EL DESARROLLO DE UN SISTEMA ABIERTO DE TRADUCCIÓN AUTOMÁTICA PIONERO (RUMANO-ESPAÑOL) A PARTIR DE UN MODELO SIMILAR (CATALÁN-ESPAÑOL) DISEÑADO EN LA UNIVERSIDAD DE ALICANTE

DELIA IONELA PRODAN
Universidad de Alicante

RESUMEN. *En el presente artículo se expone el proceso de desarrollo de un traductor automático del rumano al español (ro-es) llamado Trautorom. En primer lugar se presenta la iniciativa de crear este sistema y la acogida que tuvo. Asimismo se plantean algunos de los beneficios que este traductor puede aportar tanto a la sociedad rumana como a la Unión Europea en general y a España en particular. A continuación se describe el proceso de desarrollo del primer traductor automático (ro-es) como un sistema abierto, la base de partida –el motor y los programas informáticos desarrollados en la Universidad de Alicante por el grupo de investigación Transducens bajo el nombre de Apertium-, las principales dificultades encontradas, el estado actual de desarrollo, las estadísticas de evaluación, las estrategias de mejora a corto y a largo plazo. Al final se plantean algunas perspectivas de futuro de Trautorom.*

PALABRAS CLAVES: *traductor automático, rumano, español, Trautorom, Apertium, sistema abierto.*

ABSTRACT. *This article describes the development process of a machine translation system from Romanian to Spanish (ro-es) called Trautorom. First of all, it presents the initiative to create this system and how this initiative was received. It also describes some of the benefits this machine translator could provide both to the Romanian society and to the European Union in general and to Spain in particular. Next, it describes the development process of the first machine translator (ro-es) as an open system, its starting point –Apertium, the computer engine and programs developed at the University of Alacant by the Transducens research group-, the main difficulties it presents, current development stage, evaluation statistics and the short-term and long-term strategies for its improvement. It concludes by focusing on some of the Trautorom's future perspectives.*

KEY WORDS*: machine translator, Romanian, Spanish, Trautorom, Apertium, open system.*

1. TAUTOROM – EL PROYECTO QUE PROPONE UN TRADUCTOR DEL RUMANO AL ESPAÑOL Y DEL ESPAÑOL AL RUMANO

1.1. *Iniciativa del proyecto.*

En 2005, la Dra. Cătălina Iliescu Gheorghiu, profesora de inglés y rumano en el Departamento de Traducción e Interpretación de la Facultad de Filología y Letras de la Universidad de Alicante (U.A.) plantea una iniciativa pionera: crear un traductor automático del rumano al español y del español al rumano. Su propuesta recibe el apoyo del Vicerrectorado de Extensión Universitaria y del Departamento de Lenguajes y Sistemas Informáticos (DLSI) de la U.A. y se concede una colaboración docente para iniciar el desarrollo del traductor. En 2006, el Gobierno de Rumanía, a instancia de la propuesta formulada por la Dra. Iliescu, aprueba la financiación de un proyecto gubernamental, Trautorom, cuyo resultado final deberá concretizarse en la realización de un sistema de traducción automática bidireccional entre el rumano y el español fiable y funcional.

1.2. *Potenciales beneficiarios y beneficios.*

Los beneficios que Trautorom -en la fase actual de traductor del rumano al español- aportaría, se podrían reflejar en varias esferas del ámbito público y privado:

a) constitucional y legislativa: cualquier persona interesada podría tener acceso inmediato a la Constitución y a las leyes rumanas en vigor;
b) económica: cualquier empresario, indiferentemente de la rama de especialidad de su negocio, podría acceder por Internet a páginas que describan actividades económicas similares en España/Rumanía, contactar con empresarios españoles/rumanos, abrir sucursales en España/Rumanía o incluso montar sus propios negocios en estos países, lo que sería benéfico tanto para el mercado español (o de otros países que utilicen o manejen el español) como para el mercado rumano;
c) social: la libre movilidad de fuerza de trabajo -que constituye un tema central e inquietante en los foros de debates europeos actuales- debería verse regulada por el conocimiento riguroso y siempre actualizado de las ofertas y demandas del mercado laboral de la U.E., como también de las normas y políticas que rigen este mercado en general o en algún momento particular. Trautorom podría contribuir sustancialmente a la divulgación de estas informaciones. Asimismo, los trabajadores sociales de los centros de acogida o de los centros e instituciones dedicadas específicamente al tema de la inmigración podrían contar con un instrumento útil para conocer y comprender la situación de aquellos inmigrantes que no dominan suficientemente el español como para hacerse entender;
d) educativa: los alumnos rumanos incorporados en los colegios españoles sin conocer o dominar el idioma podrían contar con este diccionario-traductor electrónico no sólo para estudiar, sino también para plantear sus problemas de integración y adaptación escolar. Asimismo, los estudiantes españoles que reciben una Beca Erasmus a Rumania –situación cada vez más frecuente en el ámbito universitario abierto a la Unión Europea- podrían disponer de este instrumento electrónico para facilitar y aprovechar al máximo su estancia en Rumania;
e) cultural: el pleno conocimiento de otras sociedades implica *sine qua non* un acercamiento a sus valores culturales (tradiciones, ideologías, tesoros artísticos) y la cultura rumana, desafortunadamente, ha sido condenada en cierta medida a la invisibilidad en el escenario europeo y mundial sobre todo a causa de las circunstancias históricas desfavorables; etc.

2. DESARROLLO DEL SISTEMA DE TRADUCCIÓN AUTOMÁTICA TRAUTOROM

2.1. *Desarrollo del traductor*

Los cimientos de Trautorom están sentados sobre Apertium, un motor de traducción automática diseñado en la Universidad de Alicante. Apertium, fruto del trabajo de Transducens, un equipo de investigadores en lingüística e informática dirigidos por el Cat. Mikel Forcada Zubizarreta del DLSI (U.A.), funciona actualmente como un sistema abierto. La ventaja de ser un sistema abierto hace que sus herramientas y datos fundamentales sean accesibles para toda clase de usuarios[1] y sean potencialmente reutilizables y adaptables en el ámbito de la traducción a otros fines o a otros pares de idiomas.

Trautorom se ha desarrollado a partir de uno de los modelos ofrecidos por Apertium: el sistema de traducción automática del castellano al catalán y del catalán al castellano.

En este artículo se hará tan sólo un resumen del funcionamiento básico, dado que existen ya numerosas publicaciones sobre Apertium[2] realizadas por los mismos especialistas que lo han desarrollado.

2.1.1. Utensilios primordiales:

A) 2 DICCIONARIOS MONOLINGÜES (uno para la lengua origen, LO y otro para la LT lengua término)

Cada uno de estos diccionarios contiene:

a) una sección del alfabeto;
b) una sección de los símbolos gramaticales usados (ej. "n" = nombre, "ij" = interjección, "adv" = adverbio, etc.);
c) una sección de los paradigmas o los modelos de flexión;
d) una sección con las palabras del diccionario distribuidas por categorías y paradigmas.

1 DICCIONARIO BILINGÜE (que reflecta las correspondencias entre las palabras de las dos lenguas involucradas).

B) 1 ETIQUETARIO

El etiquetario es un archivo que agrupa todas las etiquetas posibles de las categorías y subcategorías utilizadas (ej. sustantivo: <nombre>, <género>, <número>, <caso>, <definido/indefinido>, etc.)

C) 1 PROCESADOR DE TRANSFERENCIA ESTRUCTURAL

Este procesador es un archivo que tiene dos bloques:

a) una sección que define algunas de las categorías y subcategorías gramaticales que se pueden ver implicadas en estructuras gramaticales complejas o distintas en LO y LT;
b) una sección donde se describen los patrones de transferencia estructural de un idioma a otro de modo que el traductor pueda generar la forma gramatical correcta en la LT.

