VNIVERSIDAD NACIONAL
AVTÓNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN LINGÜÍSTICA

ANÁLISIS LINGÜÍSTICO DE LA TRADUCCIÓN AUTOMÁTICA PARA SU EVALUACIÓN

TESIS QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN LINGÜÍSTICA APLICADA

PRESENTA:

MARINA VLADIMIROVNA FOMICHEVA

DIRECTOR DE TESIS: DR. GERARDO EUGENIO SIERRA MARTÍNEZ

INSTITUTO DE INGENIERÍA

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

CODIRECTORA DE TESIS: DRA. IRIA DA CUNHA FANEGO

INSTITUTO UNIVERSITARIO DE LINGÜÍSTICA APLICADA

UNIVERSIDAD POMPEU FABRA

MÉXICO, D.F. (DICIEMBRE) 2012

A mis padres

Agradecimientos

A todas las personas de quienes, directa o indirectamente, aprendí lo necesario para realizar este trabajo.

A mis asesores, el Dr. Gerardo Sierra y la Dra. Iria da Cunha, por haberme apoyado en todo momento y por compartir conmigo sus conocimientos. Gracias por ayudarme a poner mis ideas en orden, por sus consejos, por su paciencia y por su tiempo.

A los profesores de la UNAM por contribuir de manera profesional a nuestra formación como lingüistas. Agradezco en especial a la Dra. Teresa Peralta y a la Dra. Carmen Curcó por sus excelentes clases y por las discusiones enriquecedoras.

A la Coordinación del Posgrado en Lingüística de la UNAM por su ayuda incondicional con todo tipo de trámites, en especial a Guille y a Reina quienes siempre estuvieron dispuestas a aclarar todas mis dudas. Gracias por su paciencia.

A la Dra. Marisela Colín, a la Dra. Natalia Ignatieva-Kosminina, a la Dra. Celine Desmet y a la Mtra. Alma Ortíz Provenzal por su atención, por su comprensión y por la disponibilidad de leer mi tesis entre tanto trabajo.

A los profesores del Institut Universitari de Lingüística Aplicada a quienes conocí durante la estancia de investigación, que realicé en la Universitat Pompeu Fabra de Barcelona, por haberme ayudado a desarrollar este trabajo. En especial agradezco a la Dra. Nuria Bel y al Dr. Jordi Vivaldi por explicarme algunos aspectos importantes sobre mi área de estudio, así como a los miembros del grupo IULATERM por su cordialidad y su ayuda.

Al equipo de la empresa Lucy Software por compartir conmigo su experiencia, por escucharme y por ayudarme con la traducción de los textos de mi corpus. Muchas gracias por la información sobre el funcionamiento del sistema de traducción automática *Lucy LT* que utilizo en la presente investigación.

Agradezco al CONACYT por apoyarme económicamente durante los años de la maestría, pues así pude dedicarme exclusivamente a los estudios. Asimismo, agradezco el apoyo económico brindado por el proyecto *Sintetizador natural y emocional del*

español de México [IT116811] del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica de la UNAM, el cual me permitió realizar la última etapa del desarrollo de mi tesis.

En el plano personal quiero agradecer de la manera más especial a mi familia y a mis compañeros de la maestría. Gracias por las charlas interesantes, por la compañía y por todos los ánimos que he recibido a lo largo de estos años.

Resumen

En esta tesis se presenta un análisis comparativo de traducciones humanas y automáticas en tres niveles de la lengua: léxico-terminológico, morfosintáctico y discursivo. Dicha propuesta se aplica a un corpus paralelo inglés-español de textos especializados del ámbito médico.

El objetivo general de la investigación es estudiar las diferencias lingüísticas sistemáticas entre la traducción automática y la traducción humana a la luz de la problemática de la evaluación automática de sistemas.

Los objetivos específicos son:

a) detectar las diferencias en la distribución de unidades de análisis en la traducción humana y en las traducciones automáticas del sistema estadístico *Google Translate* y del sistema basado en reglas *Lucy LT*;

b) identificar las condiciones en las que se producen dichas diferencias teniendo en cuenta los textos originales y las estrategias de traducción humana y automática.

La metodología del estudio involucra, por un lado, el uso de técnicas estilométricas para caracterizar el lenguaje de la traducción automática frente al de la traducción humana y, por otro lado, la clasificación de las modificaciones que realizan con respecto al texto original los traductores humanos y los sistemas de traducción automática, la cual permite identificar las fuentes de las diferencias detectadas.

Los resultados de la investigación indican que las diferencias entre la traducción automática y la traducción humana relacionadas con las modificaciones opcionales realizadas por los traductores y las diferencias que se deben a la falta de modificaciones obligatorias en la traducción automática no tienen la misma relevancia para evaluar la calidad de esta última.

Índice

Lista de abreviaciones

ALPAC	*Automatic Language Processing Advisory Committee*
ARPA	*Advanced Research Projects Agency*
CAT	*Computer Assisted Translation*
CD	Coeficiente de Dominio
CT	Candidato a Término
DRAE	Diccionario de la Real Academia Española
EDU	*Elemental Discourse Unit*
FAMT	*Fully Automatic Machine Translation*
Google	*Google Translate*
HAMT	*Human-Aided Machine Translation*
LF	Lengua Fuente
LM	Lengua Meta
Lucy	*Lucy LT*
MAHT	*Machine-Aided Human Translation*
OALD	*Oxford Advanced Learner's Dictionary*
PLN	Procesamiento del Lenguaje Natural
POS	*Parts of Speech*
RST	*Rhetorical Structure Theory*
SVM	*Support Vector Machines*
TA	Traducción Automática
TEC	*Translational English Corpus*
TH	Traducción Humana
THR	Traducción Humana de Referencia
TO	Texto Original
TT	Texto Traducido
UT	Unidad Terminológica

Índice de Tablas

Índice de Figuras

1. INTRODUCCIÓN

En las últimas dos décadas el campo de la Traducción Automática [TA] ha tenido un desarrollo vertiginoso motivado por las necesidades de la comunicación multilingüe en un mundo globalizado, donde existe una demanda creciente por producir cada vez más traducciones, lo más rápido posible y a un bajo coste. Sin embargo, incluso los mejores sistemas de TA todavía están muy lejos de sustituir al traductor humano. La TA es una de las aplicaciones más ambiciosas en el ámbito de Procesamiento del Lenguaje Natural [PLN], ya que involucra la mayoría de las áreas de investigación existentes en este campo: análisis morfológico, análisis sintáctico, adquisición del léxico, desambiguación semántica, representación del conocimiento, etc., en un contexto multilingüe.

1.1. Planteamiento

Dada la complejidad y el alcance de la tarea, la TA todavía difiere mucho de la Traducción Humana [TH] en términos de calidad. En este contexto la evaluación tiene un papel estratégico, ya que permite detectar carencias, establecer prioridades y, por tanto, guiar el desarrollo de sistemas de TA. La dificultad principal para medir la calidad de la traducción estriba en que se trata de una tarea abierta, con múltiples soluciones posibles.

En términos generales, hay dos maneras de medir la calidad de la TA: la evaluación manual y la evaluación automática. La evaluación manual es lenta, costosa y subjetiva. Por ello, actualmente se usan ampliamente los métodos de evaluación automática. En este trabajo nos centramos en la problemática de la evaluación automática de la TA.

Los métodos de evaluación automática parten del siguiente supuesto: la TA sería una tarea resuelta si fuera imposible distinguirla de la TH. Por este motivo, la evaluación automática se lleva a cabo por medio de la comparación de la TA con la TH; es decir, se basa en la determinación del grado de similitud entre la TA y una o varias Traducciones Humanas de Referencia [THRs]. Las métricas de evaluación automática más utilizadas en la actualidad calculan la similitud entre la TA y la THR en términos de coocurrencia de secuencias de palabras (n-gramas de palabras), con lo cual evalúan la calidad de la TA únicamente en función de la coaparición de unidades léxicas. La evaluación

automática basada en n-gramas es parcial, ya que se enfoca únicamente en el aspecto léxico de la traducción.

Además, una o varias THRs no cubren todo el espectro de posibilidades de la traducción, lo cual implica una limitación importante para la evaluación automática. Al realizar la comparación entre la TA y la TH, las métricas de evaluación automática no son capaces de distinguir entre la variación aceptable (alternancia en la estructura sintáctica, sinonimia léxica y todo tipo de paráfrasis) y las divergencias que realmente afectan la calidad de la traducción; es decir, penalizan cualquier tipo de diferencia entre la TA y la THR de la misma manera. Sin embargo, no todas las diferencias TA-THR tienen la misma relevancia para medir la calidad de la TA.

Esta problemática se ha abordado de diversas maneras en los estudios recientes sobre la evaluación de la TA. En el presente estudio la abordamos desde la perspectiva traductológica. La investigación reciente en el ámbito de los estudios de traducción basados en corpus ha demostrado que el lenguaje de la TH posee características lingüísticas inherentes que lo distinguen de otros tipos de textos. Estas propiedades distintivas están relacionadas con las diferencias sistémicas entre la Lengua Fuente [LF] y la Lengua Meta [LM], con las particularidades del proceso traductor como una actividad cognitiva compleja y con las condiciones de la mediación cultural. Debido a los factores mencionados, la traducción implica ciertas modificaciones con respecto a la forma y al contenido del original, que en traductología se denominan *translation shifts*. Algunas de ellas son obligatorias desde el punto de vista de las relaciones tipológicas entre los sistemas lingüísticos, mientras que otras son opcionales, resultado de una elección consciente o inconsciente del traductor.

Desde nuestro punto de vista, este hecho no ha recibido atención suficiente en la evaluación o el desarrollo de sistemas de TA. Teniendo en cuenta que el objetivo de la TA, como una tarea de inteligencia artificial, es modelar el comportamiento de los traductores humanos y que la traducción es un proceso complejo que abarca múltiples operaciones y se ve condicionado por numerosos factores, debería investigarse qué aspectos de la TH se pueden y se deben modelar por medio de técnicas computacionales.

1.2. Objetivos

De cara a esta problemática nos planteamos el siguiente objetivo general: estudiar las diferencias lingüísticas sistemáticas entre la TA y la TH y sus implicaciones para la evaluación de sistemas de TA. Los objetivos específicos son: a) detectar las diferencias en la distribución de unidades de análisis a niveles léxico, morfosintáctico y discursivo en la TH y las TAs de sistemas basados en estrategias diferentes; b) identificar las condiciones en las que se dan las diferencias detectadas teniendo en cuenta los Textos Originales [TOs] y el impacto de las estrategias de traducción.

1.3. Supuestos de partida

Por un lado, algunas diferencias entre la TA y la THR están relacionadas con la naturaleza de la TA al ser producto del comportamiento de los sistemas condicionado por los principios básicos de su funcionamiento (por ejemplo, sistemas de TA basados en reglas frente a sistemas de TA estadísticos presentan errores y problemas diferentes) y por las diferencias lingüísticas entre la LF y la LM (la TA en muchas ocasiones ofrece una traducción literal y, por tanto, reproduce los patrones lingüísticos propios de la LF).

Por otro lado, algunas diferencias TA-THR se relacionan con las características de la TH al reflejar el comportamiento de los traductores humanos, que se ve afectado no solamente por las divergencias entre los sistemas lingüísticos, sino también por el propio proceso de la traducción y por las restricciones provenientes de las convenciones de uso de la LM en un contexto de situación determinado. En el contexto de la evaluación, las diferencias TA-THR que se deben a las modificaciones opcionales realizadas por el traductor humano no deben penalizarse de la misma manera que las diferencias TA-THR que resultan de los errores de la TA.

Cabe mencionar que la calidad de la traducción como objeto de evaluación tiene varios aspectos relacionados con los niveles de la lengua, y no existe una manera trivial de ponderar dichos aspectos en términos de su efecto en la calidad global de los Textos Traducidos [TTs]. Por tanto, en un primer acercamiento al análisis lingüístico de las diferencias TA-TH, conviene considerar dichas diferencias con detalle en cada nivel por separado. En esta investigación realizamos la comparación TA-TH a niveles léxico, morfosintáctico y discursivo. A nivel léxico nos centramos en el aspecto terminológico

de la traducción debido a que nuestro corpus de estudio se compone de textos especializados.

El objetivo último de la TA es lograr resultados comparables con la TH en términos de calidad. Entonces, para identificar las áreas problemáticas que requieren optimización, sería útil caracterizar la TA en oposición a la TH, con base en un análisis lingüístico. Además, dada la naturaleza literal de la TA, la comparación TA-TH permitiría observar con más claridad las diferencias entre las lenguas y las decisiones del traductor.

1.4. Delimitación y alcance

En este trabajo no proponemos una métrica de evaluación de la calidad global de la TA, sino una metodología para realizar un análisis comparativo TA-TH que proporcione información sobre la naturaleza de las diferencias entre estos dos tipos de traducción.

La presente tesis es un estudio exploratorio, en el sentido de que las pruebas se realizan a varios niveles lingüísticos partiendo de distintos tipos de representaciones de datos. Debido a esta naturaleza exploratoria, no profundizamos en el análisis de los fenómenos lingüísticos que presentan diferencias en su tratamiento en la TH y la TA. Más bien, nos enfocamos en desarrollar el procedimiento metodológico general que posteriormente podría aplicarse a problemas más específicos.

La metodología que desarrollamos está pensada para su posterior aplicación en el ámbito del PLN. Sin embargo, dada la falta de herramientas de análisis automático para el español, algunas etapas de la metodología propuesta requieren un análisis manual.

1.5. Estructura de la tesis

Para realizar la comparación TA-TH y relacionar las diferencias detectadas con las estrategias de traducción (humana vs. automática) es preciso conocer, por un lado, los principios básicos de funcionamiento de los sistemas de TA y, por otro lado, las características esenciales del comportamiento de los traductores humanos que se reflejan en las propiedades de los TTs.

Por ello, en el Capítulo 2 ofrecemos una descripción de los problemas lingüísticos con los que se enfrenta la TA y de los métodos principales que se han desarrollado para

solucionarlos: métodos simbólicos (TA basada en reglas) y métodos empíricos (TA estadística). Asimismo, revisamos las técnicas de evaluación automática existentes y las limitaciones que presentan.

En el Capítulo 3 ofrecemos una revisión crítica de las ideas que propone la teoría de la traducción con respecto a las propiedades del lenguaje de la TH. En primer lugar, discutimos las aportaciones de la perspectiva descriptiva, la cual se enfoca en las características del texto meta relacionándolas con las particularidades del proceso traductor como fenómeno *sui generis.* Compara los TTs con los textos originalmente escritos en la LM intentando identificar los rasgos del lenguaje de traducción como una variante particular de la misma. Las propuestas metodológicas desarrolladas dentro de esta perspectiva nos sirven para identificar los rasgos del lenguaje de la TA en oposición al lenguaje de la TH.

En segundo lugar, revisamos los métodos de la investigación contrastiva que busca regularidades al observar las modificaciones que realiza el traductor con respecto al texto fuente, es decir, compara los TTs con sus respectivos originales, identificando las fuentes lingüísticas de las decisiones traductológicas. Las propuestas desarrolladas en el marco de este enfoque nos sirven para identificar las fuentes de las diferencias entre la TA y la TH.

Por último, revisamos las aplicaciones computacionales de las ideas teóricas desarrolladas en el marco de los enfoques discutidos que están estrechamente relacionadas con el uso de corpus lingüísticos en los estudios de traducción.

En el Capítulo 4 describimos la metodología de la presente investigación. Por un lado, explicamos los principios en los que nos basamos para la compilación del corpus de estudio. Por otro lado, presentamos los tipos de análisis automático (o semiautomático) que aplicamos a nuestros datos: la extracción terminológica, el etiquetado de partes de la oración [*Part of Speech Tagging*, etiquetado POS] y el etiquetado de las relaciones discursivas, así como los procedimientos específicos que llevamos a cabo en cada nivel de análisis.

En el Capítulo 5, en primer lugar, explicamos la selección de dos sistemas de TA: un sistema basado en reglas lingüísticas - *Lucy LT*[1] [*Lucy*], y un sistema basado en técnicas estadísticas - *Google Translate*[2] [*Google*]. Decidimos usar dos sistemas de TA porque presentan carencias muy distintas y porque nos interesa ver en comparación qué tipo de modificaciones realizadas por el humano son capaces de modelar sistemas basados en estrategias diferentes. En segundo lugar, describimos nuestro corpus de estudio, que se compone de artículos de divulgación del ámbito médico escritos originalmente en inglés, las THs y las TAs de dichos textos al español. Finalmente, discutimos ejemplos concretos del análisis realizado y presentamos los resultados de la investigación.

En la última sección discutimos las aportaciones y las limitaciones del estudio, ofrecemos las conclusiones generales y las ideas para el trabajo futuro.

[1] http://www.lucysoftware.com/espanol/traduccin-automtica/

[2] Por facilidad se usará a partir de aquí solamente *Google* para referirse a *Google Translate* (http://translate.google.com/?hl=es&tab=mT), en tanto Google será usado para la empresa.

2. TRADUCCIÓN AUTOMÁTICA Y SU EVALUACIÓN

La TA constituye un campo interdisciplinario que puede ser abordado desde distintas perspectivas teóricas. Al ser una aplicación del PLN, la TA se apoya tanto en la lingüística como en las ciencias de la computación. Por este motivo, en el presente capítulo ubicamos la TA en el ámbito del PLN y discutimos su relación con la lingüística (apartado 2.1.). En el apartado 2.2. revisamos la problemática de la TA, ya que para identificar las fuentes de las diferencias entre la TA y la TH es necesario comprender cuáles son los fenómenos lingüísticos y traductológicos cuyo tratamiento representa dificultades para los sistemas; en otras palabras, deben conocerse las causas potenciales de las divergencias entre la TA y la TH. En el apartado 2.3. ofrecemos la descripción de las principales estrategias o métodos de la TA que se han desarrollado para lidiar con estas dificultades. De cara a la tarea de comparación TA-TH tienen que considerarse los principios básicos de funcionamiento de los sistemas que usamos en esta investigación: la TA basada en reglas (apartado 2.3.2.) y la TA estadística (apartado 2.3.3.). Finalmente, en el apartado 2.4. ofrecemos una revisión crítica de los métodos actuales de evaluación automática y sus limitaciones, esto último con el fin de poder discutir las implicaciones del análisis que realizamos para la evaluación de la TA.

2.1. Traducción automática en el contexto del procesamiento del lenguaje natural

El PLN es un ámbito multidisciplinar en cuyo marco se agrupa un conjunto relativamente heterogéneo de métodos y teorías que consideran la lengua como un objeto susceptible de tratamiento informático. Gelbukh y Sidorov (2010) mencionan que existe una relación de beneficio mutuo entre lingüística y ciencias de la computación. Por un lado, las tecnologías computacionales ofrecen a la lingüística herramientas para la comprobación de hipótesis, investigación sobre el alcance de gramáticas y diccionarios, y desarrollo de descripciones formales, completas y precisas de las lenguas. Por otro lado, el conocimiento lingüístico proporciona una base teórica para el desarrollo de las aplicaciones tecnológicas. Así, el PLN toma algunos conceptos y herramientas fundamentales de las ciencias de la computación, de la lógica, del procesamiento de señales, de la teoría de la información, mientras que otros provienen de la lingüística. Además, el PLN ha desarrollado sus propios formalismos y conceptos teóricos para la creación de sistemas que realizan tareas específicas. El principal interés

del PLN reside en el desarrollo de modelos computacionales para los fenómenos lingüísticos con base en diversas fuentes de conocimiento, tanto teóricas como empíricas.

El PLN tiene por objetivo realizar transformaciones entre distintas representaciones u objetos lingüísticos de manera automática o semiautomática (por ejemplo, transformar un texto plano a una representación enriquecida con información sobre la categoría gramatical de cada palabra, la estructura de constituyentes o la estructura semántica de las oraciones, traducir de una lengua a otra, resumir el contenido de un texto, etc.). Así, de acuerdo con Jurafsky y Martin (2009), el PLN es un conjunto de técnicas computacionales que procesan el lenguaje humano, tanto hablado como escrito. En este sentido, lo que distingue las aplicaciones de PLN de otros sistemas de procesamiento de datos es que en ellas se utiliza el conocimiento sobre el lenguaje humano.

Ahora bien, la TA se define como "the use of computer to automate translation from one language to another" (Jurafsky y Martin, 2009: 895). La investigación en la TA constituye un área interdisciplinaria que combina múltiples tareas y puede ser abordada desde diversas perspectivas teóricas. Así, tal como afirman Nyberg et al. (1994: 95):

> Machine translation is considered the paradigm task of Natural Language Processing by some researchers because it combines almost all NLP research areas: syntactic parsing, semantic disambiguation, knowledge representation, language generation, lexical acquisition, and morphological analysis and synthesis.

Existen dos enfoques generales en PLN: los métodos simbólicos y los métodos empíricos. Las herramientas desarrolladas con los métodos simbólicos tienen el conocimiento lingüístico codificado explícitamente en forma de reglas. En el caso de la TA se trata de sistemas que realizan la traducción con base en lexicones y gramáticas computacionales y se denominan sistemas de TA basados en reglas (*Rule-Based Machine Translation*). Los métodos empíricos realizan inferencias sobre el uso del lenguaje a partir de grandes cantidades de datos por medio de técnicas estadísticas. El tipo principal de sistemas de TA que hace uso de los métodos empíricos se denomina TA estadística (*Statistical Machine Translation*). Debido a su orientación hacia los datos (en oposición a la teoría) y a la estadística (en oposición a la lingüística), el enfoque empírico se ha considerado como opuesto al enfoque simbólico. Sin embargo,

actualmente en PLN no existe una división tajante entre estos dos paradigmas: "[…] it is now widely recognized that the key to automatically processing human languages lies in the appropriate combination of symbolic and nonsymbolic techniques" (Dale, 2000: 1).

De hecho, tal como afirma Giménez (2008), la distinción entre los sistemas basados en reglas y los sistemas estadísticos ya no se mantiene claramente:

> […] the expression "rule-based" is slightly inaccurate nowadays. The reason is that empirical MT systems may also use automatically induced rules. Therefore, perhaps it is more appropriate to refer to these two types of systems as knowledge-driven and data-driven (Giménez, 2008: 4).

Con todo, para evitar confusión, en esta tesis seguimos empleando la terminología tradicional.

Sea cual sea el método empleado para automatizar el proceso de la traducción, la calidad de los textos producidos por los sistemas sigue siendo baja en comparación con la TH. En el siguiente apartado mencionamos algunas de las razones por las que la TA difiere de la TH.

2.2. Problemática de la traducción automática

La TA (al igual que la TH) implica cierto análisis (o "comprensión") del texto fuente y la expresión del contenido de éste con los recursos de la LM. De cara a la tarea de análisis del texto fuente (es decir, desde la perspectiva monolingüe) se suele discutir el problema de la ambigüedad del lenguaje natural, en oposición a los códigos artificiales, en los que a un mensaje le corresponde necesariamente una única interpretación posible. Hutchins y Somers (1992) distinguen entre la ambigüedad léxica y la ambigüedad estructural[3]. De acuerdo con estos autores, existen tres tipos básicos de ambigüedad léxica: ambigüedad categorial, polisemia y ambigüedad transferencial. En el caso de la ambigüedad categorial, la misma forma superficial en ocasiones representa unidades léxicas con distintas funciones gramaticales, sobre todo tratándose de lenguas con marcación de rasgos gramaticales escasa o nula, como es el caso del inglés. El segundo

[3] En numerosas ocasiones, las expresiones cuya desambiguación resulta problemática para un sistema de TA, no resultan un problema para el traductor humano, quien hace uso del contexto, tanto lingüístico como extralingüístico, para determinar el significado de unidades léxicas y construcciones sintácticas. La codificación de la información contextual es una de las tareas más difíciles en la construcción de sistemas de PLN.

tipo de ambigüedad se presenta cuando la palabra tiene más de un significado asociado en la LM. En cuanto a la ambigüedad transferencial, se produce cuando una palabra dada tiene una única interpretación en la LF, pero tiene varias traducciones posibles a la LM.

Mientras que la ambigüedad léxica está relacionada con el análisis de palabras individuales y la codificación de sus significados en la LM, la ambigüedad estructural concierne a los problemas del análisis sintáctico y a la representación de la estructura de las oraciones. La ambigüedad estructural o sintáctica se presenta cuando existen varios análisis posibles de la misma oración, en el marco de la gramática utilizada por el módulo de análisis sintáctico del sistema de TA. Ello sucede en los casos en los que la ambigüedad de las unidades léxicas que forman parte de la oración conlleva la posibilidad de varias lecturas de la misma (ambigüedad de las funciones sintácticas de las palabras) o cuando el sistema no dispone de la información necesaria para determinar las relaciones sintácticas entre los constituyentes (por ejemplo, no logra identificar el antecedente de una oración subordinada de relativo o resolver la ambigüedad de adjunción de frase preposicional (*prepositional phrase attachment ambiguity*)). Las ambigüedades de adjunción han sido ampliamente estudiadas en el ámbito del PLN y, de acuerdo con Lin (1998), siguen siendo la fuente principal de los errores de las herramientas computacionales de análisis sintáctico. Las ambigüedades se dan también a nivel extra-oracional. La desambiguación de expresiones anafóricas (por ejemplo, la determinación del antecedente de una anáfora pronominal) es otro problema ampliamente estudiado en PLN (para una revisión detallada, véase, por ejemplo, Kehler, 1997). Actualmente, los sistemas de TA realizan la traducción oración por oración, de manera que no se lleva a cabo el tratamiento de la referencia anafórica a nivel extra-oracional, lo cual afecta la coherencia de los textos producidos por los sistemas de TA.

De cara a la fase de la transmisión del contenido del original en la LM (es decir, desde la perspectiva contrastiva), se suelen discutir las diferencias entre las lenguas que causan errores en la TA. Estas diferencias se denominan *translation divergencies* (Dorr, 1994; Jurafsky y Martin, 2009). El primer tipo de divergencias que identifican Jurafsky y Martin (2009) son las divergencias tipológicas, es decir, las diferencias sistemáticas (frente a las similitudes, también sistemáticas) que observamos con regularidad en las relaciones entre diversas lenguas (por ejemplo, el orden básico de palabras, el parámetro

de sujeto nulo, la estructura argumental, etc.). El segundo tipo de divergencias se denomina divergencias idiosincrásicas, ya que no se dan entre las lenguas de manera sistemática, sino que son propias de un par de lenguas específico (por ejemplo, la posición del adjetivo en inglés y en español). El tercer tipo son las divergencias léxicas, es decir, las diferencias entre las relaciones conceptuales, la organización de los campos semánticos, etc. en distintas lenguas. Asimismo, las divergencias léxicas se relacionan con las funciones de las clases de palabras (por ejemplo, el gerundio nominal del inglés se suele traducir al español con un nombre deverbal). La traducción con frecuencia conlleva la explicitación o la implicitación obligatoria de ciertos rasgos gramaticales o semánticos, en el sentido de que la selección léxica puede estar sujeta a más restricciones gramaticales en una lengua que en otra (por ejemplo, el español, a diferencia del inglés, marca el género en los nombres y adjetivos).

En suma, la ambigüedad y las diferencias estructurales, estilísticas y culturales entre la LF y la LM hacen de la TA una tarea compleja. Algunos estudiosos afirman incluso que la TA es un tipo de textos particular, "with a limited range of possible functions" (Sager, 1993: 263). Jurafsky y Martin (2009) mencionan que la TA puede usarse en las situaciones siguientes: a) cuando una traducción aproximada es suficiente para satisfacer las necesidades del usuario (por ejemplo, extracción de información en un contexto multilingüe); b) cuando la TA se somete a una post-edición de traductores profesionales; c) cuando los textos a traducir pertenecen a un ámbito muy restringido (por ejemplo, los pronósticos de tiempo atmosférico en los cuales se usa un vocabulario muy limitado y unos cuantos tipos de frase básicos) donde puede lograrse una TA de alta calidad. Así, la TA no pretende sustituir completamente a los traductores humanos, pero es de gran utilidad en el contexto de la comunicación multilingüe.

2.3. Sistemas de traducción automática

En relación con esta última idea, cabe mencionar que existen diversos modos de uso de la TA en función del grado de la participación del traductor humano.

2.3.1. Modos de uso de la traducción automática

Al examinar los modos de uso de los sistemas de TA, Hutchins y Somers (1992) distinguen entre *Machine-Aided Human Translation* [MAHT], *Human-Aided Machine*

Translation [HAMT] y *Fully Automatic Machine Translation* [FAMT]. Esta distinción se basa en el grado de automatización frente al grado de participación del traductor humano en el proceso de la traducción. En realidad, se trata de un continuo en cuyos extremos se encuentra, por un lado, la TH sin ningún tipo de procesamiento automático, y, por otro lado, la FAMT. La distinción entre la MAHT y la HAMT es a veces difícil de trazar, con lo cual en ocasiones se prefiere usar el término *Computer-Assisted Machine Translation* [CAT] para referirse a ambos tipos de sistemas. Esta idea se representa de manera esquemática en la Figura 1.

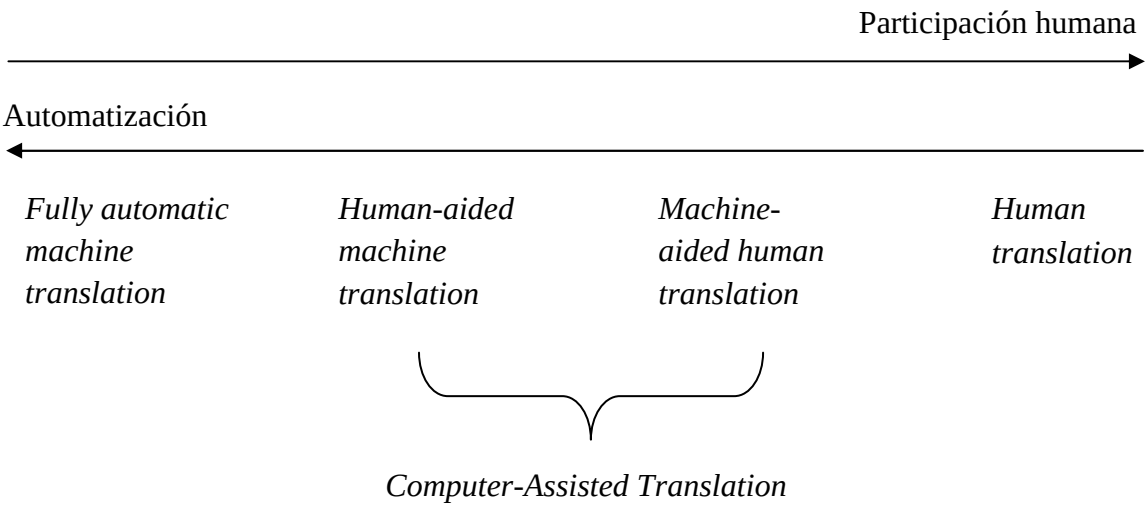

Figura 1. Traducción humana y traducción automática (Hutchins y Somers, 1992)

En los sistemas tipo MAHT, la traducción es realizada en gran parte por el traductor, que se apoya en herramientas computacionales para cumplir con su labor. Hoy en día existen numerosos programas de ayuda al traductor que le proporcionan el acceso a diccionarios electrónicos, bases de datos terminológicas y otros recursos de referencia. Asimismo, existen las bases de datos traductológicas o memorias de traducción que guardan las traducciones realizadas en una base de datos a la cual se accede de manera automática si la oración a traducir coincide total o parcialmente con las oraciones traducidas guardadas en la base de datos (para una revisión detallada de dichos programas, véase Hutchins, 1998). El ejemplo más conocido de este tipo de sistemas es *SDL Trados* (www.trados.com).

HAMT supone el uso de sistemas de TA para producir traducciones con ayuda de la pre- o post-edición manual. Uno de los enfoques desarrollados dentro de este paradigma

son los sistemas de lenguaje controlado, los cuales trabajan con textos que han pasado por un proceso de pre-edición encaminado a evitar construcciones sintácticas complejas o unidades léxicas ambiguas. Primero, se comprueba de manera automática si los TOs contienen palabras que no forman parte del vocabulario del sistema o estructuras sintácticas que el sistema no es capaz de analizar. A continuación, se le pide al usuario que sustituya dichos elementos por los elementos "aceptables" para el sistema.

Un enfoque similar se basa en la idea de sublenguaje (término empleado en el ámbito de la TA) o lenguaje de especialidad. El lenguaje usado en los textos de determinado tipo o dominio temático es naturalmente restringido en cuanto al vocabulario y la sintaxis, lo cual constituye una gran ayuda en la solución de ambigüedad tanto léxica como estructural.

Kittredge y Lehrberger (1982) destacan algunos aspectos relevantes sobre los lenguajes de especialidad de cara a la construcción de sistemas de TA:

> 1) Es más fácil crear reglas de generación de oraciones gramaticales y aceptables para un lenguaje de especialidad que para la lengua en general.
> 2) Las reglas de construcción de oraciones en un lenguaje de especialidad pueden ser diferentes a las reglas de la lengua estándar. La gramática del inglés estándar no contiene las gramáticas de todos los lenguajes de especialidad, ya que algunas reglas u operaciones gramaticales existen únicamente en ciertos sublenguajes y no juegan ningún papel en el inglés estándar.
> 3) La variante de una lengua puede ser calificada como lenguaje de especialidad con base en la sistematicidad del uso, independientemente de su complejidad o del tamaño del vocabulario. Este grado de sistematicidad del lenguaje de especialidad es el que determina qué tan apropiado es para la TA.

La detección de las construcciones más recurrentes o las construcciones ausentes en un lenguaje de especialidad posibilita la simplificación de las gramáticas o la eliminación de ambigüedades, haciendo la TA más fácil y más fiable. La idea de aprovechar esta posibilidad se originó en los trabajos del grupo de investigación TAUM de la Universidad de Montreal en los años setenta con el desarrollo de los sistemas *Météo*, enfocado en la traducción de pronósticos de tiempo atmosférico del inglés al francés

(Kittredge y Lehrenberger, 1982), y *Aviation proyect*[4], enfocado en la traducción de manuales técnicos de aviación con el mismo par de lenguas.

Los sistemas de TA aplicados en dominios restringidos han dado mejores resultados que los que se aplican a la lengua en general. Esta suposición se retoma en la construcción de sistemas estadísticos que infieren reglas lingüísticas a partir de datos empíricos de corpus textuales y producen traducciones más fiables al ser entrenados y aplicados para la traducción de un tipo textual determinado o de textos con una temática restringida (véase apartado 2.3.3.).

Los sistemas de tipo FAMT se usan cuando las traducciones de baja calidad son suficientes para el usuario con tal de que proporcionen una idea general sobre el contenido del original. De acuerdo con Giménez (2008), la mayoría de los sistemas comerciales mantienen el esquema MAHT, mientras que los sistemas disponibles en Internet de manera gratuita suelen ser de tipo FAMT.

Ahora bien, en función de la estrategia general adoptada para la traducción, se suele distinguir entre la TA basada en reglas y la TA estadística. A continuación, explicamos los principios de funcionamiento de estos dos tipos de sistemas.

2.3.2. Traducción automática basada en reglas

La TA basada en reglas ha sido influenciada en su concepción teórica por los avances en lingüística generativa e inteligencia artificial a partir de los años setenta. En este tipo de sistemas existe un conjunto de reglas desarrolladas de manera manual que describen detalladamente el proceso de traducción.

2.3.2.1. Triángulo de Vauquois

En relación con la arquitectura de los sistemas basados en reglas, se suele distinguir entre el modelo de traducción directa, el modelo de transferencia y el modelo *interlingua*. El triángulo de Vauquois (1968) (Figura 2) resume las características de los tres modelos ilustrando la relación entre la cantidad de procesamiento lingüístico

[4] Este segundo sistema no tuvo tanto éxito como *Météo*, debido a la complejidad del lenguaje de especialidad del ámbito de aviación, que presenta numerosas ambigüedades léxicas, sintácticas y pragmáticas para la traducción de los sintagmas nominales complejos y el imperativo del inglés al francés (Hutchins y Somers, 1992).

necesario para las fases de análisis y generación y la cantidad de conocimiento requerido para la fase de transferencia entre la representación producto de análisis y la representación a partir de la cual se genera el TT.

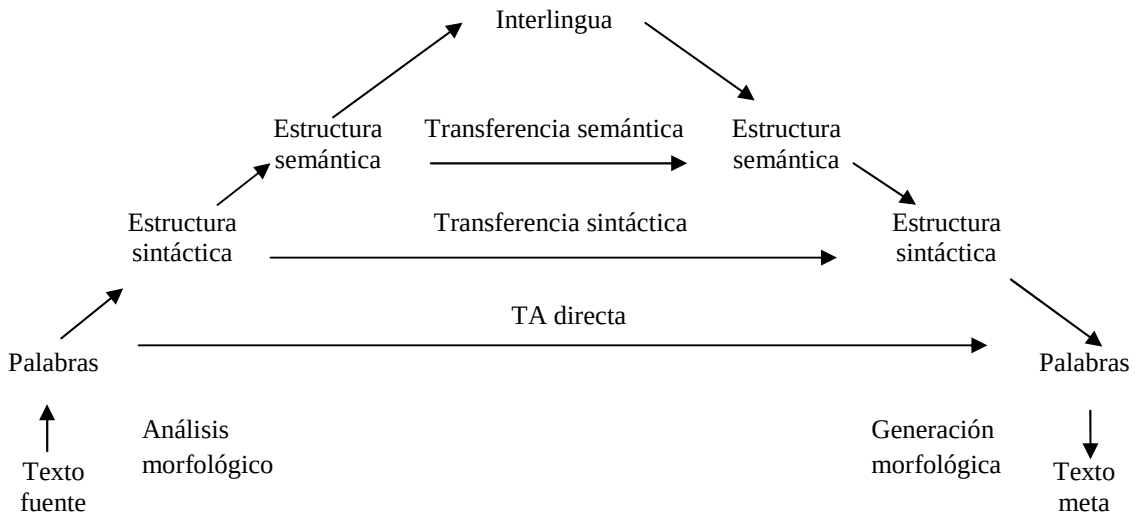

Figura 2. Triángulo de Vauquois (Jurafsky y Martin, 2009)

En la TA directa, el TO se traduce palabra por palabra o frase por frase con ayuda de grandes diccionarios bilingües sin ningún tipo de procesamiento sintáctico o semántico del texto fuente. Asimismo, pueden aplicarse simples reglas de reordenación (por ejemplo, con respecto a la posición de los adjetivos en la traducción del inglés al español). Los sistemas de transferencia se basan en el conocimiento lingüístico sobre las diferencias estructurales entre la LF y la LM, codificado explícitamente en forma de reglas. En la etapa de análisis se lleva a cabo el procesamiento sintáctico y (en ocasiones) semántico del TO. En la fase de transferencia, a la salida del módulo de análisis se aplican las reglas de transferencia léxica, sintáctica y, en ocasiones, semántica. En la fase de generación, a partir de la representación (más o menos abstracta) generada en la etapa de transferencia, se produce el texto meta. Entre los formalismos gramaticales más utilizados en el desarrollo de sistemas de TA basada en reglas, Hutchins y Somers (1992) mencionan la gramática léxica funcional, la gramática categorial, la teoría de la rección y ligamiento, la gramática de estructura de frase generalizada y la gramática de Montague. Algunos de los sistemas de TA basada en el modelo de transferencia con mayor influencia en el ámbito son *Metal* (Bennet y Slocum, 1985), *Susy* (Maas, 1987) y *Eurotra* (Allegranza et al., 1991). De acuerdo con

Jurafsky y Martin (2009), los sistemas comerciales suelen combinar ciertos rasgos de la TA directa (grandes diccionarios bilingües con información detallada sobre las propiedades de las unidades léxicas) con las características del modelo de transferencia (análisis morfosintáctico superficial) (Jurafsky y Martin, 2009). Un ejemplo clásico de dicha combinación es el sistema *Systran* (véase Hutchins y Somers, 1992, para una descripción detallada de este sistema).

