Applications
7. Open For Extensions and Supplemental Information
8. Increased Robustness
9. Document Archiving
10. Version Interoperability
11. Documented and Transparent File Content

Realmente la lectura de estas páginas es muy interesante y no tiene sentido replicarlas aquí.

1.3.1. Extensible Markup Language (XML) 1.0

En el principio fue SGML[10]. Sin SGML la *web* no existiría. XML es el hijo maduro de SGML.

Leemos en http://www.w3.org/TR/REC-xml (http://www.w3.org/TR/REC-xml)

> The «*Extensible Markup Language (XML)*» is a subset of SGML. Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML.
>
> XML was developed by an *XML Working Group* (originally known as the SGML Editorial Review Board) formed under the auspices of the World Wide Web Consortium (W3C) in 1996. It was chaired by Jon Bosak of Sun Microsystems with the active participation of an XML Special Interest Group (previously known as the SGML Working Group) also organized by the W3C.

Otros miembros célebres del SIG eran James Clark y Norman Walsh.

Notas

1. En http://www.gnu.org/philosophy/category.fig (http://www.gnu.org/philosophy/category.fig).
2. Se suele utilizar el término «propietario» para referirnos al software no libre, cuando ese software precisamente limita los derechos del usuario sobre el código.
3. La historia la cuenta él mismo en http://www.gnu.org/gnu/thegnuproject.es.html (http://www.gnu.org/gnu/thegnuproject.es.html) (publicado originalmente en el libro *Open Sources*). Traducción de César Ballardini (Argentina) <cballard@santafe.com.ar>, revisada por Ramsés Morales (Panamá) <ramses@computer.org>, César Villanueva (Venezuela) <dandel@etheron.net> y Oscar Mendez Bonilla (México) <omendez@acnet.net>; coordinación: Hugo Gayosso <hgayosso@gnu.org>.
4. Las ideas de Stallman están recogidas en http://www.gnu.org/philosophy/philosophy.es.html (http://www.gnu.org/philosophy/philosophy.es.html).
5. Ver http://www.gnu.org/gnu/gnu-history.es.html (http://www.gnu.org/gnu/gnu-history.es.html). Traducción coordinada por Hugo Gayosso <hgayosso@gnu.org> y actualizada el 9 de nov. 1999 por Conrado Alfonso Bermúdez. El significado del acróstico es una típica broma *hacker*: “GNU is not Unix”. Como resume *The Free On-line Dictionary of Computing*

   > The GNU Manifesto was published in the March 1985 issue of Dr. Dobb’s Journal but the GNU project started a year and a half earlier when Richard Stallman was trying to get funding to work on his freely distributable editor, Emacs.

6. Bibliografía: **(p)info** gettext, **man** Locale::Maketext(3pm), **man** Locale::Maketext:TPJ13(3pm), http://www.debian.org/doc/manuals/intro-i18n/ (http://www.debian.org/doc/manuals/intro-i18n/).
7. Sobre el tema se puede consultar «li18nux» (http://www.li18nux.org/ (http://www.li18nux.org/)).
8. Tomohiro Kubota, http://www.debian.org/doc/manuals/intro-i18n/ch-locale.html (http://www.debian.org/doc/manuals/intro-i18n/ch-locale.html).
9. En mensaje a la lista <debian-i18n> de 16 de mayo de 2002.
10. “ ISO 8879: Information processing - Text and office systems - Standard Generalized Markup Language (SGML) ”. Ginebra, 1986.

Capítulo 2. Lo que un informático quizás no sepa

2.1. Definiciones

Unas cuantas definiciones para entendernos.

Tom McArthur define «*Corpus*» (latinajo de uso habitual en la jerga, plural «*corpora*»)[1] como

1. A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse.
2. In linguistics and lexicography, a body of texts, utterances or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analysed by means of «*tagging*» (the addition of identifying and classifying tags[2] to words and other formations) and the use of «*concordancing programs*».

   «*Corpus linguistics*» studies data in any such corpus.

El marcado del corpus reponde a la necesidad de lo que se llama «*text annotation*»: adding linguistic information

1. Part-of-speech (POS) tagging
2. Syntactic annotation (parsed corpora)
3. Pragmatic annotation
4. Rhetorical information
5. Discourse structure

Concordancias:

índice (normalmente alfabético) de las palabras de un texto, en el cual la palabra analizada figura en el centro de una línea rodeada a derecha e izquierda de otras con las que aparece en un *contexto* determinado.

Continúa el “Tutorial: Concordances and Corpora ”[3]

> The most common form of concordance today is the «*Keyword-in-Context (KWIC) index*», in which each word is centered in a fixed-length field (e.g., 80 characters).

«*Concordance programs (concordancers)*»[4]:

> Concordance programs are basic tools for the corpus linguist. Since most corpora are incredibly large, it is a fruitless enterprise to search a corpus without the help of a computer. Concordance programs turn the electronic texts into databases which can be searched. Usually (1) word queries are always possible, but most programs also offer (2) the possibility of searching for word combinations within a specified range of words and (3) of looking up parts of words (substrings, in particular affixes, for example). If the program is a bit more sophisticated, it might also provide its user with (4) lists of collocates (colocaciones) or (5) frequency lists.

Interesante, el siguiente texto de Melamed (http://www.cs.nyu.edu/cs/projects/proteus/bma/ (http://www.cs.nyu.edu/cs/projects/proteus/bma/)):

> A «*bitext*» consists of two texts that are mutual translations. A *bitext map* is a fine-grained description of the correspondence relation between elements of the two halves of a bitext. Finding such a map is the first step to building translation models. It is also the first step in applications like automatic detection of omissions in translations.
>
> Alignments (‘alineaciones’, ‘alineamientos’, ‘emparejamientos’ o ‘correspondencias’ se lee en la literatura técnica) are “watered-down” bitext maps that we can derive from general bitext maps.

El Informe Final del proyecto POINTER se esfuerza —y creo que lo consigue— por aclarar los términos ‘lexicología’, ‘lexicografía’, ‘terminología’ y ‘terminografía’ (http://www.computing.surrey.ac.uk/ai/pointer/report/section1.html#2 (http://www.computing.surrey.ac.uk/ai/pointer/report/section1.html#2)). La cita es larga pero creo que no tiene desperdicio.

> While *lexicology* is the study of words in general, *terminology* is the study of special-language words or terms associated with particular areas of specialist knowledge[5]. Neither lexicology nor terminology is directly concerned with any particular application. *Lexicography*, however, is the process of making dictionaries, most commonly of general-language words, but occasionally of special-language words (i.e. terms). Most general-purpose dictionaries also contain a number of specialist terms, often embedded within entries together with general-language words. *Terminography* (or often misleadingly "terminology"), on the other hand, is concerned exclusively with compiling collections of the vocabulary of special languages. The outputs of this work may be known by a number of different names —often used inconsistently— including "terminology", "specialised vocabulary", "glossary", and so on.
>
> Dictionaries are word-based: lexicographical work starts by identifying the different senses of a particular word form. The overall presentation to the user is generally alphabetical, reflecting the word-based working method. Synonyms —different form same meaning— are therefore usually scattered throughout the dictionary, whereas polysemes (related but different senses) and homonyms (same form, different meaning) are grouped together.
>
> While a few notable attempts have been made to produce conceptually-based general-language dictionaries — or "thesauri", the results of such attempts are bound to vary considerably according to the cultural and chronological context of the author.
>
> By contrast, high-quality terminologies are always in some sense concept-based, reflecting the fact that the terms which they contain map out an area of specialist knowledge in which encyclopaedic information plays a central role. Such areas of knowledge tend to be highly constrained (e.g. "viticulture"; "viniculture"; "gastronomy"; and so on, rather than "food and drink"), and therefore more amenable to a conceptual organisation than is the case with the totality of knowledge covered by general language. The relations between the concepts which the terms represent are the main organising principle of terminographical work, and are usually reflected in the chosen manner of presentation to the user of the terminology. Conceptually-based work is usually presented in the paper medium in a thesaurus-type structure, often mapped out by a system of classification (e.g. UDC) accompanied by an alphabetical index to allow access through the word form as well as the concept. In terminologies, synonyms therefore appear together as representations of the same meaning (i.e. concept), whereas polysemes and homonyms are presented separately in different entries.
>
> Dictionaries of the general language are descriptive in their orientation, arising from the lexicographer’s observation of usage. Terminologies may also be descriptive in certain cases (depending on subject field and/or application), but prescription (also: "normalisation" or "standardisation") plays an essential role, particularly in scientific, technical and medical work where safety is a primary consideration. Standardisation is normally understood as the elimination of synonymy and the reduction of polysemy/homonymy, or the coinage of neologisms to reflect the meaning of the term and its relations to other terms.

