# O TRADUTOR AUTOMÁTICO INGLÉS-GALEGO DE OPENTRAD

**imaxin**|software[1]
imaxin@imaxin.com

[Recibido 16/04/09; aceptado 21/05/09]

**Resumo**

No presente artigo tentaremos explicar de xeito breve os diferentes sistemas de tradución existentes e abordaremos as actuais posibilidades de uso comercial e profesional para a tradución e localización de software tendo en conta o estado da arte e as tendencias que a nivel de investigación e desenvolvemento existen para o futuro. Igualmente faremos unha análise concisa dos sistemas actuais de tradución inglés-galego comparándoos con outros pares que inclúen o galego. Para finalizar, avaliaremos o hibridismo entre sistemas como estratexia adecuada que se debe seguir para desenvolvementos futuros no ámbito da tradución automática.

**Palabras clave:** ferramentas CAT, tradución automática, tradución humana asistida por computador (MAHT), tradución automática con intervención humana (HAMT), tradución automática baseada en regras (RBMT), tradución automática baseada en exemplos (EBMT), tradución automática estatística (SMT).

**Abstract**

In this article, we will try to explain briefly the different existing translation systems and will deal with the current commercial and professional possibilities for translation and software localisation, according to the state of the art and the research and development trends for the future. Moreover, we will concisely analyse the current English-Galician translation systems and will compare them with other pairs involving Galician. Finally, we will assess hybridism among systems as an effective strategy for upcoming developments in the Machine Translation field.

1 As áreas **imaxin**|context e **imaxin**|localiza, de **imaxin**|software, están constituídas neste momento por Antón Moura, Óscar Senra, Ángel López e José Ramom Pichel, e Bruno Ruival, Vanessa Vila Verde e Iago Bragado, respectivamente. Este artigo foi redactado de forma conxunta polos membros desas dúas áreas.

**Key Words:** CAT tools, Machine Translation, Machine-Aided Human Translation (MAHT), Human-Aided Machine Translation (HAMT), Rule-Based Machine Translation (RBMT), Example-Based Machine Translation (EBMT), Statistical Machine Translation (SMT)

## 1. Orientación da tradución automática para profesionais da tradución e localización

> *When the computer is improperly used, its effects are, of course, quite different. This happens when the attempt is made to mechanize the non-mechanical or something whose mechanistic substructure science has not yet been revealed. In other words, it happens when we attempt to use computers to do something we do not really understand. History provides no better example of the improper use of computers than machine translation.(Kay 1980:2).*

As persoas dedicadas á tradución e localización temos ao noso dispor numerosas posibilidades coa utilización de ferramentas software que facilitan o noso traballo. Existen verificadores ortográficos, verificadores gramaticais e dicionarios e glosarios en liña que permiten a revisión das traducións e a consulta e pescuda de termos. Tamén podemos contar con bases de datos terminolóxicas, de estruturas e frases traducidas e bases de datos temáticas. Todos estes recursos requiren a interacción humana nun grao que podemos clasificar como medio-alto.

O habitual no mundo da tradución é a utilización das chamadas ferramentas CAT (Computer-Aided Translation), programas de xestión de memorias de tradución, glosarios, etc. que facilitan o labor da tradución profesional e que poden ser clasificadas nun nivel medio-baixo na escala de interacción. No entanto, tamén habería un nivel baixo ou nulo, en que a persoa se limita a posteditar os textos traducidos mecanicamente ou simplemente non desempeña ningunha interacción por medio da tradución automática.

Deste modo, podemos dividir a utilización de ferramentas aplicadas á tradución e localización segundo o seu nivel de interacción humana en tres grandes grupos:

- MAHT – Machine-Aided Human Translation (tradución humana asistida por computador)
- HAMT – Human-Aided Machine Translation (tradución automática con intervención humana)
- MT – Machine Translation (tradución automática)

Evidentemente, as fronteiras entre MAHT e HAMT son difusas

dependendo de como utilicemos as ferramentas e os recursos de que dispoñamos. Por exemplo, a integración da tradución automática nunha ferramenta CAT sería unha combinación das dúas metodoloxías.

O sistema utilizado dependerá dos recursos de que dispoñamos e da súa eficacia, mais terá como obxectivo que os e as profesionais poidan centrar os seus esforzos en garantir unha maior calidade lingüística da tradución e a conxugación entre o nivel de intervención profesional no procesamento e postedición do material e a intervención mecánica.