Los modelos de transferencia requieren conjuntos de reglas lingüísticas distintos para cada par de lenguas, lo cual constituye una limitación importante en el contexto de comunicación multilingüe que involucra un número de lenguas considerable. De cara a este problema ha surgido el enfoque *interlingua.* Los modelos *interlingua* difieren de los modelos de transferencia en el grado de abstracción en la representación del conocimiento lingüístico. En los sistemas *interlingua* en la etapa de análisis se produce una representación conceptual independiente de la lengua (denominada *interlingua*), con lo cual la etapa de transferencia deja de ser necesaria (véase Figura 2), y el texto meta se genera directamente a partir de esta representación abstracta. Es decir, los modelos *interlingua* pretenden representar todas las oraciones que significan lo mismo de la misma manera, independientemente de la lengua en la que estén expresadas. El tipo de representación semántica abstracta varía en función del sistema; puede usarse, por ejemplo, la lógica de primer orden o la descomposición en primitivos semánticos para crear la representación *interlingua*. La ventaja de este tipo de sistemas es que no se necesita ningún conocimiento lingüístico idiosincrático. Además, la arquitectura de los sistemas *interlingua* constituye una modelación más fiel de los procesos involucrados en la TH, ya que en la fase del análisis se genera una representación conceptual del texto fuente, con lo cual el sistema "comprende" el TO y después lo reproduce en la LM. La investigación enfocada en los sistemas de tipo *interlingua* se inserta en el ámbito de inteligencia artificial, y los sistemas de este tipo también se denominan TA basada en conocimiento (*Knowledge-based machine translation*) (véase Nirenburg, 1989, para una revisión detallada de este enfoque).

Generar de manera automática una representación conceptual independiente de la lengua es una tarea sumamente difícil, ya que requiere una modelación del contexto extralingüístico y la codificación del conocimiento del mundo; con lo cual, la mayoría de los sistemas basados en el modelo *interlingua* no pasan de ser prototipos. Un ejemplo

conocido de los sistemas *interlingua* es el sistema *Rosetta* (Landsbergen, 1987), basado en la gramática de Montague.

2.3.2.2. Sistema de traducción automática Metal

Para ejemplificar el funcionamiento de los sistemas de TA basada en reglas, describiremos brevemente la arquitectura del sistema *Metal* (*Mechanical Translation and Analysis of Language*), ya que *Lucy*, el sistema de TA basada en reglas que utilizamos para los propósitos de esta investigación, es un desarrollo del sistema *Metal*[5].

El sistema comienza a desarrollarse a mediados de los años sesenta en el *Linguistic Research Center* de la Universidad de Texas, como un sistema *interlingua*, pero su primera versión comercial está basada en el modelo de transferencia. El sistema se construye para la traducción de grandes cantidades de documentación técnica de la empresa *Siemens* y tiene integrados programas muy complejos para facilitar la post-edición y proporcionar al usuario el acceso a las bases de datos terminológicas.

Los recursos lingüísticos del sistema son diccionarios (monolingües y bilingües) y una gramática computacional que contiene información sintáctica almacenada en forma de reglas de reescritura. Los diccionarios tienen una estructura jerárquica. Los tres módulos básicos son el vocabulario de palabras funcionales, el vocabulario de léxico general y el vocabulario técnico general, a los cuales se accede independientemente de la temática del texto. En un nivel inferior se encuentran los diccionarios terminológicos generales, los diccionarios terminológicos específicos y, finalmente, los diccionarios de usuario. El usuario puede agregar tantos glosarios específicos como desee y especificar el orden de prioridad según el cual el sistema accederá a los mismos. Los diccionarios son monolingües para las fases de análisis y generación, y bilingües para la fase de transferencia. Los diccionarios monolingües tienen la misma estructura y pueden emplearse tanto en la fase de análisis como en la fase de generación (según sea la dirección de la traducción). Cada entrada del diccionario monolingüe es una lista de pares atributo-valor. Los atributos principales son la forma canónica, la categoría morfosintáctica, el alomorfo, el emplazamiento (indica las posibles posiciones del

[5] Para realizar la descripción nos basamos en la información que ofrecen Hutchins y Somers (1992) sobre la primera versión comercial unidireccional alemán-inglés de *Metal*, implementada a mediados de los años ochenta.

morfo con respecto a otros morfos con los que puede combinarse en una palabra; estos valores se usan para restringir la aplicación de las reglas de formación de palabras), la preferencia (ponderación de significados en palabras polisémicas), la colocación léxica, el número de sentido (para distinguir entre homógrafos), el número de concepto (para establecer un índice común para las palabras semánticamente relacionadas) y el dominio temático (indica el o los dominios o áreas de conocimiento en los que la palabra o la acepción tiene más probabilidad de aparecer). Las entradas para los nombres, además de la categoría morfosintáctica, tienen codificados los rasgos semánticos (entidad, animado, etc.) que se utilizan para las restricciones colocacionales. Las entradas para los verbos incluyen los rasgos de subcategorización de los participantes en términos de papeles temáticos con las características asociadas que restringen los valores semánticos y las funciones sintácticas superficiales de los argumentos, así como la información sobre el tipo de transitividad. Aparte de estos atributos generales, existe una serie de posibles atributos que se pueden aplicar a entradas específicas o a lenguas determinadas y que contienen información adicional de tipo morfológico, sintáctico y semántico. Los diccionarios bilingües relacionan las unidades léxicas de las lenguas fuente y meta e introducen restricciones para los casos en los que hay varias traducciones posibles. Una codificación sumamente detallada del léxico es una particularidad de la arquitectura del sistema *Metal*.

El sistema *Metal* utiliza una gramática libre de contexto aumentada con transformaciones. La gramática se compone de reglas de estructura de frase libres de contexto e incrementadas con pruebas y condiciones de aplicación. Las reglas libres de contexto tienen la forma de reglas de reescritura a las que se asocian dos tipos de pruebas. Las pruebas del primer tipo especifican las condiciones morfológicas y sintácticas para la aplicación de las reglas, mientras que las pruebas del segundo tipo especifican los rasgos gramaticales que deben poseer las unidades léxicas para encontrarse en una relación sintáctica determinada.

Las fases de la traducción en la arquitectura del sistema *Metal* son el análisis morfológico, el análisis sintáctico, la transferencia léxica y estructural y, finalmente, la generación. El análisis morfológico consiste en la identificación de los morfemas base y de los afijos de cada palabra en la oración. A continuación se aplican las reglas de la gramática de estructura de frase para producir los posibles análisis sintácticos que se

ponderan en función de los valores de preferencia asociados a las categorías gramaticales de las unidades léxicas y a las configuraciones estructurales. Las reglas se aplican a partir de varias pruebas y restricciones definidas en términos de los rasgos morfosintácticos de las palabras. El proceso de transferencia se realiza a partir de diversas interacciones complejas entre las reglas de la gramática y la información del diccionario bilingüe. El módulo de transferencia recibe una representación sintáctica superficial con los papeles temáticos asignados y produce una representación superficial con el orden y la estructura morfológica de las palabras propios de la LM. De esta manera, el módulo de generación se ocupa únicamente de la síntesis morfológica. En las versiones posteriores de *Metal* el módulo de generación adquiere un papel más importante, mientras que el módulo de transferencia deja de ser sobrecargado (Trujillo, 1995).

En este apartado hemos presentado las técnicas de TA basada en reglas. A pesar de la importancia de las contribuciones al desarrollo de la TA realizadas en el marco de este paradigma, su desventaja sustancial es que los modelos basados en reglas no son capaces de dar cuenta de los casos en los que la selección de recursos lingüísticos está condicionada por los factores de uso de la lengua. Mientras tanto, tal como afirma Melamed (2001: 1):

> Natural language follows few hard and fast rules. Therefore, a good model must account for tendencies and likelihoods. Although people use language all the time, they cannot accurately assign probability distributions over linguistic data by introspection.

En el siguiente apartado discutimos los métodos empíricos de la TA que retoman esta idea.

2.3.3. Traducción automática estadística

Existen dos tipos principales de sistemas empíricos de TA: la TA estadística y la TA basada en ejemplos. Inicialmente, estos dos enfoques se diferenciaban claramente, ya que la TA basada en ejemplos hacía uso del conocimiento lingüístico, mientras que la TA estadística utilizaba únicamente técnicas estadísticas. Actualmente, estos dos acercamientos se unen bajo el paradigma de TA empírica. En esta tesis nos centramos

en la TA estadística (para un revisión de la TA basada en ejemplos véase, por ejemplo, Somers, 2000).

Los sistemas de TA estadística hacen uso de modelos probabilísticos cuyos parámetros se estiman a partir de grandes corpus bilingües. A diferencia de los sistemas basados en reglas que se centran en el proceso de traducción (el tipo de representación del TO y las etapas de la traducción), los sistemas estadísticos se enfocan en el resultado, ya que plantean el problema de traducción como una tarea de estimación de la probabilidad de que una oración en la LM sea una traducción de una oración determinada en la LF.

Los requisitos básicos para la traducción son: fidelidad con respecto al TO y aceptabilidad en el contexto de la LM[6]. La TA estadística maximiza estos parámetros por medio de modelos probabilísticos entrenados con corpus bilingües (fidelidad) y monolingües (aceptabilidad).

Así, para un sistema de TA estadística, traducir una oración *F* significa proporcionar una oración *E* que maximice cierta función que refleje la importancia de fidelidad al TO y la aceptabilidad en la LM[7]. Los sistemas estadísticos escogen la traducción más probable con base en una combinación de los modelos probabilísticos de fidelidad (modelos de traducción) y aceptabilidad (modelos de lenguaje). Para formalizar esta idea en la TA estadística se usa el modelo de canal ruidoso desarrollado en el marco de la teoría de la información por Shannon (1948). Este modelo describe la recuperación de la información que se pierde cuando un mensaje atraviesa un canal con ruido. Aplicado a la tarea de la TA, dicho modelo se formaliza de la manera siguiente. Si denominamos la oración de la LF $F = f_1, f_2,\ldots, f_n$ y la oración traducida $\hat{E}=e_1, e_2,\ldots, e_n$, entonces en un modelo probabilístico la mejor traducción es la que tiene la probabilidad P(E|F) más alta. En el modelo de canal ruidoso, esta tarea se resuelve por medio de la transformación siguiente:

$$\hat{E} = argmax_E\, P(E|F) = \; argmax_E \frac{P(F|E)P(E)}{P(F)} = argmax_E P(F|E)P(E)$$

[6] En la literatura sobre el desarrollo y la evaluación de la TA se usan los términos *fidelity, accuracy, adequacy* para hacer referencia al aspecto que aquí denominamos fidelidad. En cuanto al aspecto que nombramos aceptabilidad se utilizan también los términos *acceptability, fluency, intelligibility*.

[7] Tradicionalmente, para designar la oración fuente y la oración meta se usan las letras F y E, respectivamente, por el primer sistema moderno de TA estadística *Candide* que traducía del francés (F) al inglés (E).

De esta manera, al aplicar el modelo de canal ruidoso a la tarea de la TA, se revierte el orden natural de eventos. Es decir, se pretende que la oración original sea una versión corrupta de una oración en la LM, con lo cual el objetivo es descubrir esta oración desconocida, que generó la oración observada en la LF[8].

Con lo que hemos visto hasta ahora, para realizar la TA estadística se necesitan tres componentes: el modelo de lenguaje *P(E),* el modelo de traducción *P(F|E)* y el algoritmo de decodificación que dada una oración *F* producirá la traducción *E* más probable.

Los modelos de lenguaje representan una distribución de probabilidades que reflejan la frecuencia de aparición de secuencias de elementos en el corpus de entrenamiento. Estos modelos se construyen a partir de los n-gramas, es decir, secuencias de elementos (normalmente, palabras) de longitud *n*. El objetivo de un modelo de lenguaje consiste en estimar la probabilidad de una secuencia de palabras W=w_1, w_2,…$w_{|w|}$ en una lengua dada (en el caso de la TA, la LM). Los modelos de lenguaje calculan la probabilidad *P(W)* realizando la descomposición siguiente:

$$P(W) = \prod_{i=1}^{|W|} P(w_i|w_1^{i-1})$$

Los modelos de n-gramas simplifican la probabilidad de aparición de una palabra, dadas todas las anteriores, a una historia que sólo considera las *n-1* palabras anteriores:

$$P(W) = \prod_{i=1}^{|W|} P\left(w_i|w_1^{i-1}\right) \approx \prod_{i=1}^{|W|} P\left(w_i|w_{i-n+1}^{i-1}\right)$$

Esta probabilidad se puede estimar a partir de un corpus, por lo que se obtiene de forma automática sin necesidad de introducir conocimiento lingüístico de tipo deductivo. Para

[8] La idea de la aplicación del modelo de canal ruidoso a la TA se remonta a la famosa propuesta que expresó Warren Weaver, uno de los pioneros de la TA, con respecto a la posibilidad de usar los métodos estadísticos y las técnicas criptográficas para automatizar la traducción: "When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode"" (citado en Hutchins, 1997: 195).

la construcción de modelos de lenguaje se utilizan los corpus monolingües, de manera que la obtención de los datos para el entrenamiento del sistema es relativamente fácil.

Para la construcción de los modelos de traducción, una especie de diccionario bilingüe probabilístico, se utilizan los corpus paralelos. El primer paso en la construcción del modelo de traducción es la alineación por oraciones de los textos fuente y los textos meta. Posteriormente, a partir de las oraciones alineadas se establecen los índices de probabilidad de las alineaciones entre diversas secuencias de palabras o frases. En los primeros modelos de TA estadística (modelos IBM) desarrollados por Brown et al. (1990) en *IBM TJ Watson Research Center* para el sistema *Candide* francés-inglés, la unidad básica de traducción es la palabra. Estos modelos tienen una desventaja importante: la modelación del contexto lingüístico en el que ocurren las unidades de traducción es pobre. Las probabilidades de traducción se calculan sin considerarse las palabras que preceden o que siguen a la palabra traducida. Los sistemas de TA estadística actuales asignan la probabilidad $P(F|E)$ de que una oración traducida E genere la oración original F al considerar el comportamiento de frases, no palabras aisladas (Och et al., 1999; Koehn et al., 2003).

Uno de los problemas de la TA estadística es que los modelos basados en n-gramas no son capaces de dar cuenta de varios fenómenos de naturaleza sintáctica (por ejemplo, las dependencias sintácticas de larga distancia), razón por la que se generan numerosos errores en las traducciones producidas por los sistemas que se basan en dichos modelos. Algunos enfoques recientes toman en consideración el conocimiento sintáctico para resolver este problema, es decir, usan la anotación sintáctica de la LF y la LM para mejorar los modelos probabilísticos de traducción y de lenguaje (Gildea, 2003; Lin, 2004; Marcu et al., 2006, entre otros).

De cara a la necesidad del análisis lingüístico de los errores de TA, cabe mencionar que, a diferencia de los sistemas simbólicos cuya modularidad permite identificar el origen de los errores con relativa facilidad, en el caso de los sistemas estadísticos es difícil relacionar el error con el tratamiento de un fenómeno lingüístico particular, debido a que éste no tiene que ver con las limitaciones del análisis sintáctico o transferencia léxica, sino con el tamaño y representatividad del corpus de entrenamiento y el tipo de modelos de lenguaje y de traducción que se implementan en el sistema.

Otra limitación de la aplicación de los métodos empíricos en el PLN es su dependencia del dominio (tipo, género, temática de los textos). Dado que los parámetros de los modelos probabilísticos se estiman a partir de un corpus de textos que pertenecen a un dominio específico, la calidad de la traducción disminuye drásticamente en el caso de los textos que se encuentran fuera del dominio. Esta carencia de los métodos estadísticos queda comprobada, por ejemplo, en Callison-Burch et al. (2007), quienes discuten varios estudios comparativos de las TAs proporcionadas por diversos sistemas. Los estudios demuestran que el desempeño de los sistemas estadísticos empeora significativamente (de acuerdo con varias métricas de evaluación automática) a la hora de traducir los textos que se encuentran fuera del dominio del corpus de entrenamiento, mientras que los sistemas basados en reglas o los sistemas híbridos presentan el mismo nivel de calidad en este escenario.

La razón es que, mientras que los sistemas basados en reglas se construyen normalmente para su aplicación a los textos del dominio general, los sistemas estadísticos se adaptan a la aplicación y su desempeño es fuertemente restringido por las características del corpus de entrenamiento. El cambio de dominio implica una modificación importante de los patrones textuales (selección del léxico, orden de palabras) y, por tanto, los modelos estadísticos sufren una disminución tanto en la cobertura como en la precisión debido a la presencia de palabras, estructuras sintácticas o semánticas desconocidas o empleadas en un contexto y con un significado distinto. El interés por la adaptación al dominio radica en el hecho de que, si bien hay grandes cantidades de textos paralelos disponibles en formato electrónico, éstos pertenecen a un dominio específico que en numerosas ocasiones no coincide con el dominio de la aplicación del sistema. Para una revisión de la investigación reciente sobre este problema, véase, por ejemplo, Koehn y Schroeder (2007).

Cabe mencionar que la idea de considerar las THs como un tipo textual con rasgos lingüísticos sistemáticos e identificables, que desarrollamos en la presente investigación, ya ha sido implementada en el contexto de la TA para el desarrollo de sistemas estadísticos por medio del uso de las técnicas de adaptación al dominio para la integración del conocimiento sobre la dirección de la traducción en el corpus de entrenamiento (véase Lembersky et al., 2012).

2.4. Evaluación de la traducción automática

La evaluación es una parte importante del desarrollo de sistemas de TA. En palabras de Giménez (2008: 23):

> Automatic evaluation methods play […] a very important role in the context of MT system development. Indeed, evaluation methods are not only important but they are also an upper bound on the attainable success of the development process itself. In other words, improvements may take place as long as developers count on mechanisms to measure them.

En primer lugar, la evaluación es necesaria para realizar el análisis de errores y permite detectar las carencias de los sistemas. En segundo lugar, se usa para la comparación de diferentes versiones del mismo sistema o de diferentes sistemas de TA. Al comparar el desempeño de los sistemas basados en estrategias diferentes, es posible entender cuáles son las ventajas y desventajas de los métodos empleados y lograr una combinación eficiente de los mismos. Finalmente, la evaluación (en este caso únicamente la evaluación automática) se emplea para la optimización de sistemas de TA estadística. La optimización se lleva a cabo por medio del ajuste de parámetros de los modelos probabilísticos guiado por la maximización de la calidad de las traducciones en términos de una métrica de evaluación determinada.

La evaluación, al igual que la TA misma, es una tarea muy compleja y controversial. Desde que existe la TA, existe la problemática de su evaluación. Sin embargo, los criterios para medir la calidad de la TA aún no están claramente establecidos. A continuación discutimos algunos problemas intrínsecos de la evaluación de la TA.

La producción lingüística en el contexto de la traducción se ve restringida, por un lado, por el contenido del TO y, por otro lado, por los recursos de la LM. Sin embargo, a partir de estas restricciones es imposible definir una única solución a la tarea traductora. Así, la traducción, tanto automática como humana, es una tarea abierta, en la que existe un gran número de variantes potencialmente "aceptables"; en palabras de Giménez (2008: 15): "[…] Machine Translation is an open NLP task. Given a certain input, the set of solutions is not closed; every human subject could potentially produce a different translation, and all of them could be in principle equally valid".

Además, la traducción como objeto de evaluación tiene varios aspectos (fidelidad, aceptabilidad, estilo, etc.) y puede evaluarse a distintos niveles de la lengua (léxico, sintáctico, semántico, discursivo), con lo cual el efecto global de los resultados adquiridos con base en diversos criterios en la calidad global de la traducción es difícil de ponderar. La mayoría de las métricas de evaluación automática existentes se centran en aspectos lingüísticos específicos, los cuales tomados por separado, no suelen tener alta correlación con los juicios humanos. La importancia de estos aspectos suele cambiar en función de las lenguas, géneros y sistemas de TA, de ahí la inestabilidad de las métricas existentes. Por ejemplo, Gamon et al. (2005) proponen una métrica de evaluación que toma en cuenta únicamente el aspecto de aceptabilidad, y afirman que es adecuada para detectar las oraciones agramaticales producidas por un sistema de TA basado en ejemplos. Sin embargo, probablemente, no daría buenos resultados si se aplicara para la evaluación de un sistema estadístico.

Así, otra dificultad que se presenta a la hora de realizar la evaluación automática de la TA es que los sistemas se basan en principios y estrategias diferentes y una métrica de evaluación que tiene una alta correlación con los juicios de usuarios para un tipo de sistemas puede no tenerla para los sistemas de otro tipo. Por ejemplo, como se verá a continuación, las métricas de evaluación automática más utilizadas en la actualidad (métricas basadas en n-gramas) favorecen los sistemas estadísticos frente a la TA basada en reglas.

2.4.1. Evaluación manual

Existen dos maneras de evaluar la TA, la evaluación manual y la evaluación automática. La evaluación manual se realiza con base en los juicios de usuarios en relación con dos aspectos fundamentales de la calidad de la traducción: fidelidad al TO y aceptabilidad en la LM (White et al., 1994). Un esquema tradicional de evaluación manual es el esquema propuesto por *Automatic Language Processing Advisory Committee* [ALPAC]. Esta comisión se formó en 1964 en Estados Unidos en el marco de la investigación sobre la TA con el objetivo de determinar las perspectivas del desarrollo en el ámbito; el famoso reporte emitido por ALPAC en 1966 prácticamente condenó la TA, concluyendo que es una tarea inútil: más lenta, menos precisa y mucho más costosa que la TH (conclusión que ha sido desmentida en los años posteriores gracias al desarrollo

de la lingüística teórica y de las ciencias de la computación). El reporte se basó en una evaluación manual a partir de dos criterios: fidelidad (cuánta información del original se preservó en la traducción), e inteligibilidad (qué tan comprensible era la TA para el usuario).

Otra metodología de gran influencia fue desarrollada en el *Advanced Research Projects Agency* [ARPA] (White et al., 1994). Este marco de evaluación incluye las siguientes métricas: comprensión del TT, evaluación por traductores profesionales con un análisis de errores detallado, evaluación basada en los criterios de fidelidad y aceptabilidad, y, finalmente, comparación de TAs producidas por varios sistemas ("traducción preferida"). Un ejemplo de la evaluación de la TA con los criterios de fidelidad y aceptabilidad se ofrece en la Tabla 1.

Score	**Adequacy**	**Fluency**
5	All information	Flawless English
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensible

Tabla 1. Propuesta de evaluación manual según los criterios de fidelidad y aceptabilidad desarrollada por ARPA (White et al., 1994)

La evaluación manual es costosa, lenta, subjetiva, no reutilizable y parcial. No permite evaluar regularmente las mejoras introducidas en el sistema durante el ciclo del desarrollo. Por tanto, en los años recientes ha habido numerosos intentos de automatizar el proceso de evaluación. La evaluación automática en oposición a la evaluación manual es menos costosa, más objetiva y reutilizable.

2.4.2. Evaluación automática

Las métricas de evaluación automática parten del supuesto de que la TA sería una tarea resuelta si fuera imposible distinguirla de la TH. Así, de acuerdo con Corston-Oliver et al. (2001: 140), "Machine translation might be considered a solved problem should it ever become impossible to distinguish automated output from human translation". Esta idea tiene sus raíces en el campo de la inteligencia artificial. En el famoso artículo de

Alan Turing *Computing Machinery and Intelligence* (Turing, 1950) se discute la posibilidad de que las máquinas sean capaces de pensar. Turing (1950) afirma que, si un sujeto no pudiera distinguir entre el comportamiento del ser humano y el comportamiento de una máquina, ésta podría considerarse inteligente.

La evaluación automática consiste en la medición de similitud entre la TA y una o varias THRs. El escenario clásico de la evaluación automática es el *test suite*, un conjunto de *test cases*, cada uno de los cuales se compone de una oración original, su TA y una o varias THRs. Las métricas de evaluación existentes difieren en el tipo de comparación TA-THR.

2.4.2.1. Métricas basadas en n-gramas

Las métricas de evaluación automática más utilizadas en la actualidad (Doddington, 2002; Papineni et al., 2001; Lin y Och, 2004; Melamed et al., 2003) se denominan métricas basadas en n-gramas y calculan la similitud entre la TA y la THR en términos de coocurrencia de n-gramas de palabras, con lo cual evalúan la calidad de la TA únicamente en función de la coaparición de unidades del léxico. Así, se supone que "[…] the more shared sub-strings the MT sentence has with the human reference translation, the better the translation is" (Gamon, 2005: 103).

BLEU (Papineni et al., 2001) es la métrica más utilizada para la comparación y el desarrollo de sistemas de TA. La puntuación que proporciona esta métrica para medir la calidad global de la TA se calcula como la media geométrica de la precisión de los n-gramas (n=1...4) multiplicada por una penalización de brevedad. La precisión de los n-gramas se calcula dividiendo el número de n-gramas de la oración de la TA que aparecen en alguna de las THRs entre el número total de palabras de la TA. La evaluación se realiza oración por oración y después, para obtener la puntuación global a nivel del corpus, BLEU calcula la suma de los resultados para cada oración y la divide entre el número total de oraciones.

En los últimos años la evaluación con BLEU ha guiado el desarrollo de los sistemas estadísticos de TA, puesto que la evaluación de los cambios incrementales en el sistema y la optimización de los parámetros se hacen en numerosas ocasiones con base en los resultados ofrecidos por esta métrica. Asimismo, ha sido la medida elegida para

comparar diferentes sistemas de TA en campañas de evaluación importantes como las organizadas por NIST (*National Institute of Standards and Technology*) (Le y Przybocki, 2005). A la vez que su uso se ha generalizado, han ido surgiendo serias dudas en torno a la fiabilidad de los resultados ofrecidos por BLEU y el resto de las métricas basadas en n-gramas.

Con respecto a los problemas fundamentales de este tipo de evaluación, en primer lugar, se encuentra la disponibilidad y la representatividad de las THRs. Una sola THR no es suficiente para medir la calidad de la TA, ya que puede haber otra (u otras) traducción adecuada. Por ello, la solución es usar un conjunto de THRs (Popescu-Belis, King y Benantar, 2002), si bien en la mayoría de los casos sólo se tiene disponible una única traducción de referencia (Giménez y Amigó, 2006).

En segundo lugar, la evaluación basada en coocurrencia de n-gramas es parcial, ya que considera únicamente el aspecto léxico de la traducción. Mientras tanto, la similitud léxica no es una condición necesaria ni suficiente para que la TA se asemeje a la TH en términos de calidad: "[…] lexical similarity is not a sufficient neither a necessary condition so that two sentences convey the same meaning" (Giménez, 2008: 30). Por ejemplo, Charniak et al. (2003), quienes trabajaron en la construcción de modelos de lenguaje enriquecidos con el conocimiento sintáctico, reportan una mejora significativa en el desempeño del sistema de acuerdo con la evaluación manual. Mientras tanto, según los resultados de BLEU, la calidad de la traducción bajó un 30%. Así, el uso de las métricas basadas en n-gramas para la optimización de sistemas de TA resulta en un énfasis exagerado con respecto al nivel léxico de la traducción, al tiempo que se descuidan otros aspectos importantes.

En tercer lugar, las métricas basadas en n-gramas favorecen los sistemas estadísticos, ya que, de acuerdo con Coughlin (2003: 25), "[…] statistical systems are likelier to match the sublanguage (e.g. lexical choice and order) represented by the set of reference translations".

En cuarto lugar, las métricas basadas en n-gramas tienen un desempeño pobre a nivel de la oración: la evaluación a nivel del corpus neutraliza el "ruido" introducido por las métricas, y corresponde bien con los juicios humanos, pero es inútil para el análisis de

errores. La puntuación de BLEU es 0 si no hay al menos un 4-grama que coincida entre la TA y la THR. De manera que una puntuación alta en términos de BLEU probablemente es un indicio fiable de la calidad de la traducción, pero una puntuación baja no necesariamente implica que la traducción sea de baja calidad: "Although high n-gram scores are indicative of high translation quality, low n-gram scores are not necessarily indicative of poor translation quality" (Giménez, 2008: 30).

En quinto lugar, las métricas basadas en n-gramas presentan una falta de transparencia en los resultados de la evaluación. Los resultados que ofrecen estas métricas no tienen una clara interpretación. Así, de acuerdo con Turian et al. (2003: 1): "although the BLEU and NIST measures might be useful for comparing the relative quality of different MT outputs, it is difficult to gain insight from such measure. What does a BLEU score of 0.016 mean?".

Finalmente, las métricas de evaluación automática basadas en n-gramas no dan cuenta de la equivalencia de sentido, ya que penalizan cualquier tipo de diferencia entre la TA y la THR de la misma manera, sin hacer la distinción necesaria entre la variación aceptable (alternancia en la estructura sintáctica, sinonimia léxica y todo tipo de paráfrasis) y las divergencias que afectan la calidad de la traducción. Es difícil medir la equivalencia semántico-pragmática entre la TA y la THR a partir de una representación lingüística superficial. El hecho de que las lenguas permitan expresar el mismo contenido semántico de muchas maneras diferentes (estructura sintáctica, sinonimia léxica, todo tipo de paráfrasis) invalida la evaluación basada en una comparación de rasgos de superficie. En opinión de Padó et al. (2009), ésta es la razón fundamental por la que los resultados de la evaluación automática no se corresponden de manera estable con los juicios humanos.

Por esta misma razón, las métricas basadas en n-gramas no son capaces de distinguir entre la TA y la TH. En un estudio dedicado a la meta-evaluación, Amigó et al. (2006) indican que de acuerdo con la evaluación de BLEU, una TA puede llegar a tener más similitud con una de las THRs que las THRs entre ellas. En otras palabras: "… standard metrics are unable to identify the features that characterize human translations (as opposed to automatic ones)" (Amigó et al., 2006: 22).

Para ilustrar las limitaciones arriba mencionadas ofrecemos un ejemplo de la evaluación con BLEU de las traducciones de una oración de nuestro corpus producidas por los sistemas *Google* y *Lucy*.

TO	For example, restriction endonucleases will «search» a strand of DNA for a predetermined sequence of bases and, when found, will cut the molecule into two parts.	
THR	Por ejemplo, las endonucleasas de restricción «exploran» una hebra de ADN en busca de una secuencia determinada de bases; hallada, seccionan la molécula por ese punto, dividiéndola en dos piezas.	
		BLEU
Google	Por ejemplo, las endonucleasas de restricción se "búsqueda" una hebra de ADN para una secuencia determinada de bases y, cuando se encuentra, se cortó la molécula en dos piezas.	0.42
Lucy	Por ejemplo, endonucleases de limitación «registrarán» un hilo de ADN en busca de una secuencia predeterminada de bases y, cuándo encontradas, cortarán la molécula en dos partes.	0.25

Tabla 2. Ejemplo de evaluación de las TAs de *Google* y de *Lucy* con BLEU

En este ejemplo observamos que la TA de *Google* comparte un mayor número de n-gramas de palabras con la THR que la TA de *Lucy*, por lo tanto la TA de *Google* obtiene una puntuación más alta (0.42) que la TA del sistema basado en reglas (0.25). No obstante, la oración producida por *Google* es incomprensible, mientras que la traducción de *Lucy* es adecuada y aceptable de cara a la semántica de la oración original y a la gramática de la LM.

A pesar de los recientes avances en el área de evaluación, y de las numerosas críticas que se han hecho con respecto a las métricas basadas en n-gramas, BLEU sigue siendo la métrica que se usa por defecto en numerosas campañas de evaluación de la TA: "BLEU is still the most widely used measure by most MT research groups" (Farrús et al., 2010: 52). Con todo, ya ha habido varias propuestas encaminadas hacia una evaluación más ilustrativa en cuanto a la naturaleza de las diferencias entre la TA y la TH. A continuación, discutimos algunas propuestas encaminadas a solventar las carencias de las métricas basadas en n-gramas.

2.4.2.2. Métricas basadas en el análisis lingüístico automático

Una posible optimización de cara a la problemática que acabamos de presentar es la modelación de la variación admisible que preserva el sentido para la comparación TA-THR. Por un lado, existen métricas de evaluación que toman en cuenta la sinonimia léxica. Por ejemplo, *Meteor* es una medida basada en la alineación por unigramas (Banerjee y Lavie, 2005) que incluye un índice de fragmentación calculado con base en el orden de palabras y además permite considerar la sinonimia por medio de la consulta automática de WordNet[9]. Sin embargo, las variantes ofrecidas por el sistema y por el traductor humano no siempre van a ser "sinónimos" a nivel de la unidad léxica.

Asimismo, ya ha habido propuestas encaminadas a la consideración de las estructuras parafrásticas a la hora de comparar la TA y la THR. Por ejemplo, Russo-Lassner et al. (2005) desarrollan un método de evaluación que realiza la comparación por medio de la detección de estructuras parafrásticas. Owczarzak et al. (2006) proponen un método que deriva oraciones parafrásticas automáticamente a partir de la oración original y la THR. A continuación, las oraciones derivadas se utilizan como referencias adicionales para la evaluación de la TA.

La segunda posible dirección de la optimización de las métricas existentes consiste en realizar la comparación TA-THR a partir de una representación abstracta. En este sentido cabe mencionar los trabajos de Liu y Gildea (2005), y de Giménez y Márquez (2009). Ambos estudios mantienen el escenario clásico de la evaluación (medición de similitud TA-THR), pero, a fin de superar las limitaciones de la comparación a nivel léxico, introducen otros tipos de información. Liu y Gildea (2005) utilizan la información sintáctica para medir la similitud entre las TAs y las THRs. Giménez y Márquez (2009) se basan en la *Discourse Representation Theory* desarrollada por Kamp (1981) para capturar las diferencias de las TAs y las THRs a nivel semántico. Es interesante la relación que estos trabajos establecen entre los niveles de la lengua considerados y los aspectos de calidad de la traducción. La comparación TA-THR a nivel de la estructura sintáctica permite capturar la inteligibilidad del TT, mientras que el léxico se relaciona con el aspecto de fidelidad. Debido a estas correlaciones, Giménez y Márquez (2008) proponen un marco general para combinar las métricas de evaluación

[9] WordNet (Fellbaum, 1998) es una base de datos léxica electrónica desarrollada por la Universidad de Princeton, que ha servido como recurso léxico-semántico en numerosas aplicaciones de PLN.

a distintos niveles lingüísticos y así lograr una evaluación que cubra diversos aspectos de la calidad.

En la misma línea, Padó et al. (2009) desarrollan una propuesta encaminada hacia una combinación de diversas medidas enfocadas a aspectos particulares de la calidad de la TA con el objetivo de lograr una evaluación menos parcial. Tanto Giménez y Márquez (2008) como Padó et al. (2009) abogan por utilizar varias dimensiones del conocimiento lingüístico. La diferencia es que Giménez y Márquez (2008) usan un proceso *bottom-up* basado en un conjunto de métricas independientes y "homogéneas", cada una de las cuales mide la similitud entre la TA y la THR en un nivel lingüístico. Padó et al. (2009), en cambio, proponen un acercamiento *top-down* partiendo del paradigma de detección de entrañamiento semántico. Así, Padó et al. (2009: 304) afirman que:

> The most comparable work to ours is Giménez and Márquez (2008). Our results agree on the crucial point that the use of a wide range of linguistic knowledge in MT evaluation is desirable and important. However, Giménez and Márquez advocate the use of bottom-up development process that builds on a set of "heterogeneous", independent metrics each of which measures overlap with respect to one linguistic level. In contrast, our aim is to provide a "top-down", integrated motivation for the features we integrate through the textual entailment recognition paradigm.

El objetivo de la métrica que proponen estos autores es modelar la calidad de la TA en términos de la equivalencia semántica entre la TA y la THR. La hipótesis es que la evaluación automática que se basa en el modelo de calidad de traducción orientado a la equivalencia semántica entre la TA y la THR hará una predicción acertada sobre los juicios de los evaluadores humanos.

La propuesta de Padó et al. (2009) se enmarca en el campo de reconocimiento de entrañamiento semántico (*textual entailment recognition*). El término "entrañamiento semántico" en PLN fue introducido por Dagan et al. (2005). Esta noción se define a partir de los patrones de razonamiento de sentido común (*common sense reasoning*) más que de la noción teórica de entrañamiento proveniente de la semántica lógica. Padó et al. (2009: 298) define entrañamiento semántico como "[…] a relation between two natural language sentences (a premise P and a hypothesis H) that holds if a human reading P would infer that H is most likely true".

Así la evaluación se interpreta como una tarea de reconocimiento automático de entrañamiento semántico entre la TA y la THR. Estas dos tareas son similares, ya que ambas deben distinguir entre variación aceptable vs. inaceptable para determinar si dos oraciones transmiten la misma información: "A good translation candidate *means* the same thing as the reference translation regardless of the formulation" (Padó, 2009: 297). Los rasgos lingüísticos utilizados por el sistema de identificación automática de entrañamiento semántico están diseñados para detectar la variación que preserva el sentido y, por tanto, se consideran adecuados para evaluar la calidad de la TA.

La métrica desarrollada por Padó et al. (2009) utiliza el sistema de reconocimiento de entrañamiento semántico *Stanford entailment recognition system* (MacCartney et al., 2006). Este sistema recibe un par de oraciones, premisa e hipótesis, y realiza el procesamiento en tres etapas: análisis lingüístico, alineación de las representaciones de la estructura sintáctica de dependencias entre la premisa y la hipótesis y, finalmente, el análisis inferencial destinado para la detección del entrañamiento semántico.

En la fase de análisis se construyen las representaciones lingüísticas lo más completas posible en cuanto al contenido semántico de los fragmentos. La representación lingüística en el *Stanford entailment recognition system* tiene la forma de grafos de dependencias (*dependency graphs*) en los que cada palabra está representada con un nodo, y los arcos están marcados con las relaciones sintácticas. En la fase de alineación cada nodo del grafo de la hipótesis se relaciona con uno o cero nodos de la premisa (se calculan los índices de alineación y se selecciona el mapeo que tiene el índice más alto). El índice global de alineación es una suma de los índices de los nodos y los arcos. Para calcular el índice de alineación a nivel del nodo se mide la similitud léxica a partir de una combinación de diversos recursos, entre ellos, *WordNet*, *InfoMap*, *Dekang Lin´s thesaurus*.

En la última fase, se determina si la premisa entraña la hipótesis. Para ello, varios fenómenos sintácticos, semánticos y léxicos se modelan por medio de rasgos lingüísticos. Para cada par de oraciones alineadas (premisa - hipótesis) el sistema genera más de 100 rasgos que definen las coincidencias o las divergencias sintácticas y semánticas entre la premisa y la hipótesis. Los rasgos se agrupan en cinco clases: rasgos de alineación (calidad de la alineación), rasgos de compatibilidad semántica (polaridad,

antonimia, alcance de cuantificadores, aspecto, modo y tiempo verbal, etc.), rasgos de adición/omisión (adjuntos y aposiciones no alineados), rasgos de referencia (fechas, entidades, locativos) y rasgos de estructura (análisis de la estructura argumental). Estos rasgos se extraen a partir de los análisis realizados en las etapas anteriores. El umbral entrañamiento vs. no-entrañamiento puede ser establecido por el usuario.

