

# MANUAL DE FORMACIÓN PARA USUARIOS

# Índice

<b><u>INTRODUCCIÓN.....</u></b>	<b>3</b>
<b><u>A TRADUCCIÓN AUTOMÁTICA.....</u></b>	<b>4</b>
<b>Que é a tradución automática?.....</b>	<b>4</b>
<b>Tecnoloxías utilizadas na actualidade.....</b>	<b>5</b>
<b>Expectativas e resultados.....</b>	<b>5</b>
<b>Exemplos de tradutores automáticos en liña.....</b>	<b>7</b>
1. Google Translate.....	7
2. Opentrad.....	9
3. Gramtrans.....	10
4. UTMK.....	11
<b><u>TRABALLAR CON OPENTRAD.....</u></b>	<b>12</b>
<b>Que é Opentrad?.....</b>	<b>12</b>
5. Breve historia do proxecto Opentrad e do motor Apertium.....	12
6. Descripción do funcionamento do sistema.....	12
<b>Os diccionarios e a escolla terminolóxica.....</b>	<b>15</b>
<b>O texto que vai ser traducido.....</b>	<b>15</b>
<b>A pos-edición.....</b>	<b>15</b>
<b>Como obter mellores resultados.....</b>	<b>16</b>
<b><u>LIGAZÓN DE INTERESE.....</u></b>	<b>17</b>

## Introducción

Este curso de formación ten por obxectivo servir como introdución á tradución automática en xeral e como guía para obter os melhores resultados ao utilizar os servizos de tradución automática ofrecidos por Opentrad.

Este documento está dividido en dúas partes: “A tradución automática” e “Traballar con Opentrad”. A primeira delas tenta mostrar unha visión global sobre a tradución automática, as diferentes tecnoloxías utilizadas e as expectativas creadas así como ofrecer varios exemplos de tradutores automáticos. Na segunda parte descríbese con detalle a plataforma de servizos de tradución automática Opentrad e fanse diversas recomendacións para obter traducións con certas garantías de calidade.

## A tradución automática

Se as novas tecnoloxías facilitaron a acumulación e xestión de cantidades inxentes de información en múltiples formatos e idiomas, a internet puxo parte desa información ao alcance dun público variado e multilingüe. En moitas ocasións a información procurada polos usuarios/as destas novas ferramentas está en idiomas que non entenden. Noutros casos non se chega a saber da existencia de datos de interese porque só se realizan consultas nos idiomas que se dominan.

A tradución automática cumpre neste contexto un papel fundamental, xa que posibilita a diseminación e/ou a asimilación de grandes cantidades de información. Ese papel pode cumplilo de forma directa, é o usuario/a quen utiliza unha aplicación de tradución automática, ou indirecta, é un tradutor profesional quen a utiliza. Neste último caso convértese nunha ferramenta de grande utilidade, xa que só desta forma os profesionais poden afrontar os novos retos centrando os seus esforzos na obtención de traducións de calidade.

### Que é a tradución automática?

Chámase tradución automática (TA) aos sistemas informáticos que traducen dun idioma para outro sen requerir a intervención humana durante o proceso. Dito doutra forma, un tradutor automático é unha ferramenta computacional que se utiliza para traducir textos, documentos, webs etc. de forma que non sexa necesaria a participación de persoas con coñecementos lingüísticos. Non debe confundirse a TA con outras ferramentas relacionadas coa tradución, como, por exemplo:

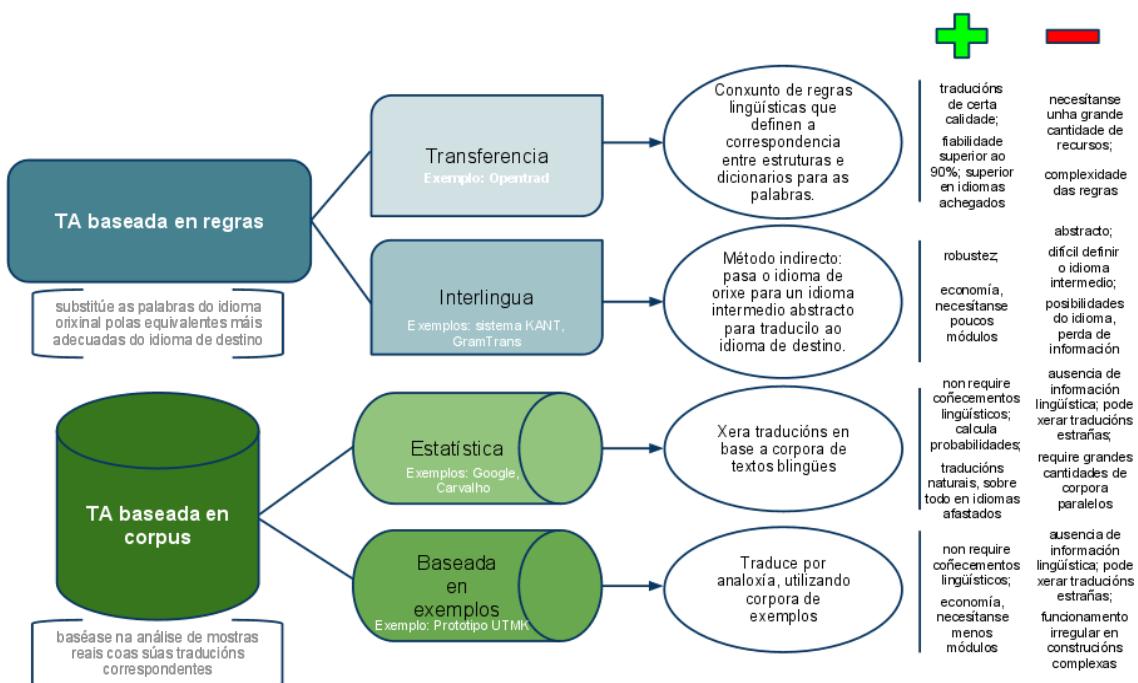
- ➔ As aplicacións de tradución asistida (CAT), ao estilo de Trados, OmegaT ou Wordfast, que ofrecen ferramentas de apoio ao tradutor humano, como bases de datos léxicas e terminolóxicas, dicionarios de sinónimos, correctores etc.
- ➔ Os dicionarios electrónicos, que ofrecen a tradución e definición de palabras soltas.

- ➔ Os correctores ortográficos e gramaticais, que corixen os errores ortográficos e gramaticais dun texto.

Por outro lado, os tradutores automáticos, como moitas outras aplicacións, poden ter diferentes formatos: webs, aplicacións instalables, *gadgets* etc.

## Tecnoloxías utilizadas na actualidade

Durante as últimas décadas coexistiron diferentes tendencias nas investigacións para mellorar o desempeño da tradución automática. En xeral pódese dicir que unha parte das solucións ten unha base lingüística e a outra parte está baseada en corpus. O seguinte cadro ofrece unha visión global sobre as diferentes tendencias e investigacións existentes:



## Expectativas e resultados

Desde o ano 54, en que se presentara publicamente un prototipo de tradutor inglés-ruso, até hoxe produciríonse moitos altibaixos na evolución da TA. Ese primeiro prototipo impulsou moitas investigacións e investimentos que se viron

reducidos a partir de 1966, cando un informe, o ALPAC, determinou que os resultados non eran os esperados e que non se obterían grandes avances a curto ou medio prazo. Só será en 1976 cando aparece unha aplicación práctica, o sistema canadense METEO, para a tradución de informes meteorolóxicos entre o inglés e o francés. Este sistema e máis o SYSTRAN, instalado pola comisión europea nese mesmo ano, provocaron un novo auxe nos investimentos e nas expectativas postas na TA que se estendería pola década dos oitenta. Nos noventa comezou a comercialización de produtos derivados, aínda que certamente de escasa calidade lingüística. Na actual década a tendencia apunta cara ao perfeccionamento das técnicas coñecidas e á exploración de técnicas híbridas que conxuguen as vantaxes das diferentes tecnoloxías.