D) 2 COMPILADORES

La barra de herramientas de Apertium cuenta con dos compiladores cuya función es transformar los datos lingüísticos (léxicos y estructurales) en códigos informáticos.

2.1.2. Módulos

El motor del sistema de traducción automática Apertium funciona a base de 8 módulos. El siguiente esquema pone de manifiesto, en grandes líneas, el trayecto de procesamiento desde la lengua origen (LO) hasta la lengua término (LT):

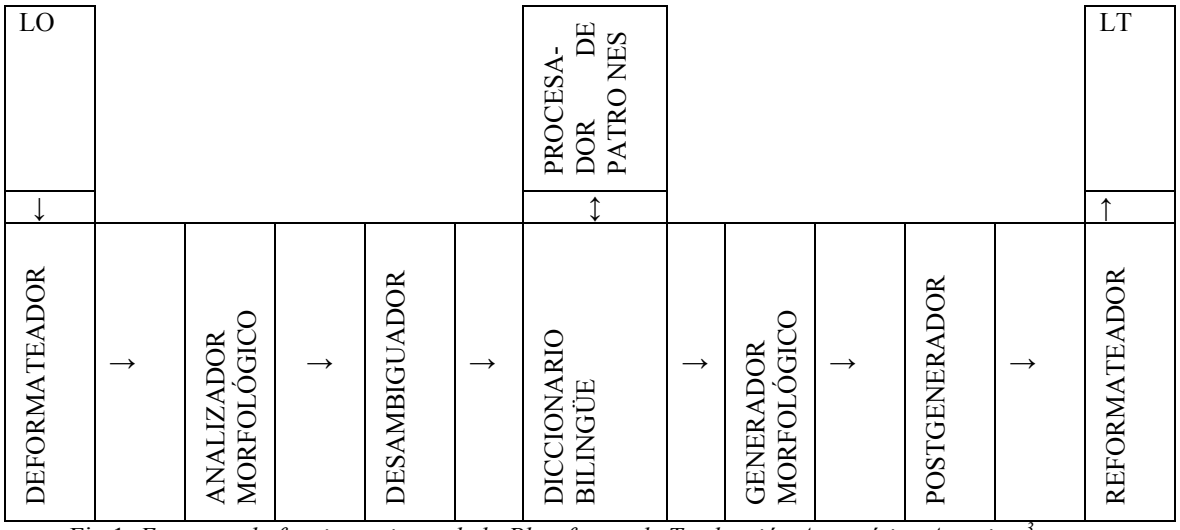

Fig.1. *Esquema de funcionamiento de la Plataforma de Traducción Automática Apertium*[3]

a. Deformateador - Reformateador

Selecciona todas las características de formato que acompañan al texto para traducir, las codifica entre corchetes transformándolas en información oculta. Esta información pasará a los siguientes módulos como espacios en blanco y no volverá a materializarse hasta finalizado el proceso de traducción cuando el último módulo, el reformatador se encargará de recuperar el formato inicial.

b. Analizador morfológico

Este módulo acciona a dos niveles:

a) fragmenta el texto para traducir en unidades FS (formas de superficie)
b) asigna a cada FS una forma léxica FL que contiene un lema (equivalente a la clásica entrada de diccionario), una categoría léxica y la información de flexión morfológica pertinente.

Evidentemente, si la FS remite a varias palabras homógrafas, se generará un número equivalente de FLs.