> «*Terminology management*», itself a neologism, was coined to emphasise the need for a methodology to collect, validate, organise, store, update, exchange and retrieve individual terms or sets of terms for a given discipline. This methodology is put into operation through the use of computer-based information management systems called «*Terminology Management Systems*» (TMS).

Dice Martínez de Sousa, sub voce *terminología*, en el *Diccionario de lexicografía práctica*

> Hoy la *terminología* es una ciencia bien estructurada que se ocupa en crear los catálogos léxicos propios de las ciencias, las técnicas, los oficios, etc., partiendo de sistemas coherentes establecidos por organismos nacionales e internacionales.

El proyecto SALT distingue entre «*lexbases*» y «*termbases*», pensadas para ser usadas en traducción automática las primeras y como recursos de ayuda a la traducción las segundas; EAGLES habla de «*termbanks*».

EAGLES-I proporciona la siguiente definición de «*Memoria de Traducción*»[6]:

a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions.

2.2. Lo que un traductor puede esperar …

2.2.1. … de un diccionario

* *[ToDo]*

2.2.2. … de un *corpus textual*

Voy a poner un ejemplo para explicar lo que creo que necesitamos. Supongamos que estás escribiendo documentación o traduciendo un texto y de pronto dudas sobre cuándo en español correcto se dice ‘delante’ o ‘adelante’. Abres la herramienta de consulta del corpus y escribes ‘delante’ y te sale una relación de ejemplos de usos del término (imaginemos el mismo caso en inglés, donde necesitamos mucha más ayuda: ¿se dice *‘angry at’* o *‘angry with’*?, ¿es *‘interesting for us’*?). Claro, el corpus tiene que generarse sobre documentos correctos *(authoritative)* y ser lo más amplio y completo posible. Un corpus supera a un diccionario estándar porque permite pasar del nivel de la lengua al de la norma; permite averiguar cuál es el uso más frecuente, cómo se dice habitualmente algo.

2.2.3. … de una memoria de traducción

* *[ToDo. Explicar lo que son y para qué sirven.]*

Notas

1. McArthur, Tom «Corpus», en: McArthur, Tom (ed.) 1992. *The Oxford Companion to the English Language*. Oxford, 265-266.
2. Se habla de *etiquetas*, *marquillas* o *anotaciones*.
3. De Catherine Ball, de la Universidad de Georgetown, http://www.georgetown.edu/cball/corpora/tutorial3.html (http://www.georgetown.edu/cball/corpora/tutorial3.html)
4. Cito http://www.uni-koeln.de/phil-fak/englisch/bald/programs.htm (http://www.uni-koeln.de/phil-fak/englisch/bald/programs.htm)
5. Abaitua habla de «lenguajes de especialidad».
6. La confusión terminológica sobre el concepto es evidente: si habla de ‘translation databases’ y de ‘catálogos’ (kbabel), ‘compendia’ (gettext), ‘learn buffers’ (gtranslator).

Capítulo 3. Cosas que saben los lingüistas computacionales

3.1. Un paseo histórico por la red

A los dos lados del charco las instituciones públicas y las universidades han desarrollado una serie de proyectos de altísimo interés para nosotros.

3.1.1. TEI

[Aquí hablar de la creación de TEI, que ya celebró sus 10 años]

3.1.2. Cíbola y Oleada

En el lado americano, *Oleada* es un desarrollo derivado de TIPSTER II[1] creado por BillOgden[2]

> «Cíbola» and «Oleada» are two related systems that provide multilingual text processing technology to language instructors, learners, translators, and analysts. The systems consist of a set of component tools that have been designed with a user-centered methodology.

Oleada proporciona

- *XAlign* y Translation Memory (http://crl.nmsu.edu/Research/Projects/oleada/tm.html (http://crl.nmsu.edu/Research/Projects/oleada/tm.html))
- *XConcord* (http://crl.nmsu.edu/Research/Projects/oleada/xcon.html (http://crl.nmsu.edu/Research/Projects/oleada/xcon.html))
- Glossary, Dictionary (http://crl.nmsu.edu/Research/Projects/oleada/gloss.html (http://crl.nmsu.edu/Research/Projects/oleada/gloss.html))

3.1.3. EAGLES I y II (1995-1999)

En Europa destacan EAGLES I y II (Expert Advisory Group on Language Engineering Standards)

El primer proyecto terminó en 1996. El segundo proyecto se extendió entre 1997 y primavera de 1999. Según la introducción (http://www.ilc.pi.cnr.it/EAGLES96/intro.html (http://www.ilc.pi.cnr.it/EAGLES96/intro.html))

> EAGLES is an initiative of the European Commission (. . .) which aims to accelerate the provision of standards for:
>
> - Very large-scale language resources (such as text corpora, computational lexicons and speech corpora);
> - Means of manipulating such knowledge, via computational linguistic formalisms, mark up languages and various software tools;

- Means of assessing and evaluating resources, tools and products.

The work towards common specifications is carried out by five working groups:

- Text Corpora
- Computational Lexicons
- Grammar Formalisms
- Evaluation
- Spoken Language

Un resultado de los trabajos fue el *Corpus Encoding Standard* (CES, http://www.cs.vassar.edu/CES/ (http://www.cs.vassar.edu/CES/)) y XCES (http://www.cs.vassar.edu/XCES/ (http://www.cs.vassar.edu/XCES/)), la versión XML.

The CES is designed to be optimally suited for use in language engineering research and applications, in order to serve as a widely accepted set of encoding standards for corpus-based work in natural language processing applications. The CES is an application of SGML compliant with the specifications of the TEI Guidelines.

The CES specifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding specifications for linguistic annotation, together with a data architecture for linguistic corpora.

In its present form, the CES provides the following:

- a set of metalanguage level recommendations (particular profile of SGML use, character sets, etc.);
- tagsets and recommendations for documentation of encoded data;
- tagsets and recommendations for encoding primary data, including written texts across all genres, for the purposes of corpus-based work in language engineering.
- tagsets and recommendations for encoding linguistic annotation commonly associated with texts in language engineering, currently including:
  - segmentation of the text into sentences and words (tokens),
  - morpho-syntactic tagging,
  - parallel text alignment.

Sin embargo lo más influyente de los resultados de los proyectos son las Directrices EAGLES. Los trabajos del grupo los prosigue el proyecto ISLE.