Neste momento podemos encontrar desenvolvimentos que ofrecen grandes resultados para idiomas próximos, como Opentrad e os pares español-galego, español-portugués ou portugués-galego. Opentrad conta co motor Apertium de tradución baseada en regras, aínda que tamén utiliza a estatística para adestrar o módulo que define a categoría gramatical das palabras en caso de existir máis dunha opción. Este tipo de tecnoloxía resulta altamente recomendable debido á súa fiabilidade, xa que, en contraste coas técnicas puramente estatísticas, dispón dos coñecementos lingüísticos que lle fornecen as regras de transferencia estrutural e os seus tres dicionarios (un por idioma e outro bilingüe). A pesar desta grande fiabilidade, as máquinas son limitadas e tampouco dispoñen da capacidade creativa dos seres humanos, e, por tanto, recoméndase a revisión humana dos textos traducidos automaticamente que vaian ser feitos públicos.

O sistemas estatísticos non dispoñen deses coñecementos e poden ofrecer resultados inesperados e non relacionados co contexto orixinal, aínda que son sistemas recomendables para pares de idiomas afastados, como o inglés-galego (Google Translate e Carvalho - desenvolvido por **imaxin|software** - son os dous sistemas existentes neste momento que utilizan esta tecnoloxía para dito par). As ferramentas de TA derivadas desta tecnoloxía son fundamentalmente de “asimilación”, é dicir, resultan de grande utilidade á hora de comprender o asunto dun texto cando o idioma de orixe resulta incomprensible para o usuario/a. Veremos exemplos destas e doutras ferramentas no seguinte punto.

## Exemplos de tradutores automáticos en liña

### Google Translate

Entre os tradutores más utilizados destaca Google Translate, sobre todo debido á grande variedade de idiomas que abrangue (51) e á popularidade doutras ferramentas de Google (Gmail, Google Earth). Só tendo en conta a calidade final o seu nivel de uso talvez sería menor, como veremos. Utiliza a tecnoloxía estatística, que non require coñecementos lingüísticos e si grandes cantidades de corpora, dos cales extrae a información para decidir a tradución do texto. Os resultados da tradución dependen da información acumulada, por iso o resultado non se corresponde sempre co contexto desexado, dando por resultados textos non totalmente relacionados cos orixinais.

No caso do galego (e doutros idiomas con escasos recursos), Google non dispón de corpora como para ofrecer unha tradución de calidade, polo cal tivo que recorrer ao portugués e ao castelán. Isto provoca que nos encontremos con traducións estrañas que non o fan fiable para usos de diseminación aínda que si de asimilación. Vexamos un exemplo de tradución feita con Google Translate:

The screenshot shows a news website layout with several articles and sections. At the top, there are links for 'Enlaces' (Links) and 'Liga adelante' (Liga ahead), followed by a section titled 'Mellores da Semana' (Best of the Week) with tabs for 'Resultados', 'Clasificación', and 'Estadísticas'. To the right, there is a digital interview with Jose Samani. Below these, there are two main columns of news articles.

- DEPORTES**
  - Un coloso no Bernabeu** (Cayetano ROS)
 

Todo o Real Madrid goleou o Villarreal por unha exhibición de Cristiano Ronaldo. - Barca non deixa dubida contra o Racing

    - Cristiano inflamar - Estadísticas - Crónicas das partes
  - Centenario Guardiola** (LUIS MARTIN)
 

Ademais do fin e títulos, a sala de curativo Barça recorda varias pequenas vitorias

    - "Cen veces Pep", RAMON BESA
  - Gran superioridade** (Juana ITURRIAGA)
 

Represión tremenda, e agora o terceiro, o Barça e Madrid

    - Ricky está en Ferrari
  - Himno vaiado** (Los Reyes)
 

Los Reyes, os espectadores forá da final da Copa do Mundo, celebrado en Bilbao, baloncesto
  - 'Abrir'** (David Trueba)
- 21 civís mortos en Afganistán**