Para adaptar este sistema de reconocimiento de entrañamiento semántico a la tarea de evaluación de la TA se realizan las modificaciones siguientes. En primer lugar, la detección de entrañamiento es una tarea asimétrica. No obstante, en el caso de evaluación de la TA, la relación de entrañamiento semántico debe mantenerse en ambas direcciones. Dadas las características de las oraciones producidas por los sistemas de TA, es conveniente tratar la TA como premisa que debe entrañar la hipótesis (THR).

En segundo lugar, *Stanford entailment recognition system* fue desarrollado para procesar oraciones bien formadas, que definitivamente no es el caso de la TA. Los autores combinan los rasgos del sistema con las puntuaciones obtenidas con ayuda de las métricas tradicionales de evaluación automática (BLEU-4, NIST y TER).

La medida en la que la TA entraña la THR proporciona una aproximación muy valiosa al aspecto de fidelidad de la traducción. La propuesta va más allá de la detección de la similitud léxica integrando aspectos composicionales de la equivalencia semántica (paráfrasis multi-palabra, relaciones argumentales y de modificación, orden de palabras y de constituyentes).

Padó et al. (2008) reportan un estudio piloto que tiene por objetivo comprobar la aplicabilidad de la técnica de detección automática de entrañamiento lógico a la evaluación de la TA. El estudio demuestra que el *Stanford entailment recognition system* (MacCartney et al., 2006) utilizado como un sistema de evaluación automática de TA da resultados muy prometedores incluso con una mínima adaptación a la nueva tarea, ya que supera las métricas estándar de evaluación automática.

2.4.2.3. Métricas basadas en la clasificación automática

Como se ha mencionado anteriormente, desde el punto de vista de la evaluación automática, el objetivo de la TA es proporcionar una traducción que sea indistinguible

de la TH. Por este motivo, otra posibilidad interesante para evaluar los sistemas de TA es por medio de la clasificación textual. Así, dado un algoritmo de clasificación que está entrenado con base en un corpus grande de traducciones humanas y automáticas para distinguir entre traducciones producidas por humano y por máquina, la probabilidad de la pertenencia de la oración objeto de evaluación al grupo de THs o TAs puede usarse como índice para la evaluación: "the higher the probability that a sentence is human-translated, the better the quality of the sentence" (Gamon et al., 2005: 105). Esta idea se desarrolla en los trabajos de Corston-Oliver et al. (2001), Kulesza y Shieber (2004), y Gamon et al. (2005). A continuación, discutimos con detalle la propuesta de Gamon et al. (2005), quienes toman en consideración los rasgos lingüísticos distintivos de la TA en oposición a la TH para evaluar la calidad de la TA.

Gamon et al. (2005) investigan la posibilidad de evaluar el aspecto de aceptabilidad de la TA en ausencia de THRs. Realizan la evaluación a partir de una combinación de los índices de perplejidad calculados con respecto al modelo de lenguaje construido con un corpus de textos escritos en la LM (los índices de perplejidad reflejan el grado en el que una palabra de la TA es "esperable" de cara al modelo de lenguaje construido), y de los índices generados por un clasificador automático entrenado para distinguir entre la TA y la TH. El clasificador se basa en el algoritmo *Support Vector Machines* [SVM] (Joachims, 1999) (ampliamente usado en la actualidad para la tarea de clasificación textual), y se entrena con rasgos producto de un análisis lingüístico automático. Así, para aplicar el método que proponen Gamon et al. (2005) se necesita un conjunto de oraciones traducidas automáticamente, un conjunto de oraciones en la LM que pertenezcan al mismo tipo de textos (pero no las traducciones de las mismas oraciones originales), y un sistema de análisis lingüístico automático.

En el experimento que llevan a cabo Gamon et al. (2005), los modelos de lenguaje se construyen a partir de 1,566,265 oraciones en francés extraídas de la documentación técnica de *Microsoft*. El clasificador SVM se entrena con 198,771 oraciones traducidas automáticamente del inglés al francés extraídas de *Microssoft Product Support Services Knowledge Base*, y 260,601 oraciones traducidas por humanos, del mismo dominio técnico. Para la selección de los rasgos lingüísticos, Gamon et al. (2005) se basan en un trabajo previo dedicado a la clasificación automática por estilo (Gamon, 2004) y en el etiquetador lingüístico automático del sistema NLPWin. "The use of this particular set

of features […] is motivated by the desire to capture linguistic generalizations that go beyond surface n-gram regularities" (Gamon, 2005: 106). En específico, los rasgos usados por el sistema para la clasificación son:

> a) medidas de extensión (oración, frase nominal, frase adjetival, cláusula subordinada);
> b) frecuencia de palabras funcionales;
> c) frecuencia de lemas de palabras funcionales;
> d) frecuencia de trigramas de etiquetas POS;
> e) producciones sintácticas de tipo NP::DETP:NOUN:PP (frase nominal que se reescribe como frase determinante más nombre más frase preposicional);
> f) rasgos semánticos binarios (por ejemplo, número y persona para nombres y pronombres, tiempo y aspecto para verbos, rasgos de subcategorización para verbos);
> g) relaciones semánticas de modificación (por ejemplo, Def Noun PrepRel indica la presencia del rasgo definido en el nodo nominal que se encuentra en la relación preposicional con el nodo gobernante, o Verb Tsub Noun Tobj Pron que indica un verbo con el sujeto lógico nominal y el objeto lógico pronominal).

Se generan los índices de perplejidad y los índices del clasificador SVM para cada oración traducida automáticamente. Para usar la salida del clasificador como medida de evaluación se parte del supuesto de que, si una oración se clasifica como TH con una alta probabilidad, entonces es más similar a las THs del conjunto de entrenamiento, y por tanto tiene una aceptabilidad alta. En cambio, si la oración se clasifica con alta probabilidad como TA, es probable que tenga cualidades similares a las de las TAs del conjunto de entrenamiento, es decir, aceptabilidad baja. De esta manera, la probabilidad de la asignación de la oración a una u otra clase puede servir como medida de evaluación.

Gamon et al. (2005) usan este método para evaluar el sistema de TA basada en ejemplos que traduce documentación técnica del inglés al francés, con dos escenarios: a) evaluación de la calidad global de la traducción y b) detección automática de casos problemáticos.

Las medidas basadas en el modelo de lenguaje para la LM y en los índices proporcionados por el clasificador SVM se evalúan en términos de su correlación con los juicios humanos, en un conjunto de 500 oraciones evaluadas a partir de los criterios de aceptabilidad y calidad global de la traducción. Dado que la métrica evalúa únicamente el aspecto de aceptabilidad de la TA, Gamon et al. (2005) miden la correlación entre la aceptabilidad y la calidad global de la traducción a partir del corpus de TAs anotados por humanos con estos dos aspectos. De acuerdo con los resultados, la correlación entre la aceptabilidad y la calidad global de la traducción para el sistema de TA basado en ejemplos es bastante alta (0.67).

La precisión de la clasificación automática sobre el test corpus es 77.59%. Los resultados de la investigación indican que para estimar la calidad global de las traducciones los resultados de BLEU tienen una correlación más alta con la evaluación manual, mientras que para la detección automática de las oraciones malformadas (es decir, para la evaluación automática a nivel de la oración) la métrica propuesta por Gamon et al. (2005) supera BLEU considerablemente.

La ventaja de la propuesta de Gamon et al. (2005) es que resuelve el problema de la disponibilidad y representatividad de las THRs, y supone una mejora de la evaluación automática a nivel oracional.

Además, a diferencia de las propuestas discutidas anteriormente (Giménez y Márquez, 2008; Padó et al., 2009), la métrica desarrollada por Gamon et al. (2005), en vez de hacer una comparación directa entre la TA y la respectiva THR, realiza una extracción de patrones lingüísticos que caracterizan las traducciones del sistema en general. Así, la evaluación se basa en una descripción del lenguaje de la TA en oposición al lenguaje de la TH.

La limitación del método consiste en que toma en cuenta únicamente el aspecto de la aceptabilidad, sin considerar la fidelidad de la TA al texto fuente. Sin embargo, de acuerdo con los resultados de la investigación realizada por Gamon et al. (2005: 110):

> In the example-based MT system, dysfluency is a very good indicator of poor overall MT quality. Instances of perfectly well-formed but semantically inadequate sentences are rare

enough in the output of our system to make a fluency-based metric appropriate - an observation that may not hold true for string-based statistical systems.

2.5. Sistemas de traducción automática y métricas de evaluación

En este capítulo hemos presentado las estrategias de la TA basadas en reglas y de la TA estadística. En la Tabla 3 resumimos las ventajas y desventajas de estos tipos de sistemas.

Ventajas	Desventajas
Traducción automática basada en reglas	
– Proporciona una modelación teóricamente fundamentada y eficiente de los fenómenos relacionados con la "competencia" lingüística – El análisis lingüístico de errores con la identificación de su origen es relativamente fácil de realizar – Es independiente del dominio de aplicación – Es adecuada para lenguas con disponibilidad de recursos (corpus lingüísticos) limitada	– Tiene un coste de construcción y optimización elevado (requiere reglas lingüísticas y diccionarios) – Tiene una capacidad pobre de desambiguación – No es robusta – Es dependiente de la lengua
Traducción automática estadística	
– Proporciona una modelación estadísticamente fundamentada y eficiente de los fenómenos relacionados con la "actuación" o el uso de la lengua en contexto – Es capaz de resolver ambigüedades léxicas – Es robusta – Es independiente de la lengua	– No es adecuada para lenguas con disponibilidad de recursos (corpus lingüísticos) limitada – Presenta problemas con lenguas con morfología rica – El análisis de errores es difícil de realizar – Es dependiente del dominio de aplicación

Tabla 3. Ventajas y desventajas de la TA basada en reglas y la TA estadística

Con todo, la manera más eficiente de procesar el lenguaje humano es a partir de una combinación de los métodos estadísticos y los métodos lingüísticos: "Although SMT systems provide, in general, good performance, it has been demonstrated in recent papers that the addition of linguistic information can be highly useful in this kind of systems" (Farrús et al., 2010: 54). Para lograr una combinación eficiente, es necesario evaluar el desempeño de diferentes tipos de sistemas. Por esta razón, la evaluación es un paso fundamental en el desarrollo de la TA.

Las métricas de evaluación automática actuales son muy valiosas para el desarrollo de sistemas de TA. Sin embargo presentan limitaciones importantes:

1) no tienen una correlación estable con los resultados de evaluación manual:

- favorecen los sistemas estadísticos frente a los sistemas basados en reglas;
- tienen un desempeño pobre a nivel de la oración;
- la diferencia en los scores de BLEU no captura la diferencia de calidad entre la traducción automática y la traducción humana;

2) no proporcionan información lingüística sobre la naturaleza de los errores de traducción.

Callison-Burch et al. (2007: 139) afirman que "while automatic measures are an invaluable tool for the day-to-day development of machine translation systems, they are imperfect substitute for human assessment of translation quality".

Las propuestas alternativas de evaluación automática pretenden solventar las carencias de las métricas basadas en n-gramas. Sin embargo, se trata de prototipos experimentales que todavía no pueden usarse ampliamente debido a la falta de recursos para el análisis lingüístico automático.

3. EL LENGUAJE DE LA TRADUCCIÓN HUMANA

Como se ha mencionado en los capítulos anteriores, la evaluación automática de sistemas de TA se basa en la comparación entre la TA y la THR, pero no todas las diferencias detectadas tienen la misma relevancia para medir la calidad de la TA. El problema de ponderación de las diferencias TA-THR puede abarcarse de diversas maneras (tomando en cuenta la sinonimia y la paráfrasis, partiendo de una representación abstracta, midiendo la equivalencia de sentido, etc.; véase apartado 2.4.2.). En la presente investigación lo abordamos desde la perspectiva de los estudios descriptivos de traducción. Con los avances recientes en este campo se ha demostrado que los TTs poseen características lingüísticas inherentes que los distinguen de otros tipos de textos. Estas propiedades distintivas están relacionadas con las diferencias sistémicas y pragmáticas entre las lenguas, las particularidades del proceso traductor como una actividad cognitiva compleja y las condiciones de la mediación cultural. Debido a los factores mencionados, la traducción implica ciertas modificaciones (*translation shifts*) con respecto a la forma y el contenido del TO. Algunas de ellas son obligatorias desde el punto de vista de las relaciones tipológicas entre los sistemas lingüísticos de la LF y la LM, mientras que otras son opcionales, es decir, no son necesarias desde el punto de vista de la gramaticalidad. Las métricas de evaluación automática basadas en n-gramas penalizan de la misma manera las diferencias TA-THR relacionadas con las modificaciones opcionales en la TH y las diferencias que se deben a la falta de modificaciones obligatorias en la TA, lo cual, en nuestra opinión, constituye una limitación importante de este tipo de evaluación.

Teniendo en cuenta dicha observación, ofrecemos una metodología para el análisis lingüístico contrastivo TA-TH relacionando las diferencias observadas con las propiedades de la TH o con las características de los sistemas de TA. Para ello, es preciso conocer los principios que subyacen el funcionamiento de sistemas, que revisamos en el apartado 2.3. De la misma manera, es necesario saber cuáles son las características esenciales del comportamiento de los traductores humanos, tema al que dedicamos el presente capítulo.

Cabe mencionar que la relación entre la TA y los estudios de traducción no ha recibido atención suficiente. Así, para el segundo Congreso Internacional de Traducción

Automática y Teoría de la Traducción (*Second International Workshop on Machine Translation and Translation Theory*) (Hauenschild y Heizmann, 1997), se establecieron los siguientes puntos de partida:

> a) la investigación en TA no se ha interesado por los estudios de traducción, aunque pretende modelar y, parcialmente, substituir la TH;
> b) esta falta de interés tiene un impacto negativo en la teoría y práctica de la TA;
> c) establecer una relación más cercana entre la TA y la traductología será de utilidad para ambas disciplinas y ayudará a encontrar nuevas soluciones para los problemas de la TA.

Por ello, en el presente capítulo revisamos las ideas desarrolladas en el marco del enfoque descriptivo sobre la naturaleza del lenguaje de traducción que, desde nuestro punto de vista, pueden constituir una aportación importante al campo de la TA y a la evaluación automática de sistemas.

3.1. Enfoque descriptivo vs. enfoque contrastivo en los estudios de traducción

A mediados del siglo XX la investigación en torno a la traducción tiene un carácter prescriptivo al ofrecer soluciones a los problemas específicos de la práctica traductora. En los años sesenta, con el desarrollo de la lingüística estructuralista y el generativismo, y con los primeros avances en la TA, los estudios de traducción se integran al campo de la lingüística contrastiva, siendo su tarea principal la identificación de las correspondencias y divergencias entre la LF y la LM a nivel de sistema. En los años ochenta, la traductología comienza a distanciarse de la perspectiva prescriptiva y de la lingüística contrastiva, enfocándose en el texto meta, la cultura de la LM y la función que desempeñan en ésta los TTs.

Dicho cambio de perspectiva se refleja en el famoso artículo de Holmes (1988), que constituye un avance fundamental en la conformación de la traductología como un campo autónomo. Holmes afirma que la traductología es una disciplina empírica que puede organizarse en tres ramas principales: aplicada, teórica y descriptiva. La Figura 3 ilustra la organización del campo de la traductología de acuerdo con Holmes (1988).

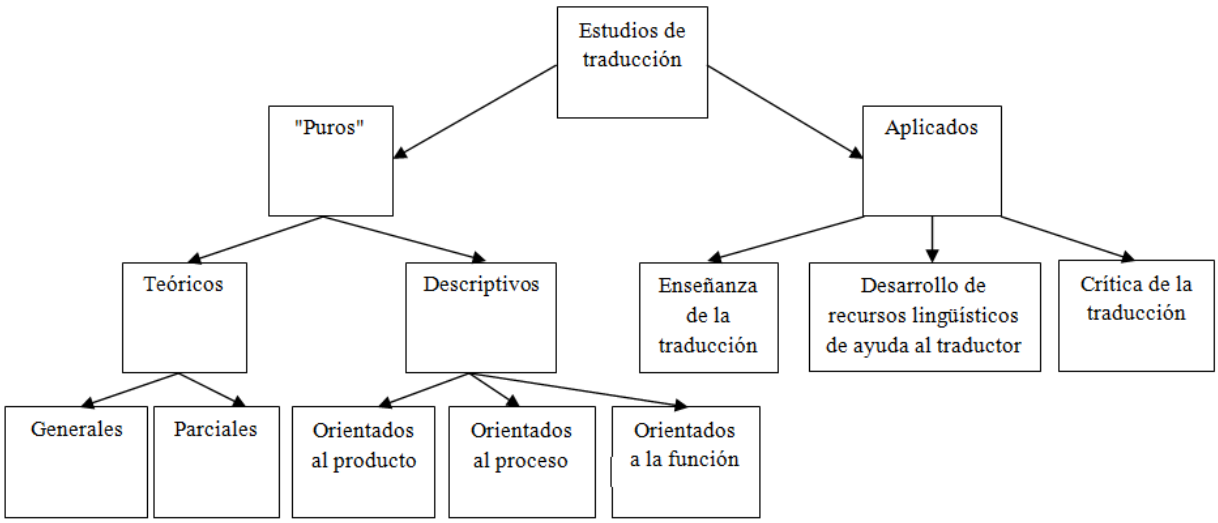

Figura 3. Estudios de traducción (Holmes, 1988)

La vertiente aplicada se concentra en la enseñanza de traducción, preparación de traductores profesionales, etc. El objetivo de los estudios de traducción en su vertiente teórica es establecer principios generales a partir de los cuales se podrían explicar los fenómenos relacionados con el proceso traductor y las características de la traducción como producto. Mientras tanto, la vertiente descriptiva se encarga de describir estos fenómenos y características a partir de la observación de datos empíricos. En el marco de los estudios descriptivos, Holmes (1988) distingue entre los estudios orientados al producto (aquellos que proporcionan una descripción de las traducciones existentes), los estudios orientados al proceso (aquellos que indagan en los procesos cognitivos involucrados en el trabajo del traductor) y los estudios orientados a la función (aquellos que investigan la función de los TTs en el contexto socio-cultural de la LM). Nuestra investigación se ubica en los estudios descriptivos de traducción orientados al producto (Toury, 1995; Baker 1993, 1995, 1996).

Los estudios descriptivos se ocupan de la observación sistemática y objetiva de traducciones reales. De acuerdo con Ulrych y Murphy (2008: 143):

> From the start, DTS [Descriptive Translation Studies] has concerned itself with systematic observation of practical case studies of translations within specific historical and cultural contexts and without any evaluative goals.

Contrariamente al escenario tradicional de la investigación traductológica, en el enfoque descriptivo se toma como punto de partida el TT y no el original (Toury, 1995).

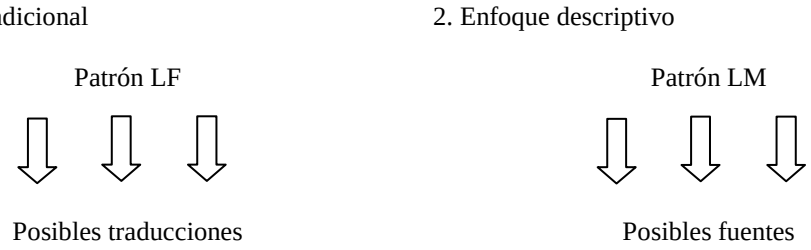

Figura 4. Metodología de investigación en el enfoque tradicional y en el enfoque descriptivo (Nilsson, 2004)

El procedimiento representado en el esquema 1 de la Figura 4 se utiliza para identificar un patrón en la LF y sus posibles traducciones a la LM. La metodología alternativa (esquema 2) tiene por objetivo detectar un patrón en la LM y sus posibles fuentes en la lengua del original; es decir, permite descubrir cuáles son las posturas regularmente adoptadas por los traductores ante un problema específico, y qué factores condicionan sus decisiones y por tanto explican su forma de actuar y los resultados finales de dicha actuación. En palabras de Nilsson (2004: 133):

> […] starting from the TL gives a picture of the multitude types of source text problems that give rise to specific types of translation solutions. In other words, a TL oriented method is well suited to describing what in original texts contributes into giving translated texts certain specific features.

3.1.1. Translationese y universales de traducción

Las regularidades en la selección de recursos por parte del traductor se manifiestan en los rasgos o patrones lingüísticos que distinguen los TTs de otros textos producidos en la LM. Esta observación ha despertado el interés por el estudio de las características distintivas del lenguaje de traducción o *translationese*[10]. Este término, de acuerdo con Puurtinen (2003: 392), se refiere a:

> […] features that distinguish translations from original target language texts; for example lexical items or syntactic structures that are not used in their normal target language functions or whose distribution is unusual.

[10] El término *translationese* fue introducido por Gellerstam (1986), quien investiga las diferencias léxicas entre los textos originalmente escritos en sueco y los textos traducidos al sueco del inglés.

Desde la perspectiva prescriptiva, el *translationese* se considera un fenómeno negativo y se estudia con el fin de enseñar a los traductores las posibles fuentes de la falta de naturalidad en sus textos. Para Duff (1981) el *translationese* es una especie de tercer código que resulta de una imposición de conceptos de una lengua a la otra por medio del uso de estructuras y colocaciones que le son ajenas, y no solamente afecta el estilo y la percepción por parte del receptor, sino que en ocasiones también distorsiona el sentido del texto. En este contexto los rasgos distintivos del lenguaje de traducción se explican únicamente como resultado de la interferencia de la LF, es decir, la influencia de sus características sistémicas, pragmáticas, estilísticas, etc. en el TT, que se pronuncia en la reproducción de la sintaxis y los patrones colocacionales del TO. Newmark (1991) indica que lo que se percibe como *translationese* es a menudo gramatical, pero viola las convenciones de uso de la LM.

En el marco del enfoque descriptivo, el *translationese* se estudia de manera más neutra con el objetivo de descubrir si los TTs difieren sistemáticamente de los originales en términos lingüísticos y qué nos dice esto sobre el proceso de traducción. Así, Puurtinen (2003: 390) afirma que "[…] translated language is a variant worth investigating like any other variant of natural language".

Los factores que afectan la toma de decisiones en el proceso traductor y que se reflejan en las características de los TTs son múltiples y de naturaleza muy variada (precisamente por eso la traducción es tan difícil de modelar y automatizar). En opinión de Duff (1981: 4) la influencia de la LF en el TT es inevitable:

> […] translator cannot help being influenced by the form of the source language. Once thoughts have been given a particular shape – set down in certain words in certain order – it is hard to conceive of them as having different shape.

Pero el grado de interferencia depende de la preparación profesional del traductor o de la estrategia de traducción (normas iniciales, en términos de Toury, 1995) que se adopte para un tipo de textos o una situación comunicativa específica: "the more the make-up of a text is taken as a factor in the formulation of its translation, the more the target text can be expected to show traces of interference" (Toury, 1995: 275).

Con todo, la interferencia no es la única fuente de los rasgos distintivos de los TTs. De hecho, de acuerdo con Toury (1995), existen dos tendencias opuestas en el comportamiento de los traductores: la ley de interferencia, de acuerdo con la cual "in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text" (Toury, 1995: 275) y la ley de estandarización por la que "the special textual relations created in the source text are often replaced by conventional relations in the target text" (Toury, 1995: 267).

Asimismo, la actividad traductora tiene ciertos aspectos inherentes, independientes de la lengua de partida, que afectan la producción lingüística (la comprensión y la interpretación del texto fuente, la situación de mediación cultural en la cual existe una gran distancia entre el autor del original y el receptor de la traducción). En el marco del enfoque descriptivo, Baker (1993, 1995, 1996) ha desarrollado varias hipótesis con respecto a las propiedades básicas de los TTs relacionadas con estos aspectos. Dichas propiedades se denominan universales de traducción y se definen como "features that typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1996: 176). Baker (1993) identifica las siguientes tendencias que podrían considerarse propias del lenguaje de traducción, independientemente de la LF: los TTs tienden a ser más explícitos, más convencionales, menos ambiguos que otros textos producidos en la lengua de llegada. Los traductores tienden a evitar la repetición que ocurre en los originales y a exagerar las convenciones de uso existentes en la LM.

Aparte de la interferencia de la LF, los rasgos del *translationese* más estudiados son la normalización, la simplificación y la explicitación (Baker, 1995).

Normalización (estandarización, en términos de Toury, 1995) se refiere al hecho de que los traductores tienden a adherirse a las reglas gramaticales y a las convenciones de uso de la LM en la misma o en mayor medida que los autores de los originalmente escritos en esta lengua. Así, los TTs muestran una preferencia por los recursos léxicos con menor expresividad o construcciones sintácticas menos marcadas (Vanderauwera, 1985; Shlesinger, 1991).

Simplificación se refiere al hecho de que los traductores tienden a usar un lenguaje más simple que los originales en términos léxicos o sintácticos. La intuición que subyace esta hipótesis es que el traductor simplifica el TO para hacerlo más accesible al receptor. De acuerdo con Baker (1995), dicha propiedad abstracta se refleja en la variedad del vocabulario y la densidad léxica (Laviosa-Braithwaite, 1996). A nivel sintáctico, la simplificación se manifiesta en una preferencia por el uso de construcciones finitas frente a construcciones no finitas o el uso de la coordinación frente a la subordinación (Vanderauwera, 1985). Con respecto a la organización textual, se observa que los traductores tienden a usar oraciones más cortas (Vanderauwera, 1985).

Explicitación significa que los traductores tienden a explicitar la información implícita del texto fuente. Blum-Kulka (1986) estudia el uso de los mecanismos de cohesión en los TTs y descubre que las relaciones discursivas implícitas en los TOs tienden a ser explicitadas en los TTs. A partir de estos resultados Blum-Kulka (1986: 19) formula la hipótesis de explicitación:

> The process of interpretation performed by the translator on the source text might lead to a TL [Target Language] text, which is more redundant than the SL [Source Language] text. […] This argument may be stated as "the explicitation hypothesis" which postulates an observed cohesive explicitness from SL to TL texts regardless of the increase traceable to differences between the two linguistic and textual systems involved. It follows that explicitation is viewed here as inherent in the process of translation.

Asimismo, de acuerdo con Blum-Kulka, la explicitación, posiblemente, es una estrategia universal propia de cualquier tipo de lenguaje de mediación: "it might be the case that explicitation is a universal strategy inherent in the process of language mediation, as practiced by language learners, non-professional translators and professional translators alike" (Blum-Kulka, 1986: 21).

La razón por la que los TTs pueden ser más explícitos que sus respectivos TOs es que, al procesar el texto fuente, el traductor enriquece su interpretación con los significados inferenciales y relaciona los fragmentos procesados con una representación más transparente, literal, congruente. Así, Steiner (2002: 218) afirma que:

> […] human translation should not be seen as a process of directly transferring features or structure on either semantic or lexico-grammatical levels, but rather as a process involving 'understanding' of the source text to a certain depth, and then re-creating that message as fully as possible.

Sin embargo, en opinión de Becher (2011: 26), esta observación no implica que la traducción necesariamente conlleve explicitación, puesto que:

> […] it depends on the assumption that translators directly verbalize their (more explicit) mental representation of the source text without applying operations that might render it more implicit, such as politeness strategies, omission of contextually inferable material, etc. There is no reason why translators – in contrast to authors of non-translated texts – should skip the application of such operations.

Steiner (2002) pone a prueba la hipótesis de explicitación estudiando el uso de la metáfora gramatical (Halliday, 1994) en los TOs y los TTs. Los resultados de su estudio apuntan a que una de las posibles propiedades de los TTs es que presentan menor número de metáforas gramaticales que los textos de partida. Sin embargo, el mismo Steiner (2002: 220) reconoce que las instancias de de-metaforización no están necesariamente condicionadas por el factor cognitivo:

> The reasons could be language-specific (i.e. because of typological-contrastive properties of the languages involved), they could be register-specific (i.e. in cases where the target language and the context suggest a lower degree of metaphoricity) and/or they could have to do with a lack of effort or ability on the part of the translator.

Así, Steiner (2002) propone tres tipos de factores que explican el comportamiento de los traductores: proceso traductor (explicitación, simplificación, normalización), relaciones tipológicas entre los sistemas lingüísticos (divergencias traductológicas; véase apartado 2.2.) y adaptación al dominio. En relación con la adaptación al dominio, Steiner (2002: 217) afirma que:

> An important source of properties of translated text is "register", or, more precisely, the fact that the preferred registers of source text and target text for a given context may or may not be exactly the same, and the translator(s) may have decided to make changes to the register of their target text. An example would be the case where, say, the German translation of an English text

> from the register of popular scientific prose shows properties of backgrounding of interpersonal meanings, which may or may not be the result of a conscious decision by the translators.[11]

En la misma línea, Klaudy (2008) propone las siguientes categorías para clasificar las instancias de explicitación en los TTs (las mismas categorías podrían aplicarse a cualquier otro tipo de modificaciones observadas en los TTs):

> a) explicitación obligatoria (aquélla que se debe a las diferencias léxico-gramaticales entre la LF y la LM);
> b) explicitación opcional (aquélla que se debe a las diferencias en las preferencias estilísticas entre la LF y la LM);
> c) explicitación pragmática (aquélla que se debe a las diferencias en el conocimiento del mundo entre los hablantes de la LF y la LM);
> d) explicitación inherente al proceso traductor (aquélla que se debe a la naturaleza del proceso traductor).

De esta manera, al identificar en el TTs los rasgos distintivos que los hacen diferentes de los originales escritos en la LM y relacionar dichos rasgos con los universales de traducción (perspectiva descriptiva), es preciso estudiar el contexto lingüístico en el que el traductor toma las decisiones que resultan en la conformación de dichos rasgos (perspectiva contrastiva). De acuerdo con Ulrych y Murphy (2008: 146), se trata de enfoques complementarios que estudian el mismo fenómeno desde puntos de vista diferentes:

> […] the CL [Contrastive Linguistics] approach and the translational approach are mirror images of each other […] While linguistics scholars will be looking at translations to see what they tell them about the languages in contact, translation scholars will be searching for ideas and making use of the experience and methodology to be found in linguistics theory to enhance their understanding of the mediation process and translated texts.

3.1.2. Translation shifts

Para estudiar las instancias concretas de la manifestación de los rasgos de los TTs e identificar las fuentes de dichos rasgos, se usa la noción de *translation shift* que, en un

[11] Cabe mencionar que en el caso de nuestro corpus de estudio se observa la misma tendencia: en las THs se produce una supresión de los recursos lingüísticos con las funciones interpersonal y metatextual. Esta tendencia se explica por las diferencias entre las convenciones de escritura científica existentes en inglés y en español.

sentido amplio, se refiere a cualquier tipo de modificación realizada por el traductor con respecto a la forma y el contenido del original. Como hemos visto en Steiner (2002) o Klaudy (2008), las modificaciones pueden clasificarse a un nivel abstracto al relacionarse con las hipótesis sobre los universales de traducción. De la misma manera, la descripción de los *shifts* puede realizarse a partir de un análisis más detallado. A continuación describimos algunas propuestas para la clasificación.

Vinay y Darbelnet (1958) realizan un análisis contrastivo del inglés y francés con el objetivo de describir de manera sistemática las diferencias y ofrecer técnicas para la solución de problemas específicos asociados con este par de lenguas. Desarrollan un sistema de siete procedimientos que se utilizan en la traducción. El préstamo, el calco y la traducción literal se consideran técnicas de "traducción directa", mientras que transposición, modulación, equivalencia y adaptación son "procedimientos oblicuos". Los procedimientos oblicuos (en otras palabras, *translation shifts*, aunque Vinay y Darbelnet (1958) no usan este término) resultan en diferencias observables entre el TO y el TT. Transposición es un cambio de categoría gramatical que no afecta el sentido global de la oración. Modulación, a diferencia de transposición, supone una modificación del significado oracional. Equivalencia conlleva la restructuración completa del mensaje y se reserva para la traducción de clichés, proverbios, etc. Adaptación se emplea cuando la situación descrita en el original es ajena a la cultura de la LM.

Nida (1964) desarrolla una clasificación de técnicas de ajuste (*techniques of adjustment*) que se basan en su idea de equivalencia dinámica: el TT debe producir en su receptor el mismo efecto que produciría en el receptor del TO. De acuerdo con Nida (1964), para lograr la naturalidad de expresión en la LM, se necesitan las técnicas de ajuste, que el autor divide en tres categorías principales: adiciones, sustracciones y alteraciones. Las adiciones se presentan en todos los casos en los que el TT tiene más material lingüístico que el TO. Son aceptables únicamente si se preserva el contenido del original y resultan en la explicitación de la información que está presente en el texto fuente de manera implícita. Por ejemplo, al traducir una construcción pasiva por una construcción activa el traductor se ve obligado a explicitar al participante con la función sintáctica de sujeto. Sustracción es un procedimiento opuesto que conlleva la implicitación de los significados expresados explícitamente en el original. Alteración es una categoría

residual que se reserva para los casos que no pertenecen a ninguna de las categorías anteriores (por ejemplo, cambios del tiempo verbal o cambios a nivel léxico relacionados con las diferencias en la organización del vocabulario en distintas lenguas).

Catford (1965), quien fue el primero en usar el término *translation shift*, lo define sobre la base de su distinción entre la correspondencia formal y la equivalencia textual. La correspondencia formal es una relación abstracta entre dos elementos de sistemas lingüísticos:

> TL category (unit, class, structure, element of structure, etc.) which can be said to occupy, as nearly as possible, the "same" place in the "economy" of the TL as given category occupies in the SL. Since every language is ultimately sui generic – its categories being defined in terms of relations holding within the language itself – it is clear that formal correspondence is nearly always approximate. (Catford, 1965: 27)

Mientras tanto, la equivalencia textual es una relación entre el TO y el TT en una situación particular. Dos elementos se consideran textualmente equivalentes si pueden servir en dos situaciones similares; por tanto, la equivalencia textual cambia en función del contexto.

De esta manera, de acuerdo con Catford (1965), los *translation shifts* se dan si la traducción textualmente equivalente no lo es desde el punto de vista formal, es decir, en los casos en los que la equivalencia textual se consigue a coste de una desviación de la correspondencia formal. Resumimos la clasificación de *translation shifts* desarrollada por Catford (1965) en la Tabla 4:

Tipo	Descripción
Level shift	cambio de nivel: un rasgo gramatical tiene su equivalente a nivel del léxico, o al revés
Category shift:	
- *class*	cambio de categoría gramatical
- *structure*	cambio de la estructura oracional en términos de función gramatical (sujeto, predicado, adjunto, etc.) o de orden de constituyentes
- *unit*	cambio del estatus de los elementos en la estructura de la oración (por ejemplo, a una frase del TO le corresponde una cláusula en el TT)
- *intra-system*	los sistemas de la lengua fuente y la lengua meta poseen elementos que se encuentran en una relación de correspondencia formal, pero el traductor selecciona una unidad diferente (por ejemplo, traduce un nombre en singular por un nombre en plural)

Tabla 4. Clasificación de *translation shifts* de Catford (1965)

Las clasificaciones desarrolladas por Vinay y Darbelnet (1958) y Nida (1988) son de corte prescriptivo y didáctico, ya que proponen técnicas o estrategias que se deben utilizar para evitar errores y lograr una "buena" traducción. La clasificación de Catford (1965) ha sido criticada por su carácter abstracto y teórico. Tal como afirma Cyrus (2009: 90), el problema fundamental de esta clasificación es que:

> it presupposes that it is actually feasible to determine those elements in two linguistic systems that are formal correspondents of each other […] For this reason, Catford's account remains purely theoretical and has never been fully applied to any actual translations.

La ventaja de la clasificación de Catford es que propone una clara definición de la noción de *translation shift* y que su clasificación es sistemática y objetiva, ya que parte de la comparación de formas, con base en la cual pueden hacerse conclusiones de corte semántico o funcional. Además, de acuerdo con Szymanska (2011: 215):

> […] Catford´s distinction between formal correspondence and textual equivalence is a concise way of accounting for the fact that formally similar structures in different languages may have different functional (semantic and pragmatic) values.

El hecho de que estructuras similares tengan valores semánticos o pragmáticos distintos en diferentes lenguas explica las desviaciones del original que se producen en mayor o menor grado en la TH y el carácter literal de la TA.

La idea de la distinción entre la correspondencia formal y la equivalencia textual como base para el estudio de los *translation shifts* se ve reflejada en la propuesta de van

Leuven-Zwart (1989), quien ofrece un análisis detallado de la traducción de la obra de Cervantes, "Don Quijote de la Mancha", al holandés. Los objetivos de su análisis son: a) ofrecer una descripción de las diferencias entre el TO y el TT; b) con base en esta descripción, formular hipótesis con respecto a la interpretación del original por parte del traductor y la estrategia o norma de traducción que subyace el proceso traductor.

La metodología que propone van Leuven-Zwart consta de dos tipos de análisis, micro-estructural y macro-estructural. El análisis micro-estructural de *translation shifts* se realiza a nivel de cláusulas y oraciones, y el análisis macroestructural pretende determinar el impacto de éstos "on the level of characters, events, time, place and other meaningful components of the text" (van Leuven-Zwart, 1989: 154). A continuación ofrecemos una breve descripción del primer tipo de análisis.

El análisis microestructural tiene cuatro etapas. En la primera, se delimitan las unidades de análisis (los "transemas"). A continuación, para cada par de transemas TO-TT se establece un denominador común (el "architransema"). En la tercera fase, se define la relación entre los transemas y el architransema. Finalmente, en función de ésta, se identifican y se clasifican los *translation shifts*.

Los transemas se delimitan con base en los criterios derivados de la Gramática Funcional de Dik (1978), y pueden ser de dos tipos: a) "state of affair transeme" (predicado y sus argumentos); b) "satellite transeme" (especificación o amplificación adverbial de los transemas del primer tipo). El architransema es el denominador común entre el transema del TO y el transema del TT, que se define a través de los aspectos del significado descriptivo o situacional que comparten los transemas. Van Leuven-Zwart (1989: 158) afirma que "[…] practice shows that in most cases an ATR [architranseme] can be identified with the help of a good descriptive dictionary in each of the two languages involved". Si entre el transema y el architransema no se observa ninguna diferencia, la relación entre ellos es sinonímica. En cambio, si difieren en algún aspecto de su significado, entonces se encuentran en una relación de hiponimia. La relación hiponímica entre el transema y el architransema se determina a partir de "form/class/mode formula: 'X is a form/class/mode of Y', in which X stands for the transeme and Y for the ATR" (van Leuven-Zwart, 1989: 159).

Al establecer la relación entre cada uno de los transemas y el architransema, hay cuatro posibilidades en cuanto a las relaciones entre los transemas del TO y el TT: a) si ambos transemas se encuentran en relación sinonímica con el architransema, entonces la relación entre ellos también es sinonímica, de manera que no ocurre ningún *translation shift*; b) si uno de los transemas se encuentra en relación hiponímica con el architransema, y el otro, en relación sinonímica, entonces la relación entre los transemas es hiponímica; c) si la relación entre cada uno de los transemas y el architransema es hiponímica, entre los transemas hay una relación de contraste; d) si no hay ninguna relación entre los transemas (no comparten ningún aspecto de su significado) es imposible identificar el architransema.

