PERSPECTIVAS DE LA TRADUCCIÓN AUTOMÁTICA CASTELLANO-GALLEGO MEDIANTE TÉCNICAS ESTADÍSTICAS Y POR TRANSFERENCIA

Gonzalo Iglesias Iglesias[1], *Leandro Rodríguez Liñares*[2], *Eduardo Rodríguez Banga*[3], *Francisco León Campillo Díaz* [3] *y Francisco Méndez Pazó*[3]

Centro Ramón Piñeiro para la Investigación en Humanidades (Santiago), E.T.S. Informática de Ourense (Universidad de Vigo), E.T.S. Telecomunicaciones (Universidad de Vigo)

RESUMEN

En esta comunicación presentamos nuestros trabajos en traducción automática entre las lenguas gallega y castellana. Los esfuerzos se centran en las dos principales líneas en que se divide este campo: traducción estadística y por transferencia.

1 INTRODUCCIÓN

En los últimos años, el avance de las tecnologías informáticas ha permitido la proliferación de productos basados en tecnologías del habla y del lenguaje natural. Entre las herramientas más útiles y revolucionarias que han aparecido figuran los traductores automáticos, que juegan un papel fundamental en la nueva forma de comunicarse que ha supuesto la aparición de Internet.

Especialmente importantes son, en este campo, los esfuerzos invertidos en el desarrollo de esta tecnología para las lenguas minoritarias, ya que amplía su visibilidad y constituye una garantía para su supervivencia.

En España podemos encontrar iniciativas de traducción automática. Podemos destacar, entre otros, productos ya maduros como el *OpenTrad Apertium* [3] y *Marie* [5].

Una buena introducción a la traducción automática puede encontrarse en [7]. En este artículo nos centraremos en la *traducción basada en reglas* o por transferencia y la *traducción estadística*.

Los traductores por transferencia se basan en la idea de que un conocimiento lingüístico lo suficientemente extenso de las dos lenguas involucradas debería permitir, mediante reglas, una traducción correcta.

La transferencia entre las estructuras de los lenguajes de origen y de destino puede tener lugar a niveles muy distintos de análisis lingüístico. El triángulo de Vauquois ilustra esta idea. En la base de este triángulo se situaría la denominada traducción directa (traducción palabra a palabra sin información lingüística contextual). En el vértice superior del triángulo tendríamos una representación conceptual (sintáctico, semántico, pragmático) de los idiomas, denominada interlingua. Entre estos dos extremos se encuentra la traducción por transferencia a distintos niveles, según la información empleada para crear las reglas entre lengua origen y lengua destino.

La *traducción estadística* ha cobrado auge desde la década de los noventa, a partir de los trabajos en el centro T.J. Watson de IBM ([1][2]). El modelo de traducción estadística más utilizado se basa en una analogía con un sistema de comunicaciones en que un mensaje codificado en la lengua destino se transmite a través de un canal ruidoso, con lo que se obtiene la versión distorsionada en el lenguaje origen. De acuerdo con este modelo, un sistema de traducción entre los lenguajes origen y destino equivale a decidir en base al mensaje recibido cuál es la secuencia de palabras transmitidas que con mayor probabilidad generaron dicho mensaje.

En esta comunicación presentamos dos traductores de castellano a gallego. Uno, basado en reglas; el segundo es un traductor estadístico. Los dos están aún en fase de desarrollo. El artículo se organiza de la siguiente manera: en la sección 2 describiremos el traductor basado en reglas; en la sección 3, se describe nuestro decodificador estadístico, inspirado en *Marie* [5]. Posteriormente, presentaremos algunos resultados relevantes y líneas futuras de trabajo.

2 TRADUCTOR AUTOMÁTICO BASADO EN REGLAS

La figura 1 representa las distintas etapas del traductor basado en reglas en el que el análisis se realiza frase a frase. En primer lugar, a cada frase del idioma origen se le aplica un preprocesado, con objeto de obtener el texto a traducir sin problemas de formato.

A continuación, en la segunda etapa se realiza un análisis de la frase para obtener las categorías morfosintácticas de cada palabra.

El análisis morfosintáctico mediante reglas y diccionarios no produce resultados unívocos. De ahí que sean necesarios criterios de desambiguación. Para este propósito se usan reglas lingüísticas fiables y un decisor estadístico basado en un modelo contextual y un modelo léxico.