El analizador morfológico resuelve también los casos más complejos como la contracción y las multipalabras (expresiones y frases hechas, locuciones de toda clase, etc.) dentro del marco de un solo idioma.

c. Desambiguador léxico

El desambiguador léxico selecciona sólo uno de todos los homógrafos posibles, evidentemente el más adecuado al contexto. El entrenamiento se realiza sea a partir de un corpus extenso (más de un millón de palabras en LO) que pasa por el analizador morfológico, sea a partir de un corpus reducido (más de diez mil palabras) que es procesado manualmente por una persona que elige en función del contexto la FL pertinente de todos los homógrafos posibles y crea, de este modo, pautas de selección que puedan guiar el desambiguador.

d. Transferencia léxica

Este módulo depende estrictamente del diccionario bilingüe, puesto que su

función es sustituir cada forma superficial de la LO por su correspondiente forma superficial en la LT.

Esta transferencia constituye sólo la base de partida del proceso de traducción, ya que la meta del sistema Apertium no es una reproducción palabra por palabra, sino, en la medida de lo posible, una reelaboración eficaz y entendedora del texto meta.

e. Transferencia estructural

Este módulo contempla todos los tipos de cambios estructurales que se deben introducir cuando aparecen diferencias de índole gramatical entre las dos lenguas. Las más frecuentes diferencias atañen a:

· género[4] y/o número distinto entre LO y LT, lo que atrae también patrones de concordancia diferentes;
· orden diferente de palabras que implica un reordenamiento de palabras en la LT;
· naturaleza y/o uso diferente de artículos[5], determinantes y preposiciones;
· formas y estructuras gramaticales distintas que exigen una reconstrucción en la LT, etc.

f. Generador morfológico

Este módulo recibe las formas léxicas (LFs) de la LT con todas sus etiquetas gramaticales y procede a generar sus correspondientes formas superficiales (FSs).

g. Postgenerador

El postgenerador interviene tan sólo en situaciones particulares como las contracciones o apóstrofos (que se producen en la LT) previamente definidas y marcadas con una señal de alarma por la persona que ha desarrollado el diccionario monolingüe de la LT.

2.2. *Fase actual*

Actualmente, Trautorom ha concluido su primer objetivo: dispone de un volumen léxico de 10.000 palabras, con todo su séquito de formas flexionadas, y de 50 reglas de transferencia gramatical.

Trautorom podrá ser accedido a partir del mes de abril del 2007 a través de la página siguiente: http://www.xixona.dlsi.ua.es/pruebas/, donde cualquier persona interesada podrá, sin compromisos, consultar palabras o probar el funcionamiento actual del traductor a nivel de textos y documentos.

Teniendo en cuenta que el sistema está todavía en fase inicial, de prototipo, las traducciones de textos y documentos que resultarán de las pruebas no serán más que meros esbozos que exigirán un esfuerzo de corrección y postedición. Queda mucho trabajo por adelante para superar los mayores obstáculos que se presentarán en un apartado ulterior y para afinar la calidad de la traducción.

2.3. *Principales dificultades*

El primer inconveniente mayor es el hecho de que un motor de traducción automática sólo puede operar sin error, por lo menos en la fase actual, con equivalencias palabra por palabra. La polisemia, la homografía, los falsos amigos parciales son retos continuos y complicados para los que se proponen entrenar máquinas para traducir. Tanto más en el caso del rumano donde la homografía representa un fenómeno bastante frecuente que, unido a una sintaxis más flexible que en cualquier otro idioma románico, dificulta mucho la distinción de palabras por categorías gramaticales.

El segundo atañe a aspectos que reflejan las estructuras mentales subyacentes a la construcción del lenguaje. Estas estructuras, bastidas en contextos sociales y culturales distintos, influyen de una manera más o menos radical –en función del grado de proximidad o lejanía- en el diseño de la arquitectura singular de cada idioma. El rumano, aparte de haberse desarrollado en la periferia más lejana del núcleo románico, conservando por ende más estrictamente las raíces gramaticales latinas, ha sufrido también una importante influencia de las lenguas y culturas de su entorno, las eslavas sobre todo.

