

*Herramientas para la construcción de sistemas de traducción automática: aplicación al par castellano-catalán*¹

**Raül Canals, Alicia Garrido, Maribel Guardiola*, Amaia Iturraspe, Sandra Montserrat*,
Hermínia Pastor i Mikel L. Forcada** (Departament de Llenguatges i Sistemes Informàtics/
*Departament de Filologia Catalana)

Resumen

La construcción de un sistema de traducción automática (STA) requiere la colaboración de lingüistas, traductores y informáticos. En el marco de un proyecto de investigación sobre traducción automática castellano-catalán, hemos aplicado esta colaboración eficientemente: los lingüistas y traductores escriben, en archivos de texto, los datos y las reglas y los informáticos, en lugar de programar los subprogramas del STA, escriben programas compiladores que, a partir de los datos lingüísticos, generan estos subprogramas. Esta distribución permite: a) la generación automática y rápida de un sistema actualizado completo cada vez que se realicen mejoras en los datos lingüísticos, y b) que la experiencia informática se concentre en mejorar el funcionamiento no lingüístico.

Hemos construido estas herramientas: (1) compilador de analizadores morfológicos a partir de diccionarios morfológicos (léxico+paradigmas de flexión); (2) compilador de programas de consulta de diccionarios bilingües a partir del vocabulario bilingüe; (3) compilador de generadores morfológicos a partir de diccionarios morfológicos; (4) compilador de módulos de tratamiento sintáctico a partir de un archivo de reglas, y (5) desambiguador de categorías léxicas entrenado automáticamente a partir de un corpus parcialmente etiquetado. Todas producen programas basados en transductores de estados finitos.

En el trabajo describimos los formatos de los ficheros (datos lingüísticos), los compiladores que los traducen a subprogramas y su aplicación a la construcción del STA castellano-catalán.

1. Introducción

La construcción de un sistema de traducción automática (TA) requiere la colaboración de personas expertas en lingüística y en traducción, por un lado, y de personas expertas en informática, por otro. En el marco de un proyecto financiado por la Caja de Ahorros del Mediterráneo para producir un sistema de TA del castellano al catalán, se ha optado por un modelo de colaboración que ha resultado especialmente eficiente: las personas expertas en lengua y traducción escriben, en archivos de texto de formato convenido y sencillo, los datos (diccionarios monolingües de la lengua origen y meta, diccionarios bilingües) y las reglas (paradigmas de flexión morfológica, reglas sintácticas) y las personas expertas en informática, en lugar de escribir ellos mismos los subprogramas del sistema de traducción automática (analizador morfológico, gestor del diccionario bilingüe, subprograma de tratamiento sintáctico, generador morfológico), escriben programas compiladores que, a partir de los datos lingüísticos, generan automáticamente los subprogramas citados del sistema de TA. Esta división del trabajo (Heyer y Waldhör 1995) permite, por un lado, que cada vez que se realicen mejoras en diccionarios o bancos de reglas se pueda generar automáticamente y en pocos minutos un

¹ Trabajo financiado por la Caja de Ahorros del Mediterráneo.

sistema actualizado completo, y por otro, que la experiencia informática se concentre en mejoras del funcionamiento no lingüístico del sistema (por ejemplo, la velocidad o la comodidad de uso).

En la actualidad, se dispone de las siguientes herramientas:

1. Un compilador de analizadores morfológicos a partir de diccionarios morfológicos (que contienen léxico y paradigmas de flexión).
2. Un compilador de programas de consulta de diccionarios bilingües a partir del vocabulario bilingüe.
3. Un compilador de generadores morfológicos a partir de diccionarios morfológicos, y
4. Un compilador de módulos de tratamiento sintáctico a partir de un archivo de reglas.
5. Un desambiguador de categorías léxicas entrenado automáticamente a partir de un texto parcialmente analizado por el analizador morfológico.

Todas las herramientas, menos la última, producen programas que utilizan autómatas de estados finitos: los productos de las tres primeras procesan texto a velocidades de decenas de miles de palabras por segundo; los de la cuarta son algo más lentos. En este trabajo se describen los formatos de los ficheros de datos lingüísticos, las herramientas que los traducen a subprogramas ejecutables del sistema de TA y su aplicación a la construcción de un sistema de TA castellano-catalán.