Relacionados con EAGLES y CES estaban los proyectos PAROLE (Preparatory Action for Linguistic Resources Organisation for Language Engineering, LE2-4017, http://www.dcs.shef.ac.uk/research/groups/nlp/funded/parole.html (http://www.dcs.shef.ac.uk/research/groups/nlp/funded/parole.html))

y

3.1.4. MULTEXT (1994-1996)

MULTEXT (Multilingual Text Tools and Corpora, LRE 62-050, 1994-96, http://www.lpl.univ-aix.fr/projects/multext/ (http://www.lpl.univ-aix.fr/projects/multext/)). Estos eran sus objetivos iniciales:

> Existing tools for NLP and MT corpus-based research are typically embedded in large, non-adaptable systems which are fundamentally incompatible. Little effort has been made to develop software standards, and software reusability is virtually non-existent. As a result, there is a serious lack of generally usable tools to manipulate and analyze text corpora that are widely available for research, especially for multi-lingual applications. At the same time, the availability of data is hampered by a lack of well-established standards for encoding corpora. Although the TEI has provided guidelines for text encoding, they are so far largely untested on real-scale data, especially multi-lingual data. Further, the TEI guidelines offer a broad range of text encoding solutions serving a variety of disciplines and applications, and are not intended to provide specific guidance for the purposes of NLP and MT corpus-based research. MULTEXT proposes to tackle both of these problems. First, MULTEXT will work toward establishing a software standard, which we see as an essential step toward reusability, and publish the standard to enable future development by others. Second, MULTEXT will test and extend the TEI standards on real-size data, and ultimately develop TEI-based encoding conventions specifically suited to multi-lingual corpora and the needs of NLP and MT corpus-based research.

Herramientas elaboradas por el proyecto MULTEXT son

- *mmorph* (Morphology tool, ftp://issco-ftp.unige.ch/pub/multext/mmorph-2.3.4_2.tar.gz (ftp://issco-ftp.unige.ch/pub/multext/mmorph-2.3.4_2.tar.gz))
- *mtag* (The multext version of the tagger, ftp://issco-ftp.unige.ch/pub/multext/tagger2.22.tar.gz (ftp://issco-ftp.unige.ch/pub/multext/tagger2.22.tar.gz))
- *tatoo* (The ISSCO TAgger TOOl, http://issco-www.unige.ch/staff/robert/tatoo/tatoo.html (http://issco-www.unige.ch/staff/robert/tatoo/tatoo.html))
- *multext_align* (Alignment program, ftp://issco-ftp.unige.ch/pub/multext/multext_align_v2.0.tar.gz (ftp://issco-ftp.unige.ch/pub/multext/multext_align_v2.0.tar.gz))

3.1.5. Corpus textual especializado plurilingüe

Un desarrollo catalán: (http://www.iula.upf.es/corpus/corpuses.htm (http://www.iula.upf.es/corpus/corpuses.htm))

El proyecto Corpus es el proyecto de investigación prioritario del IULA. Recopila textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de especialidad de la economía, el derecho, el medio ambiente, la medicina y la informática. A través del establecimiento del corpus, se intentan inferir las leyes que rigen el comportamiento de cada lengua en cada área.

Las investigaciones previstas sobre el corpus son las siguientes: detección de neologismos y términos, estudios sobre variación lingüística, análisis sintáctico parcial, alineación de textos, extracción de datos para la enseñanza de segundas lenguas, extracción de datos para la construcción de diccionarios electrónicos, elaboración de tesaurus, etc.

Los textos son marcados de acuerdo con el estándar SGML y siguiendo las directrices CES de la iniciativa EAGLES.

El procesamiento de los textos del corpus sigue los siguientes pasos:

- marcaje estructural
- preproceso (detección de fechas, números, locuciones, nombres propios…)
- análisis y marcaje morfológicos de acuerdo con los etiquetarios morfosintácticos diseñados en el IULA
- desambiguación lingüística y/o estadística
- almacenamiento en una base de datos textual

Problema

Creo que gran parte de sus desarrollos no son libres.

3.1.6. MATE

Un tanto marginal para nuestros intereses, pero interesante en lo que toca a la anotación, es MATE (Multilevel Annotation, Tools Engineering, Telematics Project LE4-8370) http://mate.nis.sdu.dk/ (http://mate.nis.sdu.dk/)

> MATE aims to develop a preliminary form of standard and a workbench for the annotation of spoken dialogue corpora. The annotation standard will:
>
> - allow multiple annotation levels, where the various annotation levels can be related to each other;
> - allow coexistence of a multitude of coding schemes and standards;
> - allow multilinguality;
> - integrate standardisation efforts in the US, Europe and Japan; and
> - be open with respect to the information levels and categories
>
> within each level.
>
> The MATE results will be of particular relevance for:
>
> - the construction of SLDS (Spoken Language Dialogue Systems) lexicons
> - corpus-based learning procedures for the acquisition of language-models, part-of-speech-tagging, grammar induction, extraction of structures to be used in the dialogue control of SLDSs;
> - lexicon and grammar development based on explicit descriptions of the interrelationships between phenomena at different descriptive levels (e.g. lexical, grammatical, prosodic clues for semantics and discourse segmentation, for inferring dialogue acts, etc).

El código producido por el proyecto, ‘The MATE Workbench’, en java y con licencia GPL, se puede descargar de http://www.cogsci.ed.ac.uk/~dmck/MateCode/ (http://www.cogsci.ed.ac.uk/~dmck/MateCode/)

3.1.7. ISLE (2000-2002)

El sitio web del proyecto se encuentra en http://lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm (http://lingue.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm).

Leemos

> The ISLE project which started on 1 January 2000 continues work carried out under the EAGLES initiative. ISLE (International Standards for Language Engineering) is both the name of a project and the name of an entire set of co-ordinated activities regarding the HLT field. ISLE acts under the aegis of the EAGLES initiative, which has seen a successful development and a broad deployment of a number of recommendations and de facto standards.[3]

> The aim of ISLE is to develop HLT standards within an international framework, in the context of the EU-US International Research Cooperation initiative. Its objectives are to support national projects, HLT RTD projects and the language technology industry in general by developing, disseminating and promoting de facto HLT standards and guidelines for language resources, tools and products.[4]

> ISLE targets the 3 areas: multilingual lexicons, natural interaction and multimodality (NIMM), and evaluation of HLT systems. These areas were chosen not only for their relevance to the current HLT call but also for their long-term significance. For multilingual computational lexicons, ISLE will:[5]
>
> - extend EAGLES work on lexical semantics, necessary to establish inter-language links;
> - design standards for multilingual lexicons;
> - develop a prototype tool to implement lexicon guidelines and standards;
> - create exemplary EAGLES-conformant sample lexicons and tag exemplary corpora for validation purposes;
> - develop standardised evaluation procedures for lexicons.

3.1.8. POINTER (-1996)

En el campo de la terminología fue muy importante el proyecto POINTER, co-financiado por la Comunidad Europea y que emitió su Informe Final (revisión nº 54) en enero de 1996. Dice http://www.computing.surrey.ac.uk/ai/pointer/ (http://www.computing.surrey.ac.uk/ai/pointer/)

> The aim of the POINTER project is to provide a set of concrete feasible proposals which will support users of terminology throughout Europe by facilitating the distribution of terminologies, as well as their re-use in different contexts and for different purposes.

El Informe Final del proyecto POINTER señala las deficiencias del campo de la terminología tal y como se daba en Europa entonces (de intercambio de terminologías, de validación y verificación, del interfaz de usuario y su «localización», de extracción de terminologías a partir de *corpora* lingüísticos, necesidad de mejorar las técnicas de almacenamiento y recuperación de información y de integrar las terminologías en el software), y recomienda direcciones para solucionarlas.

3.1.9. ELRA

The European Language Resources Association (ELRA) was founded in February 1995 and is the recipient of EU funds under the MLIS (MultiLingual Information Society) programme on a shared-cost basis. Established at the instigation of the European Commission with the active participation of the POINTER, PAROLE (corpora/lexica) and SPEECHDAT (speech data) projects in conjunction with the RELATOR project (A European Network of Repositories for Linguistic Resources), ELRA aims to validate and distribute European language resources that are offered to it for that purpose. In addition, it acts as a clearinghouse for information on language engineering, gathering data on market needs and providing high-quality advice to potential and actual funders, including the European Commission and national governments. Equally, it supports the development and application of standards and quality control measures and methodologies for developing electronic resources in the European languages. In time ELRA aims, in its own words, «to become the focal point for pressure in the creation of high-quality and innovative language resources in Europe».