A partir de esta taxonomía de las posibles relaciones entre los transemas, se establece una clasificación detallada de las diferencias entre el TO y el TT con tres categorías principales: modulación, modificación y mutación. En el caso de modulación la relación entre los transemas es hiponímica. Si el aspecto disyuntivo se encuentra en el transema del TT, el *translation shift* se clasifica como modulación/especificación y si se ubica en el TO, entonces se trata de modulación/generalización. El aspecto disyuntivo puede encontrarse a nivel semántico o estilístico, lo cual da lugar a cuatro categorías diferentes. En cuanto al aspecto sintáctico, "as every transeme appears in one or another syntactic form and as syntactic form as such can never be absent in transemes, a category such as 'syntactic modulation' does not exist" (van Leuven-Zwart, 1989: 166).

En el caso de modificación, los transemas se encuentran en una relación de contraste. La modificación puede ser semántica, estilística, sintáctico-semántica, sintáctico-estilística o sintáctico-pragmática, en función del aspecto en que difieran los transemas. Los cambios estructurales que no tienen impacto a niveles semántico, estilístico o pragmático no se toman en consideración:

> […] syntactic alterations which have no effect on any of these levels [semantic, stylistic or pragmatic] are not taken into account, as they do not give any indication of the translator's interpretation of the original text or of the strategy underlying the translation (van Leuven-Zwart, 1989: 166).

En cuanto a la mutación, esta categoría se reserva para los casos en los que es imposible establecer un denominador común para dos transemas, y se compone de tres categorías: adición u omisión de cláusulas o frases y cambio radical de significado.

Una diferencia importante entre las propuestas anteriores (Vinay y Darbelnet, 1958; Nida, 1964; Catford, 1965) y el enfoque de van Leuven-Zwart (1989) es que este último está diseñado para la descripción de traducciones reales y no para el análisis de diferencias sistémicas entre las lenguas. Por tanto, su actitud hacia los *translation shifts* difiere de la actitud manifestada en los enfoques tradicionales, en los que los *translation shifts* son aceptables únicamente como "attempts to deal with systemic differences" (Bakker et al., 1998: 226), es decir, en los casos en los que una traducción literal supondría una violación de la norma de la LM. Sin embargo, van Leuven-Zwart (1989) ofrece una clasificación demasiado detallada y ambigua, lo cual hace difícil su implementación para un análisis automático o semiautomático de los TTs.

3.1.3. Análisis del discurso en los estudios de traducción

Las modificaciones que realiza el traductor con respecto al TO se deben a varios tipos de factores: los factores cognitivos relacionados con el proceso traductor, los factores comunicativos involucrados en la mediación cultural y las diferencias sistémicas y pragmáticas entre las lenguas. Estas últimas no se pronuncian solamente a nivel del léxico, sintáctico o semántico. Algunas operan a nivel extra-oracional, hecho que ha recibido atención en los estudios de traducción en el marco del enfoque textual (Hatim y Mason, 1995; Neubert y Shreve, 1992; Larose, 1989; Papegaaij y Schubert, 1988).

El objeto de traducción no son palabras ni oraciones, sino textos. En este contexto cobra importancia el concepto de textualidad, la cual se define como un conjunto de rasgos que debe presentar un texto para ser considerado como tal (Hurtado, 2008). Castellá (1992) identifica tres características fundamentales que cualquier texto debe poseer: adecuación al contexto comunicativo; coherencia de las unidades de información que integra; cohesión de los diversos elementos que lo componen. De esta manera, se podría decir que la textualidad es una propiedad por la cual un texto tiene una continuidad en cuanto al sentido (es coherente) y a los elementos de superficie (está cohesionado), y tiene una articulación de la evolución de la información (progresión temática) (Hurtado, 2008). Así, la coherencia y la cohesión son elementos esenciales de la organización

textual. Cabe señalar que la interpretación y el uso de estos términos varía según los autores: van Dijk (1980) habla de coherencia lineal y coherencia global, ambas de carácter semántico; para Widdowson (1978), la coherencia se basa en la interpretación de los sucesivos actos ilocutivos; Halliday, el principal acuñador del término cohesión, no usa el término coherencia (Halliday y Hasan, 1976). Castellá (1992: 58) señala que estas dos características podrían resumirse en una:

> Un texto es coherente externamente, con el entorno comunicativo, e internamente, con la organización de la información. Siguiendo este planteamiento, la cohesión no sería más que una parte de la coherencia, su concreción en la materialidad lineal de la lengua.

Hurtado (2008), al realizar una revisión detallada de los enfoques traductológicos existentes, afirma que el punto indiscutible en cuanto al análisis del discurso en el contexto de la traducción es que las lenguas difieren en las preferencias por el uso de los recursos para establecer vínculos cohesivos, e incluso el grado general de cohesión puede variar, existiendo lenguas que presentan un mayor grado de cohesión explícita que otras.

En el caso del inglés y del español, Beeby (1996) señala que las diferencias más significativas están relacionadas probablemente con la referencia deíctica, la repetición léxica y el uso de conjunciones argumentativas. En español, el sistema de género es más completo, existe mayor variación pronominal y el sujeto está incorporado al verbo. En cambio, el inglés, al no señalar, generalmente, la concordancia en términos de número y género, utiliza otros medios para clarificar la referencia textual. El orden de palabras y la posición de pronombres son más rígidos en inglés y, a diferencia del español, la repetición léxica se usa ampliamente como mecanismo de cohesión. Asimismo, Beeby (1996) menciona que el inglés tiende a utilizar más conectores argumentativos que reflejan la retórica del texto.

Las diferencias entre las lenguas en cuanto a la organización textual se han estudiado también en relación con los universales de traducción. Por ejemplo, Blum-Kulka (1986) investiga las modificaciones del TO que afectan la cohesión y la coherencia con base en algunos ejemplos de las traducciones del inglés al francés y al hebreo. En su opinión, "the process of translation necessarily entails shifts both in textual and discoursal

relationships" (Blum-Kulka, 1986: 18), lo cual la lleva a postular la hipótesis de explicitación que discutimos en el apartado 3.1.1.

3.2. El uso de corpus en los estudios de traducción

Existe una relación entre el desarrollo del enfoque descriptivo en la traductología y los avances recientes en el ámbito de la lingüística de corpus. Holmes (1988) expresa una preocupación por el uso predominante de la introspección como principal herramienta de investigación en los estudios traductológicos. Toury (1980) afirma que los enfoques predominantes en la traductología no investigan traducciones reales sino que hacen especulación sobre objetos idealizados y, en este sentido, comparte las inquietudes expresadas en el marco de la lingüística de corpus (Sinclair, 1991), donde se pone de relieve la naturaleza complementaria de la introspección y la observación.

La lingüística de corpus representa un acercamiento empírico a la descripción lingüística y aboga por el estudio de instancias concretas del uso de la lengua. Biber et al. (1998) destaca las siguientes características del análisis lingüístico basado en corpus:

> a) es un enfoque empírico que estudia los patrones de uso en textos reales;
> b) utiliza grandes conjuntos de textos compilados de acuerdo con cierto propósito de investigación (corpus lingüísticos);
> c) hace uso de técnicas estadísticas para caracterizar el objeto de estudio en términos de frecuencias de aparición de las unidades de análisis;
> d) combina las técnicas de análisis cuantitativo y cualitativo.

El enfoque basado en corpus es útil de cara al estudio de la traducción, ya que en numerosas ocasiones las decisiones del traductor se ven influenciadas por los factores sutiles relacionados con el uso de la lengua en contexto (Szymánska, 2011). Además, tal como se ha mencionado en el apartado anterior, las características del lenguaje de traducción se manifiestan en la selección particular de recursos lingüísticos, que no necesariamente implica una violación con respecto a las reglas de la gramática de la LM. Así, al estudiar los rasgos gramaticales del *translationese*, Borin y Prütz (2001: 30) afirman que:

> The kind of deviance referred to here [observed in translated texts] is not to be equated with errors in the normal sense. Rather, it should reveal itself in 'odd' choices of lexical items and

> syntactic constructions, which conceivably could be the result of both assimilation and dissimilation with respect to the source language or the source text.

Es decir, los rasgos distintivos del lenguaje de traducción se manifiestan en patrones de uso y, por tanto, deberían investigarse con los métodos de lingüística de corpus. En palabras de Borin y Prütz (2001: 32): "It [translationese] ought to be seen, above all, in deviant patterns of usage, i.e. it should be eminently suited for investigation by the methods of corpus linguistics".

Baker (1993, 1995, 1996) incorpora los métodos y herramientas de lingüística de corpus a la investigación traductológica, conformándose de esta manera una nueva línea de investigación: los estudios de traducción basados en corpus (*Corpus-based Translation Studies*).

De acuerdo con Sierra (2006: 445), el corpus lingüístico se define como "la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos". Existen varios tipos de corpus que se usan en los estudios de traducción para investigar cómo los mismos significados se expresan en distintas lenguas (perspectiva contrastiva) y para identificar las características inherentes del lenguaje de traducción (perspectiva descriptiva).

De acuerdo con Ulrych y Murphy (2008), los corpus orientados a los estudios contrastivos son:

> a) el corpus multilingüe o bilingüe paralelo que contiene los TOs y sus respectivas traducciones; por ejemplo, *Canadian Hansard* (Roukos et al., 1995) o *Europarl* (Koehn, 2005);
> b) el corpus multilingüe o bilingüe comparable que se compone de TOs en distintas lenguas, que son similares con respecto a ciertos factores extralingüísticos (es decir, comparten el tema, el género discursivo, el tipo de texto, etc.); por ejemplo, *The Aarhus corpus of contract law* (Faber y Lauridsen, 1991).

Los corpus bilingües (o multilingües) representan un recurso valioso para los estudios lingüísticos en un contexto multilingüe porque permiten observar las relaciones entre las lenguas a través del uso, y se han utilizado ampliamente en la lexicografía bilingüe, la terminología, la TA, etc. Cabe mencionar que la actitud hacia los corpus paralelos es diferente en los estudios contrastivos y en los estudios traductológicos. Desde el punto de vista de la lingüística contrastiva, los corpus paralelos no pueden considerarse un recurso fiable para el estudio de las lenguas en contacto, ya que en los TTs la expresión lingüística se ve restringida por la influencia de la LM y por otros factores relacionados con la naturaleza del proceso traductor. En palabras de Granger (2003: 20):

> The main drawback of translation corpora, however, is that they often display traces of the source text and therefore cannot really be considered as reliable data as regards the target language, especially in frequency terms.

En cambio, desde el punto de vista de la traductología, los corpus paralelos constituyen una fuente de información relevante para estudiar las modificaciones (tanto obligatorias como opcionales) que realiza el traductor con respecto al TO.

Los corpus orientados a los estudios descriptivos son:

> a) el corpus monolingüe paralelo que proporciona la comparación de dos o más versiones del texto traducido a la misma lengua por diferentes traductores o en diferentes períodos históricos, etc.;
> b) el corpus monolingüe comparable que contiene los TTs de diferentes lenguas, que pueden ser comparados con los TOs escritos originalmente en la lengua de llegada; por ejemplo, *Translational English Corpus* [TEC] (Baker, 1993).

Baker (1993) arguye que este último tipo de corpus aporta información sobre los rasgos inherentes de los TTs, es decir, aquellos patrones lingüísticos que ocurren en los TTs con una frecuencia significativamente distinta frente a la frecuencia de los mismos en otros tipos de textos y que no están relacionados con la influencia de la LF:

> The most important contribution that comparable corpora can make to the discipline is to identify patterning which is specific to translated texts, irrespective of the source or target language involved (Baker, 1995: 234).

Las medidas que se usan en la traductología para la identificación de las características distintivas de los TTs a partir de los corpus comparables son: las medidas de variedad y densidad del vocabulario, la longitud de oraciones y de cláusulas, listas de palabras más frecuentes, la frecuencia de uso de las estructuras sintácticas, la frecuencia de uso de los mecanismos de cohesión (repetición, elipsis, marcadores del discurso).

A pesar de su utilidad para la investigación traductológica, la metodología de investigación del enfoque descriptivo basado en el corpus comparable presenta ciertas limitaciones importantes. En primer lugar, al realizar la comprobación de las hipótesis sobre los rasgos universales de traducción (explicitación, simplificación, etc.) no siempre se logra solventar la distancia entre dichas propiedades abstractas y la realización concreta de éstas en la superficie textual:

> Counting words with or without lemmatization, words per sentence, percentages of lexical words vs. function words, type-token ratios for lexical items, number of realized vs. non-realized conjunctions (e.g. *that* in English) and some other phenomena of this order of concreteness is methodologically too far removed from the level at which relevant models of textuality are formulated. This is what we are referring to as the 'gap' between the data in electronic corpora and the level at which hypotheses are formulated (Steiner, 2002: 215).

En segundo lugar, a la hora de construir el corpus comparable es difícil evitar el impacto de los factores externos (género discursivo, tipo de texto, tipo del lector, políticas editoriales, etc.) en las características de los TTs y los textos originalmente escritos en la lengua de llegada (Bernardini y Zanettin, 2004).

Por último, tal como se ha mencionado en el apartado anterior, los resultados de la investigación de los rasgos lingüísticos que se consideran manifestaciones de las tendencias universales de la traducción son poco fiables si no se realiza un estudio cualitativo que tenga en cuenta los TOs. Así, de acuerdo con Steiner (2002: 216):

> Monolingual (comparable corpora) allowing comparisons between translated and non-translated texts in one language are valid, and even necessary, empirical bases for investigations of properties of translated texts. But they cannot be the only basis if we want to broaden our investigation in the direction of taking into account the different sources - and in that sense, explanations - of those properties.

Antes de relacionar la cantidad de información explicitada en los TTs con una tendencia universal de la traducción, deben estudiarse las instancias concretas de explicitación / implicitación con el fin de averiguar si son resultado de modificaciones obligatorias u opcionales. Por ejemplo, el hecho de que en la traducción se realice la explicitación de uno de los participantes de la oración, puede deberse a las diferencias en los tipos de transitividad de los verbos en la LF y la LM. Otro ejemplo de explicitación obligatoria es la traducción del inglés al español del verbo copulativo *to be* como "ser" o "estar".

3.2.1. Técnicas estilométricas

El lenguaje de traducción constituye una variante estilística particular de la lengua. En palabras de Lembersky et al. (2012: 255):

> Known as *translationese*, translated texts (in any language) constitute a genre, or a dialect, of the target language, which reflects both artifacts of the translation process and traces of the original language from which the texts were translated.

Por esta razón, la identificación de los rasgos distintivos del *translationese* puede considerarse como una tarea de estilometría o estilística estadística. La estilometría es un ámbito de la lingüística aplicada que hace uso de técnicas estadísticas para la medición de las características estilísticas de los textos con el fin de su sistematización y clasificación. En los últimos años se han realizado numerosos estudios enfocados a la comparación o clasificación de textos según diversos criterios (tema, género, tipo textual, autoría, etc.) (véase Santini, 2004, para una revisión detallada), entre ellos TO vs. TT (Borin y Prütz, 2001; Koppel y Ordan, 2001; Baroni y Bernardini, 2006; van Halteren, 2008; Kurokawa et al., 2009; Lembersky et al., 2011; Lembersky et al., 2012).

Borin y Prütz (2001) estudian el aspecto gramatical del *translationese* con base en una representación superficial de las características sintácticas del texto, comparando la distribución de secuencias de categorías gramaticales (n-gramas de etiquetas POS) en los TTs y en los textos originalmente escritos en la LM.

Borin y Prütz (2001) usan el corpus creado en el marco del proyecto ETAP, que contiene los artículos de prensa originales en sueco y sus traducciones al inglés, así como dos corpus monolingües en inglés, *Flob* (corpus del inglés británico compilado en la década de los noventa en la Universidad de Freiburg) y *Frown* (una versión renovada

del Brown corpus del inglés americano desarrollado también en la Universidad de Freiburg). Los textos de los corpus se etiquetan con el etiquetador estadístico Brill (Brill, 1995). A continuación, se extraen las frecuencias de aparición de todos los unigramas, bigramas y trigramas de etiquetas POS en los subcorpus de los TTs y de los textos originalmente escritos en inglés. Para identificar los n-gramas cuya distribución es significativamente distinta en los TTs (aquellos que están sobre- o sub-representados en los TTs con respecto a los originales), Borin y Prütz (2001) usan el test estadístico de Mann-Whitney (Kilgarriff, 2001).

Los hallazgos del estudio indican que la distribución de n-gramas de etiquetas POS es un indicador útil de los efectos del *translationese* a nivel sintáctico. Los autores identifican varios rasgos del lenguaje de traducción relacionados con la interferencia de la LF. Por ejemplo, la interferencia del sueco en las traducciones al inglés se manifiesta en la sobre-representación de preposiciones y adverbios en la posición inicial de la oración. En sueco, las frases preposicionales o adverbiales se encuentran frecuentemente en esta posición, mientras que en inglés tal configuración es mucho menos usual.

En algunos casos las diferencias en la distribución de n-gramas no se deben a los efectos del *translationese,* sino a las diferencias en el tipo textual y a los principios de compilación de los corpus utilizados, lo cual, como se ha mencionado anteriormente, constituye una limitación importante de los estudios basados en corpus comparables. Por ejemplo, la distribución de algunos n-gramas indica que existen más oraciones que empiezan con un verbo en el subcorpus de los textos meta; el análisis manual de ejemplos concretos demuestra que ello se debe a que el subcorpus de los TTs incluye un número elevado de las cartas al editor que contienen numerosas oraciones interrogativas. De esta manera, para investigar los rasgos distintivos del lenguaje de traducción es necesario realizar, además del análisis cuantitativo, un análisis cualitativo con la exploración de los contextos de aparición de los n-gramas que tienen una distribución distinta en los TTs. Así, de acuerdo con Borin y Prütz (2001: 40):

> The method naturally lends itself to a working mode where we go from linguistic abstractions, i.e. POS n-grams, to increasingly concrete cases, i.e. via more specific - or longer - n-grams, to sequences of text words corresponding to particular POS n-grams.

Otro método que se aplica en estilometría es el aprendizaje automático. Así, Baroni y Bernardini (2006) utilizan el algoritmo de clasificación textual SVM para comprobar la posibilidad de detectar los TTs de manera automática. Construyen un corpus comparable que se compone de textos originalmente escritos en italiano y de textos traducidos al italiano del inglés, árabe, francés, español y ruso (se desconoce la proporción de los textos traducidos de cada una de estas lenguas). Todos los textos son artículos periodísticos sobre geopolítica y se extraen del mismo periódico italiano *Lines*, lo cual asegura la homogeneidad del corpus. Se realiza la lematización y el etiquetado POS del corpus y se entrena un clasificador SVM para distinguir entre los originales y las traducciones a partir de distintas representaciones de datos, que varían en el tamaño (unigramas, bigramas y trigramas) y en el tipo de unidades (palabras, lemas y etiquetas POS). Asimismo, se emplea una representación mixta en la cual las palabras de contenido se representan por medio de las etiquetas POS correspondientes, en tanto las palabras funcionales se dejan sin procesar.

Para investigar la naturaleza de las características distintivas de los TTs, Baroni y Bernardini (2006) evalúan el impacto de varios grupos de rasgos lingüísticos en el desempeño del clasificador: a) pronombres personales en función de sujeto; b) adverbios; c) signos de puntuación; d) formas verbales no finitas. La selección de dichos rasgos se basa en las hipótesis desarrolladas en investigaciones previas con respecto a las características del *translationese*.

En general, el clasificador realiza la tarea con una alta precisión (89.3% precisión y 83.3% cobertura). Los modelos que dan los mejores resultados se basan en unigramas de palabras, unigramas de representación mixta y bigramas de representación mixta. En cuanto a la naturaleza de las diferencias entre los TTs y los originales, se obtienen los resultados que explicamos a continuación.

La eliminación de la clase de rasgos que representan la distribución de pronombres personales en posición de sujeto tiene un impacto negativo en el desempeño del clasificador. Por consiguiente, la sobre-representación de los pronombres personales en los TTs podría considerarse un efecto del *translationese*. Es posible que esta característica se deba a la interferencia de las lenguas no *pro-drop* en las traducciones al

italiano. La eliminación de otros rasgos no tiene mayor efecto en la precisión de la clasificación automática.

Los resultados del estudio de Baroni y Bernardini son interesantes en varios sentidos. En primer lugar, comprueban la presencia de características distintivas en las traducciones de alta calidad usando como prueba la capacidad de un algoritmo de clasificación automática de distinguir entre los TTs y los TOs. Las traducciones resultan ser tan diferentes de los textos originalmente escritos en la lengua de llegada en términos de sus características lingüísticas que es posible identificarlas usando técnicas de clasificación automática que realizan la tarea con una alta precisión. En palabras de Lembersky et al. (2012: 255): "Incidentally, translated texts are so markedly different from original ones that automatic classification can identify them with very high accuracy".

En segundo lugar, Baroni y Bernardini presentan una discusión interesante sobre el tipo de representación de datos adecuada para el estudio del lenguaje de traducción. Al igual que en las tareas de clasificación textual por género y estilo, los patrones morfosintácticos tienen más relevancia para la distinción entre los TOs y los TTs que la frecuencia de palabras de contenido ampliamente utilizada para la clasificación por tema. Además, el estudio de Baroni y Bernardini demuestra que una representación sintáctica simplificada y superficial es suficiente para la identificación automática de los TTs.

El problema del método implementado por Baroni y Bernardini es que la clasificación automática con SVM no proporciona información sobre la naturaleza de los rasgos distintivos del lenguaje de traducción. La clasificación textual con técnicas de aprendizaje automático no da luz sobre las características de los TTs. Sería posible investigar la naturaleza de dichas características eliminando ciertas clases de rasgos en la fase de entrenamiento del clasificador y estimando su impacto en el desempaño del modelo. Sin embargo, tal acercamiento es especulativo y no proporciona información fiable de no realizarse un análisis cualitativo de los contextos de aparición de las unidades que presentan una distribución distinta en las traducciones y en los originales. Tal como afirman los mismos autores: "unlike with rule-based techniques, there is no

straight-forward way to interpret the models built by SVM algorithm directly, from a qualitative/linguistic point of view" (Baroni y Bernardini, 2006: 272).

Ahora bien, normalmente la dirección de la traducción no se toma en cuenta para el entrenamiento de los modelos de traducción en los sistemas de TA estadística, ya que para la mayoría de los corpus paralelos esta información no es disponible. Sin embargo, si los TTs tienen características propias que los distinguen de los textos originalmente escritos en la LM, sería lógico suponer que la dirección en la que traduce un sistema de TA debe coincidir con la dirección de traducción de los textos paralelos utilizados para su entrenamiento. Ya existen varios trabajos (Kurokawa et al., 2009; Lembersky et al., 2011; Lembersky et al., 2012) que toman en cuenta los efectos del *translationese* para la construcción de sistemas de TA estadística.

Kurokawa et al. (2009) exploran la posibilidad de la detección automática de los TTs en la línea de Baroni y Bernardini (2006) y después investigan las implicaciones que tiene el conocimiento sobre la dirección de traducción para el desempeño de un sistema de TA estadística basada en frases.

Para realizar los experimentos, Kurokawa et al. (2009) utilizan el *Canadian Hansard Corpus* (un recurso paralelo bilingüe que contiene las transcripciones de las sesiones del parlamento de Canadá de los años 1996-2007 con la información sobre la dirección de traducción), el algoritmo de clasificación automática SVM y el sistema de TA basada en frases *Portage*.

En primer lugar, el clasificador se entrena para decidir si el fragmento a clasificar es original o traducción. En la línea de Baroni y Bernardini (2006), el equipo de Kurokawa hace pruebas con diversos tipos de representación de datos (formas de palabras, lemas, etiquetas POS o representación mixta), logrando la precisión más alta con bigramas de palabras (*F-score* = 92%). Con respecto a estos resultados, Kurokawa et al. (2009: 4) afirman que:

> Although the performance of word and lemma representations may be helped by contextual or lexical cues, we also notice that the POS and mixed representations which focus solely on linguistic patterns, still reach around 85% F-score. This certainly shows that there are detectable differences in translated and original documents at the general, linguistic level.

A continuación, dividen el corpus de entrenamiento en tres partes: un subcorpus mixto que contiene textos paralelos con ambas direcciones, un subcorpus con el 80% de originales en francés, y un subcorpus con el 80% de originales en inglés. Comparan el desempeño de los modelos de traducción entrenados a partir de los subcorpus cuya dirección coincide/no coincide con la dirección en la que se realiza la TA, comprobando que la dirección de traducción del corpus de entrenamiento tiene un impacto en la calidad de la TA.

Así, una de las implementaciones de las ideas teóricas sobre la naturaleza del lenguaje de traducción es la adaptación de los modelos de traducción de sistemas estadísticos a la dirección de traducción.

3.2.2. Anotación de translation shifts

A continuación, se describen brevemente algunos trabajos que utilizan las herramientas de lingüística de corpus para el análisis de *translation shifts*. El estudio de Munday (1998) es el primer intento de aplicar los métodos básicos de lingüística de corpus y lexicografía al análisis de textos paralelos (el relato de Gabriel García Márquez "Diecisiete ingleses envenenados" y su traducción al inglés). Munday (1998) utiliza la herramienta de análisis de corpus *Wordsmith tools* (Scott, 1999). Una de las técnicas que emplea Munday es el cálculo de frecuencias de palabras (*types* y *tokens*) en el TO y el TT. El número absoluto de *tokens* es mayor en el TT, lo cual podría indicar una tendencia a la explicitación en la traducción. Sin embargo, Munday (1998) demuestra que dicha conclusión no es válida, ya que al hacer un segundo conteo de *tokens* sin tener en cuenta los pronombres personales, se obtienen resultados diferentes. Ello se debe a que el español es una lengua *pro-drop*, de manera que el traductor se ve obligado a explicitar los pronombres personales con función de sujeto, omitidos en el TO. Esta observación permite ver con claridad que antes de hacer hipótesis sobre la explicitación o implicitación como estrategias que subyacen el proceso traductor, hay que tomar en consideración las diferencias sistémicas entre las lenguas que "obligan" al traductor a explicitar o implicitar ciertos rasgos semánticos o gramaticales. Para investigar si la traducción conlleva simplificación, Munday (1998) propone comparar los *type-token ratios* del TO y el TT. Sin embargo, esta técnica presenta el mismo problema que la anterior. El TO tiene un *type-token ratio* más alto que el TT, pero no necesariamente

debido a la simplificación, sino a que el español es una lengua con morfología flexiva más desarrollada que el inglés. Este problema se soluciona por medio de la lematización de los textos precedente al análisis. Asimismo, Munday utiliza la técnica *Key Words in Context* para comparar el uso de los mecanismos de cohesión léxica en el original y en la traducción. Se extraen los contextos de aparición de unidades del léxico específicas del texto fuente y se averigua si éstas se traducen de la misma manera en el TT. En este caso, al igual que en los casos anteriores, se debe distinguir entre las diferencias que son resultado de los *translation shifts* opcionales y las que resultan de las modificaciones obligatorias debidas a las diferencias en la organización del vocabulario en español y en inglés.

Macken (2007) investiga las normas relacionadas con el grado de libertad en la traducción de diferentes tipos de textos del inglés al holandés (manuales de informática, noticias de prensa y actas de debates parlamentarios). El grado de libertad en la traducción se define con base en el tipo y la frecuencia de relaciones que se establecen entre los segmentos del TO y el TT alineados manualmente de acuerdo con el siguiente principio: "indicate the minimal language unit in the source text that corresponds to an equivalent in the target text" Macken (2007: 5). Los originales y las traducciones se alinean manualmente y se indica la unidad mínima del texto fuente que tiene un equivalente en el texto meta. En este modelo hay tres tipos de alineaciones: alineación regular para la correspondencia directa, alineación imprecisa para indicar distintos tipos de *translation shifts*, y alineación nula para los casos de omisión o adición del material del original en la traducción.

Los resultados del estudio indican que las traducciones analizadas difieren en el grado de libertad. La más cercana es la traducción de los manuales de informática (92% de alineaciones regulares), mientras que la traducción de las actas del parlamento tiene un grado de libertad mayor (81.6% de alineaciones regulares) y la traducción de las noticias de prensa está en el medio, con un 89.3% de alineaciones regulares. La propuesta de Macken es interesante, ya que permite ver que la cantidad y el tipo de *translation shifts* están relacionadas con las normas iniciales adoptadas por el traductor. Sin embargo, Macken no da indicaciones precisas que permitieran distinguir claramente entre los tipos de alineación. De la misma manera que en el estudio anterior, el trabajo

de Macken evidencia la necesidad de diferenciar los *shifts* obligatorios de los *shifts* opcionales.

Ahrenberg (2005) relaciona los tipos o estilos de traducción (traducción literal, traducción análoga, traducción semántica, traducción libre, etc.) con la tarea de evaluación automática. Este autor afirma que para la evaluación deben utilizarse THs que preserven la estructura del TO al máximo posible y, en general, incide en la importancia de la naturaleza de las THRs que se usan en la evaluación o entrenamiento de sistemas:

> While the reference translations are usually good translations, or even "expert translations", their qualities, or the requirements given to the translators, are seldom discussed in any detail […] Similarly, parallel corpora […] that are used for training statistical MT systems, are produced by human translators aiming at high quality by human standards. But this quality level may actually be beyond reach for any known system and translation approach developed to date (Ahrenberg, 2005: 13).

Ahrenberg (2005) propone establecer los requerimientos a partir de los cuales se realizará la evaluación definiéndolos a través del tipo de THR. Para ello revisa la propuesta de Newmark (1988), quien distingue ocho tipos (o modos) de traducción y los divide en dos grupos. En el primer grupo, la prioridad del traductor es preservar de la manera más cercana posible la estructura y el contenido de texto fuente: traducción palabra por palabra, traducción literal, traducción análoga y traducción semántica. En el segundo grupo, se prioriza la naturalidad de la expresión en la LM: traducción libre, traducción idiomática, traducción comunicativa y adaptación.

Ahrenberg (2005) arguye que en la evaluación o entrenamiento de sistemas de TA, se deben utilizar traducciones análogas. En este tipo de traducción no se toman en consideración los factores pragmáticos o de estilo y se preserva la estructura del original, a menos que tal preservación viole las reglas de la gramática de la LM. En otras palabras, la traducción análoga reduce al mínimo el número de desviaciones opcionales de la forma y del contenido del original.

Ahrenberg y Merkel (2000) proponen un modelo para medir los rasgos distintivos de diferentes tipos de traducción (TH, MAHT, TA) a partir de sus relaciones con los TOs. En este modelo se registran únicamente los *translation shifts* opcionales. Asimismo, se

excluyen del análisis los cambios relacionados con las palabras funcionales, a menos que afecten el tipo de cláusula o frase. Si no se introduce ningún tipo de anotación, se asume que los fragmentos del TO y del TT son formalmente equivalentes. La anotación se realiza a nivel de "segmentos de traducción" que equivalen a una unidad léxica, aunque Ahrenberg y Merkel (2000) no proporcionan una clara definición de la unidad de análisis.

Las correspondencias sintácticas entre los originales y las traducciones se establecen a nivel de oración, cláusula, frase y núcleo de frase. A nivel sintáctico, los *shifts* se clasifican de la siguiente manera:

1. cambios relacionados con la función y las propiedades de la cláusula (cambio de voz, cambio de modo, construcciones finitas vs. construcciones no finitas, cambio de nivel, cambio de función);
2. cambios relacionados con la función y la posición de constituyentes (cambio de función, cambio de nivel, transposición);
3. cambios en el número de constituyentes (adición, eliminación, divergencia);
4. paráfrasis que no entran en ninguna de las categorías arriba mencionadas.

En cuanto a la correspondencia de sentido, se distinguen tres categorías: a) la unidad del TT tiene un significado más específico; b) la unidad del TT tiene un significado menos específico; c) la unidad del TT tiene un significado diferente.

El grado de correspondencia entre los originales y las traducciones sigue la siguiente escala: isomórficas, semimórficas y heteromórficas. Después del análisis de *translation shifts,* los autores miden la proporción de las traducciones (oraciones traducidas) isomórficas. El modelo fue aplicado a un corpus de traducciones realizadas por humanos, por sistemas de TA y por sistemas de tipo MAHT, resultando, como era de esperar, en que el número de *shifts* en la TH supera significativamente el número de los mismos en las TAs:

> […] structural and semantic shifts are parts and parcel of high-quality translations and must not be mistaken for errors. In fact, the translation with the highest number of errors […] is the closest one, i.e. e. the translation produced by MT (Ahrenberg y Merkel, 2000: 45).

La afirmación de Ahrenberg (2005) sobre que la THR debe ser de tipo análogo es discutible, ya que los sistemas de TA estadística pueden aprender a partir del corpus las modificaciones que realizan con regularidad los traductores humanos. Con todo, el grado de correspondencia entre el TO y el TT (el tipo y la frecuencia de *translation shifts*), que depende de la estrategia global adoptada por el traductor y del tipo de texto, debería tomarse en consideración en el desarrollo y evaluación de sistemas de TA.

En resumen, consideramos que el estudio de las modificaciones (tanto obligatorias como opcionales) realizadas por el traductor con respecto al original puede ser de utilidad para la TA, ya que:

> Attempts to describe and systematize the "departures" or shifts are aimed at describing the factors that condition the translator's decisions and the influence of those decisions on the structure and interpretation of the TT (Szymánska, 2011: 213).

Para investigar las características del comportamiento de los traductores deben combinarse los métodos enfocados en la búsqueda de los universales de traducción y los métodos orientados a la identificación de las fuentes de las propiedades del TT a partir de su relación con el TO.

4. PROPUESTA PARA EL ANÁLISIS COMPARATIVO DE TRADUCCIONES HUMANAS Y AUTOMÁTICAS

En los capítulos anteriores se han caracterizado brevemente los principales tipos de sistemas de TA, se han revisado los métodos de evaluación automática y sus limitaciones, y se han discutido las características del lenguaje de la TH. En este capítulo presentamos la propuesta metodológica para la comparación de traducciones humanas y automáticas.

Como hemos mencionado antes, la calidad de la traducción como objeto de evaluación tiene varios aspectos relacionados con los niveles de la lengua, y no existe una manera trivial de ponderar dichos aspectos en términos de su efecto en la calidad global de la TA. Por tanto, en un primer acercamiento al análisis lingüístico de las diferencias TA-TH, conviene considerar dichas diferencias con detalle en cada nivel por separado. En la presente tesis las características de los TTs se investigan a tres niveles: léxico-terminológico, morfosintáctico y discursivo. Tal exploración no deja de ser parcial. La representación de datos, las herramientas utilizadas y el corpus de estudio limitan las posibilidades de observación. Sin embargo, dado el carácter exploratorio del estudio, consideramos que el análisis realizado puede dar luz sobre algunas características distintivas de la TA en comparación con la TH y proporcionar ideas para futuras investigaciones.

Tomando esto en cuenta, organizamos este capítulo de la manera siguiente. En primer lugar, ofrecemos un esquema general del análisis. A continuación, presentamos los criterios de compilación del corpus de estudio. Después, explicamos los procedimientos específicos realizados a nivel léxico-terminológico, morfosintáctico y discursivo, presentamos los tipos de análisis automático que llevamos a cabo y las herramientas del PLN que utilizamos. Finalmente, ofrecemos una propuesta de clasificación de las diferencias TA-TH.

4.1. Esquema general del análisis comparativo de traducciones humanas y automáticas

Uno de los objetivos principales de la evaluación es explicar en qué y por qué la TA difiere de la TH. En la presente investigación, identificamos las características distintivas de la TA frente a la TH a partir de la observación de la distribución de las

unidades de análisis en tres grupos de textos: TH, TA de *Google* y TA de *Lucy*. A continuación, relacionamos las características identificadas con el contexto dado por el TO y las estrategias empleadas por los traductores humanos y los sistemas de TA, es decir, intentamos identificar las fuentes de las diferencias detectadas.

Las fases generales del estudio son:

1. Selección de sistemas de TA.
2. Compilación del corpus.
3. Identificación de las características cuantitativas de la TH, de la TA de *Google* y de la TA de *Lucy*.
4. Detección de las diferencias significativas en la distribución de las unidades de análisis entre la TH y la TA de *Google*, y entre la TH y la TA de *Lucy*.
5. Análisis de una muestra aleatoria de contextos de aparición de las diferencias detectadas.
6. Identificación de patrones de las diferencias detectadas.
7. Clasificación de las diferencias detectadas de acuerdo con su origen.

Ilustramos las fases 3-7 en la Figura 5.

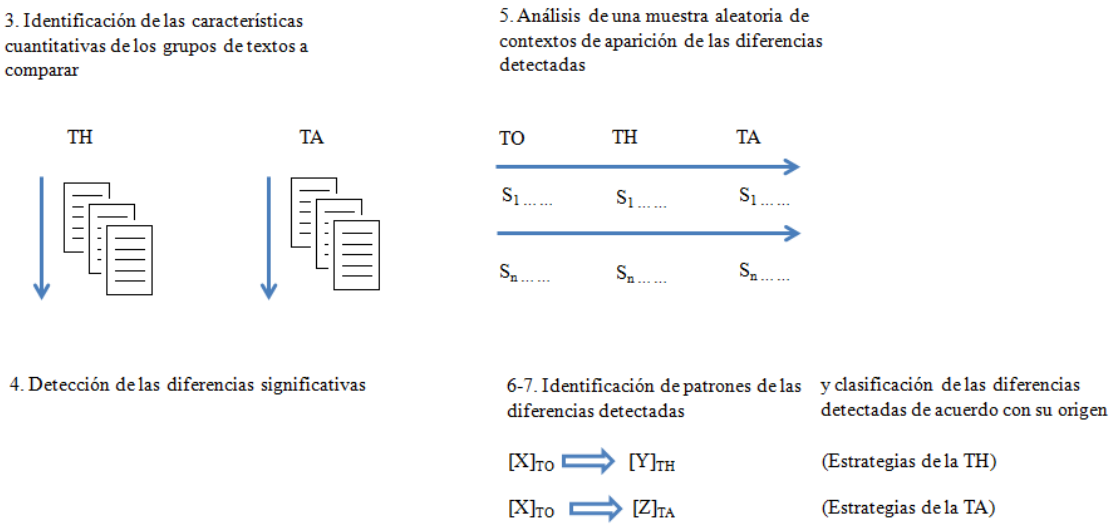

Figura 5. Análisis lingüístico comparativo TA-TH

De esta manera, combinamos el análisis cuantitativo que permite caracterizar el lenguaje de la TA en oposición al lenguaje de la TH con el análisis cualitativo encaminado a la identificación de las fuentes de las diferencias entre la TA y la TH.

Esta propuesta surge de las ideas discutidas en los capítulos anteriores. Por un lado, en el ámbito de la evaluación automática de la TA se han propuesto métodos que parten de una caracterización general de los textos producidos por traductores humanos frente a los textos generados por máquina (véase apartado 2.4.2.3.). Asimismo, se han desarrollado métricas que realizan la comparación TA-THR oración por oración con base en un análisis lingüístico (véase apartado 2.4.2.2.).

Por otro lado, el enfoque descriptivo en los estudios de traducción pretende identificar las características distintivas de los TTs frente a los textos originalmente escritos en la LM[12] por medio de un análisis cuantitativo de las características superficiales de los textos correspondientes (véase apartado 3.1.1.). Asimismo, existen estudios contrastivos que intentan descubrir regularidades en el comportamiento de los traductores a través de una comparación del texto fuente con el texto meta en términos de *translation shifts* (véase apartado 3.1.2.).