Otros impedimentos pueden surgir de aspectos de forma, como por ejemplo la grafía y su difusión a través de un soporte digital. En rumano hay dos problemas esenciales: las grafías particulares de *ă, ş* y *ţ* y la controversia sobre el uso de *î* y *â*.

El primer aspecto recuerda a los problemas que ha tenido el español con la letra *ñ*, sobre todo en Internet. Si se sustituyen esos diacríticos por los gráficamente cercanos *a*, *s* y *t* se abre paso a la ambigüedad o, aún más, a distorsiones de mensajes. Si se teclean según el gusto del usuario, muchas veces en Internet los diacríticos aparecen codificados como símbolos matemáticos truncando la información.

El segundo aspecto se refiere a la fuerte contracorriente académica que no acepta el cambio gráfico de *î* en *â* dictado por la Academia Rumana en 1993. Dada la gran variedad de autores y de opiniones en la literatura y en la prensa diaria este fenómeno ha llevado a la coexistencia de las dos grafías en libros y, aún más, en las páginas del mismo periódico o del mismo dominio electrónico.

El sistema de traducción automática necesita entrenarse sobre un corpus válido, "académico", y extenso. Y estas dos situaciones peculiares dificultan la creación de este corpus.

2.4. *Estadísticas*

Los sistemas de traducción automática se pueden someter a dos tipos fundamentales de evaluación.

El primero es de índole cuantitativo, cuando se persigue una estricta evaluación numérica: cantidad de errores, porcentaje de cobertura de léxico y coste de postedición.

El segundo, más complejo, es de naturaleza cualitativa, cuando se pone también un importante acento en la legibilidad del texto traducido, eso quiere decir la transmisión funcional –semántica y pragmática- del mensaje codificado en el texto origen al texto meta. Aparte de los medidores automáticos que ofrecen tan sólo cifras y porcentajes, hay además una persona cuya misión es clasificar los errores por categorías y frecuencia. Este tipo de evaluación, a parte de ecografiar el estado del sistema de traducción evaluado y sus resultados, facilita el proceso de corrección y mejora del sistema.

En la evaluación cualitativa se manejan dos macroindicadores fundamentales:

a) *la cobertura de los diccionarios* = la cantidad de palabras que el sistema de traducción identifica (o sea, "encuentra" en sus diccionarios) y traduce. La fórmula de cálculo de la cobertura es: "Número de palabras conocidas [x] 100 [/] Número total de palabras del texto"
b) *la tasa de error* = la cantidad de palabras que se deben de corregir en la postedición del texto traducido automáticamente. La fórmula de cálculo de la tasa de error es: "Número de palabras corregidas [x] 100 [/] Número total de palabras del texto"

Trautorom ha sido sometido a una evaluación cualitativa sobre un corpus de 10.674 palabras compilado de fuentes muy diversificadas. Su cobertura sería, según la fórmula anterior, de 88%: 9.394 x 100 / 10.674 = 88%. Si además, tenemos en cuenta que 278 de las palabras desconocidas son nombres propios de personas, animales-mascota o ciudades que no se traducen casi nunca, el porcentaje de cobertura sube con 2%, lo que sería, en total, 90%. La tasa de error, según la fórmula correspondiente, sería de 24%: 2536 x 100 / 10.674 = 23,75%.

2.5. *Estrategias de mejora*

La principal fuente de error en el caso de Trautorom es la homografía que se da en el idioma rumano con gran frecuencia. Para solucionar este impedimento es imprescindible entrenar el sistema de traducción sobre un corpus amplio que tenga sea una anotación manual de categorías gramaticales, sea una traducción menos estilística, pero correcta y respetando el original.

Aparte de la resolución de este aspecto, se deben de seguir introduciendo las reglas de transferencia estructural que faltan para crear una transposición correcta de ciertas estructuras gramaticales todavía no reflectadas en el sistema de traducción automática.

Otra estrategia de mejora es traducir intensivamente textos de documentos y páginas de Internet para detectar otros posibles errores y para ampliar el vocabulario introduciendo las palabras que aparecen marcadas como desconocidas.