3.1.10. SALT (2000-2001)

«*SALT*» (Standards-based Access to multilingual Lexicons and Terminologies) fue un proyecto integrado en el V Programa Marco (2000-2001).

Una de sus páginas *web* está en http://www.loria.fr/projets/SALT/saltsite.html (http://www.loria.fr/projets/SALT/saltsite.html). El proyecto nace de la toma de conciencia de una necesidad:

> This project responds to the fact that many organizations in the localization industry are now using both human translation enhanced by productivity tools and MT with or without human post-editing. This duality of translation modes brings with it the need to integrate existing resources in the form of (a) the NLP lexicons used in MT (which we categorize as *lexbases*) and (b) the concept-oriented terminology databases used in human-translation productivity tools (which we call *termbases*). This integration facilitates consistency among various translation activities and lever-ages data from expensive information sources for both lex side and the term side of language processing.
>
> The SALT project combines two recently finalized interchange formats: «OLIF» (Open Lexicon Interchange Format), which focuses on the interchange of data among lexbase resources from various machine translation systems, (Thurmaier et al. 1999), and «MARTIF» (ISO 12200:1999, MAchine-Readable Terminology Interchange Format), which facilitates the interchange of termbase resources with conceptual data models ranging from simple to sophisticated. The goal of SALT is to integrate lexbase and termbase resources into a new kind of database, a lex/term-base called «XLT» (eXchange format for Lex/Term-data).

XLT se basa en XML. El «*Default XLT*» se conoce como «*TBX*»: ‘TermBase eXchange format’.

> Control of TBX has been handed over from the SALT project (…) to LISA (and its OSCAR SIG).

3.1.11. LISA y OSCAR

Pendiente y urgente: TMX, TBX, SRX.

Notas

1. Cf. http://crl.nmsu.edu/~ogden/Papers/oleada.fm.pdf, sobre TIPSTER ver también http://crl.nmsu.edu/twg.annotation/.

2. Software de antes de 1997!?
3. http://lingue.ilc.pi.cnr.it/EAGLES96/isle/project_profile.htm (http://lingue.ilc.pi.cnr.it/EAGLES96/isle/project_profile.htm).
4. http://lingue.ilc.pi.cnr.it/EAGLES96/isle/objectives.htm (http://lingue.ilc.pi.cnr.it/EAGLES96/isle/objectives.htm).
5. http://lingue.ilc.pi.cnr.it/EAGLES96/isle/work_description.html (http://lingue.ilc.pi.cnr.it/EAGLES96/isle/work_description.html).

II. Proyectos de traducción de código fuente

Acerca de lo que hay.

Capítulo 4. El Proyecto de Traducción Libre y gettext

4.1. Un poquito de historia

Todo empezó en julio de 1994, cuando Patrick D'Cruze tuvo la iniciativa de internacionalizar la versión 3.9.2 de GNU fileutils. Le preguntó a Jim Meyering, el responsable del paquete, cómo incorporar esos cambios a una versión oficial. El primer esbozo estaba lleno de **#ifdefs**, y Meyering trató de encontrar un mecanismo más apropiado, lo que dio pie a varias soluciones insatisfactorias, hasta que Ulrich Drepper se implicó en el proyecto. Partiendo de cero escribió lo que primeramente se conoció como "msgutils", después 'nlsutils', y finalmente gettext; fue aceptado oficialmente por Richard Stallman hacia mayo de 1995.

Simultáneamente François Pinard adaptó media docena de paquetes GNU a gettext, proporcionando de camino un entorno de usuario efectivo para probar y afinar las nuevas herramientas. También se hizo cargo de la responsabilidad de organizar y coordinar el Proyecto de Traducción. Tras casi un año de intercambio de mensajes de correo informales entre personas de muchos países, en mayo de 1995 empezaron a existir los primeros equipos de traductores, mediante la creación y mantenimiento por parte de Patrick D'Cruze de veinte listas de correo no moderadas para veinte idiomas nativos.

La idea de François Pinard era crear un sistema de ayuda a los programadores que crearan Software Libre, de manera que les resultara fácil encontrar traductores para sus programas.

En palabras de Santiago Vila [1]:

> La FSF (el proyecto GNU) encargó inicialmente a Pinard la coordinación de la traducción de los .po de los programas de GNU. Hablo de la época en la que se crearon las listas *@li.org*. Posteriormente a esto, Pinard decidió cambiar el nombre al proyecto y llamarle "Free Translation Project" en lugar de "GNU translation project". El objetivo es muy similar pero algo más amplio: coordinar la traducción de cualquier programa libre que haya sido internacionalizado usando gettext. Esto se ofrece como un «servicio» a los autores de los programas libres, que en lugar de tener que buscar traductores por su cuenta pueden limitarse a enviar los ficheros .pot a Pinard y recoger las traducciones en un directorio habilitado al efecto. [2]

4.2. Cómo apuntarse

Para ser miembro del equipo español hay que enviar a la FSF, mediante correo electrónico y tradicional, una renuncia a los derechos de las traducciones. Aquí podéis ver una copia (sin los datos personales) de mi *Disclaimer*, de fecha 2 de mayo de 2000.

DISCLAIMER OF COPYRIGHT IN TRANSLATIONS OF PARTS OF PROGRAMS

I, [Tu_Nombre], a citizen of the Kingdom of Spain,
do hereby acknowledge to the Free Software Foundation,
a not-for-profit corporation of Massachusetts, USA,

that I disclaim all copyright interest in my works,
which I have provided or will in the future provide to
the Foundation, of translation of portions of free software
programs from one human language to another human language.
The programs to which this applies includes all programs for
which the Foundation is the copyright holder, and all other
freely redistributable software programs.

The translations covered by this disclaimer include, without
limitation, translations of textual messages, glossaries,
command or option names, user interface text, and the like,
contained within or made for use via these programs.

I currently expect to work on the Spanish translation team
(though this disclaimer applies to all translations I may
subsequently work on).

Given as a sealed instrument this [Fecha],
at [Tu_Localidad], Spain, and initially sent by e-mail,
with a manually signed copy sent by post or hand delivery to the
Foundation.

Signed

[Tu_Nombre] [Tu_email]
[Tu_Dirección] (Spain)

Poco después recibes un correo, acusando recibo de tu renuncia.

From pinard@IRO.UMontreal.CA Fri May 26 23:47:50 2000
Subject: New translation disclaimers
From: François_Pinard <pinard@IRO.UMontreal.CA>
Date: 26 May 2000 14:42:10 -0400

Hello!

The Free Software Foundation has received more translation
disclaimers. Please check below if the transcription is
accurate. Ensure your name and electronic mail address is
written the same way you intend to do it in the PO files
you produce. Please also check that the team, between square
brackets, is properly identified.

mailto:translation@iro.umontreal.ca for reporting any
error. Also use this address for sending any pending PO file
you might have. If you happen to have a home Web page, please
send me the URL, as I now maintain this information as well
in the Translation Project registry.

> TRANSLATIONS Jan Nieuwenhuizen 2000-05-05
> Disclaimer. [nl]
> janneke@gnu.org

> TRANSLATIONS Nir Bruner 2000-05-04
> Disclaimer. [he]
> xorserer@hotmail.com

> TRANSLATIONS Juan Rafael Ferna'ndez Garci'a 2000-05-02
> Disclaimer. [es]
> jrfern@bigfoot.com

This letter is also sent to your translating team, for
informing the team members of the arrival of your disclaimer,
and also as a cross-check for the electronic mail address,
augmenting the chances this letter reaches you.

Welcome to our Translation Project!

--
François Pinard <http://www.iro.umontreal.ca/~pinard>

Aparecerás en la lista de traductores del equipo español [3], y es el momento de contactar con Santiago Vila para que se te asigne una tarea (en la jerga se dirá que serás responsable de un «*dominio*»):

Translation team for Spanish

The Spanish translation team uses *es* as language code. This code is usable as the value of the *LANGUAGE* or *LANG* environment variable by users of internationalized software. It is also part of PO file names. We often use it as a short identification for the team.