El análisis morfosintáctico integra la lematización, que se puede entender como un proceso de abstracción de las variantes morfemáticas de la palabra mediante ciertos criterios o convenios. Así, el lema deducido es otra palabra que mantiene el valor semántico.

Una vez obtenido el lema del lenguaje origen se realiza la traducción al correspondiente lema del idioma

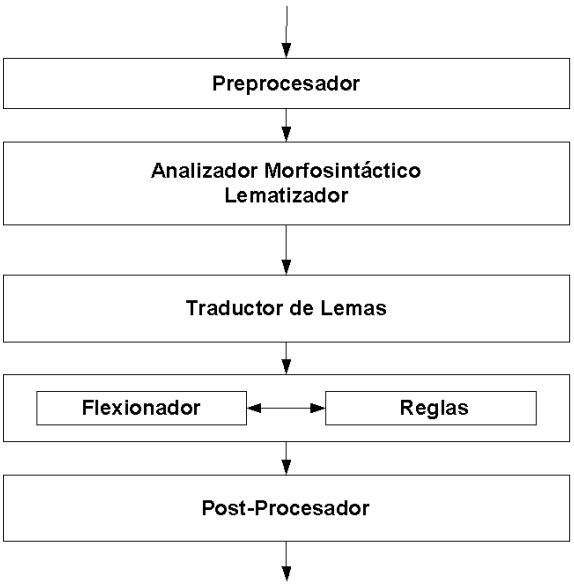

Figura 1. *Arquitectura del traductor.*

destino con la ayuda de un diccionario de traducción.

Finalmente el lema traducido es flexionado de acuerdo a características como género y número, por ejemplo, si se trata de un sustantivo, adjetivo o determinante. El proceso de flexión está supervisado por el módulo de reglas, que permite realizar además las transformaciones estructurales necesarias en la frase.

La traducción finaliza con un post-procesado, que devuelve el texto traducido en el formato inicial.

A continuación detallamos más detenidamente el funcionamiento de estos módulos.

2.1 El analizador morfosintáctico

Se trata de una etapa clave, ya que determina en gran medida el correcto funcionamiento del sistema. El analizador morfosintáctico empleado en este caso para el gallego y el castellano fue el desarrollado para nuestro conversor texto-voz Cotovía [10]. En la figura 3 se describe su estructura. En primer lugar, se asigna a cada palabra de la frase el conjunto de categorías que le pueden corresponder, por medio de diccionarios y reglas morfológicas sencillas.

A continuación, se reduce la ambigüedad aplicando un pequeño conjunto de reglas lingüísticas altamente fiables que solucionan los casos más sencillos. Posteriormente, se traducen las categorías a un conjunto de etiquetas más sencillo, y sobre ellas se emplea un modelo estadístico que combina la información resultante de una parte contextual y otra léxica para encontrar la secuencia de categorías morfosintácticas más probable para la frase de entrada. Finalmente se traducen las categorías resultantes a sus correspondientes en el conjunto de etiquetas original.

En los siguientes apartados se explican con más detalle las partes más importantes de dicho analizador.

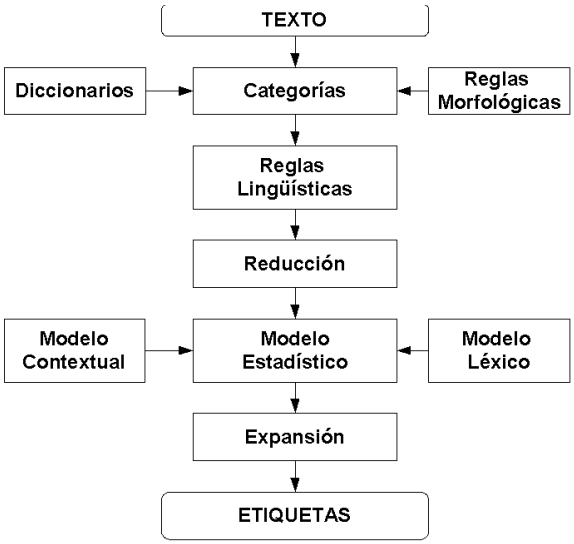

Figura 2. *Arquitectura del analizador morfosintáctico.*

2.1.1 El modelo contextual

El modelo contextual busca la categoría más probable para una palabra en función de las categorías de las palabras que la rodean. En este caso se recurre a la utilización de n-gramas de categorías gramaticales (n=5), determinando la categoría de cada palabra mediante el n-grama más probable cuando éste se halla centrado en dicha palabra.