As vítimas dos bombardeos da OTAN, aparentemente, foron confundidos cos talibás

  - PARECER.. - 'Choose Your Future War' por Paul Kennedy
- "A UE nunca vai bater a ninguén o que a nós"**

Javier Moreno - 18 comentarios

Entrevistamos o primeiro ministro turco Recep Tayyip Erdogan, que comeza hoxe unha visita oficial a España

  - O uso de carteis en árabe Crisp campaña Holanda
  - Leuthard: "Ata agora, a Suíza non debería entrar na UE"

publicidad

**324VO**

Enmarcáronse en cadros rosas as frases que ofrecen resultados que resultan evidentemente diferentes ao orixinal:

The screenshot shows a news website layout. At the top, there's a header with 'LOS VÍDEOS DE LA LIGA' and a link 'Todo sobre la liga >'. Below this are several news cards:

- Liga BBVA:** Madrid y Athletic golean, el Atlético tropieza ante Almería.
- Liga Adelante:** La Real Sociedad pincha y el Hércules es nuevo líder de Segunda.
- Lo mejor de la jornada:** Includes links to 'Resultados', 'Clasificación', and 'Estadísticas'.
- ENTREVISTA DIGITAL José Sámano:** A redactor jefe de Deportes charlará con los lectores el HOY a las 18:00.

On the left side, there's a large image of a person's hands working on a small object. Next to it is a news card:

**21 civiles muertos en Afganistán**  
Las víctimas del bombardeo de la OTAN han sido confundidas al parecer con insurgentes talibanes  
•OPINIÓN.- 'Elige tu futura guerra', por PAUL KENNEDY

Below this is another news card:

**"La UE nunca impuso a nadie lo que a nosotros"**  
JAVIER MORENO - 18 comentarios  
Entrevistamos al primer ministro turco, Recep Tayyip Erdogan, que inicia hoy una visita oficial a España  
•El uso de carteles en árabe crispa la campaña holandesa  
•Leuthard: "De momento, Suiza no debe entrar en la UE"

On the right side, there are more news cards:

**DEPORTES**  
**Un coloso en el Bernabéu** CAYETANO ROS  
El conjunto madrileño vapulea al Villarreal con una exhibición de Cristiano Ronaldo. El Barça no deja lugar a las dudas ante el Racing  
•Cristiano entra en combustión  
•Estadísticas •Crónicas de los partidos

**Centenario Guardiola** LUIS MARTÍN  
Más allá de finales y títulos, el vestuario del Barça recuerda varios triunfos menores  
•'Cien veces Pep', RAMÓN BESA

**Gran superioridad** JUANMA ITURRIAGA  
Tremendo varapalo, y ya van dos, del Barça al Madrid  
•Ricky va en Ferrari

**Abucheo al himno** LOS Reyes, espectadores excepcionales de la final de la Copa de baloncesto celebrada en Bilbao

**'Open'** DAVID TRUEBA

At the bottom right, there's a logo for '324VO' and a small 'publicidad' label.

Podemos percibir que encontramos unha grande cantidade de erros ou inexactitudes que fan do texto traducido diferente do orixinal e pouco fiable, xa que nalgúns casos mesmo poden entenderse cousas opostas. A nivel lingüístico na tradución podemos encontrar lusismos (*tropeçar*), frases traducidas para o inglés (*Real Sociedad > Royal Society | Elige tu futura guerra > Choose Your Future War*), alteracións do texto orixinal que poden dificultar ou impedir a comprensión do texto (*Tremendo varapalo, y ya van dos, del Barça al Madrid > Represión tremenda, e agora o terceiro, o Barça e Madrid*) e identificacións erradas de nomes propios (*El uso de carteles en árabe crispa la campaña holandesa > O uso de carteis en árabe Crisp campaña Holanda*).