El procedimiento general descrito en la Figura 5 varía en función del nivel de análisis, ya que no podemos seguir todos los pasos de la misma manera a nivel léxico-terminológico, morfosintáctico y discursivo, debido a las diferencias en el tipo de representación de datos.

4.2. Diseño del corpus

Los criterios que suelen tomarse en consideración para la construcción de corpus paralelos son la homogeneidad, la representatividad, la calidad de las traducciones y el tamaño. Con respecto al primer criterio, consideramos, siguiendo a Toury (2004), que para identificar las regularidades en el comportamiento de los traductores, los TTs objeto de estudio deben compartir el contexto de situación. Así, para llegar a conclusiones válidas sobre las características de las traducciones debemos partir de un corpus homogéneo en términos de género y tipo textual.

[12] Nuestro corpus no contiene textos originalmente escritos en la LM, lo cual supone una limitación importante del presente estudio.

El segundo criterio está relacionado con el primero en el sentido de que, si se pretende construir un corpus representativo de cierto género textual, los textos deben proceder de distintas fuentes y alcanzar un número relativamente grande para cubrir todas las variantes del tipo de producción lingüística determinado. Sin embargo, tal escenario no permite controlar los factores externos que afectan las propiedades de los TTs. Por ello, dada la necesidad de realizar una comparación lo más objetiva posible entre la TH y la TA, es preferible que los textos del corpus sean del mismo tipo textual y del mismo ámbito temático, de manera que la distribución de las unidades de análisis en los TTs no se vea afectada por factores irrelevantes para la investigación. Asimismo, es importante que los TTs sean de alta calidad y constituyan ejemplos representativos del tipo y género textual en la LM.

En cuanto al tamaño del corpus, Biber (1990) indica que 10 textos por categoría son suficientes de cara a los propósitos de comparación por rasgos lingüísticos. En el mismo estudio menciona que las frecuencias de aparición de unidades del léxico son relativamente estables en muestras de 1000 palabras por texto.

Ahora bien, para realizar un análisis lingüístico a partir del corpus existen diversos tipos de procesamiento automático. Los sistemas de análisis lingüístico automático o *parsing* asignan a las unidades de análisis algún tipo de información lingüística (morfológica, sintáctica, semántica o discursiva). Entre los procedimientos más comunes del análisis automático de textos se encuentra la tokenización (segmentación de texto en palabras), la lematización (identificación de las formas base de las palabras), el análisis morfológico (reconocimiento automático de la estructura morfológica de las palabras), el etiquetado POS (asignación de las categorías gramaticales a las palabras del texto), el análisis sintáctico (reconocimiento automático de la estructura sintáctica de las oraciones), diversos tipos de análisis semántico (por ejemplo, asignación automática de papeles temáticos) y de análisis discursivo (por ejemplo, reconocimiento automático de la estructura discursiva del texto). Los aspectos de la descripción lingüística más desarrollados en el PLN son el sintáctico y el semántico, mientras que los ámbitos de la pragmática y del discurso todavía representan grandes dificultades para el procesamiento automático, ya que requieren la codificación del contexto extralingüístico.

A continuación, se discuten con detalle las técnicas que utilizamos para el análisis de los TTs (la extracción automática de términos, el etiquetado POS y el análisis discursivo), las herramientas específicas correspondientes y los procedimientos que llevamos a cabo en cada nivel de análisis.

4.3. Nivel léxico-terminológico

En este trabajo se estudia el comportamiento de los traductores humanos y los sistemas de TA en el contexto de la traducción de textos especializados. El discurso especializado tiene por objetivo la transmisión del conocimiento, que en este tipo de textos se vehicula a través de las unidades terminológicas [UTs]. Por ello, dado el carácter exploratorio del presente estudio, en el plano léxico se analiza únicamente el uso de términos en la TA y en la TH. Teniendo en cuenta que la metodología está pensada para su posterior aplicación en el ámbito del PLN, realizamos la extracción de términos de manera automática.[13]

Los términos son unidades del lenguaje natural que poseen forma y significado y son susceptibles de todos los procesos que afectan a otras unidades lingüísticas. La diferencia entre los términos y otras unidades del léxico reside en su valor pragmático, su uso en condiciones discursivas determinadas. De acuerdo con Cabré (2003), un término es una unidad del léxico que activa un significado preciso (su valor terminológico) cuando se usa en el contexto de la comunicación especializada. Adquirir el valor terminológico significa entrar en un esquema de significación preciso elaborado intelectualmente y consensuado por una comunidad científica. El carácter terminológico de las unidades léxicas está condicionado por su uso en un contexto y situación adecuados. Así, cuando las unidades léxicas se usan en un ámbito de especialidad, su contenido se adecúa a la situación comunicativa. En palabras de Cabré (1999: 133), "Los términos no pertenecen a un ámbito, sino que son usados en un ámbito con un valor singularmente específico". Si una denominación que posee el valor terminológico potencial se utiliza fuera del contexto de especialidad, pierde su carácter de término.

[13] Debido a que no existe un sistema de extracción automática de términos libre de errores, se considera que la salida de un extractor está conformada por candidatos a términos [CTs] en lugar de UTs.

En la terminología tradicional se niega la existencia de sinonimia en textos especializados y se postula la biunivocidad de los términos, lo cual implica que a un concepto, a modo de etiqueta, le corresponde una única denominación. En cambio, los planteamientos más recientes en el ámbito de la terminología (Cabré, 1999, 2003) consideran la variación como un hecho real en la comunicación especializada. Freixa (2002: 54) define la variación denominativa como "el fenómeno por el cual a una misma noción le corresponden diversas denominaciones". La noción de variación denominativa es relevante para el análisis de la traducción especializada y para la evaluación de las traducciones, ya que permite, dada una UT del TO, delimitar el grado de variación aceptable en el TT. Si las UTs que ofrecen el sistema de TA y el traductor humano son variantes denominativas, la diferencia no debe penalizarse en la evaluación.

4.3.1. Extracción automática de términos

La extracción de términos responde a las necesidades de investigadores en diferentes ámbitos y contribuye al desarrollo de numerosas aplicaciones de PLN, tales como la construcción de glosarios, vocabularios y diccionarios de especialidad, la indexación de textos, la TA, el análisis de corpus, etc. Por un lado, la extracción automática de términos es una tarea de extracción de información, una de las aplicaciones importantes en el campo del PLN. Por otro lado, de acuerdo con Vivaldi y Rodríguez (2011: 66), la extracción terminológica puede considerarse una tarea de anotación semántica: "a term extractor can be viewed as performing a semantic annotation task because it intends to provide machine-usable information based on meaning".

Vivaldi y Rodríguez (2011: 66) afirman que las características fundamentales de una UT desde el punto de vista de la extracción automática son "a) unithood; b) termhood and c) specialized usage". *Unithood* se refiere al grado de fuerza o estabilidad de las combinaciones sintagmáticas o colocaciones. *Termhood* se define como el grado en que una secuencia de unidades lingüísticas está relacionada a conceptos usados en algún dominio específico. Finalmente, *specialized usage* se refiere al uso de las unidades lingüísticas en los textos especializados.

Como se ha mencionado en el capítulo anterior al discutirse las propiedades de los TTs, una de las mayores dificultades de la aplicación de las técnicas del análisis de corpus al estudio de la traducción es la distancia entre las características abstractas del objeto de

estudio y la realización concreta de éstas en la superficie textual. De la misma manera que en otras aplicaciones del PLN, en la extracción automática de términos la identificación de tales características tiene que abordarse de modo indirecto, con ayuda de rasgos que son más fáciles de definir y de medir: las frecuencias de aparición de las palabras, las medidas de asociación, la exploración del contexto sintáctico, la posición del CT en una ontología, etc. El objetivo de la investigación es encontrar las medidas adecuadas y el modo de combinarlas.

La manera de abordar la tarea de extracción terminológica varía en función de la disponibilidad de recursos requeridos. Si para una lengua dada existen diccionarios electrónicos de especialidad, glosarios o bases de datos terminológicas etc., la extracción se realiza por medio de la asociación de las palabras del texto con una lista de términos del dominio existente (Krauthammer y Nenadic, 2004). En el caso de las lenguas, para las cuales no existen suficientes recursos de este tipo, se tiene que recurrir a métodos más complejos, que pueden ser lingüísticos, estadísticos o híbridos. Los últimos son los que más se han utilizado en la extracción de términos, ya que la aplicación exclusiva de los métodos lingüísticos o estadísticos no da buenos resultados. De hecho, prácticamente todos los sistemas de extracción terminológica combinan estos dos paradigmas en diferentes etapas del procesamiento. Por esta razón, Vivaldi (2001: 24) considera más apropiado hablar de sistemas "mayoritariamente lingüísticos" y sistemas "mayoritariamente estadísticos".

La mayoría de los extractores terminológicos se basan, al menos en alguna de las etapas del procesamiento, en la explotación del conocimiento lingüístico. La fuente de información más utilizada en los sistemas lingüísticos son los patrones morfosintácticos típicos de las UTs. La desventaja de este método es su baja precisión, es decir, la sobregeneración de CTs que no pertenecen al ámbito de especialidad.

En cuanto a los extractores estadísticos, suelen hacer uso de las frecuencias de aparición de las unidades del léxico como parámetro para la estimación de *termhood*. Vivaldi (2001: 25) indica que "la idea de frecuencia de un CT puede tomar formas diferentes pero muy similares entre sí: una unidad frecuente en un dominio, una unidad que sólo aparece en un dominio y una unidad que aparece más frecuentemente en un dominio específico que en un dominio general". La desventaja de este método es que no

detectará los términos que tienen la frecuencia de aparición baja. Otro método estadístico ampliamente utilizado para la detección de UTs poliléxicas se basa en el criterio de *unithood,* al medir la fuerza asociativa entre las palabras (las medidas típicas utilizadas para esta tarea son información mutua y *association ratio*). Un ejemplo de los extractores estadísticos es el sistema TermoStat Web 3.0 (Drouin, 2003).[14]

Algunos sistemas lingüísticos utilizan la información semántica al recurrir a los recursos que codifican la estructura conceptual del ámbito de especialidad. Entre ellos destacan los sistemas TRUCKS (Maynard, 1999), YATE (Vivaldi, 2001) y el extractor basado en *Wikipedia* desarrollado por Vivaldi y Rodríguez (2011). Este último es el que se usa en la presente investigación. Por ello, a continuación ofrecemos una descripción detallada de la propuesta de Vivaldi y Rodríguez (2011).

4.3.2. Extractor terminológico basado en Wikipedia

La extracción terminológica del sistema desarrollado por Vivaldi y Rodríguez (2011) se realiza de la manera siguiente. En la primera fase de la extracción se lleva a cabo el pre-procesamiento automático de los textos del corpus en el que se detectan las fechas, números, locuciones, nombres propios y abreviaturas; se realiza el análisis morfológico (identificación de los posibles lemas y categorías gramaticales de las palabras del texto) y, finalmente, se realiza una desambiguación lingüística y estadística, con la cual a cada palabra termina correspondiéndole un solo lema y una sola etiqueta POS.

A partir de esta anotación previa, el extractor produce una lista de CTs con base en los siguientes patrones sintácticos: N, N+Adj, N+Adj+Adj, N+Prep+N, N+Prep+N+Adj.[15] A continuación, a cada CT se le asigna un coeficiente de dominio [CD] que indica su *termhood* (su grado de pertenencia al dominio de especialidad) y que se obtiene a partir de la estructura de páginas y categorías de *Wikipedia*, como fuente de conocimiento especializado.

La unidad básica de información en *Wikipedia* es la página o el artículo que se identifica por su título y, supuestamente, corresponde a un término. La estructura de *Wikipedia*

[14] http://www.mapageweb.umontreal.ca/drouinp/#termostat

[15] De acuerdo con el planteamiento teórico (Cabré, 1999) en el que se basa el extractor, las UTs son, en esencia, unidades con función referencial, de ahí que los verbos no se tomen en consideración para la extracción terminológica.

para una lengua dada puede representarse en forma de dos grafos conectados: el grafo de categorías y el grafo de páginas (o artículos). Cada artículo está relacionado con una o varias categorías por medio de *category links*, de tal manera que las categorías representan clases asociadas a las páginas de *Wikipedia*. Al mismo tiempo, cada categoría se relaciona con una o varias categorías que se estructuran en forma de clases constituyendo a su vez un grafo conectado. Los arcos que relacionan las categorías normalmente tienen una relación semántica de hiponimia o meronimia.

Además, existen páginas de redireccionamiento que solamente contienen un enlace a otra página con el artículo completo (las páginas de redireccionamiento incluyen variaciones ortográficas, morfológicas o abreviaturas de los títulos de los artículos) y de desambiguación (para títulos polisémicos). Estos dos tipos de páginas proporcionan información adicional al extractor terminológico basado en *Wikipedia*. La estructura de grafos de *Wikipedia* se representa de manera esquemática en la Figura 6.

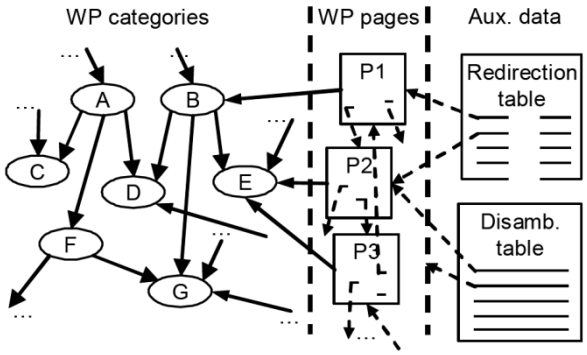

Figura 6. Estructura de grafos de *Wikipedia* (Vivaldi y Rodríguez, 2011: 67)

Para calcular el CD de un CT a partir de estos datos, se identifican de manera automática los sub-grafos de *Wikipedia* (el sub-grafo de categorías y el sub-grafo de páginas) que representan el dominio de especialidad. A continuación el CD se calcula con base en el número o longitud de caminos que relacionan el CT y la categoría que representa el dominio de especialidad (frontera del dominio). Vivaldi y Rodríguez (2011) proponen varias maneras de calcular el CD. En esta tesis empleamos la fórmula siguiente:

$$CD_{nc(t)} = \frac{NC_{dominio(t)}}{NC_{total(t)}}$$

Donde $CD_{nc(t)}$ es el CD, t es el CT, $NC_{dominio(t)}$ es el número de caminos a la página asociada a la categoría que representa el dominio de especialidad y $NC_{total(t)}$ es el número de caminos a la categoría máxima de *Wikipedia*. Los valores de los CDs asociados a los CTs se encuentran en la escala de 0 a 1, en la cual 0 indica que el CT no tiene ninguna relación con el ámbito de especialidad. La anotación -1 se reserva para los casos en los que el CT no se encuentra en *Wikipedia*.

En la Figura 7 presentamos un ejemplo de representación a partir de la que se calcula el CD para el término "ratón de laboratorio".

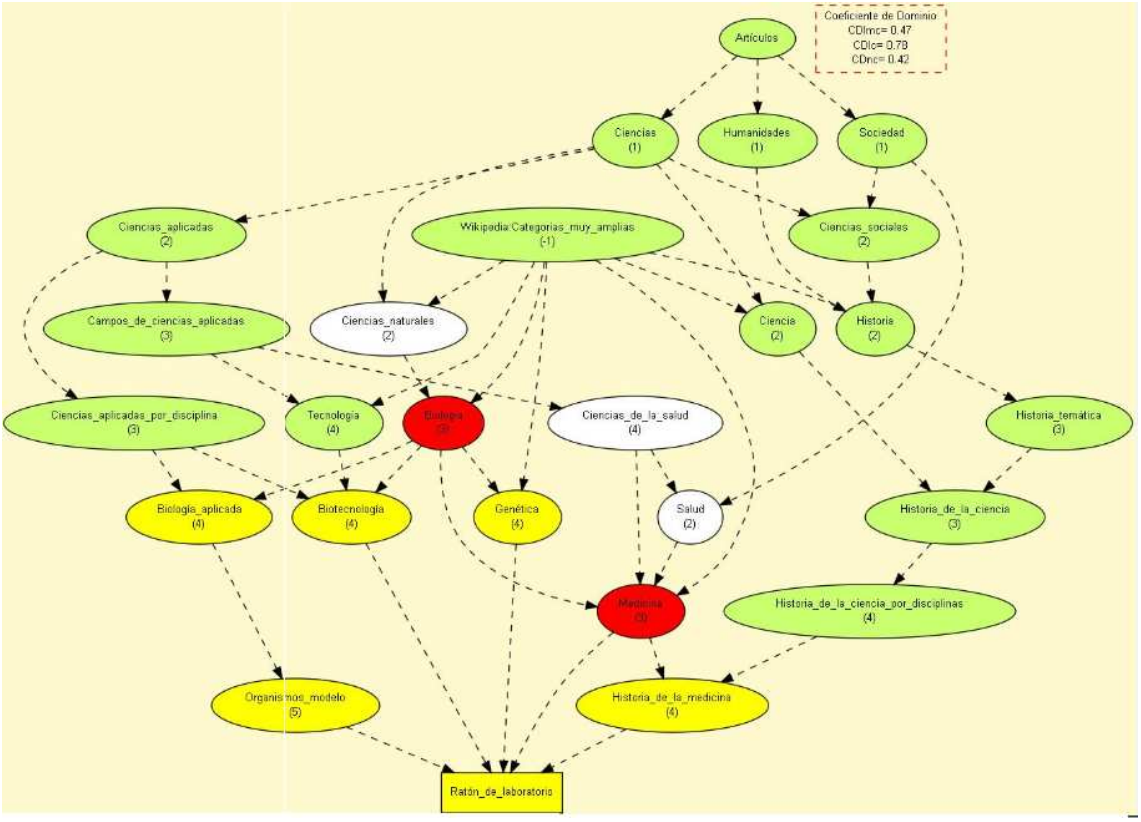

Figura 7. Grafo de *Wikipedia* para el término "Ratón de laboratorio"

El rectángulo en amarillo indica el término a buscar. Los óvalos amarillos marcan las categorías que se encuentran entre el término y la frontera de dominio. Los óvalos rojos son las fronteras de dominio, que para nuestro experimento definimos como "Biología" y "Medicina". Los óvalos verdes, son las categorías que llevan a la categoría máxima o tope pero no pasan por las fronteras de dominio. Finalmente, los óvalos blancos

representan las categorías por las que pasa el camino que va únicamente de la frontera de dominio a la categoría tope.

Vivaldi y Rodríguez (2011) mencionan que la ventaja de utilizar *Wikipedia* como recurso de información semántica es su disponibilidad y su amplio alcance. Sin embargo, el desempeño del extractor no es igual para todas las lenguas, ya que la cantidad de información que proporciona *Wikipedia* varía en función de la lengua. Además, la arquitectura de *Wikipedia* como recurso de información terminológica presenta varios errores. Por ejemplo, en ocasiones los artículos no están vinculados a la categoría correcta o no están unidos a ninguna categoría.

4.3.3. Procedimientos del análisis comparativo del tratamiento de la terminología en traducciones humanas y automáticas

A continuación presentamos los procedimientos del análisis del tratamiento de la terminología en las TAs y en la TH. En primer lugar, realizamos la extracción automática de CTs con la herramienta de Vivaldi y Rodríguez (2011). Dada la temática de los textos que conforman nuestro corpus de estudio, definimos como categorías representantes del dominio de especialidad las categorías "Medicina" y "Biología". Con ayuda del extractor terminológico se obtienen tres listas de CTs: la lista de los CTs de la TH, de la TA de *Google* y de la TA de *Lucy*.

Para cada CT el extractor proporciona la información siguiente: lema, patrón morfosintáctico y CD. Sometemos las listas de CTs a un proceso de filtrado manual en el cual: a) se eliminan todos los CTs con un CD menor de 0.5[16]; b) se eliminan las unidades cuyo significado identificado por el sistema no coincide con el significado que activan en el contexto (por ejemplo, "echar *mano* de", "tener en *mente*"). No realizamos ningún otro tipo de validación o corrección manual con respecto a estas listas para lograr la máxima sistematicidad en el análisis cuantitativo. Las listas de los CTs extraídos junto con la información correspondiente se presentan en el Anexo A.

[16] Establecemos este umbral con base en una exploración preliminar de los datos, la cual indica que los CTs con el coeficiente menor a este número no se relacionan con el dominio de especialidad en la mayoría de los casos.

En segundo lugar, identificamos las características cuantitativas del uso de la terminología en los TTs. Una de las estrategias ampliamente utilizadas en los estudios de traducción basados en corpus para medir la variedad del vocabulario y comprobar si la traducción conlleva simplificación léxica es el *type-token ratio*, el cual se calcula de acuerdo con la fórmula siguiente:

$$Type\text{–}token\ ratio = \frac{No.\,de\ types \times 100}{No.\,de\ tokens}$$

El No. de *types* es el número de palabras diferentes en el corpus. El número de *tokens* es el número total de palabras en el corpus. Aplicamos esta medida a las listas de los CTs identificados por el extractor terminológico con la finalidad de medir la variedad del vocabulario de especialidad en las TAs y en la TH. Asimismo, calculamos la proporción de los CTs con los CDs altos (CD = 1) en cada grupo de textos para comprobar el grado de pertenencia al dominio de los términos usados por los traductores humanas y por los sistemas de TA.

En tercer lugar, calculamos el número total de diferencias TA-TH a nivel léxico-terminológico de la siguiente manera. Desarrollamos un script, por medio del cual realizamos la extracción de todas las ocurrencias de los CTs junto con los números de las oraciones en las que aparecen. Después, con ayuda de la función BUSCARV de *Microsoft Excel* detectamos las ocurrencias de los CTs en la TH que no se encuentran en las TAs, y las ocurrencias de los CTs en las TAs que no aparecen en la TH, calculando las proporciones correspondientes.

Por último, seleccionamos una muestra aleatoria de 50 oraciones, en las que realizamos una alineación manual de los CTs identificados en los TTs y las unidades correspondientes de los TOs[17]. Al organizar de esta manera los CTs, clasificamos las diferencias TA-TH en términos de su origen especificando el tipo de modificaciones que realizan los traductores humanos y los sistemas de TA con respecto a las UTs del

[17] Si es imposible establecer la equivalencia a nivel del CT detectado por el extractor, la unidad de análisis se amplía. Por ejemplo, en la frase "agente etiológico del SIDA" el extractor asigna el CD a la unidad "agente etiológico". La unidad correspondiente en el original es "*AIDS-causing agent*", por tanto en este caso tomamos como unidad de análisis la frase "agente etiológico del SIDA".

original[18] (la clasificación se lleva a cabo con base en la taxonomía que presentaremos en el apartado 4.6.).

Para realizar el análisis comparativo y para comprobar el estatus terminológico de las unidades de análisis utilizamos diversos recursos terminológicos: diccionarios de lengua general, diccionarios de especialidad, ontologías y bases de datos terminológicas, entre ellos:

- *Oxford Advanced Learner's Dictionary* [OALD][19]
- The Dictionary of cell biology (1995)
- Diccionario de la lengua española, Real Academia Española, 22ª edición [DRAE][20]
- Diccionario de bioquímica y biología molecular (2000)
- Base de datos terminológica *InterActive Terminology for Europe* [IATE][21]
- Base de datos terminológica TERMCAT[22]

4.4. Nivel morfosintáctico

A nivel morfosintáctico, en la línea de Borin y Prütz (2001) (véase pp. 60-61), identificamos los rasgos distintivos del lenguaje de la TA en oposición a la TH en términos de sobre- o sub-representación de las secuencias de categorías gramaticales (n-gramas de etiquetas POS) en estos grupos de textos.

4.4.1. Etiquetado POS

El etiquetado POS es el proceso de asignación de una categoría gramatical u otros rasgos morfosintácticos a cada una de las unidades léxicas del texto. A este proceso le preceden varias etapas de pre-procesamiento como la tokenización y el análisis morfológico automático. Las palabras se forman a través de mecanismos de flexión, derivación o composición a partir de sus formas base. La tarea de descomposición de

[18] La lista de los CTs alineados y la clasificación de las diferencias TA-TH a nivel de UT se ofrece en el Anexo B.
[19] http://oald8.oxfordlearnersdictionaries.com
[20] http://lema.rae.es/drae
[21] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load
[22] http://www.termcat.cat/

una palabra en su forma base y sus afijos se denomina *parsing* morfológico o análisis morfológico automático. Un analizador morfológico normalmente consta de tres módulos: diccionario o lexicón que contiene una lista de lemas (formas base de las palabras) con la información sobre su categoría gramatical; conjunto de reglas ortográficas que indican las modificaciones que se tienen que realizar en la forma gráfica de la palabra al marcarse un rasgo gramatical; lista de afijos y reglas morfotácticas de su combinación, las cuales se representan por medio de los modelos de autómatas de estados finitos.

Ahora bien, a la misma forma lingüística superficial, en función del contexto, le pueden corresponder distintas etiquetas POS; por tanto, la asignación de una etiqueta única a cada palabra puede considerarse una tarea de desambiguación. La mayoría de los algoritmos de etiquetado POS pertenecen a una de las siguientes clases: los etiquetadores basados en reglas, los etiquetadores estadísticos y los etiquetadores híbridos. Los primeros se basan en un gran número de reglas de desambiguación elaboradas de manera manual (por ejemplo, el etiquetador *EngCG* de Karlsson et al., 1995). Los etiquetadores estadísticos se basan en un corpus previamente etiquetado de manera manual para calcular la probabilidad de que una palabra dada tenga una u otra etiqueta en un contexto morfosintáctico determinado (por ejemplo, el etiquetador basado en los Modelos Ocultos de Markov de Brants, 2000). Asimismo, existen etiquetadores híbridos que combinan los rasgos de los dos tipos anteriores (por ejemplo, el etiquetador de Brill, 1995).

4.4.2. Freeling

En esta investigación el etiquetado POS se realiza con ayuda de la herramienta de análisis lingüístico automático *FreeLing*[23] (Carreras et al., 2004). *FreeLing* es una librería de código abierto para el procesamiento multilingüe automático, desarrollada en la Universidad Politécnica de Catalunya, que ofrece una amplia gama de servicios de análisis lingüístico para diversos idiomas.

El etiquetador morfológico de *Freeling* se basa en un diccionario de palabras en español que proporciona la información sobre las categorías gramaticales posibles de cada

[23] En esta investigación usamos la versión *Freeling-2.2* para *Windows*. Para obtener más información sobre esta herramienta, puede consultarse la página http://nlp.lsi.upc.edu/freeling

palabra. Este diccionario se ha generado por medio de la obtención, a partir del análisis de varios corpus, de una lista de raíces de palabras. Con dichas raíces, se han producido todas las inflexiones posibles de las palabras de manera automática, así como las probabilidades de las posibles etiquetas. Las palabras del texto a analizar se relacionan con las entradas del diccionario para determinar las etiquetas que les corresponden. En el etiquetado POS los casos de ambigüedad se resuelven en la línea del trabajo de Brants (2000) con base en los Modelos Ocultos de Markov.

Los diferentes tipos de etiquetadores utilizan etiquetas distintas en función de la postura teórica subyacente sobre la morfología y también en función de las necesidades de análisis (representación más o menos detallada). En el caso de *Freeling* se emplea el conjunto de etiquetas del estándar EAGLES (para la descripción de los estándares del etiquetado morfosintáctico existentes, véase Leech y Wilson, 1999).

La codificación de EAGLES se basa en un conjunto de letras para cada categoría (atributos obligatorios) y los rasgos morfosintácticos correspondientes (atributos recomendados), ordenadas en posiciones consecutivas. Un ejemplo de la codificación morfosintáctica en formato EAGLES para el español se ofrece en la Tabla 5 (la tabla completa se presenta en el Anexo C).

Posición	Atributo	Valor	Designación
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5 y 6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

Tabla 5. Codificación de rasgos morfosintácticos para nombres según el estándar EAGLES

Cada rasgo gramatical tiene una letra asociada que lo representa en la secuencia final de la etiqueta. La ausencia de un rasgo es indicada por un 0. Así, para un nombre común masculino singular sin caso se obtendría la codificación NCMS000.

A continuación proporcionamos un ejemplo del etiquetado POS de *Freeling* de una oración de nuestro corpus.

```
Las el DA0FP0 0.97051
células célula NCFP000 1
de de SPS00 0.999919
nuestro nuestro DP1MSP 0.933333
organismo organismo NCMS000 1
contienen contener VMIP3P0 1
unas uno DI0FP0 0.925926
redes red NCFP000 0.875
de de SPS00 0.999919
comunicación comunicación NCFS000 1
interna interno AQ0FS0 0.75
sorprendentes sorprendente AQ0CP0 1
. . Fp 1
```

Cada línea es ocupada por una palabra. Primero se indica la forma de la palabra, después el lema, a continuación la etiqueta POS correspondiente y, finalmente, la probabilidad de asignación de la etiqueta.

4.4.3. Procedimientos del análisis comparativo de traducciones humanas y automáticas a nivel morfosintáctico

A fin de poder realizar el conteo de frecuencias de aparición de unigramas, bigramas y trigramas de etiquetas POS en nuestro corpus, transformamos[24] la salida de *Freeling* en una secuencia de etiquetas de categorías gramaticales. Así, la oración anterior quedaría representada de la manera siguiente:

```
DA0FP0 NCFP000 SPS00 DP1MSP NCMS000 VMIP3P0 DI0FP0 NCFP000
SPS00 NCFS000 AQ0FS0 AQ0CP0 .
```

Usamos una representación más o menos detallada en función del tamaño de las secuencias de etiquetas POS (unigramas, bigramas o trigramas). Para los unigramas se utiliza una representación detallada a fin de poder obtener información sobre el tratamiento de la morfología en las TAs y en la TH. Se dejan fuera algunos rasgos que no tienen relevancia para la identificación de las regularidades en el comportamiento de los traductores y los sistemas de TA (los rasgos que no se toman en cuenta para el análisis están marcados en gris en el Anexo C). En el caso de los bigramas y trigramas se toma en consideración únicamente la categoría principal y el primer rasgo morfosintáctico de las etiquetas POS de *Freeling*, ya que nos interesan las combinaciones de categorías gramaticales que reflejan, aunque de manera superficial, el uso de las estructuras sintácticas en los TTs.

Calculamos las frecuencias de aparición de unigramas, bigramas y trigramas de etiquetas POS con ayuda de la herramienta *Jaguar*[25] (Nazar et al., 2008). Los signos de puntuación no forman parte de las secuencias identificadas, pero sí se toman en cuenta en el sentido de que si las palabras están separadas por un signo de puntuación no pueden formar un n-grama. Ante la necesidad de llevar a cabo una prueba de

[24] De aquí en adelante para realizar las transformaciones necesarias entre diferentes representaciones de datos usamos las expresiones regulares en el editor de textos *EditPlus*.
[25] http://melot.upf.edu/cgi-bin/jaguar/jaguar.pl

significación estadística, se eliminan todos los n-gramas cuya frecuencia es igual o menor a 5 en todos los grupos a comparar (TH, TA de *Google* y TA de *Lucy*).

Para establecer la significación estadística de las diferencias en las frecuencias de aparición de las secuencias de etiquetas POS, usamos el test estadístico de la X^2. La prueba de la X^2 ha sido ampliamente utilizada en lingüística de corpus. De acuerdo con McEnery y Wilson (1996: 70), el test de la X^2 tiene varias ventajas de cara al análisis de datos lingüísticos:

> (a) it is more accurate than, for example, the t-test; (b) it does not assume that the data is normally distributed (quite frequent with linguistic data); (c) it is easy to calculate, even without a computer statistics package; and (d) disparities in corpus size are unimportant.

El test de la X^2 comprueba la independencia de dos grupos de frecuencias categóricas para las cuales no podemos suponer la distribución normal. Estima la diferencia entre las frecuencias observadas en el corpus y las frecuencias esperadas (aquellas que se darían si el único factor fuera el azar). Si la diferencia entre los valores observados y los valores esperados no es grande, quiere decir que, muy probablemente, las diferencias entre las variables son azarosas. En cambio, si la diferencia entre las frecuencias observadas y las frecuencias esperadas es grande, entonces es probable que las diferencias entre las variables sean significativas.

El estadístico de la X^2 se calcula a partir de las tablas de contingencia que representan la distribución de las frecuencias de las unidades de análisis en el corpus. La Tabla 6 es una tabla de contingencia para el trigrama "nc vs vm" (nombre común, verbo semiauxiliar, verbo principal) en el subcorpus de la TH y en el subcorpus de la TA de *Google.*

	TH (i)	TA (j)	Total
POS = nc vs vm (i)	5	16	21
POS != nc vs vm (j)	18559	20546	39105
Total	18564	20562	39126

Tabla 6. Frecuencias observadas de aparición del trigrama "nc vs vm" en la TH y en la TA de *Google*

Para calcular las frecuencias esperadas para cada celda de la tabla, el total de la fila de la celda se multiplica por el total de la columna de la celda y se divide por el total de las observaciones. Así, obtendríamos las frecuencias esperadas para el trigrama "nc vs vm" incluidas en la Tabla 7:

	TH (i)	TA (j)	Total
POS = nc vs vm (i)	9.96	11.04	21
POS != nc vs vm (j)	18554.04	20550.96	39105
Total	18564	20562	39126

Tabla 7. Frecuencias esperadas de aparición del trigrama "nc vs vm" en la TH y en la TA de *Google*

El estadístico de la X^2 se calcula a partir de la fórmula siguiente:

$$X^2 = \sum_{ij} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

Donde *n* representa las frecuencias observadas, μ representa las frecuencias esperadas, los subíndices *i* y *j* corresponden a las columnas y las filas de la tabla de contingencia. Con ayuda de la tabla de distribución de la X^2 (que puede encontrarse en cualquier libro de texto de estadística, por ejemplo, Agresti, 2007) se identifica el p-valor (establecemos el nivel de significación estadística p-valor < 0.05); el valor del estadístico de la X^2 que corresponde al p-valor < 0.05 es $X^2 > 3.81$.

Así, al realizar el cálculo para el ejemplo anterior obtendríamos el valor $X^2 = 4.68$, con lo cual concluiríamos que la diferencia en la distribución del trigrama "nc vs vm" en el subcorpus de la TH y en el subcorpus de la TA es significativa. La lista completa de los n-gramas de etiquetas POS, sus frecuencias de aparición y el estadístico de la X^2 para las diferencias entre las TAs y la TH se ofrece en el Anexo D.

A continuación, seleccionamos 10 n-gramas cuyas frecuencias de aparición difieren significativamente en la TH y en las TAs de alguno (o ambos) de los sistemas y que se relacionan claramente con una construcción sintáctica determinada. Después, buscamos en el corpus de estudio 10 contextos de aparición para cada uno de estos n-gramas

siguiendo el patrón establecido por las frecuencias. Por ejemplo, si el trigrama "nc vs vm" está sobre-representado en la TA de *Google*, buscamos 10 oraciones en las que éste aparece en la TA de *Google* y no aparece en la TH. De esta manera, obtenemos una muestra de 100 casos de diferencias en la distribución de n-gramas de etiquetas POS. A partir de estos datos, detectamos las regularidades en el comportamiento de los traductores y los sistemas de TA al enfrentarse con una construcción sintáctica del original determinada y clasificamos las diferencias de acuerdo con la tipología que presentaremos en el apartado 4.6.

4.5. Nivel discursivo

Una de las posibles perspectivas lingüísticas para acercarse al problema de la calidad de la TA es el análisis discursivo. Este acercamiento permite caracterizar el TT en términos de una de las propiedades fundamentales que éste debe poseer, la coherencia. En numerosas aplicaciones del PLN, tales como la generación automática de textos, el resumen automático y la TA, se tienen que tratar fenómenos lingüísticos que sobrepasan los límites de la oración. Sin embargo, tanto las métricas de evaluación como los sistemas de TA mismos operan a nivel oracional, lo cual, en definitiva, afecta a la calidad global de la traducción. Tal como indica Wilks (2009: 121):

> In order to produce MT of superior quality that existing systems, one of the most powerful key ideas is the use of discourse-related and pragmatic terms. Most MT systems operate on a sentence-by-sentence basis only; they take no account of the discourse structure […] structural information should be taken into account and can be used to improve the quality of the translation.

4.5.1. Rhetorical Structure Theory

Existen varias propuestas para la anotación automática de la estructura discursiva (véase, por ejemplo, Webber, 2004). En este estudio usamos la Teoría de la Estructura Retórica [*Rhetorical Structure Theory,* RST] desarrollada por Mann y Thompson (1988). A nuestro conocimiento, es la única propuesta de anotación con base en la cual se está desarrollando un sistema de anotación discursiva automática para el español (da Cunha et al., 2012).

La RST se creó de cara a la tarea de generación automática de textos y ofrece un marco descriptivo para el análisis discursivo. El propósito del aparato metodológico que proponen Mann y Thompson es describir la estructura del texto en términos funcionales, con base en las relaciones discursivas que se establecen entre las unidades que lo componen. De acuerdo con el planteamiento de Mann y Thompson, el texto posee una estructura jerárquica en la que cada unidad tiene una función determinada, con lo cual se pone de relieve el papel comunicativo de la estructura discursiva del texto: "an RST analysis always constitutes a plausible account of what the writer wanted to achieve with each part of the text. An RST analysis is thus a functional account of the text as a whole" (Mann y Thompson, 1988: 258).

La primera etapa del análisis de la estructura discursiva en términos de la RST es la segmentación del texto en unidades discursivas elementales o mínimas [*Elementary Discourse Units*, EDUs]. De acuerdo con Tofiloski et al. (2009: 77), "Discourse segmentation is the process of decomposing discourse into elementary discourse units (EDUs), which may be simple sentences or clauses in a complex sentence". La regla comúnmente aceptada para segmentar el texto en EDUs es que cada cláusula independiente del texto, que muestra una relación discursiva (o, dicho con otras palabras, cuya función discursiva, en principio, no es ambigua), junto con diversos elementos que dependen de ella, constituye una unidad mínima del análisis. La RST establece dos tipos de EDUs: núcleos (elementos gobernantes) y satélites (elementos que dependen de los núcleos y aportan cierta información sobre ellos). Los núcleos son considerados las unidades más importantes del texto de cara a los propósitos del autor, mientras que los satélites se caracterizan por contribuir al significado de los elementos nucleares y se consideran secundarios con respecto a éstos. De acuerdo con Mann y Thompson (1988), la nuclearidad es uno de los principios fundamentales de la organización textual.

Para definir una relación discursiva entre dos elementos el analista hace suposiciones (*plausibility judgments*) sobre las intenciones del autor, de manera que los criterios para la identificación de las relaciones de coherencia son funcionales y semánticos. La RST establece cuatro criterios para definir las relaciones discursivas: restricciones para el núcleo, restricciones para el satélite, restricciones para la combinación núcleo-satélite y

efecto. Por ejemplo, Mann y Thompson (1988: 253) ofrecen la definición siguiente para la relación de Antítesis:

> *relation name*: ANTITHESIS
> *constraints on N* [Nucleus]: W [Writer] has positive regard for the situation presented in N
> *constraints on S* [Satellite]: none
> *constraints on the N + S combination:* the situations presented in N and S are in contrast (cf. CONTRAST, i.e. are (a) comprehended as the same in many respects, (b) comprehended as differing in a few respects and (c) compared with respect to one or more of these differences); because of an incompatibility that arises from the contrast, one cannot have positive regard for both the situations presented in N and S; comprehending S and the incompatibility between the situation presented in N and S increases R's [Reader's] positive regard for the situation presented in N
> *effect*: R's positive regard for N is increased.