The team uses es@li.org (mailto:es@li.org) for an official email address, which reaches either a mailing list, or someone who broadcasts information to all other team members. Santiago Vila Doncel (mailto:sanvila@unex.es) currently acts as the team leader, and you may write to him or her for all matters related to team coordination. Team members expressed a preference towards using the *ISO-8859-1* charset. You might want to consider using it whenever you send email to the team list or members, or if you produce any translation file meant for this team.

You may get more information about the Spanish effort by visiting some team links, according to the following table.

Topic	**URL**

Tabla 4-1. Enlaces

Topic	**URL**
Team site	ftp://ftp.unex.es/pub/gnu-i18n/spanish-gnu (ftp://ftp.unex.es/pub/gnu-i18n/spanish-gnu)
Status	http://homepage.iprolink.ch/~justine/estado.html (http://homepage.iprolink.ch/~justine/estado.html)
In works	ftp://ftp.unex.es/pub/gnu-i18n/spanish-gnu/revisar (ftp://ftp.unex.es/pub/gnu-i18n/spanish-gnu/revisar)
PO mirror	ftp://ftp.unex.es/pub/gnu-i18n/po (ftp://ftp.unex.es/pub/gnu-i18n/po)

The Translation Project registry knows about the following translators for the Spanish team.

Tabla 4-2. Traductores

Translator	**Web home**	**Disclaimer**	**Autosend**	**Count**
Andrés Felipe Mancipe Galvis (mailto:)		Yes		
...	...	...	...	...
Jordi Mallach Pérez (mailto:jordi@sindominio.net)		Yes		2
...	...	...	...	...
Juan Rafael Fernández García (mailto:jrfern@bigfoot.com)		Yes		1
...	...	...	...	...
Santiago Vila Doncel (mailto:sanvila@unex.es)		Yes		14
...	...	...	...	...
Vicente E. Llorens (mailto:vllorens@mundofree.com)		Yes		3

The Autosend column is for translators who want PO files sent to them on generation, while new POT files are being uploaded to the project. Some translators want both the notice *and* the file in their mailbox, instead of fetching it through the Web. Just ask (mailto:translation@iro.umontreal.ca) if you want this service for yourself.

Here is the current list of assignments of textual domains to translators, as known to the Translation Project registry. If no current version is listed, the information is identical to the most recent submission. The Translation Project robot relies on this information for directly accepting submissions from translators. If there is an error or an omission in this table, please write to Santiago Vila Doncel (mailto:sanvila@unex.es) to get it corrected.

Tabla 4-3. Asignaciones

Domain	Assigned translator	Version	Translated	Current Version	Translated
a2ps (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=a2ps)	Miguel Ángel Varó Gaines (mailto:mvaro@dlsi.ua.es)			4.13b	0 / 179
bash (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=bash)	Cristian Othón Martínez Vera (http://eniac.rhon.itam.mx/~cfuga/)	2.0	840 / 840		
...	...	...	...	...	...
jwhois (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=jwhois)	Cristian Othón Martínez Vera (http://eniac.rhon.itam.mx/~cfuga/)	3.2.0	64 / 64		
kbd (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=kbd)	Juan Rafael Fernández García (mailto:jrfern@bigfoot.com)			1.06	215 / 215
ld (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=ld)	Cristian Othón Martínez Vera (http://eniac.rhon.itam.mx/~cfuga/)	2.12-pre020121	378 / 378		
libc (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=libc)	Santiago Vila Doncel (mailto:sanvila@unex.es)	2.2.5	1172 / 1172		
...	...	...	...	...	...
wget (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=wget)	Salvador Gimeno Zamora (mailto:salgiza@jazzfree.com)	1.8.1	189 / 189		

Last recomputed on 2002-05-08 12:21 -0400

Your comments (mailto:translation@iro.umontreal.ca) are welcome.

4.3. Cómo se trabaja

En mi caso se me asignó el «dominio» kbd. [4]

The kbd textual domain

Here is a short description for the textual domain kbd. The current template for this domain is kbd-1.06.pot.

The maintainer does not require special papers prior to accepting translations.

The following URL information may help translators for this package, for if they need finer translation context, but the distributions might well be experimental, and might not even compile. Be well aware that the URLs given here are not necessarily official.

ftp://ftp.win.tue.nl/pub/linux-local/utils/kbd/kbd-1.06.tar.gz
(ftp://ftp.win.tue.nl/pub/linux-local/utils/kbd/kbd-1.06.tar.gz)

The following table gives some information about PO files which are available for that textual domain.

Tabla 4-4. The kbd textual domain

Code	Language	Version	Last Translator	Translated
fr	French	1.06	Michel Robitaille	215 / 215
es	Spanish	1.06	Juan Rafael Fernández García	215 / 215
sv	Swedish	1.06	Martin Sjögren	215 / 215
tr	Turkish	1.06	Nilgün Belma Bugüner	215 / 215

Last recomputed on 2002-05-08 12:22 -0400.

Your comments are welcome.

Lo primero es obviamente descargar el fichero que hay que traducir, en este caso `kbd-1.06.pot`, y «visitarlo» con emacs. Podemos verlo en la figura 2.

Emacs en modo po

Emacs en modo po

Salva el fichero de trabajo como `kbd-1.06.es.po`. Traduce cada mensaje con la ayuda de emacs, como en la figura 3.

Emacs en modo po, fichero traducido

Cuando hayas terminado

```
[Mi_maquina]$
msgfmt -v -o /dev/null kbd-1.06.es.po
215 translated messages.
[Mi_maquina]$
```

No ha devuelto errores, luego la traducción ha terminado.

Sólo falta enviarla al robot del proyecto. Basta un correo electrónico, en la cabecera

```
To Translation Project Robot <translation@IRO.UMontreal.CA>
Subject TP-Robot kbd-1.06.es.po
```

En el cuerpo del mensaje se hallará el fichero .po. El robot responderá aceptando la traducción (si pasa una serie de pruebas) o señalando los errores que ha encontrado.

4.4. Qué hay en un fichero .po

Un fragmento nos servirá de ejemplo:

```
#: openvt/openvt.c:67
#, c-format
msgid "openvt: %s: illegal vt number\n"
msgstr "openvt: %s: número de term. virt. ilegal\n"
```

Notas

1. Santiago es el coordinador del equipo español y recuerda estar en el proyecto desde finales de 1995.
2. e-mail personal con fecha de 22 de abril de 2000.
3. http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?team=es (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?team=es).
4. http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=kbd (http://www2.iro.umontreal.ca/~pinard/po/registry.cgi?domain=kbd).

Capítulo 5. KDE y KBabel

El proyecto KDE[1] se preocupa también de la internacionalización de sus programas y librerías y recoge información pertinente en «*The KDE Translators' and Documenters' Web Site*»[2] . Uno de los documentos principales que allí aparecen[3] es «*The KDE Translation HOWTO*». Por lo pronto nos interesa lo que allí se llama la traducción del 'Interfaz Gráfico de Usuario'.

La herramienta de traducción del proyecto es KBabel (http://i18n.kde.org/tools/kbabel (http://i18n.kde.org/tools/kbabel)).

Kbabel en acción

Podemos ver en la figura 4 que se ha cotejado la traducción francesa del fichero. Dos aportaciones destacan en KBabel: el gestor de catálogos y la posibilidad de utilizar diccionarios, sean archivos .po auxiliares (como la citada traducción francesa) sean Compendios PO (http://i18n.kde.org/translation_archive/kde-i18n-es.tar.bz2 (http://i18n.kde.org/translation_archive/kde-i18n-es.tar.bz2) contiene el conjunto de las traducciones del equipo español).