2.1.2 El modelo léxico

El modelo léxico considera la probabilidad de que cada palabra concreta pertenezca a una categoría gramatical, independientemente del contexto que la rodee. Para calcular dichas probabilidades de forma fiable se necesitaría un corpus de entrenamiento excesivamente grande, por lo que se optó por el uso de *clases de ambigüedad* [6], ya que han demostrado ser una aproximación razonable [10] cuando no se dispone de suficiente texto etiquetado.

Esta estrategia consiste en agrupar palabras del corpus de entrenamiento en función del conjunto de

categorías que, a priori, se les pueden asignar. Por ejemplo, una clase de ambigüedad podría ser artículo-preposición-nombre. Dicha clasificación reduce de forma considerable el tamaño del espacio considerado, lo que supone que la estimación de las probabilidades sea más fiable.

2.1.3 El conjunto reducido de etiquetas

El número de categorías consideradas influye notablemente en el tamaño del corpus etiquetado que es preciso para obtener una estimación fiable de las probabilidades de los modelos contextual y léxico. Por otra parte, a medida que se vayan considerando menos categorías, el análisis generado irá perdiendo en poder descriptivo. La solución adoptada en este caso para evitar dicho problema consiste en considerar dos conjuntos diferentes de etiquetas: uno principal, y otro reducido.

El conjunto de etiquetas principal es más descriptivo e incluye información sobre género, número, persona, tiempos verbales, tipos de adverbios, tipos de determinantes..., mientras que el conjunto reducido agrupa las categorías del principal según su comportamiento gramatical en la oración (por ejemplo, artículos determinados e indeterminados se incluyen en la misma clase), respetando la información de género y número en aquellos casos en que tenga sentido.

De esta manera, el grupo de categorías principal puede ser tan amplio y descriptivo como se desee, debido a que en el modelo estadístico se construye sobre el conjunto de categorías reducido, que consta únicamente de 49 etiquetas [10]. Se requiere así de menor cantidad de texto etiquetado para el entrenamiento, lo que permite trabajar con un modelo de lenguaje más pequeño y manejable sin afectar a la capacidad descriptiva del análisis realizado.

2.1.4 El lematizador

El analizador morfosintáctico integra tanto en castellano como en gallego un lematizador. Antes de aplicarse el modelo estadístico, el problema de análisis y lematización se aborda por dos vías: mediante un diccionario de palabras y mediante un diccionario de derivaciones.

2.1.4.1 Diccionario de palabras.

Los diccionarios de palabras del analizador contienen información sobre el lema, directa o indirectamente, junto con sus posibles categorías. Por ejemplo, a un sustantivo como *actriz* el diccionario le asigna directamente el lema *actor*. Por otra parte, cualquier verbo constituye un buen ejemplo de búsqueda indirecta. Así, si en el análisis nos encontramos con la palabra *comieras*, se ha de determinar primero si esta palabra puede ser verbo, eliminando las desinencias y pronombres enclíticos, si los hay, y buscando la raíz *com-* en la lista de raíces verbales. Además, de nuestro diccionario de modelos de conjugación se obtiene que esta raíz está relacionada con el verbo *comer* mediante el modelo regular de la segunda conjugación. Un diccionario de desinencias proporciona la relación entre *-ieras* y dicho modelo regular y, como resultado, la categoría buscada: segunda persona singular del pretérito de subjuntivo.

2.1.4.2 Diccionario de derivaciones.

En caso de que no se pueda analizar y lematizar mediante los diccionarios de palabras, se acude al diccionario de derivaciones, que contiene información sobre los morfemas que permiten derivar una palabra de otra y cubre las ausencias de los diccionarios previos Así, se puede determinar típicamente si una palabra es un diminutivo, un aumentativo o un peyorativo; si la palabra en cuestión es un adverbio terminado en *-mente* o si se trata de una forma sustantivada de un verbo, etcétera.

El empleo del diccionario de derivaciones implica también la utilización del diccionario de palabras. Es decir: para que se considere una forma derivada debe verificarse la existencia de la palabra de la que procede.

2.1.5 Resultados

En [10] se realiza una descripción pormenorizada de resultados. Como muestra, en la tabla 1 presentamos los resultados del análisis sobre un texto en gallego. Algunas características de este texto se resumen en la tabla 2. La columna *ambiguas* se refiere a palabras con inicialmente más de dos etiquetas de categorías gramaticales asignadas; *etiquetas* representa el número medio de etiquetas por palabra; por último, *etiquetas ambiguas* representa el número medio de etiquetas en palabras donde inicialmente hay ambigüedad.