## Opentrad

Opentrad, que será descrito más adiante, ofrece unha tradución más literal e respectuosa co orixinal, se o comparamos con Google Translate. Podemos consultar a seguir a mesma tradución do xornal El País realizada con Opentrad:

The screenshot shows a section of the El País website with the following content:

- Liga BBVA**: Madrid e Athletic golean, o Atlético tropeza ante Almería
- Liga Adiante**: A Real Sociedad pica e o Hércules é novo líder da Segunda
- O mellor da xornada**: Resultados, Clasificación, Estatísticas
- ENTREVISTA DIXITAL José Sámano**: O redactor xefe de Deportes charlará cos lectores o HOXE ás 18:00
- DEPORTES**
- Un coloso no Bernabéu**: CAYETANO ROS. O comunio madrileño vapulea ao Villarreal cunha exhibición de Cristiano Ronaldo. O Barça non deixa lugar ás dúbidas ante o Racing.
- Centenario Guardiola**: LUIS MARTÍN. Mais aló de finais e títulos, o vestuario do Barça lembra varios triunfos menores.
- Gran superioridade**: JUANMA ITURRIAGA. Tremendo pau, e xa van dous, do Barça ao Madrid.
- Apuro ao himno**: OS REYES, espectadores excepcionais da final da Copa de baloncesto celebrada en Bilbao.
- 'Open'**: DAVID TRUEBA.
- O cóctel Vázquez Montalbán**: JUAN CRUZ. Unha antoloxía xornalística traza un retrato da figura do escritor que mesturou a alta cultura coa cultura popular.
- O modelo de Arco toca fondo**: I. LAFONT / R. BOSCO. A feira de arte de Madrid cerra a súa 29.ª edición con incógnitas sobre a súa vocación.
- publicidad**
- 324VO**

## "A UE nunca impuxo a ningún o que a nós"

JAVIER MORENO - 18 comentarios  
Entrevistamos ao primeiro ministro turco, Recep Tayyip Erdogan, que inicia hoxe unha visita oficial a España  
• O uso de carteis en árabe crispera a campaña holandesa  
• Leuthard: "De momento, Suíza non debe entrar na UE"

Enmarcáronse en laranxa o que podemos considerar como erro: a tradución dun nome propio, Cristiano indica que o tradutor podía escolher entre varias categorías gramaticais para a palabra “Cristiano”, e escolleu erradamente a forma do adjetivo en vez da do nome propio. O abuso no uso da preposición débese á falta do verbo “vapulear” no diccionario nun caso e á falta de información sobre o comportamento do verbo “entrevistar” no outro, xa que si que existe unha regra de transferencia estrutural que indica o comportamento correcto. Á parte destes dous casos (xa está corrixido o caso das preposiciones nos novos diccionarios), podemos percibir que a tradución resultante é de grande corrección e calidade, obtendo un texto perfectamente comprensible e correcto.

## Gramtrans

Este é un exemplo de tradutor que serve só para a asimilación, xa que pode axudar a definir o asunto dun texto áinda que talvez non ao comprender na súa totalidade. Exemplo de uso (neste caso de dinamarqués para galego):

The screenshot shows the GramTrans website interface. At the top, there is a navigation bar with links to Home, Products, Languages, DeepDict, QuickDict, and WebPa Lexical. Below the navigation bar, there is a banner with the word "GRAMTRANS" in large, stylized letters. On the left side, there is a sidebar with a "Navigation" section containing links to Home, News, Products, Free, Personal, Commercial Lite, Commercial Standard, Commercial Direct, Schools, Languages, System Features, Donate Words, QuickDict, GramTrans, Contact, Feedback, and Research. There is also a "Tools / Other" section with links to Document Translation, ... via Email, Mozilla Firefox Extension, Texting (SMS), and Linking to GramTrans. The main content area has a "Translated Result" section showing three lines of Danish text translated into Galician. Below that is a "Text Translation" section where a Danish sentence is input and translated into English. The "Web Page Translation" section shows a URL input field and a "Translate" button. The "Cross Language Search" section has a dropdown menu set to "Danish to Galician" and a "Translate & Search" button.