Notas

1. http://kde.kde.org/.
2. http://i18n.kde.org/.
3. Está en http://i18n.kde.org/translation-howto/index.html (http://i18n.kde.org/translation-howto/index.html).

Capítulo 6. Gnome y gtranslator

La documentación sobre nuestro tema del proyecto Gnome, liderado por el mejicano Miguel de Icaza, está en http://www.gtranslator.org/leftern_index.html (http://www.gtranslator.org/leftern_index.html).

La herramienta del proyecto es gtranslator [1], creada inicialmente por Fatih Demir (kabalak) y Gediminas Paulauskas, y que últimamente ha experimentado un gran impulso.

6.1. Cómo se trabaja

La figura 5 nos muestra la apariencia del programa.

gtranslator funcionando

Notas

1. http://www.gtranslator.org (http://www.gtranslator.org).

III. Diccionarios

Capítulo 7. Diccionarios para humanos

Existen varios diccionarios consultables en línea [1] y hay varias interfaces de consulta disponibles. Nos vamos a centrar en varios de los que pueden ser descargados libremente. [2]

7.1. El tesauro de ORCA

El objetivo de este glosario[3] no es explicar el significado de los términos de informática usados en inglés, sino dar una lista de sugerencias para su traducción al español, para quien ya tenga suficientes conocimientos de informática en inglés.

La principal fuente para este glosario ha sido la comunidad hispano-parlante que desarrolla y usa software libre, participando directamente en la edición del glosario a través de su interfaz web http://quark.fe.up.pt/orca (http://quark.fe.up.pt/orca), o indirectamente a través de sus discusiones en las listas de correo sobre el tema. Distingue entre «colaboradores» y «editor»: un colaborador no puede borrar lo que ya ha sido escrito por otro, pero puede escribir comentarios; el editor después va a leer esos comentarios, y altera la definición respectivamente. Los números de versiones terminados en *.0* quieren decir que el glosario acaba de ser revisado por el editor; si el último número no es cero, indica el número de contribuciones que han sido introducidas desde la última revisión.

Para muestra, un botón; consultamos «*driver*» en la versión 2.0.178, de 16 de marzo de 2002. Esta es la salida:

```
driver
controlador, manejador, gestor, driver video
```

En la salida de dict tenemos otro ejemplo de consulta a ORCA, con un comentario de un colaborador.

7.2. Glosario de la ATI

Se trata del «glosario» de la Asociación de Técnicos de Informática. [4]

Veamos por ejemplo la entrada «FSF» en la versión HTML 4.0 (julio 2001) de la cuarta edición (mayo 2001):

```
FSF Ver: "Free Software Foundation"
```

Seguimos el enlace

```
Free Software Foundation -- FSF (Fundación para
el Software Libre)
```

```
Fundación norteamericana creada en 1996 por Richard
M. Stallman cuyo objetivo es promover el desarrollo
y el uso de software libre en todas las áreas de la
Informática. Ver también: "Free Software".
[Fuente: RFCALVO].
```

Me recuerda Jaime Villate que esta entrada tiene un error manifiesto; la FSF no fue creada en el ’96 sino en 1985.

7.3. El protocolo DICT

El *Grupo de Desarrollo de DICT* pretende dar solución a un problema: ¿cómo estandarizar el acceso a los múltiples diccionarios disponibles?

RFC 2229 describe el *protocolo DICT* como un protocolo de consulta/respuesta sobre TCP que permite a un cliente acceder a un diccionario de definiciones utilizando un conjunto de bases de datos de diccionarios de lenguajes naturales.

El grupo de desarrollo tiene su página en http://www.dict.org. Los servidores y clientes son libres (licencia GPL). Existen diccionarios y tesauros que pueden instalarse localmente. La distribución que utilizo, GNU Debian Woody, incluye los siguientes diccionarios de interés en nuestro campo:

dict-gcide

: the GNU version of the Collaborative International Dictionary of English, which is based on the Webster’s Revised Unabridged Dictionary (G & C. Merriam Co., 1913, edited by Noah Porter), supplemented by many definitions from WordNet, the Century Dictionary, 1906, and by numerous definitions contributed by volunteers.

dict-wn

: WordNet 1.7, A Lexical Database for English from the Cognitive Science Laboratory at Princeton University. WordNet defines only nouns, verbs, adjectives and adverbs. Other parts of speech, such as pronouns and articles, are omitted. Definitions in this dictionary are more concise that in the 1913 Webster. This is a 2001 edition, so it fills in many of the gaps left by the latter.

dict-foldoc
dict-jargon

: the Free On-line Dictionary of Computing, and the Hacker’s Jargon file. There is a great deal of overlap between the Jargon file and the FOLDOC. Although the FOLDOC is much larger than the Jargon file, there are numerous entries in the Jargon file that are not found in FOLDOC.

dict-vera

: a dictionary of acronyms used in the computer field.

i2e

: diccionario inglés-español de Alfredo Casademunt, a su vez basado en el trabajo de José Luis Triviño.

Además, he añadido

El Glosario de Orca

en su versión .dict

eng-spa
spa-eng

descargados de http://www.freedict.d (http://www.freedict.de)

leo_ftp

English-German dictionary (ftp://ftp.leo.org/pub/comp/doc/dict/ (ftp://ftp.leo.org/pub/comp/doc/dict/))

Un ejemplo nos mostrará su uso:

```
[Mi_maquina]$ dict font
```

da la siguiente salida

```
  8 definitions found

  From WordNet (r) 1.7 [wn]:

  font
  n 1: a specific size and style of type within a type
    family [syn: {fount}, {typeface}, {face}]
  2: bowl for baptismal water
        [syn: {baptismal font}, {baptistry},
    {baptistery}]

  From The Collaborative International Dictionary of English
  [gcide]:

  Font \Font\, n. [F. fonte, fr. fondre to melt or cast. See
     {Found} to cast, and cf. {Fount} a font.] (Print.)
     A complete assortment of printing type of one size,
     including a due proportion of all the letters in the
     alphabet, large and small, points, accents, and whatever
     else is necessary for printing with that variety of types;
     a fount.
     [1913 Webster]

  From The Collaborative International Dictionary of English
  [gcide]:

  Font \Font\, n. [AS. font, fant, fr. L. fons, fontis, spring,
     fountain; cf. OF. font, funt, F. fonts, fonts baptismaux,
     pl. See {Fount}.]
     1. A fountain; a spring; a source.
     [1913 Webster]

     Bathing forever in the font of bliss. --Young.
     [1913 Webster]
```

```
2. A basin or stone vessel in which water is contained for
   baptizing.
   [1913 Webster]

   That name was given me at the font.   --Shak.
   [1913 Webster]

From The Free On-line Dictionary of Computing (09 FEB 02)
[foldoc]:

font

   A set of {glyphs} ({images}) representing the
   {characters} from some particular {character set} in a
   particular size and {typeface}.  The image of each character
   may be encoded either as a {bitmap} (in a {bitmap font}) or
   by a higher-level description in terms of lines and areas
   (an {outline font}).

   There are several different computer representations for
   fonts, the most widely known are {Adobe Systems, Inc.}'s
   {PostScript} font definitions and {Apple}'s {TrueType}.
   {Window systems} can display different fonts on the screen
   and print them.

   [Other types of font?]

   (2001-04-27)

From i2e [i2e]:

   font : tipo (de letra)

From i2e [i2e]:

   font : fuente

From ORCA - Glosario de Informática Inglés-Español
[glosario]:

font
   fuente, tipo de letra, (TIPO DE LETRA, fuente, en español,
   tiene que ver con agua, no con tipografía)

From LEO ftp collection [leo_ftp]:

font
   Schrift
   Schriftart
```

7.4. Diccionarios. Apuntes para continuar

* *[ToDo. Proyecto Dino (http://boadicea.rediris.es/Dino/), de José Manuel Macías Luna <macias@rediris.es>]*

* *[ToDo. WorldWideLexicon (http://picto.weblogger.com/)]*

* *[ToDo. Ismael Olea <ismael@olea.org> me recuerda rl-dicc (http://cvs.hispalinux.es/cgi-bin/cvsweb/rl-dicc): «Sé que no está publicado del todo, pero sólo le queda un hervor y alguien que se lo dé. Y con todo no deja de ser un recurso extraordinario.»]*

Notas

1. Hay bastantes más que los que vamos a examinar, de interés especializado: pydict de inglés y chino; skk y edict de japonés, mueller para inglés-alemán…
2. Es bastante discutible que el Glosario de la ATI sea libre. Según las «normas»

   1. © 1994—2002 Rafael Fernández Calvo

   2. El autor autoriza la reproducción y difusión de este documento, por cualquier medio, sea en su totalidad o parcialmente, si es realizado sin ánimo de lucro por organizaciones sin ánimo de lucro. Estas organizaciones pueden también enlazar este glosario desde sus sitios web, si bien se agradece a los enlazantes que lo comuniquen al autor.