Existen varias listas de relaciones propuestas por diferentes investigadores. La lista original de Mann y Thomson incluye 24 relaciones. En opinión de estos autores (Mann y Thompson, 1988: 256), "no single taxonomy seems suitable", ya que el número de relaciones depende de los propósitos de la investigación. Tanto las reglas de segmentación como el número y el tipo de relaciones deben ser adaptados a las necesidades del análisis.

La última etapa del análisis con la RST es la construcción de árboles discursivos. Dado el supuesto sobre la organización jerárquica del texto, la representación arbórea es la que más se ha utilizado en el marco de la RST.

La interpretación de las relaciones de coherencia (relaciones semánticas abstractas que no se establecen a nivel de lo dicho, sino a nivel de lo inferido) que realiza el lector con base en su conocimiento del mundo y del contexto de situación es extremadamente difícil de modelar computacionalmente. A veces, las relaciones de coherencia no están marcadas explícitamente en el texto. Incluso en aquellos casos en los que es posible identificar los rasgos que indican cierta relación discursiva, éstos a menudo son ambiguos desde el punto de vista computacional, ya que pueden estar asociados a varios tipos de relaciones discursivas (para resolver la ambigüedad el lector suele recurrir al contexto, tanto lingüístico como extralingüístico). Dado que el aparato descriptivo de la RST inicialmente estaba destinado para la generación de textos, Mann y Thompson

(1988) no proponen ningún método para la detección automática de relaciones discursivas con base en las características lingüísticas de los textos a analizar.

Posteriormente se han propuesto varios tipos de rasgos lingüísticos que podrían servir como marcadores de las relaciones discursivas en el contexto de análisis discursivo automático. Corston-Oliver (1998) propone incluir, además de los marcadores del discurso, otras fuentes de información como el orden de los segmentos, la presencia de ciertos adverbios y pronombres, etc. Le y Abeysinghe (2003) combinan el análisis de marcadores discursivos, relaciones sintácticas y mecanismos de cohesión. Schauer y Hahn (2001) recurren a las relaciones anafóricas como un indicador de la estructura discursiva. Reitter y Stede (2003) usan los marcadores discursivos, las etiquetas POS y las cadenas léxicas. Marcu (2000) explora la posibilidad de crear un algoritmo que derive la estructura discursiva del texto con base en varias características superficiales como puntuación, marcadores discursivos y cadenas léxicas.

Ya existen algunos analizadores automáticos de la estructura discursiva basados en la RST para distintas lenguas. Véase Marcu (1998) para el inglés, Pardo et al. (2004) para el portugués, Sumita et al. (1992) para el japonés y da Cunha et al. (2012) para el español.

4.5.2. Posibles aplicaciones del análisis discursivo al desarrollo y evaluación de los sistemas de traducción automática

A diferencia de los aspectos léxico, sintáctico o semántico, el análisis discursivo no se ha implementado en el desarrollo o en la evaluación de los sistemas de TA. La gran mayoría de sistemas operan oración por oración. De la misma manera, las métricas empleadas para la evaluación automática realizan la comparación entre la TA y la TH a nivel oracional. Ello supone una limitación importante de los sistemas actuales, ya que afecta la calidad global de la traducción.

Una manera de aplicar el análisis discursivo a la evaluación automática de TA es medir la coherencia de los TTs generados por los sistemas. En el PLN se suele distinguir entre dos tipos de coherencia textual: la coherencia referencial (*entity-based coherence*) y la coherencia relacional (*relation-based coherence*). La coherencia referencial describe las relaciones entre el texto y los eventos que introduce, mientras que la coherencia

relacional caracteriza el texto en términos de relaciones de sentido que se establecen entre los enunciados que lo componen.

Lapata y Barzilay (2005) investigan la posibilidad de la evaluación automática de la coherencia referencial local, es decir, realizan la evaluación a nivel de transiciones de oración a oración. El objetivo de su trabajo es proponer un modelo cuantitativo que mida la coherencia del texto a partir de sus características superficiales. Los autores utilizan dos modelos de la distribución de entidades en el discurso: el modelo sintáctico y el modelo semántico. El modelo sintáctico describe cómo están distribuidas en el texto las menciones de la misma entidad en diversas posiciones sintácticas. Este modelo, basado en *Centering theory* de Grosz et al. (1995), representa el texto como un conjunto de secuencias de transiciones de entidades y define un modelo probabilístico sobre su distribución. El modelo semántico cuantifica la coherencia como el grado de similitud semántica entre las oraciones adyacentes.

Lin et al. (2011) mejoran los resultados de Lapata y Barzilay al combinar el aspecto referencial y el aspecto relacional de la coherencia textual. La idea fundamental de su trabajo es que los textos coherentes muestran una preferencia por un orden de segmentos determinado. Lin y sus colaboradores anotan los textos con las relaciones discursivas al estilo de la RST. Los textos se representan por medio de una matriz en la que se registran todas las ocurrencias de las palabras de contenido junto con el papel discursivo que desempeñan, definido a través de la función de la unidad discursiva en la que se encuentran. El modelo de Lin et al. (2011) obtiene mejores resultados que la métrica de Lapata y Barzilay (2005), demostrando la importancia del análisis de la estructura discursiva para la evaluación de la coherencia. Las métricas propuestas por Lapata y Barzilay (2005) y Lin et al. (2011) podrían aplicarse a la evaluación automática de la TA en el aspecto textual.

Asimismo, algunos autores opinan que el hecho de que las lenguas muestren preferencias distintas en la organización textual debe tomarse en cuenta para el desarrollo de sistemas de TA. Debido a que los sistemas de TA actuales procesan el texto fuente oración por oración, no son capaces de reagrupar frases u oraciones para conseguir la organización textual propia de la LM. Este problema se aborda en el estudio de Marcu et al. (2000), quienes demuestran que la estructura discursiva se

modifica en el paso del TO al TT y proponen un modelo para el módulo de transferencia discursiva que, al ser integrado a un sistema de TA, se encargaría de reescribir la estructura discursiva del texto fuente tomando en cuenta las restricciones propias de la LM.

Para ello, el equipo de Marcu compila un corpus paralelo que se compone de 40 TOs en japonés y sus traducciones al inglés seleccionados de manera aleatoria del ARPA corpus (White et al., 1994). Después realizan la anotación discursiva de los textos del corpus al estilo de la RST y proceden a comparar las estructuras discursivas de los TOs y los TTs. Los resultados de la comparación indican que las estructuras discursivas del texto fuente y del texto meta difieren tanto a nivel de oración como a nivel de párrafo y de texto. Los traductores reagrupan cláusulas, oraciones, párrafos, reorganizando la información del TO a fin de reflejar las restricciones propias de la LM. De esta manera, para que el TT no sólo sea gramatical sino también coherente, la estructura discursiva del TT debe corresponder a las particularidades de la organización textual propias de la LM, no de la LF. Así, en palabras de Marcu et al. (2000: 16):

> If a translation system is to produce text that is not only grammatical but also coherent, it will have to ensure that the discourse structure of the target text reflects the natural renderings of the target language, not that of the source language.

Con base en los TOs y las THRs anotados con relaciones discursivas, Marcu et al. (2000) entrenan un modelo de transferencia que aprende las transformaciones que deben realizarse en la traducción con respecto a la estructura discursiva del TO.

Sin embargo, es poco lo que sabemos sobre las diferencias entre las estructuras textuales preferidas en lenguas distintas. Asimismo, el hecho de que la estructura discursiva del TO deba ser modificada en el TTs es un punto discutible. Así, Hatim y Mason (1995: 247) afirman que:

> […] la secuencia de las relaciones de coherencia habrá de resistir, en circunstancias normales, el paso del texto original a la versión traducida. Y es que relaciones básicas como causa-efecto, el problema y su solución, la secuencia temporal, etc., son fundamentos universales del significado y de la estructuración de éste en un texto. Ahora bien, habrá más probabilidad de que los modos en que se refleje esta coherencia de base en los elementos superficiales, es decir, la cohesión o

conectividad secuencial de dichos elementos, sean específicos de los distintos idiomas o incluso de los distintos textos.

Determinar el grado de variación aceptable de la estructura discursiva al traducir de una lengua a otra sería un tema interesante para futuras investigaciones.

4.5.3. Procedimientos del análisis comparativo de traducciones humanas y automáticas a nivel discursivo

La finalidad del análisis discursivo en el contexto de la comparación TA-TH es doble. Por un lado, sirve para comparar los patrones de organización textual en la TA y en la TH[26]. Por otro lado, da luz sobre aquellos rasgos distintivos de la TA que afectan la coherencia textual y, de esta manera, permite estimar la gravedad de los errores cometidos por los sistemas.

Debido a que de momento no existe una herramienta de anotación discursiva automática para el español, en el presente estudio realizamos únicamente el análisis cualitativo de 9 ejemplos del corpus que ilustran las diferencias en la estructuración del discurso en la TH y en las traducciones de *Google* y de *Lucy*. Dejamos para el trabajo futuro el análisis cuantitativo que consistiría en la anotación discursiva y el conteo del número de las EDUs y de los diferentes tipos de relaciones discursivas en todos los TTs.

Realizamos la anotación manual de los ejemplos escogidos para el análisis, con las relaciones discursivas de la RST, siguiendo la metodología desarrollada en da Cunha e Iruskieta (2010). En cuanto a la segmentación, en la línea de los trabajos de Tofilosky et al. (2009) y da Cunha e Iruskieta (2010), se consideran EDUs los elementos del texto que: a) presenten claramente una relación discursiva y b) contengan un verbo conjugado, un gerundio o un infinitivo. No se consideran EDUs las cláusulas subordinadas de relativo ni de complemento. En cuanto a la detección de relaciones discursivas, se utiliza la lista extendida de relaciones de Mann (2005).

El aparato metodológico desarrollado por Mann y Thompson (1988), en principio, no está destinado para el análisis de textos mal formados. Por tanto, a fin de adaptarlo a la

[26] Mejor dicho, ilustra las diferencias en la estructuración discursiva propia de la LF y de la LM, ya que en la TA no se realiza ningún tipo de modificaciones a nivel extra-oracional y, por tanto, se reproduce de manera literal la estructura discursiva del original.

tarea de anotación de TAs, introducimos la marca "?" para los casos en los que es imposible interpretar una relación discursiva entre las EDUs debido a una falta de coherencia en la TA. Por ejemplo:

> TO. [Thus, [even if gene vaccines did activate immunity in these individuals,]S_CONCESIÓN(2) the responses might not be easily noticeable.]N(1)
> TH. [Por tanto, [aun cuando las vacunas génicas activaran la inmunidad en estos individuos,]S_CONCESIÓN(2) la respuesta emitida apenas si resultaría perceptible.]N(1)
> *Google*. [Por tanto, [incluso si las vacunas de genes de inmunidad hizo activar en estos individuos,]?(2) las respuestas pueden no ser fácilmente perceptible.](1)[27]

En este caso en la TA se produce una alteración de la estructura argumental de la oración original. En el original y en la TH "la inmunidad" es argumento del verbo "activar", mientras que en la TA es modificador del sustantivo "genes". Debido a esta alteración, es imposible interpretar la relación discursiva entre las EDUs en la traducción de *Google*.

Ahora bien, las diferencias en la estructura discursiva pueden darse tanto en la segmentación como en las relaciones discursivas. Se producen, por un lado, a causa de las modificaciones que realiza el traductor con respecto a la estructura discursiva del original para reflejar las restricciones de organización textual de la LM o por los procesos de simplificación o explicitación propios de la actividad traductora, que los sistemas de TA no son capaces de llevar a cabo. Por otro lado, dichas diferencias pueden deberse a errores léxicos o sintácticos de los sistemas, a causa de los cuales a veces es imposible interpretar las relaciones de coherencia en la TA.

Así, al comparar las estructuras discursivas de la TH con las de la TA, pueden darse las situaciones siguientes:

1. No coincide la *segmentación* discursiva:
    a. debido a una modificación realizada por el traductor humano;
    b. debido a los errores cometidos por los sistemas de TA.
2. No coinciden las *relaciones* discursivas:

[27] Los corchetes indican los límites de las EDUs, la letra N se refiere al núcleo, y la letra S denomina el satélite.

a. debido a una modificación de la estructura discursiva del TO en la TH;
b. debido a los errores de la TA (en este caso la diferencia entre las estructuras discursivas indicaría una falta de coherencia en la TA).

3. Las relaciones discursivas coinciden, pero no están marcadas de la misma manera (consideramos que la relación entre las EDUs está marcada sólo si está presente un elemento cuya función primaria es codificar dicha relación).

De la misma manera que en los niveles léxico-terminológico y morfosintáctico, a nivel discursivo relacionamos las diferencias TA-TH con las estrategias de la traducción y las clasificamos de acuerdo con la taxonomía que presentamos en el siguiente apartado.

4.6. Clasificación de las diferencias entre la traducción humana y la traducción automática

En la última fase del análisis, a partir de una muestra aleatoria de oraciones de nuestro corpus, clasificamos las diferencias TA-TH en términos de las modificaciones que realizan los traductores humanos frente a los sistemas de TA con respecto al TO.

La clasificación nos sirve para identificar las regularidades en el comportamiento de los traductores y de los sistemas de TA que aportan a la conformación de los rasgos distintivos de la TA detectados mediante el análisis cuantitativo.

Para llevar a cabo esta tarea, nos basamos en la propuesta de clasificación de *translation shifts* desarrollada por van Leuven-Zwart (1989) (véase apartado 3.1.2.). De cara a la comparación TA-TH introducimos las modificaciones siguientes a la propuesta original de esta autora. En primer lugar, tal como se ha mencionado en el apartado 3.1.2., la propuesta de van Leuven-Zwart es demasiado detallada para implementarla en el presente estudio, ya que la cantidad de datos que analizamos de manera manual no es grande, y tener una clasificación detallada no nos permitiría llegar a generalizaciones sobre el comportamiento de los traductores y los sistemas de TA. Por esta razón, tomamos en cuenta únicamente las categorías generales de la clasificación de van Leuven-Zwart: modulación (generalización vs. especificación), modificación y mutación (omisión vs. adición).

En segundo lugar, para las categorías "modulación/generalización" y "modulación/especificación" de van Leuven-Zwart (1989) usaremos los términos

explicitación e implicitación, respectivamente, y los emplearemos en un sentido amplio, siguiendo a Becher (2011: 18), quien define estas categorías de la manera siguiente:

> Implicitness is the non-verbalization of information that the addressee might be able to infer.
>
> Explicitness is the verbalization of information that the addressee might be able to infer if it were not verbalized.
>
> Explicitation is observed where a given target text is more explicit tan the corresponding source text.
>
> Implicitation is observed where a given target text is less explicit (more implicit) than the corresponding source text.

En tercer lugar, a diferencia de van Leuven-Zwart (1989), tomaremos en cuenta no solamente los *translation shifts* opcionales, sino también las desviaciones obligatorias que se deben a las diferencias sistémicas entre las lenguas. Consideramos que esta distinción es relevante para la evaluación de los sistemas de TA, ya que las diferencias TA-TH relacionadas con las modificaciones opcionales realizadas por el traductor humano y las diferencias que se deben a la falta de modificaciones obligatorias en la TA no reflejan la calidad de la TA de la misma manera.

En cuarto lugar, la unidad de análisis que en el estudio de van Leuven-Zwart (1989) es el transema, en nuestro caso dependerá del nivel de análisis. Así, a nivel léxico-terminológico realizaremos la clasificación únicamente con respecto a las UTs, es decir, hablaremos de explicitación, implicitación o modificación con respecto al significado léxico codificado por medio de las UTs. A nivel morfosintáctico, aplicaremos dichas categorías al significado gramatical de palabras, frases o cláusulas. A nivel discursivo, hablaremos de explicitación o implicitación con respecto a la marcación de las relaciones de coherencia entre las EDUs. En cuanto a la modificación, usaremos esta categoría para los cambios en la segmentación o en la estructura discursiva del original en la TH o la TA. Siguiendo a van Leuven-Zwart (1988), distinguimos entre explicitación/implicitación como uso de unidades más o menos informativas y adición/omisión de las unidades de análisis (términos a nivel léxico-terminológico, frases o cláusulas a nivel morfosintáctico y EDUs a nivel discursivo).

Finalmente, introducimos dos categorías adicionales relacionadas con la naturaleza de los textos a analizar y con los propósitos de la clasificación. Primero, usaremos la

categoría "alteración" para los casos en los que el significado de la unidad de análisis del original es alterado debido a los errores de la TA. Segundo, anotaremos como "variación" a nivel léxico-terminológico aquellos casos en los que el humano y la máquina ofrecen variantes, que, de acuerdo con los recursos de referencia que usamos, son posibles equivalentes del término.

Así, la clasificación se realizará a partir de la taxonomía incluida en la Tabla 8.

Nivel de análisis	**Origen de la diferencia**		
	TH	**TA de *Google***	**TA de *Lucy***
Léxico-terminológico	explicitación implicitación modificación variación omisión adición	explicitación implicitación modificación variación omisión adición alteración	explicitación implicitación modificación variación omisión adición alteración
Morfosintáctico	explicitación implicitación modificación omisión adición	explicitación implicitación modificación omisión adición alteración	explicitación implicitación modificación omisión adición alteración
Discursivo	explicitación implicitación modificación omisión adición	explicitación implicitación modificación omisión adición alteración	explicitación implicitación modificación omisión adición alteración

Tabla 8. Clasificación de las diferencias TA-TH en términos de *translation shifts*

5. ANÁLISIS Y RESULTADOS

En este capítulo se presentan los resultados de la investigación. En primer lugar, se describen brevemente los sistemas de TA utilizados (apartado 5.1.). En segundo lugar, se ofrecen las características del corpus de estudio (apartado 5.2.). En tercer lugar, se presenta el análisis y los resultados de la investigación a nivel léxico-terminológico (apartado 5.3.), morfosintáctico (apartado 5.4.) y discursivo (apartado 5.5.).

5.1 Selección de sistemas de traducción automática

Las características de los textos traducidos de manera automática están condicionadas por el tipo de sistema y por las diferencias lingüísticas entre la LF y la LM. Como se ha mencionado anteriormente (véase apartado 2.2.), existen varios tipos de sistemas de TA. Para los fines de esta investigación seleccionamos un sistema basado en reglas y un sistema estadístico, ya que sus traducciones presentan carencias y errores muy distintos y, además, una de las preguntas que nos interesa discutir de cara a la comparación TA-TH es qué tipo de modificaciones realizadas por el traductor humano son capaces de reproducir los sistemas de TA basados en estrategias diferentes.

5.1.1. Lucy LT

Lucy es un sistema de TA comercial creado por la empresa *Lucy Software*. El sistema está basado en el modelo de transferencia y se ha desarrollado a partir de la arquitectura del sistema *Metal* (véase el apartado 2.3.2.2.).

Lucy tiene la información léxica almacenada en forma de vocabularios. El sistema dispone de vocabularios monolingües para la LF y la LM (el mismo vocabulario se utiliza en las fases de generación y de análisis), así como de un vocabulario bilingüe para la fase de transferencia. Las entradas del vocabulario monolingüe contienen la información morfológica, sintáctica y semántica. Las unidades formadas por medio de conversión tienen una entrada aparte si están lexicalizadas y las unidades poliléxicas tienen una entrada aparte con la información sobre su estructura morfosintáctica. El vocabulario bilingüe indica los equivalentes de las palabras de la LF en la LM. Los vocabularios están organizados de acuerdo con el criterio temático: vocabulario general, vocabulario social general (que a su vez se divide por subtemas: arte, humanidades, etc.), vocabulario técnico general (que se divide en informática, medicina, etc.). Así,

para traducir los textos de nuestro corpus, escogimos la opción vocabulario técnico general. El sistema resuelve la ambigüedad léxica por medio de pruebas de contexto en las que se toman en consideración las categorías gramaticales y las clases semánticas de las palabras adyacentes. Si no es capaz de resolver la ambigüedad, ofrece al usuario todas las traducciones posibles de la palabra ambigua[28].

En cuanto a la información sintáctica, ésta está representada en forma de reglas de reescritura. En la etapa de análisis se realiza el etiquetado POS del TO y, a continuación, las reglas de reescritura se aplican de manera recursiva para generar una representación sintáctica bien formada de la oración. La selección entre todos los análisis sintácticos posibles se realiza con base en principios heurísticos. En general, el sistema siempre da preferencia a la lectura con más evidencia explícita (la información semántica no se toma en consideración para la resolución de ambigüedades sintácticas). En el caso de una ambigüedad de relaciones sintácticas el sistema selecciona por defecto la lectura hundida de los complementos (por ejemplo, los complementos preposicionales se ensamblan al nivel más bajo posible).

5.1.2. Google Translate

Google es un sistema de TA desarrollado por el grupo de investigación de la empresa Google. El sistema es de libre acceso a través del Internet. *Google* está basado en técnicas estadísticas y genera traducciones a partir de los patrones identificados en un gran número de textos disponibles en Internet. De acuerdo con Hoang et al. (2009), *Google* utiliza los modelos de TA estadística basados en frases (véase apartado 2.3.3.)[29]. La calidad de la traducción es dependiente de lengua, ya que varía en función del tamaño del corpus de entrenamiento disponible para un par de lenguas determinado.

Los sistemas de TA que empleamos para la traducción de textos especializados son de uso general (es decir, no están adaptados a la tarea de traducción de textos especializados en el ámbito médico), lo cual supone una limitación importante del presente estudio. Además, como ya hemos mencionado, en numerosas ocasiones *Lucy* no realiza la desambiguación léxica ofreciendo al usuario varias traducciones posibles

[28] Dado que se trata de un sistema comercial, existe la posibilidad de integrar a la base de datos los glosarios requeridos o consensuados por el usuario. Nosotros no recurrimos a esta posibilidad.
[29] No hemos podido encontrar una descripción detallada de este sistema.

para las palabras ambiguas. Para poder realizar el análisis necesitamos seleccionar una única opción entre las variantes que ofrece *Lucy*. Decidimos hacerlo con base en la comparación de las opciones que propone el sistema basado en reglas con la TA de *Google*. Así, escogemos la opción más adecuada, si *Google* también realiza una selección apropiada. De lo contrario, se selecciona la opción equivocada. En el siguiente ejemplo *Lucy* ofrece tres variantes para la traducción de la palabra *right* (una palabra ambigua desde el punto de vista del sistema). Escogemos la opción adecuada - "correctas", ya que *Google* ofrece esta misma opción.

> TO. Those messages elicit the right responses only because they are transmitted accurately far into a recipient cell and to the exact molecules able to carry out the directives.
> *Lucy*: Esos mensajes obtienen las respuestas <A[**correctas**|de derechas|derechas]> solamente porque se transmiten con precisión lejos en una célula de receptor y a las moléculas exactas capaces de realizar las directivas.
> *Google*: Esos mensajes obtener las respuestas **correctas** sólo porque se transmiten con precisión el momento en una célula receptora y las moléculas exactas capaces de llevar a cabo las directivas.

5.2. Constitución del corpus

Nuestro corpus de análisis está compuesto de textos que forman parte del Corpus Técnico del IULA. Este corpus fue diseñado en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra de Barcelona. Es un corpus multilingüe especializado que contiene documentos en cinco lenguas (catalán, español, inglés, francés, alemán) en varias áreas de especialidad (economía, derecho, medioambiente, medicina, información), así como documentos paralelos. Todos los textos del corpus fueron seleccionados por especialistas y clasificados según el área y el sub-área al que pertenecen, esto último para facilitar el procesamiento y la extracción automática de información a partir del corpus.

5.2.1. Género de divulgación científica

En el presente trabajo se analizan las traducciones de textos especializados, en específico, artículos de divulgación científica. El género de divulgación es interesante desde el punto de vista de la investigación en la TA, puesto que en él se combinan los rasgos prototípicos de los textos de alto nivel de especialización, que tienden a ser sistemáticos y evitar la ambigüedad al máximo posible, y de los textos periodísticos,

que presentan más diversidad en la selección del léxico y el uso de construcciones sintácticas.

A continuación se resumen brevemente algunas de las características principales de los artículos de divulgación científica. Un texto especializado es aquella producción lingüística que sirve para expresar y transmitir conocimiento especializado; que tiene una serie de rasgos lingüísticos que le dan una especificidad en el conjunto de textos producidos en la lengua; y que, además, presenta un conjunto de características pragmáticas determinadas por los elementos específicos del proceso de comunicación (el tema, los interlocutores y la situación comunicativa) (Cabré, 2003).

Los artículos de divulgación pertenecen al discurso especializado, pero poseen ciertas características lingüísticas particulares relacionadas con el objetivo general de este tipo de textos: familiarizar al público no experto con los avances científicos, lo cual implica una reformulación del mensaje original (artículos científicos) para una nueva audiencia. La particularidad del destinatario y de los objetivos de la divulgación científica se refleja en sus características lingüísticas en comparación con los artículos científicos. A nivel léxico, se produce una reducción del vocabulario de especialidad y su sustitución por el léxico de uso más extendido, así como la utilización de diversas estrategias para aclarar el significado de los términos. A nivel sintáctico, el paso del discurso profesional al discurso de divulgación implica varios cambios encaminados a aumentar la legibilidad del texto (por ejemplo, reducción de la longitud y complejidad de las oraciones y de las estructuras sintácticas, la sustitución de las construcciones pasivas e impersonales por activas, la reducción del número de las nominalizaciones típicas de los textos profesionales).

También se dan algunos cambios importantes a nivel de la organización global del texto y en el uso de mecanismos de cohesión. La adaptación del discurso para los nuevos lectores supone la pérdida de la organización textual convencional del artículo científico en torno al esquema "introducción-métodos-resultados-discusión". Los textos de divulgación tienden a seguir un orden cronológico: la narración de los distintos experimentos llevados a cabo por los investigadores o de las actividades del organismo objeto de estudio. En cuanto a los mecanismos de cohesión, Myers (1991) comenta las distintas preferencias por diversas formas de cohesión léxica en los textos profesionales

y de divulgación, tal que los primeros tienden a hacer uso abundante de la repetición léxica, en tanto los segundos prefieren la cohesión mediante sinónimos u otros tipos de expresiones referenciales. De acuerdo con Myers (1991), la cohesión mediante sinónimos permite a los lectores de textos divulgativos realizar inferencias acerca del significado de los términos que desconocen y, por tanto, facilitan la comprensión de los textos para lectores no expertos. Uno de los criterios importantes de la legibilidad es el diseño de párrafos cortos y bien ligados con aquellos que les siguen y les preceden. En este sentido, es importante en los textos divulgativos el empleo de conectores textuales que expresan de manera explícita la relación entre diversas oraciones o párrafos del texto.

5.2.2. Descripción del corpus de estudio

Para fines de esta investigación compilamos un corpus de 40 textos: 10 artículos publicados en inglés en la revista *Scientific American* entre los años 1994 y 2000, 10 traducciones de estos textos al español publicadas en la versión española de dicha revista, *Investigación y Ciencia*, 10 TAs inglés-español realizadas por el sistema estadístico *Google*, y 10 TAs inglés-español realizadas por el sistema basado en reglas *Lucy*.

Scientific American es una prestigiosa publicación mensual editada en Estados Unidos. Todos los artículos publicados en la revista están escritos por especialistas en el tema específico del artículo. La divulgación de *Scientific American* suele caracterizarse como una divulgación sofisticada o de alto nivel, tanto por la condición de especialistas de los divulgadores, como por las características de los lectores previstos, a quienes se les supone una cierta formación científica y un interés por el tema del que trata el artículo.

Investigación y Ciencia es la versión española de *Scientific American.* La traducción de los artículos de *Scientific American* es encargada a un investigador experto en la materia. No se trata de un equipo fijo de traductores, ya que éstos no forman parte del personal de la revista, sino que están adscritos a un centro de enseñanza o de investigación y son invitados a colaborar esporádicamente con *Investigación y Ciencia.* Los traductores apenas reciben indicaciones de ningún tipo por parte de la editorial sobre cómo llevar a cabo su labor. La intención fundamental al confiar la traducción a un experto en la materia es que los términos empleados en la traducción sean los que se

usan en la comunidad científica. Fernández Polo (1999) realiza una encuesta entre los traductores que han colaborado regularmente con la revista, en la cual se les pide que valoren la importancia de una serie de factores relacionados con su actividad para descubrir sus prioridades a la hora de llevarla a cabo. De acuerdo con dicha encuesta, los traductores conciben las finalidades de su trabajo, en orden descendiente de importancia, como:

1. Conservar toda la información del original.
2. Velar por la legibilidad de la versión traducida.
3. Emplear un léxico y sintaxis correctos.
4. Emplear una terminología apropiada.
5. Hacer que el texto suene como un original en castellano.
6. Procurar que la traducción sea lo más amena posible.

La importancia concedida por los traductores al componente informativo y la poca importancia que conceden al componente de entretenimiento se debe a que la versión en español prevé unos lectores ligeramente distintos de los correspondientes a la versión inglesa. Myers (1990) caracteriza a los lectores de esta última como público general, por oposición a los especialistas previstos para los textos profesionales. En la misma encuesta se sondeó la opinión de los traductores acerca del tipo de lectores que preveían para la revista, con el fin de comprobar si coincidían con el perfil de los lectores de la versión inglesa. En orden descendiente de importancia, los lectores prototípicos de *Investigación y Ciencia* serían, según la opinión de los traductores encuestados:

1. Investigadores o profesionales expertos en el tema del artículo.
2. Estudiantes universitarios expertos.
3. Investigadores o profesionales no expertos en el tema del artículo.
4. Profesionales de enseñanza secundaria.
5. Estudiantes universitarios no expertos.
6. Otros.

El mayor grado de especialización previsto para los lectores de la versión en español, según esta escala, pudiera explicar algunos de los cambios que observaremos en las traducciones: en concreto, la supresión de ciertas señales destinadas a facilitar la lectura

de los textos, que serían previsiblemente redundantes o inútiles para los lectores a los que se les presupone un alto grado de familiarización con el tema.

Dentro de las revistas de divulgación, *Scientific American* es la única publicación traducida sistemáticamente del inglés al castellano. El propio hecho de que los textos hayan sido aceptados para su publicación en dos revistas prestigiosas como *Scientific American* e *Investigación y Ciencia* supone una garantía de que dichos textos responden a las expectativas que tienen los hablantes nativos de ambas lenguas sobre la forma de este tipo de materiales escritos.

Hemos descrito los TOs y las THs que forman parte de nuestro corpus. En cuanto a las TAs, son realizadas por los sistemas *Google* y *Lucy*, cuya descripción ofrecimos en el apartado anterior.

En la Tabla 9 presentamos las estadísticas del corpus.

	No. de textos	**No. de oraciones**	**No. de palabras**
TO	10	1134	24 053
TH	10	1156	25 305
Google	10	1134	27 547
Lucy	10	1134	26 206

Tabla 9. Estadísticas del corpus de estudio

Tal como indica la Tabla 9, el número de oraciones es el mismo en el original y en las TAs, ya que los sistemas de TA no realizan ningún tipo de modificaciones a nivel extra-oracional. Mientras tanto, en la TH el número de oraciones es mayor, lo cual podría explicarse a partir de la hipótesis de simplificación, desarrollada en el marco de los estudios descriptivos de traducción: en los TTs suelen emplearse oraciones más cortas que en los originales. En cuanto al número de palabras, éste es menor en la TH debido, posiblemente, a que, como veremos a continuación, las THs tienden a ser más concisas que sus respectivos TOs.

5.3. Análisis léxico-terminológico

A nivel léxico-terminológico comparamos el tratamiento de la terminología en la TH y en las TAs de *Google* y de *Lucy*. Para ello, de acuerdo con el procedimiento descrito en el apartado 4.3.3., se realiza la extracción automática de términos con ayuda del extractor terminológico basado en *Wikipedia* de Vivaldi y Rodríguez (2011) y se lleva a cabo el proceso de filtrado manual. Así, se obtienen tres listas de palabras: los CTs extraídos de la TH, de la TA de *Google* y de la TA de *Lucy* (la tabla con todos los CTs extraídos de los tres grupos de textos se ofrece en el Anexo A).

En la Tabla 10 se presentan las características cuantitativas del tratamiento de la terminología en la TH y en las TAs.

	TH	***Google***	***Lucy***
No. total de tipos de CTs	200	183	136
No. total de ocurrencias de CTs	1528	1635	1612
Type-token ratio	13%	11%	8%
No. de CTs con el CD = 1	27%	22%	24%
No. total de diferencias TA-TH		17%	26%

Tabla 10. Características cuantitativas del tratamiento de la terminología en la TH y en las TAs de *Google* y de *Lucy*

En las primeras dos filas de la Tabla 10 se indica el número total de tipos (*types*) y de ocurrencias (*tokens*) de CTs. En la tercera fila se indica el *type-token ratio*. En la cuarta fila se presenta el porcentaje de tipos de CTs con el grado de pertenencia al dominio alto (CD = 1). En la quinta fila se indica el número total de diferencias TH-TA de *Google* y TH-TA de *Lucy*.

El número de tipos de CTs identificados por el extractor terminológico en la TH es mayor que en las TAs. Al realizar la traducción de las UTs, los sistemas de TA ofrecen variantes que no pertenecen al dominio de especialidad a causa de los problemas de ambigüedad o la falta de cobertura del vocabulario. Además, el extractor de Vivaldi y

Rodríguez (2011) sólo detecta las unidades con función referencial y no asigna ningún coeficiente a los verbos, de manera que el número elevado de nominalizaciones en la TH tiene un impacto en los resultados. Por último, en la TH se observa una tendencia a cambiar el nivel de especialización, al usarse UTs incluso en aquellos casos en los que en los TOs se utilizan palabras de uso generalizado (de ahí que el número de CTs con el CD más alto sea mayor en la TH, tal como se observa en la cuarta fila de la Tabla 10).

El número de CTs extraídos de la TA de *Google* es mayor que el número de CTs identificados en la TA de *Lucy*. Además, el número de diferencias TA-TH es mayor en el caso del sistema basado en reglas. Ello se debe a que, como era de esperar, dada la naturaleza de los sistemas estadísticos, las traducciones de *Google* se acercan más a la selección léxica de la TH. El número total de las diferencias confirma la hipótesis ya comprobada en otros trabajos (Coughlin, 2003), de que los sistemas de TA estadística superan los sistemas basados en reglas en cuanto a la selección del léxico.

Type-token ratio es una medida de la variedad del léxico que se aplica, normalmente, a todas las palabras de los textos a caracterizar. En esta investigación la aplicamos a las listas de CTs extraídos automáticamente de los textos del corpus a fin de proporcionar una caracterización cuantitativa de la variedad del vocabulario especializado en los TTs. La variedad terminológica es mayor en la TH debido al uso de sinónimos, hipónimos e hiperónimos para la traducción del mismo término original. A continuación ofrecemos un fragmento de la tabla con los CTs alineados (la tabla completa se ofrece en el Anexo B), que nos sirve para ilustrar las diferencias en el grado de la variedad del vocabulario entre la TH, la TA de *Google* y la TA de *Lucy*.

No.	**TO**	**TH (CD)**	***Google* (CD)**	***Lucy* (CD)**
909	stroke	accidente cerebrovascular (0.65)	carrera (0.00)	golpe (0.00)
911	stroke	accidente cerebro vascular (0.65)	derrame cerebral (0.62)	golpe (0.00)
915	stroke	ictus (0.62)	accidente cerebrovascular (0.65)	golpe (0.00)
852	seizure	convulsión (1.00)	convulsión (1.00)	toma (-1.00)
911	seizure	convulsión (1.00)	embargo (0.00)	toma (-1.00)
915	seizure	convulsión (1.00)	ataque (0.85)	toma (-1.00)

Tabla 11. CTs extraídos de los TTs y las unidades correspondientes de los TOs

En la primera columna registramos el número de la oración. En la columna "TH" encontramos los CTs extraídos de la TH. En las columnas "TO", "*Google*" y "*Lucy*" se registran las unidades correspondientes de los TOs y las TAs. Para cada unidad objeto de análisis se registra la información sobre su CD.

En la Tabla 11 observamos que *Lucy*, aun realizando una selección poco apropiada, al ofrecer "golpe" como equivalente de *stroke*, y "toma" como equivalente de *seizure*, lo hace siempre de la misma manera, de acuerdo con el diccionario bilingüe que utiliza. En cambio, *Google* realiza la traducción en función del contexto lingüístico inmediato y ofrece varias opciones para el mismo término ("derrame cerebral", "accidente cerebrovascular" y "carrera" para *stroke* y "convulsión", "embargo" y "ataque" para *seizure*), algunas de las cuales son adecuadas, mientras que otras son completamente erróneas, con lo cual el TT pierde la coherencia a nivel del léxico.

Somos conscientes de que el análisis cuantitativo que realizamos tiene limitaciones importantes, debido a los errores del extractor terminológico y al tamaño reducido de las listas de CTs, a las que aplicamos la medida *type-token ratio*. El CD proporcionado por el extractor basado en *Wikipedia* no siempre es indicador fiable de la pertenencia al dominio o de la terminologicidad de las unidades del léxico. Por este motivo, en el futuro tenemos pensado realizar experimentos con otras estrategias de extracción de términos y de medición de las características de los textos a nivel léxico-terminológico.

Ahora bien, con respecto a la naturaleza de las diferencias detectadas, a partir del análisis de una muestra aleatoria de 50 oraciones de nuestro corpus, obtenemos los resultados incluidos en la Tabla 12:

Translation shift	**TH**		***Google***		***Lucy***	
	Cantidad	**%**	**Cantidad**	**%**	**Cantidad**	**%**
Adición	3	6	0	0	0	0
Omisión	2	4	0	0	0	0
Explicitación	8	15	2	8	0	0
Implicitación	8	15	0	0	0	0
Modificación	17	32	5	21	0	0
Variación	15	28	10	42	11	31
Alteración	0	0	7	29	25	69
Total	**53**		**24**		**36**	

Tabla 12. Clasificación de las diferencias TH-TA de *Google* y TH-TA de *Lucy* a nivel léxico-terminológico

En primer lugar, la Tabla 12 indica que el número de diferencias que se originan en la TH es mayor que el número de diferencias que se originan en las TAs. Ello se debe, por un lado, a que a nivel de la UT la calidad de la TA no es baja y, por otro lado, a que en la TH se producen numerosos *translation shifts* con respecto a las UTs de los originales.

En segundo lugar, el número de diferencias relacionadas con la TA de *Lucy* es mayor que el número de diferencias asociadas a la TA de *Google.* Asimismo, el número de alteraciones en la TA de *Lucy* es mayor que en la TA de *Google*. Ello se explica por el hecho de que, como se ha mencionado anteriormente (véase p. 108), los sistemas estadísticos superan a los sistemas basados en reglas en cuanto a la selección del léxico.

En tercer lugar, en el Anexo B (donde registramos la naturaleza obligatoria u opcional de los *translation shifts*) se observa que el número de *shifts* opcionales es mayor que el número de *shifts* obligatorios tanto en la TH como en la TA, debido a que en la traducción de términos son pocas las ocasiones en las que es necesario modificar el original a causa de las divergencias léxicas entre los sistemas de la LF y la LM.