## UTMK

Aínda que non dispón do galego e aínda está en fase de probas, xa podemos ver un exemplo de tradución de inglés para malaio deste tradutor baseado en exemplos:

**Example-Based Machine Translation  
(EBMT) Prototype**

Input Text: (max. 800 words)

My name is Uxia.

Output Text:

nama saya Uxia.

Time taken: 438ms

Domain:

Proper Noun Detection  
(25,286 translation examples from general texts)

Last Updated: Friday, July 11, 2008  
©2001-2008 UTMK Research Group. All rights reserved.

## Traballar con Opentrad

### Que é Opentrad?

Opentrad é unha plataforma de servizos de tradución automática en código aberto que traduce entre 21 pares de idiomas. Estes pares encóntranse en distinto nivel de desenvolvemento, e de entre eles destacan: español-galego, español-catalán, español-éuscaro, español-portugués, portugués-galego, español, francés, inglés-español ou inglés-galego.

Esta plataforma utiliza como base o motor de tradución automática [Apertium](#), motor de transferencia superficial baseado en regras.

### Breve historia do proxecto Opentrad e do motor Apertium

Levado a cabo por un consorcio de empresas e universidades, no cal participa **imaxin|software**, Opentrad naceu no ano 2004 como un proxecto de tradución automática de código aberto. Financiado nun principio polo Ministerio de Industria, Turismo e Comercio a través das axudas PROFIT, o proxecto Opentrad recibiu tamén financiamento da Generalitat de Cataluña, da Xunta de Galicia, da Universitat d'Alacant e das empresas que conforman o consorcio Opentrad.

Apertium é dos dous motores de tradución automática de código aberto desenvolvidos dentro do proxecto Opentrad, e foi inicialmente deseñado para traducir entre idiomas relacionados, aínda que nestes últimos anos está a ser utilizado para a tradución de pares como o inglés-galego. O outro motor, Matxin, foi deseñado exclusivamente para o par español-éuscaro.

### Descripción do funcionamiento do sistema

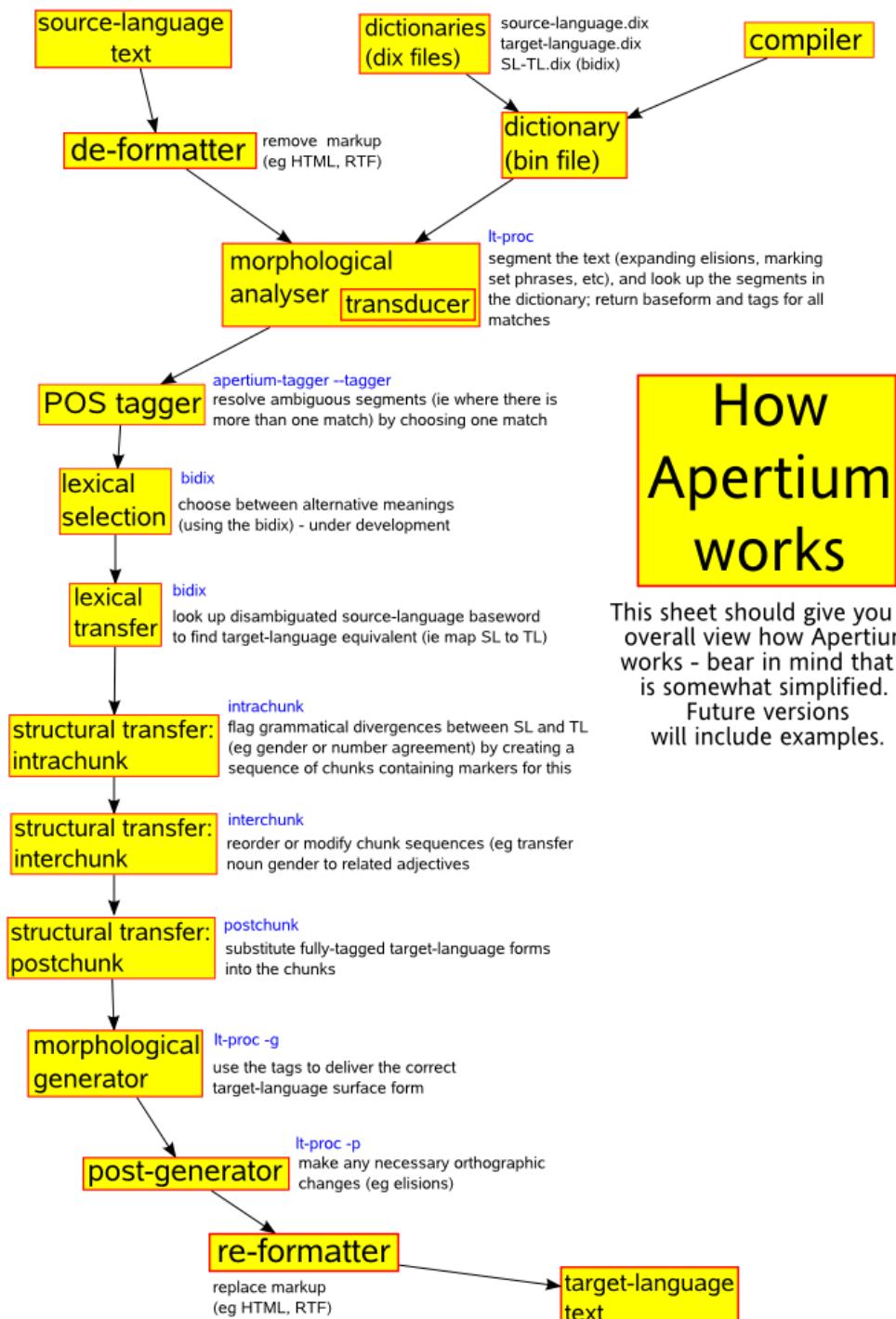
O tradutor Opentrad funciona grazas ao motor Apertium, o cal está composto por unha serie de módulos polos cales vai pasando o texto introducido polo usuario/a de forma que o último deses módulos devolve o texto na lingua de