2. *Arquitectura del sistema de traducción automática*

2.1. *Traducción automática por transferencia*

La *traducción automática* (TA) es la traducción (normalmente aproximada) de textos informatizados de una lengua a otra por parte de un programa adecuado que se ejecuta sobre un ordenador. El programa y el ordenador forman el *sistema de traducción automática*. Una de las aproximaciones clásicas a la traducción automática la constituyen los denominados sistemas de *TA indirecta por transferencia* (Arnold 1993; Arnold et al. 1994; Hutchins y Somers 1995; Vandooren 1993). En estos sistemas, la tarea de traducción se divide en tres fases bien diferenciadas:

1. *Análisis*: en esta fase, el texto origen (TO) se analiza de manera que se genera a partir de él una representación abstracta (RATO) adecuada para su traducción a alguna lengua meta (LM). La fase de análisis se diseña de manera que sólo precisa de información sobre la lengua origen (LO).
2. *Transferencia*: esta es la fase verdaderamente bilingüe del sistema; en ella, la RATO se transforma en una representación abstracta del texto meta (RATM) mediante el uso de reglas y de un diccionario bilingüe.
3. *Generación*: esta tercera fase es básicamente la inversa del análisis pero para la lengua meta; en ella, la RATM se transforma en texto meta (TM). La fase está diseñada de manera que sólo precisa de información sobre la LM.

Los sistemas de TA por transferencia se clasifican de acuerdo con la profundidad del análisis: así, se habla de sistemas de transferencia *morfológica*, *sintáctica*, y *semántica*. Cuando el análisis es tan profundo que la RATO y la RATM son idénticas e independientes de la lengua y por tanto no se precisa de fase de transferencia, se habla de sistemas de *interlingua*.

2.2. *Sistemas de transferencia morfológica avanzada*

La mayor parte de los sistemas de TA comerciales (disponibles para ordenadores personales a bajo precio –entre 30 y 300 euros– y que traducen entre el inglés y otras lenguas como el español (Mira-Giménez y Forcada 1998)) se pueden clasificar como sistemas de *transferencia morfológica avanzada*; es decir, realizan un análisis morfológico y algunas operaciones que podríamos considerar como un análisis sintáctico parcial para resolver algunos

problemas de la transferencia morfológica pura entre los que destacan la *ambigüedad léxica categorial* (que presentan palabras como *vino* que pueden pertenecer a más de una categoría morfológica), la *concordancia de género y número* (cuando estos no existen en la LO o son diferentes en la LM y en la LO para algunas palabras) y *el orden de las palabras*, cuando varía de una lengua a otra.

Las herramientas que se presentan aquí sirven para generar sistemas de este tipo, en los que:

- el subprograma de análisis realiza un análisis morfológico de todas las palabras (*formas superficiales*) del TO (es decir, las convierte en *formas léxicas* que contienen un *lema* o *forma canónica*, información sobre la categoría léxica e información sobre la flexión) y elige, en el caso de los *homógrafos*, una de las formas léxicas usando un modelo estadístico sencillo que considera las probabilidades de aparición de secuencias de dos o tres categorías morfológicas en corpus representativos de la LO (elección que se puede considerar que se basa en un análisis sintáctico parcial);
- el subprograma de transferencia sustituye el lema en LO por el lema en LM consultando un diccionario bilingüe, y, en general, construye una forma léxica añadiendo a este lema la información sobre la flexión proporcionada por el subprograma de análisis y envía esta forma léxica al subprograma de generación, excepto cuando detecta una secuencia de palabras (un remedo de sintagma) que, por las categorías morfológicas que la componen, necesita de un tratamiento especial (flexión diferente de la original por concordancia, cambio del orden de palabras) antes de ser enviadas al subprograma de generación;
- el subprograma de generación sencillamente genera la forma superficial correspondiente a partir de las formas léxicas que envía el subprograma de transferencia.

2.3. Tareas básicas

Como puede verse, en este modelo se detectan cinco tareas básicas, cada una de las cuales se basa en información lingüística detallada.

1. análisis morfológico de las palabras del texto original, usando la siguiente información sobre la LO: un diccionario de lemas, una descripción de los paradigmas de flexión y la relación entre lemas y paradigmas. Si una palabra no está recogida en el diccionario, el analizador morfológico la marcará con un asterisco para indicar que es desconocida para el sistema;
2. desambiguación categorial de los homógrafos que se encuentren, usando información estadística sobre la aparición conjunta de categorías léxicas en textos del idioma (en secuencias de dos o tres palabras);
3. consulta del diccionario bilingüe, guiada por el lema y la categoría gramatical elegida, usando un diccionario de correspondencia de lemas en LO y lemas en LM;
4. detección y tratamiento de secuencias de palabras que constituyen ámbitos de concordancia o que deben ser reordenadas, usando reglas adecuadas.
5. generación de las formas superficiales que constituyen el texto meta, usando información sobre la LM análoga a la que usa el análisis morfológico sobre la LO.