Ahora bien, tal como se indica en la Tabla 12, el *translation shift* más frecuente en la TH es la **modificación**. Al analizar ejemplos concretos observamos que este procedimiento está relacionado con los factores de uso. Por ejemplo[30]:

1. TO. The tiny cells in our bodies harbor amazing internal communication networks.
   TH. Las células de nuestro organismo contienen unas redes de comunicación interna sorprendentes.
   Google. Las pequeñas células en nuestros cuerpos albergan increíbles redes de comunicación interna.
   Lucy. Las células diminutas en nuestros cuerpos albergan cadenas de comunicación internas asombrosas.

Suponemos que en los textos especializados la UT "célula" tiene mayor fuerza asociativa al relacionarse con el término "organismo" que con la palabra "cuerpo" (que en este caso actúan como sinónimos contextuales), por ello el traductor se permite la licencia de ofrecer el término "organismo" como equivalente de *body*.

Asimismo, la modificación en la traducción de UTs se produce en el contexto de las diferencias sintácticas entre el español y el inglés. Ilustramos esta observación con el ejemplo 2:

2. TO. On a simplistic level, neurobiologists associate the brainstem with the most basic functions: breathing, eating, balance, motor coordination and so forth.
   TH. Solemos asociar el tronco cerebral con funciones básicas, desde la respiración hasta la deglución, pasando por el equilibrio, la coordinación motora y otros.
   Google. En un nivel simple, los neurobiólogos asocian el tronco cerebral con las funciones más básicas: respirar, comer, el equilibrio, la coordinación motora y así sucesivamente.
   Lucy. En un nivel simplista, neurobiologists asocian el brainstem con las funciones más básicas: respirando, comiendo, equilibrio, coordinación de motor y así sucesivamente.

Las unidades léxicas *breathing*, *eating*, *balance*, *motor coordination* son UTs, ya que hacen referencia a las "funciones básicas" del organismo. De la misma manera que en el ejemplo anterior, en el original se utilizan denominaciones más comunes, mientras que en la TH se observa un incremento en el grado de especialización (*eating* > deglución). No obstante, la traducción del término *eating* como "deglución" no necesariamente está relacionada con una tendencia hacia el cambio del nivel de especialización en la TH,

[30] Las TAs presentan muchos errores, pero de aquí en adelante solamente discutiremos las diferencias TA-TH relacionadas con el tratamiento de los fenómenos objeto de análisis.

sino que podría explicarse desde el punto de vista estilístico: los términos de una enumeración tienen que ser similares en cuanto a su función sintáctica y su significado.

En cuanto a la TA, en la traducción de *Lucy* se produce una oración agramatical al no realizarse la modificación obligatoria (al gerundio nominal del inglés le corresponde en español el infinitivo o el nombre deverbal) debido a la ambigüedad de las formas en *-ing* del inglés. Mientras tanto, *Google* selecciona una opción estilísticamente imperfecta: "respirar, comer, el equilibrio, la coordinación motora".

Este ejemplo también ilustra uno de los problemas de la TA en el ámbito de la terminología, la traducción de los sintagmas nominales complejos del inglés al español (Maxwell, 1992). Esta tarea resulta problemática para el sistema basado en reglas, debido a que en inglés no hay marcas explícitas que permitan determinar la categoría gramatical del modificador en la frase *motor coordination* (ambigüedad categorial). *Lucy* asigna a esta frase nominal el patrón "Nombre + Nombre" y, de acuerdo con las reglas sintácticas del módulo de transferencia, la traduce por defecto como "Nombre + Frase Preposicional", cuando en realidad en este caso *motor* es un adjetivo que significa, de acuerdo con OALD, "connected with movement of the body that is produced by muscles; connected with the nerves that control movement".

En el ejemplo 2 se presenta otra alteración típica del sistema de TA basado en reglas, que se debe a la falta de cobertura de su vocabulario. Las palabras *neurobiologists* o *brainstem* no se encuentran en el diccionario del sistema y, por ello, *Lucy* no es capaz de traducirlas. Hay que mencionar que en otros ejemplos de nuestro corpus en los que el término *brain stem* está escrito por separado, *Lucy* ofrece una traducción adecuada, "tallo de cerebro". Así, tal como mencionamos en el apartado 2.5., al discutir las ventajas y desventajas de la TA estadística y de la TA basada en reglas, ésta última no es robusta y presenta errores en la traducción si en el original hay erratas o variantes ortográficas que no aparecen en el diccionario del sistema.

La siguiente categoría más frecuente en la TH es la **variación.** Las diferencias TA-TH que pueden caracterizarse como variación denominativa con frecuencia están relacionadas con el proceso de adaptación al dominio en la TH, que se manifiesta en la

preferencia de los traductores por el uso de UTs con mayor grado de especialización. Por ejemplo:

3. TO. Blood-forming stem cells, for example, give rise to every other type of blood cell (red cells, white cells of the immune system, and so on) and reconstitute the blood as needed; they also make more copies of themselves.

   TH. Las células hematopoyéticas, por ejemplo, originan todos los tipos celulares de la sangre (eritrocitos, leucocitos del sistema inmunitario, etc.) y reconstituyen la sangre cuando es necesario; también producen copias de sí mismas.

   Google. Que forman la sangre células madre, por ejemplo, dar lugar a cualquier otro tipo de células sanguíneas (glóbulos rojos, glóbulos blancos del sistema inmunológico, etc.) y la reconstitución de la sangre como sea necesario, sino que también hacen más copias de sí mismos.

   Lucy. Formando de sangre las células indiferenciadas, por ejemplo, dan lugar a cada dos tipos de célula sanguínea (células rojas, células blancas del sistema inmunitario, y así sucesivamente) y rehidratan la sangre como necesitada; también hacen más copias de ellos mismos.

En el OALD, *stem cell* se define como "a basic type of cell which can divide and develop into cells with particular functions". De acuerdo con la base de datos terminológica IATE, el equivalente de este término en español es "célula madre" (cf. la definición del término "célula madre" que ofrece el DRAE: "célula indiferenciada que puede dar lugar a distintos tipos de tejidos, como los constituidos por células hepáticas, nerviosas, epiteliales o a las diversas estirpes de células sanguíneas"). En la terminología médica del inglés existe la denominación *hematopoietic stem cell*, que significa "cell that can develop into any type of specialized blood cell". Dado el carácter divulgativo de los textos, en el original se usa una variante más transparente, más accesible para el lector. En cambio, en la TH se emplea el término *células hematopoyéticas*, lo cual conlleva, por un lado, la implicitación de la información que concierne a las características de las células madre, y, por otro lado, el aumento en el grado de pertenencia al dominio de especialidad.

Ahora bien, antes de relacionar este caso con la adaptación al dominio como uno de los universales de traducción, habría que pensar en las diferencias lingüísticas entre el inglés y el español que, probablemente, condicionan el *translation shift* presente en la TH. El empleo del participio presente con función adjetival es usual en inglés, pero el gerundio en español no posee esta función; por tanto, para traducir el modificador

blood-forming, el traductor se vería obligado a recurrir a otros recursos de modificación, por ejemplo, introducir una cláusula subordinada de relativo (una opción menos concisa). Precisamente, *Google* realiza esta modificación obligatoria, sin embargo, no logra cambiar el orden de palabras. Mientras tanto, *Lucy* ofrece una traducción literal de la estructura sintáctica del inglés, que resulta en una oración agramatical.

El cambio en el nivel de especialización en la TH se refleja con más claridad en la traducción de los términos *red cells* y *white cells*. En este caso no hay ninguna explicación tipológica que dé cuenta de la selección del traductor. En inglés existen los términos *erythrocyte* y *leukocyte*, y, al igual que en el caso anterior, en el original se opta por una denominación más accesible para el lector. Mientras tanto, el traductor prefiere usar las UTs con un grado de pertenencia al dominio mayor, debido, posiblemente, al tipo de lectores a quienes va dirigida la versión española de la revista *Scientific American* (véase el apartado 5.2.) Las opciones que ofrecen los sistemas de TA no coinciden con la TH. El traductor estadístico selecciona una opción más adecuada, más frecuente en el contexto de los textos especializados ("glóbulos rojos" y "glóbulos blancos"), mientras que *Lucy* ofrece una traducción literal ("células rojas" y "células blancas") que en este caso tampoco es inaceptable.

La **implicitación** (y otros tipos de *translation shifts*) en la TH se produce con frecuencia cuando una traducción literal conlleva la repetición léxica. En el ejemplo 4, a la segunda ocurrencia de la UT *cancer* le corresponde en la TH el término "patología". Cáncer es un tipo de patología, de manera que en este caso el traductor usa el hiperónimo para evitar la repetición léxica presente en el original:

4. TO. They are also finding that microsatellites change in length early in the development of some cancers, making them useful markers for early cancer detection.
   TH. Se está comprobando que los microsatélites cambian de longitud en fases precoces de ciertos cánceres, lo que les convierte en valiosos marcadores para el diagnóstico precoz de tales patologías.
   Google. Ellos también están encontrando que el cambio microsatélites de longitud temprana en el desarrollo de algunos tipos de cáncer, lo que los marcadores útiles para la detección temprana del cáncer.
   Lucy. Están encontrando también que los microsatélites cambian en longitud temprano en el desarrollo de algunos cánceres, haciéndoles marcadores útiles para la primera detección de cáncer.

Asimismo, la implicitación ocurre en aquellos casos en los que ciertos rasgos del significado de la UT original pueden inferirse con facilidad a partir del contexto lingüístico inmediato. En el ejemplo 5 observamos que el traductor prefiere usar una denominación más concisa al traducir la UT *phagocytic cells* como "fagocitos":

5. TO. At other times, it is strategically more advantageous for the bacterium not to interact with host cells - particularly phagocytic cells, which engulf and destroy bacteria.
   TH. En otras ocasiones le resulta a ésta más ventajoso no interaccionar con la célula huésped; en particular si se trata de fagocitos, que destruyen las bacterias tras atraparlas en su interior.
   Google. En otras ocasiones, es estratégicamente más ventajoso para la bacteria que no interactúan con las células huésped, particularmente las células fagocíticas, que se tragan y destruyen las bacterias.
   Lucy. En otros tiempos, es estratégicamente más ventajoso que la bacteria no interactúe con células de anfitrión - células especialmente fagocitarias, que inundan y destruyen bacterias.

El *translation shift* de **omisión** se produce en la TH en el contexto de la traducción de las UTs originales que designan a los investigadores. En numerosas ocasiones los términos que hacen referencia a los participantes de la investigación quedan implícitos en la TH, se omiten dejando lugar al uso del verbo en primera persona del plural, lo cual podría explicarse por el carácter divulgativo de los textos en los que es importante destacar el protagonismo de los investigadores. Por ejemplo:

6. TO. From this population of mutagenized bacteria or yeast, the geneticists can identify individuals not capable of replicating their DNA.
   TH. A partir de esta población de bacterias o levaduras que han experimentado mutagénesis, identificaremos individuos incapaces de replicar su ADN.
   Google. De esta población de bacterias o levaduras mutadas, los genetistas pueden identificar a los individuos que no son capaces de replicar su ADN.
   Lucy. De esta población de bacterias de mutagenized o levadura, los genetistas pueden identificar individuos no competentes de reproducir su ADN.

La **explicitación** en la TH se relaciona con la preferencia de los traductores por los términos con mayor grado de pertenencia al dominio, más específicos. Así, en el ejemplo 7 a la frase *the AIDS-causing agent* le corresponde en la TH la unidad "el agente etiológico del sida":

7. TO. Early clinical experiments using this strategy are now under way in cancer patients, as well as in those infected with HIV, the AIDS-causing agent, and other pathogens.

TH. Y se han puesto en marcha los primeros ensayos clínicos que emplean este método en el tratamiento no sólo del cáncer, sino también en infectados por el VIH, el agente etiológico del sida, y por otros patógenos.

Google. Los primeros experimentos clínicos utilizando esta estrategia están en marcha en pacientes con cáncer, así como en las personas infectadas con el VIH, el agente causante del SIDA, y otros agentes patógenos.

Lucy. Los primeros experimentos clínicos que utilizan esta estrategia están ahora en curso en pacientes de cáncer, así como en esos infectados con HIV, el agente que causa el SIDA, y otros pathogens.

De acuerdo con el DRAE, "etiológico" significa "perteneciente o relativo a la etiología", y "etiología" se refiere a "1. f. Fil. Estudio sobre las causas de las cosas. 2. f. Med. Estudio de las causas de las enfermedades. 3. f. Med. Estas causas", siendo esta última la acepción con la que se emplea el término en la TH. Así, el traductor utiliza una unidad léxica más informativa y en cierto sentido redundante (etiológico <causante de enfermedades> del SIDA).

La **adición** en la TH está condicionada por los factores pragmáticos. En el ejemplo 9, el traductor, además de dar el equivalente del término *enhancer* en español, ofrece, entre paréntesis, la versión original con el fin de informar al lector sobre la denominación del concepto en inglés:

8. TO. Walter Schaffner and Steven Lanier McKnight, among others, had additionally identified an unusual set of regulatory elements called enhancers, which facilitate transcription.

   TH. Walter Schaffner, Steven Lanier McKnight y otros identificaron, además, unos elementos reguladores nuevos, los intensificadores (enhancers), que estimulan la transcripción.

   Google. Walter Schaffner y Steven McKnight Lanier, entre otros, han identificado, además, un inusual conjunto de elementos reguladores llamados potenciadores, que facilitan la transcripción.

   Lucy. Walter Schaffner y Steven Lanier McKnight, entre otros, habían identificado adicionalmente un conjunto inusual de elementos legales llamados potenciadores, que facilitan transcripción.

Tal como observamos en la Tabla 12 en las traducciones de *Google* (pero no en las traducciones de *Lucy*) también se realizan *translation shifts* que en ocasiones coinciden con la TH. En el ejemplo 9, tanto en la TH como en la TA de *Google* se usa el término "cicatrización", lo cual supone una explicitación con respecto al significado léxico de la UT *healing*, ya que, de acuerdo con el OALD, el verbo *heal* significa "to become healthy again; to make something healthy again", mientras que el término "cicatrizar"

en español significa "completar la curación de las llagas o heridas, hasta que queden bien cerradas" (DRAE). En la TA de *Lucy* el significado de la UT *healing* es alterado, debido al problema de la ambigüedad de las formas en *-ing* del inglés, que ya hemos discutido anteriormente.

9. TO. These reactions ultimately stimulate proteins in the nucleus to activate genes that cause the cells to divide, an action that promotes wound healing.
   TH. Ras pone en marcha una serie de procesos enzimáticos que terminan por estimular proteínas del núcleo que activan genes promotores de la división celular, lo que redunda en la cicatrización de la herida.
   Google. Estas reacciones en última instancia, estimular las proteínas en el núcleo para activar los genes que causan que las células se dividen, una acción que promueve la cicatrización de heridas.
   Lucy. Estas reacciones en el fondo estimulan proteínas en el núcleo activar genes que causan las células para dividirse, una acción que promueve herida que se cura.

Como hemos visto en los ejemplos anteriores, la **alteración** en la TA de *Lucy* se produce ya sea en el contexto de ambigüedad léxica o estructural o a causa de la falta de cobertura del vocabulario. En cuanto a la TA de *Google*, los errores de los sistemas estadísticos son difíciles de explicar a partir de un análisis lingüístico (véase p. 22). Lo que observamos es que se manifiestan con frecuencia en una ordenación inadecuada de los elementos. En el ejemplo 10 (véase también ejemplo 3), *Google*, al traducir el sintagma terminológico *blood type*, no cambia el orden palabras, con lo cual se altera el significado del término original:

10. TO. Although many other genes appear in several forms - for example, the genes that encode eye color or blood type - highly conserved genes are not commonly found in multiple versions (also known as polymorphic alleles, or allelic variants).
    TH. Asimismo, muchos otros genes aparecen en formas diversas; por ejemplo, los que determinan el color de los ojos o el grupo sanguíneo. No ocurre así con los genes muy conservados, de los que no suele haber alelos polimórficos, o variantes alélicas.
    Google. Aunque muchos otros genes aparecen en varias formas, por ejemplo, los genes que codifican el color de ojos o la sangre de tipo altamente conservadas genes no se encuentran comúnmente en varias versiones (también conocido como alelos polimórficos, o variantes alélicas).
    Lucy. Aunque muchos otros genes aparecen en varias formas - por ejemplo los genes que codifican color de ojo o tipo de sangre - los genes altamente conservados no se encuentran comúnmente en versiones múltiples (también conocido como alleles polimorfos, o alelomorfos).

5.4. Análisis morfosintáctico

A nivel morfosintáctico identificamos los rasgos distintivos del lenguaje de la TA en oposición a la TH en términos de sobre- o sub-representación de las secuencias de categorías gramaticales (n-gramas de etiquetas POS) en estos grupos de textos. En la Tabla 13 presentamos los resultados del análisis cuantitativo: el porcentaje de las diferencias significativas TH-TA de *Google* y TH-TA de *Lucy.*

	TH - *Google*	**TH - *Lucy***
Unigramas	45%	51%
Bigramas	29%	33%
Trigramas	21%	28%

Tabla 13. Diferencias significativas TH-TA de *Google* y TH-TA de *Lucy* a nivel morfosintáctico

El número de diferencias es mayor en el caso de *Lucy*, debido, probablemente, a que el sistema basado en reglas se enfrenta con numerosos problemas de ambigüedad categorial y estructural en el análisis, ya que el inglés es una lengua con poca evidencia explícita de las funciones sintácticas de las unidades léxicas y de las relaciones sintáctico-semánticas entre ellas.

En la Tabla 14 se presentan las diferencias estadísticamente significativas[31] en las frecuencias de aparición de secuencias de etiquetas POS en la TH, la TA de *Google* y la TA de *Lucy*, que seleccionamos para realizar el análisis manual (la tabla completa se encuentra en el Anexo D).

[31] Las diferencias son significativas (p-valor < 0.05) si el valor de la X^2 es mayor a 3.81.

N-grama POS	**Frecuencia TH**	**Frecuencia *Google***	**X^2 TH-*Google***	**Frecuencia *Lucy***	**X^2 TH-*Lucy***
pp30[32]	594	486	*19.91*	451	*23.74*
pt00	28	56	*7.52*	83	*25.97*
nc sp nc	579	710	*3.37*	1160	*180.09*
pr vs	14	32	*5.62*	23	*1.82*
vmsp	121	21	*77.75*	84	*7.65*
vmii	216	52	*112.63*	374	*38.95*
vmis	138	178	*2.71*	1	*138.83*
vmg0	71	87	*0.71*	165	*35.32*
vs vm	12	92	*55.18*	83	*50.16*
rg rg	26	51	*6.16*	89	*31.94*

Tabla 14. Diferencias significativas en las frecuencias de aparición de secuencias de etiquetas POS en la TH, la TA de *Google* y la TA de *Lucy*

[32] Las etiquetas POS utilizadas por *Freeling* y su significado están indicados en el Anexo C. Tal como explicamos en el apartado 4.4.3., usamos una representación más o menos detallada en función de las categorías gramaticales y del tamaño de los n-gramas.

De acuerdo con el procedimiento descrito en el apartado 4.4.3., analizamos 10 ejemplos de las ocurrencias de cada n-grama e identificamos los patrones presentados en la Tabla 15.

TO	**TH**	***Shift* TO-TH**	***Google***	***Shift* TO-*Google***	***Lucy***	***Shift* TO-*Lucy***
FN::nc	FN::pp30	implicitación (opcional)	FN::nc	-	FN::nc	-
Cláusula de complemento	FN	implicitación (opcional)	Cláusula de complemento	-	Cláusula de complemento	-
FN::nc:nc	FN::nc:aq	modificación (obligatoria)	FN::nc:aq	modificación (obligatoria)	FN::nc:FP	alteración
Cláusula de relativo con predicación atributiva	FADJ	implicitación (opcional)	Cláusula de relativo con predicación atributiva	-	Cláusula de relativo con predicación atributiva	-
Cláusula de complemento no finita	Cláusula de complemento finita	explicitación (obligatoria)	Cláusula de complemento no finita	alteración	Cláusula de complemento finita	explicitación (obligatoria)
Verbo: Pasado Simple	vmis	explicitación (obligatoria)	vmis	explicitación (obligatoria)	vmii	alteración
Verbo: Pasado Simple	vmii	explicitación (obligatoria)	vmis	alteración	vmii	explicitación (obligatoria)
Cláusula circunstancial con gerundio	FN	implicitación (opcional)	Cláusula circunstancial con gerundio/infinitivo	-	Cláusula circunstancial con gerundio	-
Construcción pasiva	Construcción activa	modificación (opcional)	Construcción pasiva	-	Construcción media	modificación (opcional)
FADV	0	implicitación (opcional)	FADV	-	FADV	-

Tabla 15. Clasificación de las diferencias TH-TA de *Google* y TH-TA de *Lucy* a nivel morfosintáctico

En las columnas "TO", "TH", "*Google*" y "*Lucy*" indicamos las construcciones sintácticas del original, TH, TA de *Google* y TA de *Lucy* con las que se relacionan los n-gramas de etiquetas POS presentados en la Tabla 14. En las columnas "*Shift* TO-TH", "*Shift* TO-*Google*" y "*Shift* TO-*Lucy*" indicamos el tipo de *translation shift* realizado en los textos correspondientes. Dado el tamaño reducido de la muestra, no detectamos

todas las regularidades asociadas a las diferencias significativas en las frecuencias de aparición de etiquetas POS. Los resultados del análisis, más que ofrecer una caracterización completa de las tendencias que se observan en la TA frente a la TH, demuestran qué tipo de información podemos obtener por medio de la metodología desarrollada en el presente estudio.

Con respecto a los recursos que se utilizan en la TH en relación con la función referencial, observamos una preferencia por la referencia anafórica frente a la referencia nominal en los casos en los que esta última conlleva la repetición léxica. Ello se manifiesta en la frecuencia de aparición de los pronombres personales, la cual es significativamente más baja en las TAs. En estos casos, en la TH se observa el *translation shift* de implicitación (opcional), que está condicionado tanto por el proceso traductor como por la adecuación a la LM. De acuerdo con Baker (1993) (véase apartado 3.1.1.), uno de los universales de traducción es la tendencia a evitar la repetición. Dicha tendencia se manifiesta con mayor claridad en las traducciones del inglés al español debido a que estas lenguas difieren en sus preferencias en cuanto al uso de los mecanismos de cohesión. La repetición léxica frente al uso de los pronombres es más frecuente en inglés que en español, debido a las diferencias sistémicas entre estas lenguas: "En el discurso inglés, a diferencia del español, debido a la ausencia de elementos desambiguadores como el género, a la menor capacidad de flexión verbal y del sistema deíctico, se tiende a repetir los términos de forma muy frecuente en oraciones cortas" (Rodríguez Medina, 2003: 97). Así, en el ejemplo 11 a la frase nominal *that information* le corresponde el pronombre personal "ella" en la TH, mientras que los sistemas de TA reproducen este constituyente de manera literal:

11. TO. To build a computer, only two things are really necessary - a method of storing information and a few simple operations for acting on that information.
    TH. Para construir una computadora bastan un método para almacenar información y unas cuantas operaciones simples para actuar sobre [ella/***pp30***]FN[33].
    Google. Para crear un equipo, sólo dos cosas son realmente necesarias - un método de almacenamiento de la información y de unas pocas operaciones simples para actuar sobre [esa/dd información/nc]FN.

[33] En los ejemplos ofrecemos solamente la etiqueta POS de los elementos que nos interesa discutir.

Lucy. Para construir un ordenador, solamente dos cosas son realmente necesarias - un método de información de almacenaje y unas cuantas operaciones sencillas para actuar sobre [esa/dd información/nc]FN.

Otro caso de implicitación opcional es la nominalización, operación que se da con frecuencia en la TH. Como se ha mencionado en el apartado del análisis léxico-terminológico, la nominalización es un procedimiento común en la traducción de los gerundios nominales del inglés al español, pero también ocurre en la traducción de los verbos conjugados. Por ejemplo:

12. TO. Understanding how those circuits are organized could help scientists develop new therapies for many serious disorders.
    TH. De la comprensión [de/sp [la/da organización/nc]FN]FP [de/sp [estos/dd circuitos/nc]FN]FP depende la creación de nuevas terapias para muchas enfermedades graves.
    Google. La comprensión [de/sp [cómo/***pt00*** [estos/dd circuitos/nc]FN [se/p0 organizan/vm]FV]CL]FP podría ayudar a los científicos a desarrollar nuevos tratamientos para muchas enfermedades graves.
    Lucy. Entender [cómo/***pt00*** [esos/dd circuitos/nc]FN [se/p0 organizan/vm]FV]CL podría ayudar científicos a desarrollar nuevas terapias para muchos desórdenes graves.

La reducción de la cláusula "*how those circuits are organized*" a una frase preposicional en la TH conlleva la implicitación de las relaciones sintáctico-semánticas entre sus elementos. Tal como afirma Halliday (2004: 171), "A great deal of semantic information is lost when clausal expressions are replaced by nominal ones". La nominalización resulta en un aumento de la ambigüedad sintáctica, creando secuencias que permiten múltiples lecturas y cuya legibilidad depende del conocimiento especializado del lector. Debido a que los textos pertenecen al género de divulgación científica, se esperaría que en la TH se redujera el número de nominalizaciones, pero por las razones mencionadas en el apartado 5.2.2., observamos una tendencia contraria que concuerda con la conclusión de Fernández Polo (1999), quien estudia el uso de los mecanismos cohesivos en la traducción a partir de la revista *Scientific American* y su versión en español, de que las traducciones de *Investigación y Ciencia* en algunos aspectos pierden el carácter divulgativo propio de los originales. En cuanto a las TAs, los sistemas ofrecen una traducción literal de la secuencia original. Ésta es una de las razones por las que los pronombres interrogativos están sobre-representados en las TAs en comparación con la TH.

En relación con los mecanismos de modificación, el recurso más frecuente en la TA de *Lucy* es la frase preposicional, mientras que en la TH y la TA de *Google* su uso es menos frecuente. Dicha diferencia se produce en el contexto del tratamiento de los sintagmas nominales del inglés de forma FN::nc:nc, que los traductores humanos y el sistema estadístico traducen de maneras diversas. En cambio, el sistema basado en reglas los traduce por defecto por medio de frases preposicionales, selección que en algunos casos coincide con la TH pero en otros conlleva agramaticalidad o falta de naturalidad en el discurso (la traducción de los sintagmas nominales complejos del inglés al español ilustra la dificultad que representa modelar la traducción únicamente con base en reglas rígidas). Por ejemplo:

13. TO. Overcoming the Obstacles to Gene Therapy
    TH. Problemas de [la/dd [terapia/nc génica/aq]FN]FN
    Google. La superación de los obstáculos a [la/dd [terapia/nc génica/aq]FN]FN
    Lucy. Superando los obstáculos a [terapia/***nc*** [de/***sp*** gen/***nc***]FP]FN

Otro ejemplo de diferencias TA-TH a nivel morfosintáctico es la sobre-representación en las TAs del bigrama "pr vs", que corresponde a las subordinadas de relativo con predicación atributiva. Al realizar el análisis cualitativo de una muestra de los contextos de aparición de este bigrama, observamos que dicha diferencia se produce en la traducción de la misma construcción del original. Los traductores humanos muestran una preferencia por la modificación adjetival (implicitación opcional) debido, probablemente, a que ésta tiene un mayor grado de concisión, mientras que los sistemas de TA ofrecen una traducción literal. Así, en el ejemplo 14, a la cláusula de relativo *that could potentially be helpful*, que modifica a la frase nominal *a gene*, le corresponde en la TH la frase adjetival "potencialmente beneficioso", mientras que las TAs reproducen de manera literal la estructura sintáctica de la oración original.

14. TO. Under those conditions, a gene that could potentially be helpful would have little chance of affecting a disease process.
    TH. En esas condiciones, un gen potencialmente beneficioso/aq tendría pocas posibilidades de influir en el desarrollo de una enfermedad.
    Google. En esas condiciones, un gen [que/***pr*** [podría/***vm*** [ser/***vs*** útil/aq]FV]FV]CL que tienen pocas posibilidades de afectar a un proceso de la enfermedad.
    Lucy. Bajo esas condiciones, un gen [que/***pr*** [podría/***vm*** [ser/***vs*** [potencialmente/rg útil/aq]FADJ]FV]FV]CL tendría poca posibilidad de afectar a un proceso de enfermedad.

En cuanto a los mecanismos de predicación, dado que el inglés es una lengua con una morfología verbal pobre, ambos sistemas de TA presentan problemas en su tratamiento. Así, las formas verbales del futuro, subjuntivo e imperfecto están sub-representadas en la TA de *Google*, lo cual concuerda con la observación confirmada en numerosos estudios sobre los problemas que presentan los sistemas estadísticos al traducir a una lengua con morfología flexiva rica (Lee, 2004). Así, en el siguiente ejemplo *Google* no realiza la explicitación obligatoria en la traducción de la cláusula no finita "*to take up sugar*":

15. TO. Pancreatic cells, for instance, release insulin to tell muscle cells to take up sugar from the blood for energy.
    TH. Las del páncreas, por ejemplo, segregan insulina y, con ello, ordenan a las células musculares [que/cs [capten/***vmsp*** [el/da azúcar/nc]FN [de/sp [la/da sangre/nc]FN]FP]FV]CL y produzcan energía.
    Google. Las células del páncreas, por ejemplo, liberar insulina para contar las células musculares [para/sp [absorber/vmn0 [azúcar/nc]FN [de/sp [la/da sangre/nc]FN]FP]FV]CL para obtener energía.
    Lucy. Las células pancreatic, por ejemplo, liberan insulina para decir a células musculares [que/cs [empiecen/***vmsp*** [azúcar/nc]FN [de/sp [la/da sangre/nc]FN]FP]FV]CL para energía.

La TA de *Lucy* también presenta diferencias con la TH en la traducción de la morfología verbal, pero debido, sobre todo, a la necesidad de explicitar los rasgos gramaticales que no se marcan en inglés de la misma manera que en español, es decir, debido a la ambigüedad transferencial de las formas verbales de esta lengua. Un ejemplo de ello es la sobre-representación de los verbos en imperfecto y la sub-representación de los verbos en pretérito indefinido que se da en el contexto de la traducción de las formas verbales en pasado simple. En la TA de *Google* se detecta la tendencia contraria (la sub-representación de los verbos en imperfecto), la cual, suponemos, se debe a la baja frecuencia de estas formas en su corpus de entrenamiento. Por ejemplo:

16. TO. All three received Nobel Prizes for their discoveries.
    TH. Los tres recibieron/***vmsi*** el premio Nobel por sus descubrimientos.
    Google. Los tres recibieron/***vmsi*** el premio Nobel por sus descubrimientos.
    Lucy. Los tres recibían/***vmii*** premios de Nobel por sus descubrimientos.

En cuanto a la sobre-representación del unigrama que corresponde a los verbos en gerundio en las traducciones de *Lucy* y, en menor medida, en las traducciones de

Google, uno de los contextos en los que se produce es la traducción de las construcciones de gerundio con valor circunstancial, que el traductor humano sustituye por frases preposicionales realizando una implicitación opcional. Así, en el siguiente ejemplo a la cláusula *thus inducing the synthesis of a protein* le corresponde en la TH el sintagma preposicional "con la inducción consiguiente de la síntesis de una proteína":

17. TO. James E. Darnell Jr., of Rockefeller showed that when one of these proteins attaches, through its linker module, to an activated receptor kinase, the interaction spurs the bound protein to detach, move to the nucleus and bind to a particular gene, thus inducing the synthesis of a protein.

    TH. James Darnell Jr., de la Universidad Rockefeller, demostró que, cuando una de estas proteínas se une, por medio de su módulo de conexión, a una quinasa receptora activada, la interacción provoca el desprendimiento de la proteína, su tránsito al núcleo y su asociación con un gen determinado, [con/sp [la/da [inducción/nc consiguiente/aq]FN]FN [de/sp [la/da síntesis/nc]FN]FP [de/sp [una/di proteína/nc]FN]FP]FP.

    Google. James E. Darnell Jr., de Rockefeller mostró que cuando una de estas proteínas se conecta, a través de su módulo de enlazador, a un receptor quinasa activada, estimula la interacción de la proteína unida a separar, mover al núcleo y se unen a un gen en particular, [[induciendo/***vmg0*** así/rg]FV [la/da síntesis/nc]FN [de/sp [una/di proteína/nc]FN]FP]]CL.

    Lucy. James E. Darnell Jr., de Rockefeller mostraba que cuando una de estas proteínas se engancha, a través de su módulo de montador, a un kinase de receptor activado, la interacción impulsa la proteína de salto para separar, cambiar al núcleo y unirse a un gen particular, [[[de/sp [esta/dd forma/nc]FN]FP induciendo/***vmg0***]FV [la/da síntesis/nc]FN] [de/sp [una/di proteína/nc]FN]FP]]CL.

Otra diferencia en cuanto a las construcciones con función predicativa es la sobre-representación en la TA de *Google* del bigrama "vs vm", que corresponde a la construcción pasiva analítica, frente al uso de las construcciones medias y activas en la TA de *Lucy* y en la TH. En este caso la traducción que ofrece el sistema estadístico no es agramatical, pero es inadecuada desde el punto de vista del aspecto pragmático-funcional de estas construcciones y de su frecuencia de uso en inglés y en español (la construcción media es más frecuente en español que la pasiva analítica). Por un lado, de acuerdo con Rodríguez Medina (2003: 97), la diferencia en el uso de la pasiva en español y en inglés se da por las razones siguientes:

> la flexibilidad del orden de palabras en español frente a la relativa rigidez del inglés, […] el hecho de que el sujeto español esté simplemente representado en la desinencia verbal - y la existencia en español de la pasiva refleja […]

Por otro lado, uno de los rasgos lingüísticos más evidentes de los registros científicos es el uso de la voz pasiva, que contribuye a crear la impresión de impersonalidad de los textos científicos. De acuerdo con Fernández Polo (1999), se trata de una de las manifestaciones de la cortesía propia de la interacción entre el científico y el resto de la comunidad investigadora que conlleva la mitigación del protagonismo en la investigación. Ante los nuevos lectores previstos para los textos de divulgación, atribuirse un papel activo en la investigación deja de ser una acción que amenaza la imagen pública, por lo que la impersonalidad y la voz pasiva son menos comunes en los artículos de divulgación. El uso de la voz pasiva en la TA de *Google* frente a la voz activa (TH) y la voz media (TA de *Lucy*) queda ilustrado en el siguiente ejemplo:

18. TO. At that time, cells were viewed as balloonlike bags filled with a soupy cytoplasm containing floating proteins and organelles (membrane-bound compartments, such as the nucleus and mitochondria).
    TH. En esos años, [representábamos/vm [las/da células/nc]FN]FV a la manera de bolsas hinchadas y rellenas de un caldo citoplásmico que contenía proteínas y orgánulos flotantes (compartimientos ceñidos por membranas, como el núcleo o las mitocondrias).
    Google. En ese momento, [las/da células/nc]FN [fueron/***vs*** vistos/***vm***]FV como bolsas a globos llenos con un citoplasma espesa que contiene proteínas flotantes y orgánulos (unidas a la membrana compartimentos, tales como el núcleo y mitocondrias).
    Lucy. En aquella época, [las/da células/nc]FN [se/p0 veían/vm]FV como bolsas balloonlike llenadas de un citoplasma soupy que contenía proteínas flotantes y orgánulos (compartimientos decididos de membrana, como el núcleo y las mitocondrias).

El bigrama "vs vm" también corresponde a la perífrasis, estar + gerundio cuyo uso es poco frecuente en la TH, y que está sobre-representada en las TAs debido a la interferencia del inglés. Asimismo, corresponde a la perífrasis estar + participio la cual, en numerosas ocasiones, se construye en la TH por medio de los verbos "hallarse" o "encontrarse". *Freeling* no etiqueta dichos verbos como auxiliares, por tanto su uso en la TH también afecta la frecuencia de aparición del bigrama "vs vm". Por ejemplo:

19. TO. Such receptors, the functional equivalent of antennae, are able to relay a messenger's command into a cell because they are physically connected to the cytoplasm.
    TH. Estos receptores, cuya función evoca la de una antena, transmiten la orden del mensajero a la célula porque [[se/p0 hallan/vm]FV [físicamente/rg conectados/vm]FV]FV con el citoplasma.
    Google. Tales receptores, el equivalente funcional de las antenas, son capaces de transmitir comandos a un mensajero en una célula, ya que [están/***vs*** conectados/***vm***]FV físicamente al citoplasma.

Lucy. Tales receptores, el equivalente funcional de antenas, pueden transmitir el comando de un mensajero en una célula porque [se/p0 conectan/vm]FV físicamente con el citoplasma.

En cuanto a la frecuencia del bigrama "rg rg", que corresponde a las frases adverbiales, en la TA de *Lucy* están sobre-representados los adverbios en -mente, mientras que en la TA de *Google* y en la TH, éstos se omiten y su significado se pierde por completo o se codifica con otros recursos. Por ejemplo[34]:

20. TO. Tumors would probably result only rarely, but even the remote chance of increasing cancer risk must be taken seriously.
    TH. [La/da posibilidad/nc]FN de que se produzca un tumor [puede/vm [ser/vs remota/aq]FV]FV, pero aun así tal riesgo no debe despreciarse.
    Google. Los tumores probablemente/rg se produciría sólo/rg [en/sp [raras/aq ocasiones/nc]FN]FP, pero incluso la remota posibilidad de aumentar el riesgo de cáncer debe ser tomado en serio.
    Lucy. Los tumores resultarían probablemente/***rg*** solamente/***rg*** raramente/***rg***, pero incluso la posibilidad remota de riesgo de cáncer creciente se debe llevar gravemente.

La comparación TA-TH basada en la frecuencia de aparición de etiquetas POS presenta limitaciones importantes. La simple comparación de frecuencias de aparición de unigramas, bigramas y trigramas de categorías gramaticales deja fuera el tratamiento de diversos fenómenos sintácticos o semánticos en la TA (por ejemplo, las diferencias relacionadas con la traducción de las construcciones con dependencias de larga distancia o el uso de la negación), con lo cual muchos errores de los sistemas no se reflejan en los resultados del análisis. El método de comparación de textos por frecuencias de aparición de n-gramas de etiquetas POS se ha utilizado con éxito para observar las diferencias estilísticas en los textos producidos por diferentes autores, o en distintos géneros, así como para identificar los rasgos propios del lenguaje de la traducción; sin embargo, no es del todo adecuado para identificar errores en la TA.

Además, los resultados se ven afectados por los errores de anotación automática de *Freeling*. Por ejemplo, suponíamos que todas las palabras no traducidas en la TA serán

[34] Si evaluáramos las TAs de este ejemplo con BLEU, ambos sistemas obtendrían puntuaciones bajas que, más que reflejar la calidad de la TA, se deberían a la reestructuración global de la oración original en la THR. Así, tal como mencionamos en el Capítulo 3 (véase p. 66) en la evaluación automática o el entrenamiento de sistemas de TA, deben utilizarse THRs análogas, es decir, aquellas en las que la estructura del original se preserva al máximo posible, a menos que tal preservación viole las reglas de la gramática de la LM (en la traducción análoga se reduce al mínimo el número de los *translation shifts* opcionales).

etiquetadas como nombres propios, lo cual nos permitiría demostrar que el número de las palabras no traducidas es mayor en el caso de *Lucy*, debido al problema de cobertura del vocabulario. Sin embargo, *Freeling* asigna a las palabras no traducidas diversas etiquetas POS en función del contexto lingüístico inmediato, con lo cual el número elevado de las palabras no traducidas en la TA de *Lucy* no se refleja en los resultados.