destino indicada. O primeiro dos módulos separa o texto das posibles marcas de formato e envíao para o segundo, que o analiza morfoloxicamente e este para os seguintes dous, os cales, entre outras cousas, resolven posibles ambigüidades de significado e detectan o equivalente na lingua de destino consultando os seus diccionarios (ten tres: un por idioma e outro bilingüe). Após esas análises o texto pasa ao módulo de transferencia estrutural, que reordena as palabras ou secuencias de palabras segundo certas regras gramaticais que resolven as diverxencias entre a lingua de orixe e a de destino. Por último, xérase o texto xa traducido e devólveselle o formato de orixe.

A seguir pode verse un esquema explicativo do funcionamento de Apertium<sup>1</sup> e que mostra todos os módulos que o compoñen:

---

<sup>1</sup> Extraido de: [http://wiki.apertium.org/wiki/Apertium\\_for\\_Dummies](http://wiki.apertium.org/wiki/Apertium_for_Dummies)



## Os dicionarios e a escolha terminolóxica

Tal e como se pode observar no cadro anterior, os dicionarios son a orixe do coñecemento terminolóxico do tradutor. Os dicionarios están construídos en XML e son tres: un por idioma e outro bilingüe. O dicionario bilingüe relaciona o termo dun idioma co equivalente no outro. Vexamos, por exemplo, como está almacenada a palabra “*trivialidad*”/“*trivialidade*”:

- > No dicionario de español: <e lm="trivialidad"> <i>trivialidad</i> <par n="a\_n"/> </e>
- > No dicionario de galego: <e lm="trivialidade"> <i>trivialidade</i> <par n="aba\_n"/> </e>
- > No dicionario bilingüe: <e> <p> </>trivialidad<s n="n"/></l> <r>trivialidade<s n="n"/></r> </p> </e>

Desta forma, “*trivialidad*” tradúcese por “*trivialidade*” e viceversa. Este é un caso simple. Os problemas aparecen, por exemplo, cando un termo non ten unha tradución unívoca. É o caso de “*aire*” ou “*loco*” que traducirse por “*aire*” ou “*ar*” e por “*louco*” ou “*tolo*”. A escolha terminolóxica defínese tendo en conta:

- os contextos que se prevén predominantes á hora de utilizar o tradutor,
- o libro de estilo da entidade ou organización que o vai utilizar,
- os termos consolidados polo uso.