   3. Si las actividades citadas en 2. las realizan organizaciones con ánimo de lucro, o si las realizan con ánimo de lucro organizaciones sin ánimo de lucro, se requiere siempre el permiso previo por escrito del autor.

   4. En todos los casos es obligatoria la mención completa de la fuente.

      Según Javier Fernández-Sanguino Peña, uno de los desarrolladores principales del equipo español de Debian, en mensaje a <debian-l10n-spanish@lists.debian.org> de 24 de mayo de 2002 en respuesta a una consulta mía

   No se llegó a un acuerdo con el autor con respecto a las condiciones/licencia de distribución (…) No sé si las condiciones han variado.

      En el mismo hilo y mismo día Jaime Villate confirma

   El glosario de Rafael Fernández Calvo no ha sido convertido a formato .dict y no se puede crear un paquete Debian con él, por no ser un glosario libre. Mi plan actual es dejar que ORCA se convierta en glosario, en vez de tesauro, y ya comienzan a aparecer algunas explicaciones.
3. Proyecto ORCA — Herramientas de ayuda para los traductores y productores de software libre en español (programas y documentación), http://quark.fe.up.pt/orca/index.es.html (http://quark.fe.up.pt/orca/index.es.html); proyecto responsabilidad de Jaime E. Villate bajo los términos de la Licencia GNU Para Documentación Libre. Realmente es un *tesauro*, como reconoce el propio Villate en correo a la lista <debian-l10n-spanish> de fecha 29 de marzo de 2000: un «glosario» explica con detalle los términos, mientras que un «tesauro» sugiere simplemente sinónimos. Recordemos también que la intención del coordinador es convertirlo en un verdadero glosario.
4. http://www.ati.es/novatica/glointv2.txt (http://www.ati.es/novatica/glointv2.txt) ó http://www.ati.es/PUBLICACIONES/novatica/glointv2.html (http://www.ati.es/PUBLICACIONES/novatica/glointv2.html), coordinado por Rafael Fernández Calvo . Hay que recordar las objeciones que desde el mundo del software libre se ponen a su licencia.

Capítulo 8. Diccionarios para máquinas

8.1. Estándares relativos a la terminología

Interesante lista de estándares relativos a la terminología, en un mensaje de KaraWarburton al foro de discusión sobre «localización» y terminología de LISA,
http://www.lisa.org/sigs/phpBB/viewtopic.php?topic=69&forum=1&1
(http://www.lisa.org/sigs/phpBB/viewtopic.php?topic=69&forum=1&1) Como no tiene desperdicio lo parafraseo:

> Here is a list of Terminology Standards that I am familiar with and which I find useful. This is a starting point for a SIG list. Please post as a reply any additional ones which you find useful.
>
> 1. *TBX* - TermBase eXchange format. This is the XML terminology markup format proposed by the LISA/OSCAR standards group as an standard for the localization industry. More info here: http://www.lisa.org/tbx/ (http://www.lisa.org/tbx/)
> 2. *OLIF2* - Open Lexicon Interchange Format. This is an interchange format specifically for machine-readable lexicographical data, such as for machine translation systems. OLIF data will be recordable in TBX.
>
> *ISO TC 37 Standards*
>
> 3. ISO DIS 16642:2002 - *TMF* - Terminology Markup Framework. This is a high-level standard framework for defining individual *TMLs* (terminology markup languages). It covers basic structure and architecture of TMLs and terminology databases.
> 4. ISO 12620:1999 (under revision) - Terminology Data Categories. This standard is being revised into 2 parts. The first part describes a standard generic method for defining data categories for terminology collections (what standard properties they should have, etc.). The second part is an inventory of all the possible types of data categories in lexicology and terminology (term, part of speech, definition, context, variants, etc., etc.). This part is very useful as a catalog for picking data categories for your own terminology database.
> 5. ISO 12200:1999 - *MARTIF* - If you're still into SGML, this is a good established standard markup format. MARTIF is the basis for TBX and is also supported by a number of terminology tools. MARTIF will be integrated as an appendix in the final version of ISO 16642.
> 6. ISO 17241 - *GENETER*. Another standard SGML format for terminology, used by some databases in Europe. GENETER will be integrated as an appendix in the final version of ISO 16642.
> 7. ISO 704:2000 - Terminology Work - Principles and methods. Provides all kinds of useful information about terminology research methods and writing definitions, and other process-oriented tasks.
> 8. ISO 12616 - Translation-oriented terminography. Provides useful guidelines for terminology work specifically to support the translation process.
> 9. ISO 1087-1 and 1087-2 - Terminology Work - Vocabulary. These basically contain definitions of standard terms in terminology. A useful reference point for the SIG.

Problema

¿Cómo conseguir los caros estándares del Comité Técnico 37 de ISO, esenciales en este campo?

8.2. Estandarizando lexicones computacionales: OLIF2

«OLIF», the «*Open Lexicon Interchange Format*» [1]

is a user-friendly vehicle for exchanging terminological and lexical data.

What is special about OLIF?

OLIF is XML-compliant and offers support for *natural language processing* (NLP) systems, such as machine translation, by providing coverage of a wide and detailed range of linguistic features.

Notas

1. http://www.olif.net/ (http://www.olif.net/).

IV. Memorias de traducción

Capítulo 9. Estandarizando las memorias de traducción: TMX

http://www.lisa.org/tmx/ (http://www.lisa.org/tmx/)

> «*TMX*» stands for *Translation Memory eXchange*. «*OSCAR*» (Open Standards for Container/Content Allowing Re-use) is the LISA Special Interest Group responsible for its definition.
>
> The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process.

Muy interesante, en español, “ Una guía al TMX. De la Traducción Automática a la Traducción Asistida”, de Josu Gómez, del Grupo DELi, Universidad de Deusto.

Dice Gómez [1]

> «*TMX*» es un lenguaje que cumple las especificaciones XML, y cuyo propósito es proporcionar un estándar para el intercambio de las memorias de traducción. Cuando se esté trabajando con una utilidad y se desee pasar a trabajar con otra manteniendo la TM que se había ido recopilando, bastará con exportarla a formato TMX, e importarla en la nueva utilidad. Para ello es necesario que todas las utilidades soporten dicho formato: a 2001, puede decirse que ya hemos llegado a esta situación, puesto que actualmente las herramientas más importantes del mercado admiten la importación y exportación de memorias en TMX, si bien en distintos grados.

Aplicaciones que implementan TMX

- Deja Vu — Atril
- Eurolang Optimizer — LANT
- Foreign Desk — Lionbridge
- King Memo — Wolfgang Abele
- Logos Translation Control Center — Logos Corp
- Multitrans — Multicorpora
- Okapi Framework — OpenTag
- Prolyphic — Prolyphic
- ProMemoria 2.5 — Bridgeterm
- Sakhr Translator Workbench — Sakhr
- RC-WinTrans Software Localizer — schaudin.com
- SDLX 4.0 — SDL International
- Transit 3.0 — Star
- Trados 5 — Trados
- Trans Suite 2000 — Cypresoft
- Wordfast — Champollion & Partners
- WordFisher — Környei Tibor

Notas

1. http://sirio.deusto.es/abaitua/deli/xtrabi-e341.htm (http://sirio.deusto.es/abaitua/deli/xtrabi-e341.htm).