	Categoría	***Género***	***Número***
Texto	97.42%	97.96%	98.90%

Tabla 1. *Tasas de acierto del análisis.*

	Palabras	***Ambiguas***	***Etiquetas***	***Etiquetas Ambiguas***
Texto	24321	7528	1.68	3.08

Tabla 2. *Complejidad del texto de prueba.*

Por lo que respecta al castellano, todavía no se ha realizado una evaluación formal ya que el corpus disponible no es aún lo bastante amplio. En breve dispondremos de un corpus con una cobertura adecuada

y esperamos que los resultados serán similares a los obtenidos para gallego. Evidentemente, que el analizador morfosintáctico en castellano no sea aún todo lo preciso que debiera tiene una incidencia negativa en el traductor.

2.2 Diccionario de traducción

El traductor consta de un diccionario de lemas en sentido castellano → gallego (unas 20.000 entradas).

2.3 Flexionador

Se basa en los mismos diccionarios del analizador, optimizados adecuadamente para esta tarea. La flexión permite, a partir de un lema, generar la o las palabras que cumplan las características deseadas como pueden ser género, número, persona o tiempo verbal.

2.4 Módulo de reglas

El módulo de reglas procesa las palabras teniendo en cuenta la información morfosintáctica del entorno. De esta manera se pueden realizar las transformaciones necesarias para el idioma destino, en este caso el gallego. Este módulo controla también el flexionador; siempre a través de reglas en un diccionario.

La notación es bastante intuitiva y flexible. Cada regla consta de dos campos obligatorios: *condición* y *acción.* La acción tiene lugar si se verifica la condición, que puede contener uno o más elementos separados por espacios. Cada elemento de la condición puede ser una palabra, una categoría o un lema con una categoría. La *acción* indica la transformación que sufren las palabras involucradas. Una regla para transformar el antepretérito de subjuntivo castellano en pretérito de subjuntivo en gallego podría formularse de la siguiente manera (condición y acción se separan con una coma):

haber/VPS VT0SM,%1/VPS**

Esta regla busca el verbo *haber* en la forma pretérito de subjuntivo (*VPS*). La regla especifica que a continuación del verbo tiene que existir un participio (*VTOSM*). De cumplirse la condición, se sustituye con el lema del participio (referenciado con *%1*), flexionado en este caso a pretérito de subjuntivo en gallego.

De esta forma, el módulo de reglas recorre cada palabra de la frase y decide si se dan las condiciones en cada caso para aplicar alguna regla. Si es así, la aplica y, si no hay más reglas aplicables, avanza hasta la siguiente palabra. El orden en que se definen las reglas es relevante, ya que los cambios efectuados por una pueden ser modificados por otra.

Con el módulo de reglas desarrollado se resuelven adecuadamente muchos de los problemas más frecuentes que plantea la tarea de traducción. Exponemos a continuación algunos de los casos más relevantes.

2.4.1 Contracciones

Mientras que en castellano actual sólo se emplean dos contracciones, el número de contracciones en gallego es muy elevado. Así son muy frecuentes contracciones entre preposiciones y determinantes, preposiciones y pronombres, etc., además de los problemáticos conglomerados de pronombres átonos que según en qué caso van en posición enclítica o proclítica respecto al verbo. Como ejemplo sencillo, *de estas* en castellano debe traducirse a *destas* en gallego, que se podría especificar simplemente por la regla:

de estas, destas

Pero si deseamos resumir en una regla los determinantes demostrativos de la primera persona, escribiremos:

de este/DD,deste/CDD**

El * de la regla propaga en este ejemplo el género y el número desde la condición hasta la acción: con ello se resuelven cuatro casos con la misma regla.

2.4.2 Transformación de tiempos verbales

En gallego no existen los tiempos verbales compuestos, por lo que los tiempos compuestos del castellano deben transformarse en el tiempo gallego más adecuado.

A modo de ejemplo, las siguientes dos reglas: la primera resuelve la transformación de antecopretérito de indicativo en castellano, tiempo compuesto, a antepretérito en gallego, que es una forma simple. La segunda resuelve la regla de antepresente de indicativo a pretérito de indicativo.

haber/VCP VT0SM,%1/VAP**
haber/VR VT0SM,%1/VP**

Y de nuevo, el * propaga número y persona de la condición a la acción.