Cada una de las tareas transforma la información producida por la anterior y por tanto, se puede considerar que trabajan secuencialmente; sin embargo, no es necesario que una tarea haya procesado el texto completo antes de que opere la siguiente, ya que los resultados parciales de cada tarea están disponibles muy a menudo y pueden servir a las siguientes.

2.4. Separación de métodos y datos: herramientas

Todas estas tareas las realizan módulos o subprogramas de ordenador en los cuales se puede establecer la separación efectiva clásica entre:

- los *métodos* o *algoritmos* (la mecánica general de cada tarea, que no depende de la lengua o lenguas implicadas), y

- los *datos* (la información lingüística concreta pertinente a la tarea sobre la lengua o lenguas implicadas, codificada de manera que pueda ser utilizada por el programa).

Esta separación permite la escritura de herramientas informáticas que construyen estos subprogramas a partir de información lingüística codificada de forma que sea a la vez comprensible para una persona y tratable por un programa de ordenador; estas herramientas transforman esta información lingüística en datos que pueden ser usados directamente por los métodos correspondientes.

3. Generación automática de los subprogramas del sistema a partir de los datos lingüísticos

En esta sección se describen las herramientas que se usan para generar cada uno de los subprogramas del sistema de traducción automática, el formato de los archivos de datos que leen y la naturaleza y el funcionamiento de los módulos que generan.

3.1. Compilador de analizadores morfológicos

Como ya se ha dicho, este compilador (Garrido et al. 1999) es un programa que genera automáticamente, a partir de un fichero que contiene un diccionario morfológico, un programa analizador morfológico muy eficiente. Se dan detalles técnicos de este programa en otra comunicación presentada en este mismo congreso (Garrido et al. 2000). El fichero que contiene el diccionario morfológico contiene tres secciones bien diferenciadas:

- la sección donde se declaran los símbolos que se usarán para representar las categorías léxicas (<n>, <adj>, etc.) y los rasgos flexivos (<m>, <sg>, <pp>, etc.);
- la sección donde se declaran los paradigmas de flexión, de los cuales un ejemplo sencillo es el siguiente:

```
[gen_num]>(a:<f><sg>)  
| (as:<f><pl>)  
| (o:<m><sg>)  
| (os:<m><pl>);
```

en el que se define el análisis morfológico de las desinencias de sustantivos y adjetivos;

- la sección de diccionario, donde se relacionan las formas superficiales con las formas léxicas correspondientes, usando los paradigmas de flexión definidos o, en el caso de irregularidades muy específicas, dando todas las formas. Un ejemplo de la utilización del paradigma anterior en el diccionario podría ser la entrada siguiente:

```
(alumn:alumno<n>)[gen_num];
```

3.2 Entrenamiento del desambiguador léxico

La desambiguación léxica categorial se realiza usando técnicas estadísticas que consideran secuencias de categorías léxicas. La probabilidad de una determinada secuencia de categorías léxicas se calcula, de manera aproximada, como el producto de las probabilidades de todas las secuencias de tres categorías léxicas que contiene (lo que constituye un modelo de trigramas, Jelinek 1998). Así, la probabilidad de la secuencia *art-n-adj-vb* (como la que se observaría en la frase "el muchacho inglés canta") se aproximaría como el siguiente producto de probabilidades:

$$P(\text{art}/\# \#) \times P(\text{n}/\# \text{art}) \times P(\text{adj}/\text{art n}) \times P(\text{vb}/\text{n adj}) \times P(\#/\text{adj vb})$$

donde $P(c/a b)$ representa la probabilidad de observar la categoría c después de las categorías a y b , y $\#$ representa la frontera entre oraciones. Estas probabilidades se aproximan a partir de las frecuencias observadas en un corpus suficientemente representativo de textos de la lengua.

Cuando una oración contiene homógrafos, p.ej. "el ahorro doméstico conviene", se calcula la probabilidad de cada desambiguación posible de la oración, en el ejemplo *art-n-adj-vrb* o *art-vrb-adj-vrb*, y se elige la desambiguación más probable (pesándola convenientemente en cada caso con parámetros estadísticos como la probabilidad $P(\{n, vrb\}, n)$ de que un sustantivo n se manifieste como un homógrafo sustantivo-verbo $\{n, vrb\}$.) El cálculo de la desambiguación ganadora se puede organizar de manera muy eficiente, de manera que si la palabra p del texto es homógrafa, se habrá desambiguado cuando se haya procesado la palabra $p+2$.