Capítulo 10. gtranslator

Según la FAQ de gtranslator [1]

> «*UMTF*» is a quite effective translation memory format currently used by gtranslator.

Sigue explicando la FAQ:

P: What is this learn buffer you're talking so often?

R: The «*learn buffer*» is the gtranslator specific implementation of a «translation» to help translators easyfying their work by translating always re-occuring messages with already learned and saved translations. gtranslator's learn buffer is a quite simple but effective kind of translation memory in the *UMTF XML format*.

After learning some of your already well translated po files you can simply use the auto translating capabilities of gtranslator to perform a good level of auto translation for already in the learn buffer saved message & translation pairs. A learn buffer just makes your life as a translator much easier as if you did build up a quite normal-sized (effectively 300+ KB) learn buffer gtranslator performs an auto translation level of ca. 15 %.

P: Will gtranslator support enhanced translation memories (in formats like *TMX/OpenTag*)?

R: Surely it will; support of translation memories is one of the biggest TODO's for future gtranslator releases. OpenTag support will occur surely one day, TMX also, but I think the order might be so that we would integrate first OpenTag and then TMX — but we will do it.

Este punto podemos actualizarlo con una referencia al archivo de la lista <`lucas-desarrollo@listas.hispalinux.es`>, hilo 'gtranslator, learn buffer etc.' que se inicia el 16 de enero de 2003 (y que continúa una conversación privada entre kabalak, Ismael Olea y Juan Rafael Fernández). Puede consultarse en https://listas.hispalinux.es/pipermail/lucas-desarrollo/2003-January/000256.html (https://listas.hispalinux.es/pipermail/lucas-desarrollo/2003-January/000256.html). En él kabalak pide consejo sobre qué formatos de TM implementar y si hacerlo de forma nativa o simplemente la capacidad de importar y exportar los formatos estándar. Nosotros le hablamos de la necesidad de adoptar los estándares abiertos, del mismo modo que la industria ha comprendido que son necesarios. La conclusión de kabalak es esta:

> I think I'll code for a UMTF-backboned gtranslator further on but with support for Opentag and TMX which is also somehow included in gtranslator's code (good news in this field, as the TMX and OpenTag fields are "very easy" to handle as XML files…)

> **Nota:** Ismael Olea ha puesto en http://www.olea.org/gtranslator-es/ (http://www.olea.org/gtranslator-es/) «una memoria de traducción de español construida con gtranslator con las actuales traducciones del proyecto Gnome».

Notas

1. http://www.gtranslator.org/faq.html (http://www.gtranslator.org/faq.html).

Capítulo 11. Mimers brunn

El 23 de febrero de 2002 VeronicaLoell anuncia en la lista <Translation-i18n@lists.sourceforge.net> (el anuncio puede consultarse en http://sourceforge.net/mailarchive/forum.php?thread_id=520105&forum_id=7939 (http://sourceforge.net/mailarchive/forum.php?thread_id=520105&forum_id=7939)) su versión 1.0.0a (pre-alpha) de «*Mimers brunn TM*» (el pozo de Mimer, gigante de la mitología nórdica dueño del pozo de la sabiduría), como parte del proyecto «Mimers brunn Translator tools»

> I have just released a very simple GUI Translation Memory. So far it has only very basic search facilities of glob-type. But there will eventually be more features and also APIs in various languages to communicate directly with the TM. The format before import is TMX 1.3. There is a facility (*PoToTMX*) to automatically convert a directory of po-files into tmx-format and then of course to import it into the TM.

La *url* del proyecto es http://mimersbrunn.sourceforge.net/TM.html (http://mimersbrunn.sourceforge.net/TM.html). Me temo que está algo parado porque la última modificación de su sitio *web* es del 23 de febrero de 2002.

V. Un paso adelante

En esta parte del documento pretendíamos hablar de lo que no tenemos y deberíamos de tener, de los estándares que está desarrollando la industria (¡y las instituciones!) y de las tendencias actuales. Ha crecido y se ha independizado; ahora constituye mi propuesta de especificación de TLDP http://es.tldp.org/especificaciones/herramientas-linguisticas/herramientas-linguisticas/ (http://es.tldp.org/especificaciones/herramientas-linguisticas/herramientas-linguisticas/).

Capítulo 12. Miscelánea

- Interesante, desde nuestra perspectiva (aunque algo anticuado, es de 1996), el trabajo de TomazErjavec “Public Domain Generic Tools: An Overview” (http://citeseer.nj.nec.com/430552.html (http://citeseer.nj.nec.com/430552.html))
- Paai’s text utilities: A set of utilities consisting of unix-scripts and c-programs for frequency-counts and lexical cohesion. De J.J.Paijmans (http://pi0959.kub.nl:2080/Paai/Publiek (http://pi0959.kub.nl:2080/Paai/Publiek)). Last additions: 23 December 2000.
- tea (a KWIC —KeyWord In Context— tool), de Masao Utiyama mutiyama@crl.go.jp (mailto:mutiyama@crl.go.jp), última versión de mayo de 2002 (http://www2.crl.go.jp/jt/a132/members/mutiyama/software.html (http://www2.crl.go.jp/jt/a132/members/mutiyama/software.html)) It displays keywords along with their contexts. Tea allows you to: search multiple text files, list search-words in a tree structure, sort retrieved contexts in various ways, etc.
- textseg (http://www2.crl.go.jp/jt/a132/members/mutiyama/software.html (http://www2.crl.go.jp/jt/a132/members/mutiyama/software.html))
- openNLP (http://opennlp.sourceforge.net/ (http://opennlp.sourceforge.net/))
- GATE (General Architecture for Text Engineering, http://gate.ac.uk/)

Apéndice A. Pequeño glosario de acrónimos

CAT

Computer Assisted Translation

CES

Corpus Encoding Standard

CL

Computational Linguistics

DXLT

Default XLT. Ver TBX

EAD

Encoded Archival Descriptions

EAFT

European Association For Terminology

EAGLES

Expert Advisory Group on Language Engineering Standards

EBMT

Example-Based Machine Translation

ELDA

European Language Resources Distribution Agency

ELRA

European Language Resources Association

ETIS

European Terminology Information Server

HLT

Human Language Technology

ISLE

International Standards for Language Engineering

IULA

Institut Universitari de Lingüística Aplicada

KWIC

KeyWord In Context

LISA

Localisation Industry Standard Association

MARTIF

Machine-Readable Terminology Interchange Format

MATE

Multilevel Annotation, Tools Engineering, Telematics Project

MLIS

MultiLingual Information Society

MRD

Machine Readable Dictionary

MT

Machine Translation

MULTEXT

Multilingual Text Tools and Corpora

NIMM

Natural Interaction and MultiModality

NLP

Natural Language Processing

OLIF

Open Lexicon Interchange Format

OSCAR

Open Standards for Container/Content Allowing Re-use

PAROLE

Preparatory Action for Linguistic Resources Organisation for Language Engineering

POS

Part of Speech

SALT

Standards-based Access service to multilingual Lexicons and Terminologies

SLDS

Spoken Language Dialogue System

SRX

Segmentation Rules eXchange Format

TBX

: TermBase eXchange (sometimes called DXLT)

TEI

: Text Encoding Initiative

TLDP

: The Libre (antes Linux) Documentary Project

TM

: Translation Memory

TMF

: Terminological Markup Framework

TMS

: Terminology Management System

TMX

: Translation Memory Exchange

XLIFF

: XML Localisation Interchange File Format

XLT

: XML representation of Lexicons and Terminologies (ver DXLT y TBX)