El típico problema de la perífrasis *ir a + infinitivo* se puede resolver también con la siguiente regla:

ir/V a VN,%0 %2

2.4.3 Enclisis

Los pronombres átonos en gallego sufren a menudo el proceso de enclisis con el verbo; salvo que alguna

partícula lo impida, como puede ser un pronombre interrogativo o un exclamativo. Por el contrario, en castellano es usual que los pronombres átonos vayan delante del verbo. Este hecho se controla nuevamente mediante una serie de reglas. Por ejemplo, para traducir adecuadamente formas del tipo *me dijiste* a *dixéchesme* basta con la siguiente regla:

me/PA V,%1 +me

Esta regla busca pronombre atóno *me* seguido de un verbo y lo sustituye por el mismo verbo seguido del pronombre. El símbolo + indica que debe realizarse además una operación de fusión con la palabra anterior. La fusión revisa la acentuación de la nueva palabra (*dixéchesme)*.

2.4.4 Expresiones y otros problemas.

De lo expuesto en las subsecciones precedentes es fácil deducir que las reglas permiten resolver la traducción de expresiones hechas o el artículo del posesivo en gallego.

Además, ofrecen una solución parcial para los casos en que el sustantivo en castellano y su correspondiente traducción en gallego tienen distinto género.

2.5 Preprocesador y postprocesador

Estos dos módulos se ocupan de mantener el formato de entrada en el texto de salida. Además de texto plano y sus características típicas (tabulaciones, retornos de carro, ciertos signos de puntuación, etc.), reconoce el formato *xml/html,* permitiendo un filtrado de datos no traducibles, bien sean las propias etiquetas, texto comentado o, típicamente, código *javascript*. De esta manera se mantiene íntegro el formato y la funcionalidad de cualquier *página web*.

3 TRADUCTOR ESTADÍSTICO

La traducción estadística parte de la idea, poco intuitiva lingüísticamente, de que la traducción puede ser vista como la transformación de una frase de un lenguaje a otro mediante un proceso estocástico. Por tanto, según este enfoque, traducir una frase f en el lenguaje origen consistirá en buscar una frase e en el lenguaje destino que maximice la fórmula:

$$\arg\max_{e} p(f/e) * p(e)$$

donde $p(f/e)$ representa el modelo de traducción y $p(e)$ corresponde al modelo de lenguaje destino.

Mientras que para entrenar el modelo de lenguaje destino es suficiente con un conjunto de texto de tamaño considerable en la lengua destino, para entrenar el modelo de traducción hace falta un corpus bilingüe revisado y alineado a nivel de frase.

3.1 Corpus de entrenamiento

Disponemos de un corpus de entrenamiento bilingüe creado a partir de texto periodístico. La cifra total de palabras presentes en este corpus es superior a un millón. Hemos utilizado el algoritmo LCS para alinear el corpus y rechazar pares de frases que no cumplen determinados umbrales de parecido.

Como se ha comentado con anterioridad, es deseable que el corpus bilingüe esté revisado manualmente. En nuestro caso, a falta de una revisión exhaustiva, hemos realizado una validación informal de material extraído del corpus de manera aleatoria, obtieniendo un resultado satisfactorio.

Por motivos de disponibilidad de material, hemos usado el material en lengua gallega presente en este corpus para la estimación del modelo de lenguaje destino.

3.2 Entrenamiento de los modelos

De entre todas las alternativas para la estimación del modelo de traducción, los denominados modelos IBM 1-5 [2] son probablemente los más populares. Estos modelos son válidos para lenguas origen y destino muy dispares, ya que la fertilidad o el reordenamiento de frases está contemplado. En este caso, el entrenamiento se realiza en fases de complejidad creciente, siendo el resultado de cada fase el punto de partida para la siguiente.

Nosotros hemos usado GIZA++ [8] para realizar el alineamiento del corpus de entrenamiento bilingüe, para a continuación extraer las unidades de traducción (tuplas) mediante el método expuesto en [4]. Para entrenar los modelos de N-gramas, hemos utilizado el conjunto de herramientas SRILM [9]. Estas herramientas también las hemos utilizado para la obtención del conjunto de N-gramas correspondiente al modelo de lenguaje destino.