El programa entrenador usa el analizador morfológico para etiquetar un corpus (de un millón de palabras en nuestros experimentos), recoge las estadísticas observadas en las partes no ambiguas del mismo, estima las probabilidades mencionadas, y las escribe en un archivo, listas para ser usadas por un subprograma etiquetador genérico.

3.3. *Compilador de diccionarios bilingües*

El compilador de diccionarios bilingües no es excesivamente diferente del de analizadores morfológicos, sólo que en este caso lee una lista de correspondencias entre lemas origen y lemas meta (un diccionario bilingüe) y genera un programa muy eficiente que lee un lema origen y entrega el lema meta correspondiente. En aquellos casos en los que se produce cambio de número o género gramatical de una lengua a otra, esta circunstancia se puede hacer constar de manera muy sencilla en el diccionario bilingüe; por ejemplo, *señal* es femenino en castellano pero su traducción al catalán, *senyal*, es masculina, por tanto la entrada que habrá en el diccionario bilingüe para este caso será:

(señal<n><f>:senyal<n><m>);

El programa que gestiona el diccionario bilingüe es invocado por el subprograma de transferencia.

3.4. *Compilador del subprograma de transferencia*

El subprograma de transferencia detecta determinadas secuencias o patrones de categorías léxicas (por ejemplo, *art-n-adj*, *una señal inequívoca*), invoca al diccionario bilingüe, para obtener su traducción en la que constará si se ha producido algún cambio significativo en el núcleo del sintagma, propaga el género y el número del núcleo a los modificadores si es diferente en la lengua meta y la lengua origen, y genera la misma secuencia de categorías léxicas o una secuencia reordenada, si fuera necesario.

En el ejemplo *una señal inequívoca*, el núcleo (sustantivo *señal*) cambia de género al traducirlo al catalán, por tanto la regla correspondiente propagará el género del núcleo a los modificadores (artículo *una* y adjetivo *inequívoca*), generando la secuencia *un senyal inequívoc*.

Este módulo lee el texto analizado de izquierda a derecha, intenta siempre detectar la secuencia de categorías más larga posible, opera sobre ella, y continúa inmediatamente detrás, sin operar dos veces sobre la misma palabra.

Un compilador genera automáticamente este módulo a partir de un archivo que contiene reglas de la forma *patrón-acción* que indican para cada secuencia detectada (patrón) las operaciones (acción) que deben realizarse para construir las formas léxicas de su traducción y el orden en que deben escribirse, que puede ser diferente del original.

Un ejemplo sencillo de regla sería la que contempla la concordancia de género y número entre el artículo y sustantivo:

REGLA articulo sustantivo

```
{
  SI (sustantivo.meta.genero = "<No_definido>")
      sustantivo.meta.genero ← articulo.meta.genero
  EN_OTRO_CASO
  {
    SI (sustantivo.meta.numero = "<No_definido>")
        sustantivo.meta.numero ← articulo.meta.numero
    EN_OTRO_CASO
    {
      SI (sustantivo.meta.genero ≠ articulo.meta.genero)
          articulo.meta.genero ← sustantivo.meta.genero
      SI (sustantivo.meta.numero ≠ articulo.meta.numero)
          articulo.meta.numero ← sustantivo.meta.numero
    }
  }
}
```

Esta regla tiene la siguiente interpretación:

- Si el sustantivo tiene en la lengua origen una única forma para el masculino y el femenino pero formas diferenciadas en la lengua meta (por ejemplo, castellano *estudiante*, catalán *estudiant/estudianta*), se propaga el género del artículo al sustantivo.
- Si el sustantivo tiene en la lengua origen una única forma para el singular y el plural pero formas diferenciadas en la lengua meta (por ejemplo, castellano *análisis*, catalán *anàlisi/anàlisis*), se propaga el número del artículo al sustantivo.
- En los demás casos, el género y el número del sustantivo en la lengua meta están definidos; si son distintos al género y el número en la lengua origen, se propagan al artículo.

3.5. *Compilador del generador morfológico*

El generador morfológico genera formas superficiales en lengua meta a partir de las formas léxicas producidas por el subprograma anterior; en este sentido, es básicamente el inverso de un analizador morfológico, pero para la lengua meta.