3. PERSPECTIVAS DE FUTURO

El principal objetivo a corto plazo, después de eliminar los errores identificados en la última evaluación, es ampliar el vocabulario y las estructuras gramaticales para convertir el actual prototipo en un traductor automático completo y funcional.

A largo plazo, está previsto crear el sistema de traducción inversa, del español al rumano, adaptando debidamente el ya existente traductor del rumano al español. Asimismo, a partir del sistema de traducción rumano-español se pueden crear, reutilizando los datos y las herramientas ya construidos, otros sistemas similares del rumano al francés, inglés o cualquier otro idioma europeo.

Además, el desarrollo de un traductor no se limita exclusivamente a crear las herramientas materiales (los diccionarios y los datos lingüísticos) y funcionales (los programas informáticos específicos de procesamiento), sino que supone también un esfuerzo continuo de mantenimiento y mejora.

NOTAS

1. Para descargar los programas y archivos pertinentes, los usuarios deben consultar el dominio http://apertium.sourceforge.net/
2. Véase referencias número 1,2, 4 y 5 para consultar algunos artículos sobre el desarrollo de estos sistemas. Para consultas más detalladas o específicas, véase http://www.dlsi.ua.es/~mlf/publ_en.html y las referencias número 3 y 6 de este artículo.
3. Apud M.L. Forcada y otros, “El sistema de traducción automática castellano <-> catalán interNOSTRUM”, en http://www.internostrum.com/docum/iN_sepln2001.pdf
4. En rumano aparece también un tercer género, el neutro, que concuerda con sus determinantes como masculino en singular y femenino en plural
5. En rumano existen dos tipos de artículos más que en español: el artículo adjetival (cel, cea, cei, cele) y el artículo posesivo-genitival (al, a, ai, ale).

REFERENCIAS BIBLIOGRÁFICAS

Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., y Scalco, M. (2006) “Open-Source Portuguese-Spanish Machine Translation”.

[Documento de Internet disponible en http://transducens.dlsi.ua.es/repositori/transducens/pubs/213/armentano06.pdf]

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., and Sadler, L. (1994). *Machine translation: An introductory guide.* NCC Blackwell, Oxford. [Documento de Internet disponible en http://www.essex.ac.uk/linguistics/clmt/MTbook/PostScript/]

Canals-Marote, R., Esteve-Guillén, A., Garrido-Alenda, A., Guardiola-Savall, M.I., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Pérez-Antón P.M. y Forcada, M.L. (2000). "El sistema de traducción automática castellano <-> catalán interNOSTRUM", presentado en *Jornades sobre Productes de Suport a la Traducció* organizado por el Instituto Joan Lluís Vives [Documento de Internet disponible en http://www.internostrum.com/docum/iN_sepln2001.pdf]

Forcada, M.L. (2006) "Open-source machine translation: an opportunity for minor languages", en *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages).* [Documento de Internet disponible en http://www.dlsi.ua.es/~mlf/docum/forcada06p2.pdf]

Gilabert-Zarco, P., Herrero-Vicente, S., Ortiz-Rojas, S., Pertusa-Ibáñez, A., Ramírez-Sancho, G., Sánchez-Martínez, F., Samper-Asensio, M., Scalco, M. y Forcada, M. "Construcción rápida de un sistema de traducción automática español-portugués partiendo de un sistema español-catalán", en *Procesamiento del Lenguaje Natural*, XIX Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, Alcalá de Henares, España. [Documento de Internet disponible en http://www.dlsi.ua.es/~mlf/docum/gilabert03p.pdf]

Hutchins, W. John; and Harold L. Somers (1992). *An Introduction to Machine Translation.* Academic Press, London. [Documento de Internet disponible en http://ourworld.compuserve.com/homepages/WJHutchins/IntroMT-TOC.htm]