Hoxe en día, os avances na investigación e no desenvolvemento de sistemas de tradución automática aínda non permiten obter resultados totalmente satisfactorios sen intervención humana e, canto maior for a distancia entre as linguas en cuestión, peores serán os resultados. A MT actualmente constitúese como unha ferramenta útil para a tradución e, como tal, debe ser integrada xunto coas restantes ferramentas existentes, aínda que será sempre precisa a intervención humana.

Entre linguas distantes é difícil baixar de 20% de erros, o que equivale a unha intervención de 1/5 por parte do tradutor ou tradutora. A nivel profesional este sistema podería non ser rendíbel e os seus usos serían máis de asimilación, isto é, a súa función sería facer comprensíbeis textos en linguas afastadas só até o punto de ter unha noción básica do tema que traten.

Porén, entre linguas próximas, a tradución automática integrada nos sistemas de tradución asistida por computador pode aforrarmos moito tempo, sendo preciso intervir o necesario para resolver problemas de desambiguación contextual.

## 2. Sistemas de MT

Na actualidade, o desenvolvemento da MT deu como resultado tres tipos fundamentais de sistemas:

- RBMT – Rule-Based Machine Translation (Tradución automática baseada en regras)
- EBMT – Example-Based Machine Translation (Tradución automática baseada en exemplos)
- SMT – Statistical Machine Translation (Tradución automática estatística)

RBMT foi o sistema predominante até a década de 1980 e, basicamente, caracterízase pola utilización de dicionarios bilingües e monolingües xunto con regras de transferencia estrutural, isto é, regras de tradución que permiten modelar as variacións entre a estrutura sintáctica e morfolóxica das linguas de orixe e de destino. Os sistemas RBMT poden recorrer en certos casos (como o Opentrad) á utilización do coñecemento estatístico para a análise morfolóxica, para a selección léxica, etc.

EBMT consiste na utilización dun corpus paralelo de exemplos nun proceso de tradución por analoxía. O sistema consiste na descomposición das frases, na tradución das unidades sobre a base da analoxía cos exemplos e na recomposición adecuada dos fragmentos nas frases.

SMT consiste na xeración da tradución sobre a base de modelos estatísticos extraídos da análise de *corpora* bilingües de traducións humanas.

Os sistemas RBMT contan cunha serie de problemas: as regras gramaticais poden chegar a ser moi complexas e deben ser formuladas por especialistas; os dicionarios non dan cobertura á multiplicidade de significados e ás súas restricións; a colocación e orde das frases e das perífrases é problemática; as estruturas complexas con frases longas son dificilmente analizábeis e sintetizábeis; e a dificultade para amplialos e mantelos fainos custosos. No entanto, teñen como vantaxes seren de fácil construción inicial, estaren relacionados co coñecemento lingüístico e seren eficientes nos fenómenos lingüísticos máis simples.

No caso de EBMT os problemas derivan da selección dos exemplos, a adición de exemplos innecesarios que poden sobrecargar o sistema, os custos relativos á pescuda e a difícil representación do coñecemento. Con todo, o seu lado positivo é que a súa base de datos é extraída dun corpus, está baseada en padróns de tradución e a intervención humana redúcese sensibelmente.

Os problemas dos sistemas SMT veñen provocados polo exhaustivo adestramento a que teñen que ser sometidos a través de grandes *corpora*, a ausencia de control da calidade do *corpus*, a escaseza de datos suficientes para determinadas linguas (inclusive as máis faladas), a inexistencia dun fondo ou base de coñecemento lingüístico e os caros custos de pescuda. Os beneficios teñen que ver coa extracción do coñecemento a partir dos *corpora,* a fácil realimentación con máis *corpora* e a posibilidade de readestramento, e tamén a fundamentación matemática do modelo.

### 3. Os sistemas de tradución para EN-GL e outros pares

Actualmente os tradutores automáticos existentes para a tradución inglés-galego son o Opentrad e o Google Translate. O primeiro utiliza o sistema RBMT e fai parte dun proceso de investigación en curso. O de Google utiliza o sistema SMT. Os resultados para o par bidireccional inglés-galego son bastante semellantes, se ben que por seren linguas afastadas o único que permite é termos unha aproximación para podermos comprender o conxunto dos temas dos textos traducidos. Isto é interesante para certos tipos de usos da MT que, no caso do Google, ten a utilidade de fornecer unha noción básica dun texto escrito en finlandés, por exemplo, mais a nivel profesional aínda non é posíbel tirar rendemento para o par inglés-galego.