Un compilador muy similar al que se usa para generar el analizador morfológico lee un diccionario morfológico de la lengua meta, en el que se especifican los lemas y los paradigmas que sirven para generar las formas superficiales correspondientes. Siguiendo con el ejemplo expuesto para el compilador del analizador morfológico, en este diccionario se incluiría el paradigma inverso para el catalán:

```
[gen_num]>( <f><sg>:a
            | <f><pl>:es
            | <m><sg>:e
            | <m><pl>:es );
```

el cual genera a partir del análisis morfológico de las desinencias la correspondiente forma superficial. Igualmente la entrada en el diccionario utilizando este paradigma, sería la inversa:

```
(alumne<n>:alumn)[gen_num];
```

El compilador genera, a partir de esta información, un subprograma de ordenador que realiza la generación morfológica.

El generador morfológico se complementa con un postgenerador, compilado a partir de un archivo de reglas muy sencillo, que realiza las operaciones necesarias para producir las formas apostrofadas de artículos, preposiciones y pronombres proclíticos.

4. Conclusiones

En este artículo se ha presentado un sistema de traducción automática por transferencia, detallando cómo se construye automáticamente cada uno de sus subprogramas a partir de ficheros con datos lingüísticos, usando herramientas informáticas apropiadas. Estas herramientas son:

1. *Compilador de analizadores morfológicos*: A partir de un diccionario, confeccionado por personas expertas en lengua, se genera un programa que analiza morfológicamente textos a una velocidad de unas 10.000 palabras por segundo.
2. *Compilador de diccionarios bilingües*: Utilizando la tecnología anterior, genera un programa que traduce un texto, en la representación abstracta producida por el analizador morfológico, a una velocidad aproximada de 10.000 palabras por segundo.
3. *Compilador del subprograma de transferencia*: A partir de un fichero que contiene reglas, realizadas por lingüistas, genera un programa que reconoce patrones sintácticos sobre el texto (en la representación que proporciona el analizador morfológico), y utilizando el programa generado por el diccionario bilingüe, traduce el fragmento de texto reconocido por el patrón y opera sobre él según indique la regla correspondiente. Por tanto, este programa traduce un texto (en representación abstracta), aplicando sobre él análisis sintáctico parcial, a una velocidad aproximada de 5.000 palabras por segundo.
4. *Compilador de generadores morfológicos*: Utiliza la misma tecnología que el compilador de analizadores morfológicos para generar un programa, que a partir de un texto, en la representación abstracta que proporciona el subprograma de transferencia, obtiene un texto legible. Es decir, a partir de cada forma léxica meta, genera la forma superficial de la palabra. La velocidad de proceso de este programa es de 10.000 palabras por segundo aproximadamente.
5. *Desambiguador de categorías léxicas*: El programa entrenador usa el analizador morfológico para etiquetar un corpus y recoge las estadísticas observadas en las partes no ambiguas del mismo para ser usadas por un subprograma etiquetador genérico.

Las velocidades de proceso que se mencionan son sobre un ordenador (o procesador) Pentium III a 400MHz.

Todos estos subprogramas pueden utilizarse de forma independiente, para obtener análisis intermedios del texto, o conjuntamente para obtener la traducción del mismo. La velocidad de traducción del sistema completo es de 5.000 palabras por segundo aproximadamente.

Referencias bibliográficas

- Arnold, D (1993): "Sur la conception du transfert", Bouillon, P., Clas, A. (eds.), *La traductique*, Montreal: Presses Univ. Montréal, pp. 64-76.
- Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L., Sandler, L. (1994): *Machine Translation: an Introductory Guide*, Oxford: NCC Blackwell.
- Garrido, A., Iturraspe, A., Montserrat, S., Pastor, H., Forcada, M.L. (1999): "A compiler for morphological analysers and generators based on finite-state transducers", XV Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, pp. 93-98.

Garrido, A., Iturraspe, A., Montserrat, S., Forcada, M.L. (2000): "Generación automática de lematizadores de textos catalanes antiguos basados en técnicas de estados finitos", IV Congreso de Lingüística General.

Jelinek, F. (1998): *Statistical Methods for speech recognition*, Cambridge, Massachusetts: MIT Press.

Mira-Giménez, M., Forcada, M.L. (1998): "Understanding PC-based machine translation systems for evaluation, teaching, and reverse engineering: the treatment of noun phrases in Power Translator", *Machine Translation Review* 7.

Vandooren, F. (1993): "Divergences de traduction et architectures de transfert", Bouillon, P., Clas, A. (eds.), *La traductique*, Montreal: Presses Univ. Montréal.

Heyer, G., Waldhör, K. (1995): "General Language Resources: Lexica", Kugler, M., Ahmad, K., Thurmair, G. (eds.), *Translator's Workbench*, Berlin: Springer-Verlag, pp. 40-48.