## O texto que vai ser traducido

Os resultados da tradución automática dependen en grande medida da calidade do texto que vai ser traducido. Se o texto contén erros estes serán transmitidos á tradución, xa que o tradutor non dispón dun corrector. As palabras mal escritas, incorrectas ou non aceptadas pola norma do idioma de orixe non van ser traducidas, xa que o tradutor, como os dicionarios correntes, só garda palabras na súa forma correcta.

## A pos-edición

Os resultados do tradutor automático deben revisarse sempre que sexa posible. Hai que lembrar que o tradutor non dispón dos mesmos coñecementos que

os seres humanos ou da súa capacidade creativa e moito menos da súa sensibilidade e intuición.

Hai contextos que poden resultar ambiguos e provocar que o tradutor automático escolla a opción errada, xa que, mediante estatísticas de frecuencia de uso nun corpus concreto, o computador pode escoller unha opción mais en ningún caso ten coñecemento ou comprende o que está a traducir. Trata as palabras como datos, como números, o que tende a empobrecer o idioma.

## Como obter mellores resultados

Existen certos trucos que poden axudar a obter unha tradución automática de alta calidade. Estes son os consellos más relevantes:

- I. Corrixir o texto orixinal antes de realizar a tradución. Un tradutor automático non é un corrector, polo cal os errores que teña o texto orixinal afectarán o seu desempeño. Por exemplo, traducir a frase "*el tendría la libertad*" dá por resultado "*o \*tendría a libertade*", xa que "él" (pronome) e "tendría" (verbo) son palabras acentuadas en castelán e o tradutor identificou, respectivamente, como artigo ("el" sen acento) e como palabra descoñecida.
- II. As frases curtas e simples favorecen que as traducións sexan correctas, xa que é máis fácil evitar desta forma posibles ambigüidades. Un texto ambiguo resultará nunha tradución ambigua.
- III. Coidar tanto a puntuación como a sintaxe, elementos fundamentais para a comprensión daquilo que se quere expresar e para a súa correcta transmisión.
- IV. Evitar as expresións coloquiais e idiomáticas xa que talvez causen confusión e dificulten a boa comprensión por parte da persoa tradutora/lectora e, ao mesmo tempo, provoquen errores na tradución.
- V. A revisión da tradución é fundamental xa que os procesos automatizados tenden a empobrecer o idioma.

## Ligazóns de interese

---

- [Tradutor de transferencia Opentrad](http://www.opentrad.com/) (<http://www.opentrad.com/>)
- [Tradutor baseado en exemplos de UTMK](http://utmk.cs.usm.my/) (<http://utmk.cs.usm.my/>)
- [Sistema KANT, tradución interlingua](http://www.lti.cs.cmu.edu/Research/Kant/) (<http://www.lti.cs.cmu.edu/Research/Kant/>)
- [Tradutor estatístico de Google](http://translate.google.com/) (<http://translate.google.com/>)
- [Tradutores de transferencia e interlingua de GramTrans](http://gramtrans.com/) (<http://gramtrans.com/>)
- [Motor Apertium](http://wiki.apertium.org/wiki/Main_Page) ([http://wiki.apertium.org/wiki/Main\\_Page](http://wiki.apertium.org/wiki/Main_Page))
- [Apertium for Dummies](http://wiki.apertium.org/wiki/Apertium_for_Dummies) ([http://wiki.apertium.org/wiki/Apertium\\_for\\_Dummies](http://wiki.apertium.org/wiki/Apertium_for_Dummies))
- [imaxin|software](http://www.imaxin.com) (<http://www.imaxin.com>)