	Modelo de lenguaje destino	***Modelo de traducción***
Unigramas	64858	74213
Bigramas	468925	484251
Trigramas	156250	166739

Tabla 3. *Número de N-gramas de los modelos.*

En la tabla 3 se pueden ver las características de los conjuntos de N-gramas obtenidos, tanto para el modelo de traducción como para el modelo de lenguaje destino.

3.3 Decodificador

Para la traducción, se utilizó una herramienta de elaboración propia que realiza una búsqueda basada en el algoritmo de Viterbi.

A partir de los N-gramas del modelo de traducción, se recorre la frase en el lenguaje origen de izquierda a derecha de tal modo que se exploran las posibles alternativas de traducción. Para evitar la aparición de un número desmesurado de posibilidades, en cada paso se realiza una poda N-best, quedándose el decodificador con las N mejores posibilidades. En nuestros experimentos, N=20 fue un valor que resultó ser adecuado para alcanzar un compromiso entre tamaño de las estructuras de datos y exhaustividad de la exploración.

El modelo de lenguaje destino es aplicado en nuestro traductor sobre las frases candidatas obtenidas con el modelo de traducción.

El algoritmo de búsqueda da lugar a traducciones monótonas, ya que no permite cubrir partes de la frase origen de manera desordenada. De todas maneras, la proximidad lingüística entre las lenguas origen y destino (castellano y gallego) hace que los reordenamientos lejanos sean innecesarios.

El decodificador fue implementado mediante una arquitectura cliente/servidor en el lenguaje de programación Java. La utilización de Java conlleva las ventajas de facilidad de programación (tanto algorítmica como de interfaces gráficas) y portabilidad, pudiendo ser utilizado en distintas plataformas o a través de Internet (como applet o servicio web).

A priori podría pensarse que la utilización de Java iba a reducir las prestaciones del decodificador; sin embargo, las pruebas realizadas demostraron que, una vez lanzado el servidor, los tiempos de ejecución para n=20 eran sólo ligeramente superiores a programas implementados en lenguaje C.

4 CONCLUSIONES Y LÍNEAS FUTURAS

En este artículo hemos presentado nuestros trabajos en traducción automática entre gallego y castellano. Son muchas las mejoras pendientes de realizar en ambos traductores, que todavía están en fase de desarrollo.

Además de todos los aspectos concretos que mejorarán las prestaciones de ambos traductores, en este momento, sólo se realiza una traducción castellano a gallego. En breve plazo desarrollaremos versiones que realicen traducción gallego → castellano y más a medio plazo entre estas dos lenguas e inglés.

5 BIBLIOGRAFÍA

[1] P. Brown,J. Cocke, S.A. Della Pietra, V.J. Della Pietra., F. Jelinek, J. Lafferty, R. Mercer y P. Roossin. "A statistical approach to machine translation". Computational Linguistics, vol 16-2, pp. 79-85. 1990.

[2] P. Brown, V.J. Della Pietra, S.A. Della Pietra y R. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation Computation". Computational Linguistics,vol 19-2, pp. 263-311. 1993.

[3] A.M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, Aingeru Mayor and Kepa Sarasola. "An open-source shallow-transfer machine translation engine for the romance languages of Spain." Proceedings of the European Association for Machine Translation, 10th Annual Conference, Budapest, pp. 79-86. 2005.

[4] J.M. Crego, J.B. Mariño, A. Gispert. "Finite-state-based and Phrase-based Statistical Machine Translation." Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04, pp. 37-40. 2004.

[5] J.M. Crego, ,J.B. Mariño, Adrià Gispert. "An Ngram-based Statistical Machine Translation Decoder." 9th European Conference on Speech Communication and Technology. Interspeech, Lisboa. 2005.

[6] D. Cutting,,J. Kupiec, J. Pedersen y P. Sibun. "A practical part-of-speech tagger." Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP), pp. 133–140. 1992

[7] Ismael García. *Traducción Automática Estadística: Modelos de Traducción basados en Máxima Entropía y Algoritmos de Búsqueda.* Tesis Doctoral. Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación. 2003

[8] F.J. Och, "An Efficient Method for Determining Bilingual Word Classes." Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics. pp. 71-76, EACL'99, Bergen, Norway. 1999.

[9] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit". Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado. 2002.

[10] F. Méndez, F. Campillo, E. R. Banga y E. F. Rei, "Análisis morfológico estadístico en lengua gallega". Procesamiento del lenguaje natural nº 31, pp 71-76, 2003.