No entanto, o tradutor do Google conta con numerosos erros lingüísticos debido tanto ao uso incorrecto da norma ortográfica de 2003 como á imposibilidade de distinguir en determinados casos a ortografía do

galego e do portugués. Probabelmente, cando o corpus galego non fornece unha solución o termo en cuestión é apañado do portugués e, tamén, do español. Tampouco existe un control da coherencia para as duplas solucións morfolóxicas, de xeito que se pode obter nun mesmo texto tanto formas en «-bel» como en «-ble», por exemplo. Contrariamente, cando o Opentrad non encontra un determinado termo, por defecto, este fica na lingua de orixe. Isto fai que teñamos a sensación de o desempeño do Google ser maior, mais unicamente podemos concluír que esta sensación é debida a que os lectores galegos teñen unha competencia adicional para leren en portugués, isto é, se todas as palabras que aparecen escritas en ortografía portuguesa co tradutor do Google permanecesen en inglés ou estivesen en alfabeto cirílico, por exemplo, non se daría este caso. Por tanto, a nivel técnico podemos asegurar que o sistema SMT conta cunhas posibilidades limitadas a curto prazo para un uso profesional da tradución automática para este par de linguas.

Se compararmos o par inglés-galego do Opentrad co par español-galego ou portugués-galego os resultados serán moi diferentes. A proximidade lingüística xa permite un uso profesional do tradutor automático que facilitará o labor da tradución, e será unicamente necesaria unha simple revisión da lingua de destino para obter uns bos resultados. Neste caso, a análise que realiza o motor é máis superficial ca no caso inglés debido á dita distancia lingüística. Os resultados que podemos obter con estes pares utilizando o Google Translate son moito máis precarios, polo que podemos concluír que, para usos profesionais, a tecnoloxía SMT non é funcional para estes pares de linguas, probabelmente polos problemas arriba indicados que os ditos motores poden provocar (sobre todo a ausencia dun corpus suficientemente grande). Por ese motivo, unicamente pode ser útil para a asimilación de textos.

Para aproveitar ao máximo os puntos fortes de cada sistema a liña de investigación iniciada por **imaxin**|software co prototipo inglés-galego de Opentrad pretende combinar os dous motores RBMT e SMT, o cal poderá dar como resultado unha maior calidade nas traducións ao suplir cada sistema as carencias do outro. Se este for o resultado final, sería factíbel a utilización de Opentrad como apoio á tradución humana ou, inclusive, a tradución automática inglés-galego con intervención humana.

### 4. Futuro: hibridismo entre sistemas

As novas perspectivas achan como natural a pescuda de métodos de hibridación de sistemas apañando o mellor de cada un para suplir as carencias existentes cando só utilizamos un.

Os sistemas híbridos poderían consistir en aproveitar as análises morfolóxicas e dependencias do RBMT, as diverxencias existentes entre o RBMT e o EBMT na súa estrutura bilingüe, a construción, colocación de palabras e selección de exemplos do EBMT e, do SMT, a selección das

formas máis frecuentes en dominios lingüísticos específicos e modelos fluídos de lingua de destino.

## REFERENCIA BIBLIOGRÁFICA

KAY, Martin. 1980. «The Proper Places of Man and Machines in Language Translation». Xerox Corporation. [http://www.mt-archive.info/Kay-1980.pdf].

## LIGAZÓNS DE INTERESE

http://www.periodicos.ufsc.br/index.php/traducao/article/viewFile/5392/4936
http://www2.dc.ufscar.br/~helenacaseli/pdf/2007/TeseDoutorado.pdf
http://www.hutchinsweb.me.uk/Leeds-2006-ppt.pdf
http://www.euromatrix.net/internal/minutes-of-project-meetings/euromatrix-review-meeting-02-2008/EM_WP6.pdf
http://www.google.com/intl/pt-PT/help/faq_translation.html#whatis
http://www2.dc.ufscar.br/~helenacaseli/pdf/2007/TeseDoutorado.pdf
http://www.hutchinsweb.me.uk/Leeds-2006-ppt.pdf
http://www.euromatrix.net/internal/minutes-of-project-meetings/euromatrix-review-meeting-02-2008/EM_WP6.pdf