5.5. Análisis discursivo

Como se ha mencionado en el apartado 4.5.3., debido a que de momento no existe una herramienta de anotación discursiva automática para el español, en la presente investigación no realizamos el análisis cuantitativo a nivel discursivo y nos limitamos a analizar algunos ejemplos que ilustran, en nuestra opinión, las situaciones prototípicas que pueden darse en relación con las diferencias en la estructuración del discurso en las TAs y la TH.

Con respecto a la *segmentación* del discurso, suponemos que las diferencias que se originan en la TH se deben principalmente a las diferencias sintácticas entre el inglés y el español. El inglés muestra una preferencia por la yuxtaposición y la coordinación (parataxis), mientras que el español, la subordinación (hipotaxis) (Rodríguez Medina, 2003). Así, en el siguiente ejemplo a la EDU (3) del fragmento original, que desempeña el papel discursivo de CAUSA, le corresponde una oración de relativo en la TH:

21. TO. [Even before therapy became a goal,]S_CIRCUNSTANCIA(1) [transcription had long captivated scientists for another reason:]N(2) [knowledge of how this process is regulated promises to clarify some central mysteries of life.]S_CAUSA(3)

    TH. [Mucho antes de que la terapia se convirtiera en objetivo,]S_CIRCUNSTANCIA(1) [la transcripción había cautivado el interés de los científicos que buscaban desentrañar los mecanismos que regulan este proceso, en la esperanza de que aclarase algunos puntos del misterio de la vida.]N(2)

    Google. [Incluso antes de la terapia se convirtió en una meta,]S_CIRCUNSTANCIA(1) [la transcripción había cautivado a los científicos a largo por otra razón:]N(2) [el conocimiento de cómo este proceso está regulado se compromete a aclarar algunos misterios centrales de la vida.]S_CAUSA(3)

Lucy. [Incluso antes de que la terapia se convirtiera en un objetivo,]S_CIRCUNSTANCIA(1) [la transcripción había cautivado mucho tiempo científicos por otra razón:]N(2) [el conocimiento de cómo este proceso se regula promete aclarar algunos misterios esenciales de vida.]S_CAUSA(3)[35]

También se dan los casos contrarios en los que a una EDU del original le corresponden 2 EDUs en la TH, lo cual puede deberse a la tendencia a la simplificación. Así, en el ejemplo 22, a la EDU (1) del TO le corresponden las EDUs (1) y (2) en la TH:

22. TO. [The molecules that form signaling circuits in cells are often modular-built from components that carry out distinct tasks.]N(1) [This discovery emerged in part from studies of molecules known as receptor tyrosine kinases.] S_ELABORACIÓN(2)
    TH. [Las moléculas que integran los circuitos celulares de señalización son, a menudo, modulares.]N(1) [Sus elementos componentes realizan distintas tareas.]S_ELABORACIÓN(2) [A ese descubrimiento se llegó desde el estudio de las tirosinaquinasas receptoras.]S_ELABORACIÓN(3)
    Google. [Las moléculas que forman circuitos de señalización en las células suelen ser modular, construido a partir de componentes que realizan tareas distintas.]N(1) [Este descubrimiento surgió, en parte, a partir de estudios de moléculas conocidas como receptores de tirosina quinasas.]S_ELABORACIÓN(2)
    Lucy. [Indicando las moléculas que se forman por señas que los circuitos en células están a menudo modularmente construía de componentes que realizan tareas claras.]?(1) [Este descubrimiento emergía en parte de estudios de moléculas sabidas como kinases de tirosina de receptor.]S_ELABORACIÓN(2)

Además, en este ejemplo es imposible llegar a una interpretación coherente de la TA de *Lucy* debido a la alteración de la estructura sintáctica del original en la primera EDU. El sistema no logra determinar la función sintáctica del participio presente *signaling* que modifica al nombre *circuits* debido a la ambigüedad estructural de las formas en *-ing* del inglés.

Las diferencias en las *relaciones discursivas* entre la TA y la TH, que se originan en la TH, reflejan la interpretación del TO por parte del traductor y la adecuación a los patrones de organización textual propios de la LM. Este proceso afecta el orden de las EDUs y la selección de marcadores del discurso, y conlleva modificaciones en la estructura discursiva del original en el TT. El ejemplo 23 ilustra esta situación:

[35] En los ejemplos del nivel discursivo marcamos el tipo de elemento (N - núcleo, S - satélite) y el tipo de relación. Asimismo indicamos entre paréntesis el número del segmento para registrar los cambios en el orden de las EDUs.

23. TO. [Relatives of people with autism may fail to meet all the criteria for the disorder]N(1) [but still have some of its symptoms.]S_CONCESIÓN(2)
    TH. [Los parientes de autistas pueden presentar algunos síntomas,]N(2) [aunque no el cuadro completo que justifique el diagnóstico de la enfermedad.]S_CONCESIÓN(1)
    Google. [Los familiares de las personas con autismo pueden no cumplir todos los criterios para el trastorno,]N(1) [pero todavía tienen algunos de sus síntomas.]S_CONCESIÓN(2)
    Lucy. [Los familiares de gente con el autismo pueden fracasar en encontrar todos los criterios para el desorden]N(1) [pero todavía tener algunos de sus síntomas.]S_CONCESIÓN(2)

En la TH se presentan con frecuencia los casos de la omisión de marcadores discursivos, lo cual supone una implicitación con respecto a la marcación de las relaciones discursivas. De acuerdo con Loureda Lamas (2010: 81), "los marcadores del discurso se consideran como unidades lingüísticas que por su significado de procesamiento guían de acuerdo con sus propiedades morfosintácticas, semánticas y pragmáticas las inferencias que se realizan en la comunicación". Suponemos que la supresión de dichas unidades en la TH está relacionada con el cambio en el tipo de lectores previstos para la versión española de la revista *Scientific American*, ya que estos elementos podrían ser redundantes para los especialistas en el tema. La importancia concedida por los traductores al componente informativo y la poca importancia que conceden a los materiales que procuran una contribución escasa en términos de información (véase p. 105) también pudiera explicar las supresiones de los elementos que cumplen con las funciones de carácter interpersonal o metatextual. Finalmente, tal como se ha mencionado en el Capítulo 3 (véase p. 55), el uso de los conectores argumentativos es más frecuente en inglés que en español, con lo cual la omisión de estas unidades podría considerarse una adaptación de los textos a las restricciones en el uso de los mecanismos cohesivos propias de la lengua de llegada. Así, en el siguiente ejemplo se omite la conjunción causal *because*, lo cual conlleva la implicitación con respecto a la marcación de la relación discursiva de CAUSA entre las EDUs (3) y (2).

24. TO. [All cells in a body carry the same genes in the chromosomes of the nucleus.]N(1) [But neurons, say, behave unlike liver cells]S_ANTÍTESIS(2) [<u>because</u> different cells use, or express distinct subsets of genes]S_CAUSA(3) [and hence make separate sets of proteins.]S_RESULTADO(4)
    TH. [Todas las células del cuerpo portan los mismos genes en los cromosomas que están en el núcleo.]N(1) [Pero las células nerviosas, por ejemplo, no se comportan igual que las hepáticas.]S_ANTÍTESIS(2) [Células distintas utilizan, o expresan, subgrupos diferentes de genes]S_CAUSA(3) [y, por tanto, fabrican grupos diversos de proteínas.]S_RESULTADO(4)

Google. [Todas las células en un cuerpo llevan los mismos genes en los cromosomas del núcleo.]N(1) [Sin embargo, las neuronas, por ejemplo, se comportan como las células del hígado]S_ANTÍTESIS(2) [debido a que diferentes células utilizan, o expresan, subconjuntos de genes distintos]S_CAUSA(3) [y por lo tanto hacer que distintos conjuntos de proteínas.]?(4)

Lucy. [Todas las células en un cuerpo llevan los mismos genes en los cromosomas del núcleo.]N(1) [Pero las neuronas, digamos que, se comportan a diferencia de células de hígado]S_ANTÍTESIS(2) [porque las células diferentes usan, o expresan subconjuntos distintos de genes]S_CAUSA(3) [y por consiguiente hacen conjuntos separados de proteínas.]S_RESULTADO(4)

Además, en este ejemplo es imposible interpretar una relación discursiva entre las EDUs (3) y (4) en la TA de *Google*, debido a la alteración de la estructura sintáctica del original que se produce a causa de la adición del pronombre relativo "que". En la oración original el sintagma nominal *sets of proteins* es el complemento del verbo *make*, mientras que en la TA de *Google* éste desempeña el papel de sujeto de una oración de relativo deficiente.

A continuación ofrecemos otros ejemplos de las diferencias en la estructuración del discurso relacionadas con la alteración de la segmentación o de las relaciones discursivas en las TAs de *Google* y de *Lucy*. En el ejemplo 25 la *segmentación* discursiva del original es alterada en la TA de ambos sistemas debido a la ambigüedad categorial de la palabra *design*: en la oración original *design* es un verbo que desempeña el papel de predicado, mientras que en las TAs forma parte de la frase preposicional "drogas de diseño":

25. TO. [To answer these questions]S_PROPÓSITO(1) [and design drugs able to modulate transcription,]S_PROPÓSITO(2) [investigators need to know something about the makeup of the apparatus that controls reading of the genetic code in human cells.]N(3)

    TH. [Para responder a estas preguntas]S_PROPÓSITO(1) [y diseñar drogas capaces de modular la transcripción,]S_PROPÓSITO(2) [había que conocer el funcionamiento de la maquinaria que controla la lectura del código genético en la célula humana.]N(3)

    Google. [Para responder a estas preguntas y las drogas de diseño capaces de modular la transcripción,]S_PROPÓSITO(1) [los investigadores necesitan saber algo acerca de la composición del aparato que controla la lectura del código genético en las células humanas.]N(2)

    Lucy. [Para responder a estas preguntas y las drogas de diseño capaces de modular transcripción,]S_PROPÓSITO(1) [los investigadores necesitan saber algo sobre la composición del aparato que controla lectura del código genético en células humanas.]N(2)

En el ejemplo 26, en la TA de *Google* la sustitución del verbo copulativo *to be* por el pronombre relativo "que", con la cual se pierde la relación atributiva entre "las anomalías simples" y "el diagnóstico de la enfermedad", conlleva la alteración de la estructura discursiva, ya que no existe una interpretación coherente de la primera EDU, de manera que es imposible relacionarla con ninguna otra EDU del texto.

26. TO. [If simpler behavioral abnormalities could be shown to be diagnostic of the disorder,]S_CONDICIÓN(1) [researchers might have a better chance of identifying their source in the nervous system.]N(2)
    TH. [Si se pudiera comprobar el valor diagnóstico de algunas disfunciones elementales del comportamiento autista,]S_CONDICIÓN(1) [avanzaríamos con paso más firme en la búsqueda del origen nervioso de las mismas.]N(2)
    Google. [Si más simples anomalías de comportamiento puede ser demostrado que el diagnóstico de la enfermedad,]?(1) [los investigadores podrían tener una mejor oportunidad para identificar su origen en el sistema nervioso.]N(2)
    Lucy. [Si las anormalidades behavioral más sencillas se podrían enseñar a ser diagnóstico del desorden,]S_CONDICIÓN(1) [los investigadores podrían tener una posibilidad mejor 'de identificar su fuente en el sistema nervioso.]N(2)

Tal como hemos observado en los ejemplos anteriores, la alteración de la estructura discursiva en la TA de *Google* se produce a causa de la omisión o adición injustificada de palabras funcionales, que representan dificultades para el aprendizaje automático. En el ejemplo 27, la alteración de la estructura discursiva del original se debe a la adición del adverbio negativo "no":

27. TO. [Biology was now the study of information stored in DNA-strings of four letters: A, T, G and C for the bases denine, thymine, guanine and cytosine - and of the transformations that information undergoes in the cell.]N(1) [There was mathematics here!]S_INTERPRETACIÓN(2)
    TH: [La biología consistía ahora en el estudio de la información almacenada en ADN - ristras de cuatro letras, A, T, G y C, símbolos de las bases adenina, timina, guanina y citosina - y de las transformaciones que esa información experimenta en el interior de la célula.]N(1) [¡Aquí había matemáticas!]S_INTERPRETACIÓN(2)
    Google. [Biología era ahora el estudio de la información almacenada en las cadenas de ADN de cuatro letras: A, T, G y C para las bases denine, timina, guanina y citosina - y de las transformaciones que sufre la información en la célula.]N(1) [No era la matemática aquí!]?(2)
    Lucy. [La biología era ahora el estudio de información almacenada en cadenas de ADN de cuatro letras: un, T, G y C para el denine de bases, thymine, guanina y cytosine - y de las transformaciones que la información sufre en la célula.]N(1) [¡Había matemáticas aquí!]S_INTERPRETACIÓN(2)

Una de las posibles inferencias que se podrían derivar del primer enunciado es que la biología se asemeja en algunos aspectos a las ciencias exactas. Esta inferencia es explicitada en el segundo enunciado: "Aquí había matemáticas". En la traducción de *Google* no es posible interpretar una relación de coherencia, ya que lo que se deriva del primer enunciado se contradice en el segundo. De hecho, si hubiera un marcador contra-argumentativo en el segundo enunciado, el fragmento en sí constituiría un texto coherente.

En el siguiente ejemplo la alteración de la estructura discursiva en la traducción de *Google* se debe a la modificación de la estructura argumental de la oración original.

28. TO. [This brings up an important fact about biotechnologists:]S_PREPARACIÓN(1) [we are a community of thieves.]N(2) [We steal from the cell.]S_JUSTIFICACIÓN(3)
    TH. [Lo cual nos lleva a una confesión.]S_PREPARACIÓN(1) [Los biotecnólogos somos una comunidad de ladrones.]N(2) [Robamos de la célula.]S_JUSTIFICACIÓN(3)
    Google. [Esto trae a colación un hecho importante acerca de los biotecnólogos:]S_PREPARACIÓN(1) [somos una comunidad de ladrones.]N(2) [Nos roban de la célula.]?(3)
    Lucy. [Esto plantea un hecho importante sobre biotecnólogos:]S_PREPARACIÓN(1) [somos una comunidad de ladrones.]N(2) [Robamos la célula.]S_JUSTIFICACIÓN(3)

Steal es un verbo transitivo que requiere un complemento directo. Éste está omitido en el TO, lo cual resulta en un análisis fallido del original en el caso de *Lucy* que ofrece una oración gramatical que difiere del original en el plano semántico. Sin embargo, esta diferencia no afecta la estructura discursiva del fragmento, ya que robar la célula bien puede ser una razón para considerarse ladrones. En cambio, en el caso de la TA de *Google*, la estructura discursiva del fragmento es alterada, ya que el participante "biotecnólogos" en la EDU (3) desempeña el papel de paciente, no de agente.

En algunos casos la alteración de la estructura discursiva del original en las TAs se produce debido a una traducción inapropiada de los marcadores del discurso. Por ejemplo:

29. TO. [The shock was not generated by the unexpected result]N(1) [but by the realization that I had seen this pattern of shortening before, in a paper that showed pictures of abnormal mouse brains.]S_REFORMULACIÓN(2)

TH. [Mi excitación no la había provocado un resultado inesperado,]N(1) [sino la clara conciencia de que había observado antes ese tipo de acortamiento, en un artículo donde aparecían fotografías de cerebros anormales de ratón.]S_REFORMULACIÓN(2)

Google. [El choque no fue generado por el resultado inesperado,]N(1) [pero al darse cuenta de que había visto este patrón de acortamiento antes, en un artículo que mostraba imágenes de los cerebros de ratones anormales.]?(2)

Lucy. [La sorpresa no era generada por el resultado inesperado]N(1) [sino por la realización de que había visto este patrón de acortamiento antes, en un papel que mostraba imágenes de cerebros de ratón anormales.]S_REFORMULACIÓN(2)

Desde el punto de vista semántico, "sino" se considera el nexo prototípico de la oposición exclusiva y exige la negación en el primer miembro para apoyar o acentuar contrastivamente el segundo. La función pragmática del nexo "sino" es reformulativa-rectificativa. Los reformuladores rectificativos, de acuerdo con Portolés (1998: 142), "sustituyen un primer miembro, que presentan como una formulación incorrecta, por otra que la corrige o, al menos, la mejora". A pesar de poseer valor semántico adversativo, "sino" no funciona como conector contra-argumentativo. "Pero", en cambio, es un conector contra-argumentativo, que "vincula dos miembros del discurso, de tal modo que el segundo se presenta como supresor o atenuador de alguna conclusión que se pudiera obtener del primero" (Portolés, 1998: 140). Así, en el ejemplo 29 la diferencia en la estructura discursiva de la TA de *Google* y la TH se debe al uso inadecuado de los marcadores del discurso (agravado por el hecho de que en la traducción de *Google* los miembros de la coordinación adversativa no tienen la misma función sintáctica: "por el resultado inesperado" vs. "al darse cuenta"). Este error puede deberse a que la diferencia que se marca en español entre "pero" y "sino" no tiene un equivalente formal en el sistema del inglés. Al parecer, en términos de frecuencia el conector *but* se asocia con "pero" más que con "sino" o con "sino que", de ahí la selección poco apropiada de *Google*.

6. CONCLUSIONES

En esta sección se presentan las principales conclusiones de la tesis, se evalúa el cumplimiento de los objetivos y se discuten las limitaciones y las posibles líneas de trabajo futuro.

En esta investigación desarrollamos una propuesta metodológica para la comparación TA-TH en tres niveles de la lengua (léxico-terminológico, morfosintáctico y discursivo) y la aplicamos a un corpus paralelo inglés-español de textos especializados del ámbito médico que incluye TOs, THs y TAs. El objetivo general de la investigación era estudiar las diferencias lingüísticas sistemáticas entre la TA y la TH y sus implicaciones para la evaluación automática de sistemas. Específicamente, nos propusimos detectar las diferencias en la distribución de unidades de análisis (UTs, n-gramas de etiquetas POS y unidades discursivas) en la TH y en las TAs de un sistema estadístico (*Google*) y de un sistema basado en reglas (*Lucy*), e identificar las condiciones en las que se producen dichas diferencias teniendo en cuenta los TOs y las estrategias de traducción humana y automática.

Aunque nuestro objetivo no era desarrollar una métrica de evaluación de sistemas, la motivación de esta investigación radica en la necesidad de mejorar las métricas de evaluación automática existentes. Al indagar en la problemática de la evaluación, establecimos que, a pesar de las críticas y del desarrollo de técnicas alternativas, las métricas basadas en n-gramas son las que se usan por defecto para evaluar la TA. Una de las limitaciones fundamentales de dichas métricas es que no toman en cuenta el tipo y las fuentes de las diferencias entre las traducciones producidas por humano y por máquina. Si el objetivo último de la TA es lograr resultados comparables con la TH en términos de calidad, la evaluación debe indicar no solamente el grado de similitud, sino también los aspectos en los que la TA difiere de la TH.

Para llevar a cabo el presente estudio, partimos del supuesto de que las diferencias TA-TH se asocian tanto con los errores de los sistemas como con las modificaciones que realizan los traductores humanos con respecto a la forma y el contenido del TO. Por este motivo, para realizar la descripción de las mismas, fue preciso tomar en consideración las características de los sistemas de TA, así como las propiedades de la TH.

En relación con la TA, discutimos, en primer lugar, los fenómenos lingüísticos y traductológicos cuyo tratamiento representa dificultades para los sistemas (y que, por ello, constituyen las causas potenciales de las diferencias). En este sentido los estudios en el ámbito de la TA demuestran que la principal dificultad reside en las divergencias sistémicas, estilísticas y pragmáticas entre las lenguas, así como en la ambigüedad de la expresión lingüística. La interferencia de la LF confiere a la TA su carácter literal, mientras que la ambigüedad léxica, estructural y transferencial en numerosas ocasiones conlleva la producción de oraciones agramaticales. En segundo lugar, revisamos los principios de funcionamiento de la TA para poder explicar el comportamiento de los sistemas. Esto nos permitió establecer que la principal limitación de la TA basada en reglas consiste en que no es capaz de realizar una selección de recursos lingüísticos apropiada cuando ésta se ve condicionada por factores de uso; mientras que los errores de la TA estadística se relacionan con el tratamiento de fenómenos sintácticos.

Para describir las diferencias TA-TH asociadas al comportamiento de los traductores humanos, nos apoyamos en la traductología, específicamente en los estudios de traducción basados en corpus. En el marco de esta perspectiva se postula que la TH posee características lingüísticas inherentes que la distinguen de otros tipos de textos y, en opinión de algunos autores, constituye una variante particular de la lengua. De esta manera, el *translationese* y el lenguaje de la TA representan tipos particulares de expresión lingüística, cuya descripción implica, además de la identificación de sus rasgos distintivos, el estudio de las estrategias empleadas por los sistemas y por los traductores humanos que contribuyen a la conformación de dichos rasgos.

Teniendo en cuenta estas observaciones, desarrollamos una metodología para la descripción de las diferencias TA-TH que involucra, por un lado, el uso de técnicas estilométricas para caracterizar el lenguaje de la TA frente a la TH en términos de las diferencias cuantitativas en la distribución de las unidades de análisis y, por otro lado, el análisis de las traducciones en términos de *translation shifts* que permite identificar las fuentes de las diferencias detectadas. Con base en la revisión de las propuestas de clasificación de *translation shifts*, establecimos que la clasificación desarrollada por van Leuven Zwart (1989) es la que más se adecua a los propósitos de la presente investigación y la complementamos con las categorías que consideramos necesarias dada la naturaleza de los textos a analizar.

Entre los resultados relevantes del análisis, encontramos que a **nivel léxico-terminológico** las diferencias asociadas a la TH son más frecuentes que las diferencias relacionadas con la TA, debido a que a nivel de la UT la traducción que ofrecen los sistemas es cercana al original, mientras que en la TH se producen modificaciones opcionales que están condicionadas por factores estilísticos o pragmáticos y no por las divergencias léxicas entre los sistemas de la LF y la LM. Las diferencias asociadas a la TH se dan en el contexto de la adaptación del TO a las convenciones de uso en la LM o a las expectativas de los lectores previstos, la cual se manifiesta en el cambio del nivel de especialización.

La TA de *Google* se asemeja más a la TH que la TA de *Lucy*, puesto que el sistema estadístico es capaz de reproducir aquellas modificaciones que los traductores realizan con regularidad en un contexto lingüístico determinado. Mientras tanto, la selección léxica en la TA de *Lucy* está limitada por la información presente en los diccionarios del sistema.

A **nivel morfosintáctico** las diferencias relacionadas con la TH se manifiestan, por un lado, en la explicitación obligatoria de los rasgos gramaticales que no se marcan en inglés, pero cuya marcación es necesaria en español y, por otro lado, en la implicitación opcional relacionada con una preferencia por el uso de construcciones más concisas propias de los textos de alto nivel de especialización.

Los sistemas de TA suelen reproducir de manera literal las estructuras sintácticas de la LF, lo cual en algunos casos resulta en la generación de oraciones agramaticales y en otros conlleva una falta de naturalidad en el discurso. El número de diferencias es mayor en el caso de *Lucy*, debido a los problemas de ambigüedad categorial y estructural en la fase análisis.

A **nivel discursivo** las diferencias que se originan en la TH, reflejan la interpretación del TO por parte del traductor y la adecuación a los patrones de organización textual en la LM. La alteración de la estructura discursiva en la TA de *Google* se produce a causa de la omisión o adición injustificada de palabras funcionales. En el caso de *Lucy*, se debe a los errores de traducción provocados por la ambigüedad léxica, estructural y transferencial.

La modificación de la estructura discursiva del original en la TH indica el grado de adaptación del original a las convenciones de la organización textual propias de la lengua de llegada. Mientras tanto, el cambio de las relaciones discursivas en la TA es indicio de errores, ya que los sistemas no realizan ningún tipo de modificaciones a nivel extra-oracional.

Con respecto a las implicaciones del análisis realizado para la evaluación de los sistemas de TA, las diferencias TA-TH relacionadas con las modificaciones opcionales en la TH y las diferencias que se deben a la falta de modificaciones obligatorias en la TA no tienen la misma relevancia para evaluar la calidad de esta última. Aunado a ello, el estudio demuestra que en la evaluación automática de sistemas de TA deben utilizarse THRs de tipo análogo, es decir, aquellas en las que el número de los *translation shifts* opcionales se reduce al mínimo.

Somos conscientes de que el presente trabajo tiene varias limitaciones. En primer lugar, utilizamos sistemas de TA generales, no adaptados a la tarea de traducción de textos especializados del ámbito médico. En segundo lugar, no realizamos la comparación de las THs con textos originalmente escritos en español, que aportaría datos interesantes sobre las características distintivas de los TTs. En tercer lugar, la exploración de las diferencias TA-TH que llevamos a cabo es parcial. La representación de datos, las herramientas utilizadas y el corpus de estudio limitan las posibilidades de observación. Así, la comparación TA-TH basada en la frecuencia de aparición de etiquetas POS deja fuera el tratamiento de diversos fenómenos sintácticos o semánticos en la TA, con lo cual muchos errores de los sistemas no se reflejan en los resultados del análisis. Además, no logramos implementar todas las fases de la metodología propuesta en relación con los tres niveles del análisis, debido a las diferencias en el tipo de representación de datos y a la disponibilidad de los recursos de análisis lingüístico automático.

Nos planteamos como trabajo futuro solventar estas carencias, así como aplicar la metodología a otro tipo de textos. Además, trataremos de automatizar la metodología el máximo posible y ahondar en la investigación de la aplicabilidad del criterio discursivo a la evaluación de la TA.

BIBLIOGRAFÍA

Referencias

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* New Jersey: John Wiley & Sons.

Ahrenberg, L. (2005). Codified Close Translation as a Standard for MT. En *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, 13-22.

Ahrenberg, L. y Merkel, M. (2000). Correspondence measures for machine translation. En *Proceedings of the Language Resources and Evaluation Conference Workshop on Machine Translation Evaluation*, 41-46.

Allegranza, V., Krawer, S. y Steiner, E. (eds.) (1991). Eurotra Special Issue. *Machine Translation*, 6(2/3).

Amigó, E., Giménez, J., Gonzalo, J. y Márquez, Ll. (2006). MT Evaluation: Human-Like vs. Human Acceptable. En *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Lingiustics*, 17-24.

Baayen, H., van Halteren, H. y Tweedie, F. (1996). Outside the cave of shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing* 11(3): 121-131.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. En M. Baker, G. Francis y E. Tognini-Bonelli (eds.), *Text and Technology. In Honour of John Sinclair,* Amsterdam: Benjamins, 233–250.

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* 7(2): 223-243.

Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. En H. Somers (ed.) *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager,* Amsterdam: John Benjamins, 175-186.

Bakker, M., Koster, C. y van Leuven-Zwart, K. (1998). Shifts of Translation. En M. Baker (ed.), *Routledge Encyclopedia of Translation Studies*, London: Routledge, 226-231.

Banerjee, S. y Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. En *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 65-72.

Baroni, M. y Bernardini, S. (2006). A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3): 259-274.

Becher, V. (2011) *Explicitation and implicitation in translation.* Tesis doctoral. Hamburgo: Universidad de Hamburgo.

Beeby, L. A. (1996). *Teaching translation from Spanish to English, Didactics of Translation Series 2.* Otawa: University of Ottawa Press.

Bennet, W.S. y Slocum, J. (1985). The LRC machine translation system. *Computational Linguistics*, 11: 11-121.

Bernardini, S. y Zanettin, F. (2004). When is a Universal not a Universal? En A. Mauranen y P. Kujamäki (eds.), *Translation Universals. Do they exist?* Amsterdam: Benjamins, 51–62.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5: 257-269.

Biber, D., Conrad, S. y Reppen, R. (1998). *Corpus Linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

Blum-Kulka, S. (1986). Shifts of Cohesion and Coherence in Translation. En J. House y S. Blum-Kulka (eds.), *Interlingual and intercultural communication: discourse and cognition in translation and second language acquisition studies*, Tübingen: Narr, 17-35.

Borin, L. y Prütz, K. (2001). Through a Glass Darkly: Part of Speech Distribution in Original and Translated Text. En W. Daelemans, K. Simaan, J. Veenstra y J. Zavrel (eds.), *Computational Linguistics in the Netherlands 2000*, Amsterdam: Rodopi, 30-44.

Brants, T. (2000). Tnt - a statistical part- of-speech tagger. En *Proceedings of the 6th Association for Computational Linguistics Conference on Applied Natural Language Processing*, 224-231.

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4): 543-565.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., y Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2): 76–85.

Cabré, T. (1999). *La Terminología. Representación y comunicación*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cabré, T. (2003). Theories of terminology. Their description, prescription and explanation. *Terminology*, 9(2): 163-200.

Cabrera-Diego, L., Sierra, G., Vivaldi, J. y Pozzi, M. (2011). Using Wikipedia to Validate Term Candidates for the Mexican Basic Scientific Vocabulary. En *LaRC 2011: First International Conference on Terminology, Languages, and Content Resources*, 76-85.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. y Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. En *Proceedings of the Association for Computational Linguistics Workshop on Statistical Machine Translation*, 136-158.

Carreras, X., Chao, I., Padró, L. y Padró, M. (2004). Freeling: An open-source suite of language analyzers. En *Proceedings of the 4th International Language Resources and Evaluation Conference*.

Castellá, J. M. (1992). *De la frase al text.* Barcelona: Empuries.

Catford, J. C. (1965). *A Linguistic Theory of Translation: an Essay on Applied Linguistics*. London: Oxford University Press.

Charniak, E., Knight, K. y Yamada, K. (2003). Syntax-based Language Models for Machine Translation. En *Proceedings of Machine Translation Summit IX*.

Corston-Oliver, S. (1998). *Computing representations of the structure of written discourse.* Technical Report MSR-TR-98-15, Microsoft Research, Redmond, WA.

Corston-Oliver, S., Gamon, M. y Brockett, C. (2001). A Machine Learning Approach to the Automatic Evaluation of Machine Translation. En *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 140-147.

Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. En *Proceedings of Machine Translation Summit IX*, 23–27.

da Cunha, I. e Iruskieta, M. (2010). Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies*, 12(5): 563-598.

da Cunha, I., San Juan, E., Torres-Moreno, J.-M., Cabré, T. y Sierra, G (2012). A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-sentence Discourse Segments in Spanish. *Lecture Notes in Computer Science 7181*, 462-474.

Cyrus, L. (2009). Old concepts, new ideas: approaches to translation shifts. *MonTi. Monografías de Traducción e Interpretación*, 1: 87-106.

Dagan, I., Glickman, O. y Magnini, B. (2005). The PASCAL recognising textual entailment challenge. En *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Dale, R. (2000). Symbolic Approaches to Natural Language Processing. En D. Robert, M. Hermann y Somers, H. L. (eds), *Handbook of Natural Language Processing*, New York: M. Dekker, 1-9.

Dik, S. (1978). *Functional Grammar.* Amsterdam: North-Holland.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. En *Proceedings of the 2nd International Conference on Human Language Technology*, 138–145.

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4): 597-634.

Drouin P. (2003). Term-extraction using Non-technical Corpora as Point of Leverage. *Terminology*, 9(1): 99-117.

Duff, A. (1981). *The third language: recurrent problems of translation into English*. Oxford: Pergamon Press.

Faber, D. y Lauridsen, K. (1991). The compilation of a Danish-English-French corpus in contract law. En S. Johansson y A.-B. Stenström (eds.), *English computer corpora. Selected papers and research guide*, Berlin: Mouton de Gruyter, 235-243.

Farrús, M., Costa-jussá, M. R., Mariño, J. B. y Fonollosa, J. A. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. En *Proceedings of the Annual Conference of the European Association for Machine Translation*, 52–57.

Fellbaum, C. (ed.). (1998). *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fernández Polo, F. J. (1999). *Traducción y retórica contrastiva: A propósito de la traducción de textos de divulgación científica del inglés al español*. Santiago de Compostela: Servicio de Publicacións da Universidade de Santiago de Compostela.

Freixa, J. (2002). *La variació terminológica: análisis de la variació denominativa en textos de diferent grau d'especialització de l'área de medi ambient.* Tesis doctoral. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Gamon, M. (2004). Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. En *Proceedings of the Joint International Conference on Computational Linguistics*, 611- 617.

Gamon, M., Aue, A. y Smets, M. (2005). Sentence-Level MT evaluation without reference translations: beyond language modeling. En *Proceedings of the Annual Conference of the European Association for Machine Translation*, 103-111.

Gelbukh, A. y Sidorov, G. (2010). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. México: Instituto Politécnico Nacional.

Gildea, D. (2003). Loosely Tree-Based Alignment for Machine Translation. En *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 80-87.

Giménez, J. (2008). *Empirical Machine Translation and its Evaluation*. Tesis doctoral. Barcelona: Universidad Politécnica de Cataluña.

Giménez, J. y Amigó, E. (2006). IQMT: A Framework for Automatic Machine Translation Evaluation. En *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 685–690.

Giménez, J. y Márquez, Ll. (2008). Heterogeneous automatic MT evaluation through non-parametric metric combination. En *Proceedings of the Third International Joint Conference on Natural Language Processing*, 319-326.

Giménez, J. y Márquez, Ll. (2009). On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation. En *Proceedings of the 4th Workshop on Statistical Machine Translation of the European Chapter of the Association for Computational Linguistics*.

Granger, S. (2003). The corpus approach: a common way forward for contrastive lingusitics and translation studies. En S. Granger, J. Lerot y S. Petch-Tyson (eds.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Amsterdam/Atlanta: Rodopi, 17-29.

Grosz, B., Joshi, A. K. y Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2): 203–225.

Joachims, T. (1999). Making Large-scale SVM Learning Practical. En B. Schölkopf, C. Burges y A. Smola, (eds.), *Advances in Kernel Methods – Support Vector Learning*, Cambridge (MA): MIT Press.

Jurafsky, D. y Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M. A. K. (2004). *The Language of science*. London: Continuum.

Halliday, M. A. K. y Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hatim, B. y Mason, I. (1995). *Teoría de la traducción. Una aproximación al discurso.* Barcelona: Ariel.

Hauenschild, C. y Heizmann, S. (eds.) (1997). *Machine Translation and Translation Theory.* Berlin: Mouton de Gruyter

Hoang, H., Koehn, P. y Lopez, A. (2009). A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. En *Proceedings of the International Workshop on Spoken Language Translation*, 152–159.

Holmes, J. S. (1988). *Translated! Papers on literary translation and translation studies*. Amsterdam: Rodopi.

House, J. (1997). *Translation quality assessment. A model revisited.* Tübingen: Gunter Narr.

Hurtado, A. A. (2008). *Traducción y traductología. Introducción a la traductología*. Madrid: Cátedra.

Hutchins, J. (1997). From first conception to first demonstration: the nascent years of machine Translation, 1947-1954. A chronology. *Machine Translation*, 12(3): 195-252.

Hutchins, J. (1998). The origins of the translator's workstation. *Machine Translation*, 13: 287-307.

Hutchins, J. (2000). The IAMT Initiative in Defining Translation System Categories. En *Proceedings of the Fifth Workshop of the European Chapter of the Association for Computational Linguistics*.

Hutchins, W. J. y Somers, H. L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. En *Formal Methods in the Study of Language*, Amsterdam: Mathematisch Centrum, 277–322.

Karlsson, F., Voutilainen, A., Heikkilä, J. y Anttila, A. (Eds.) (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, Berlin: Mouton de Gruyter.

Kehler, A. (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3): 467-475.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1): 1–37.

Kittredge, R. I. y Lehrberger, J. (1982). (eds.) *Sublanguage: studies of language in restricted semantic domains*. Berlin: De Gruyter.

Klaudy, K. (2008). Explicitation. En M. Baker (ed.), *Routledge Encyclopedia of Translation Studies*, London: Routledge, 80-84.

Koehn, P. (2005). A parallel corpus for statistical machine translation. En *Proceedings of Machine Translation Summit XI.*

Koehn, P. (2009). *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Koehn, P., Och, F. J. y Marcu, D. (2003). Statistical Phrase-Based Translation. En *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics*.

Koehn, P. y Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. En *Proceedings of the Second Workshop on Statistical Machine Translation of the Association for Computational Linguistics,* 224–227.

Koppel, M. y Ordan, N. (2011). Translationese and its dialects. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1318–1326.

Krauthammer M. y Nenadic, G. (2004). Term identification in the Biomedical Literature. *Journal of Biomedical Informatics,* 37(6): 512-526.

Kulesza, A. y Shieber, M. (2004). A learning approach to improving sentence-level MT evauation. En *Proceedings of the 10th International Conference on Theoretical and Methodoogical Issues in Machine Translation*, 75-84.

Kurokawa, D., Goutte, C. y Isabelle, P. (2009). Automatic Detection of Translated Text and its impact on Machine Translation MT. En *Proceedings of the Machine Translation Summit XII*.

Landsbergen, J. (1987). Montague grammar and machine translation. En P.J. Whitelock, M.M. Wood, H.L. Somers, R. Johnson y P. Bennet (eds.) *Linguistic theory and computer applications,* London: Academic Press, 113-147.

Lapata, M. y Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. En *Proceedings of the 19th International Joint Conference on Artificial Intelligence.*

Larose, R. (1989) *Théories contemporaines de la traduction*. Universidad de Quebec.

Laviosa-Braithwaite, S. (1996). *The English Comparable Corpus (ECC): A Resource and a Methodology for Empirical Study of Translation.* Tesis doctoral. Universidad de Manchester.

Le, A. y Przybocki, M. (2005). NIST 2005 machine translation evaluation official results. *Official release of automatic evaluation scores for all submissions*.

Le, H. T y Abeysinghe, G. (2003). A study to improve the efficiency of a discourse parsing system. En A. Gelbukh (Ed.), *Proceedings of 4th International Conference on Intelligent Text Processing and Computational Linguistics*, Vol. 2588, 101-114.

Lee, Y.-S. (2004). Morphological analysis for statistical machine translation. En *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Assocciation for Computational Linguistics.*

Leech, G. y Wilson, A. (1999). Standards for Tagsets. En H. van Halteren (ed.), *Syntactic Wordclass Tagging*, Dordrecht, Netherlands: Kluwer Academic, 55-80.

Lembersky, G. Ordan, N. y Winter S. (2011). Language models for machine translation: Original vs. translated texts. En *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 363-374.

Lembersky, G., Ordan N. y Winter, S. (2012). Adapting Translation Models to Translationese Improves SMT. En *Proceedings of the European Association for Machine Translation Conference*, 255-266.

Lin, C.-Y. y Och, F. J. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. En *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Lin, D. (1998). Dependency-based evaluation of Minipar. En *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

Lin, D. (2004). A Path-Based Transfer Model for Machine Translation. En *Proceedings of the 20th International Conference on Computational Linguistics*.

Lin, Z., Tou Ng, H. y Kan, M-Y. (2011). Automatically evaluating text coherence using discourse relations. En *Proceedings of the Joint Conference on Human Language Technology and the Association for Computational Linguistics*, 997–1006.

Liu, D. y Gildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. En *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 25–32.

Loureda Lamas, O. (2010). Marcadores del discurso, pragmática experimental y traductología: horizontes para una nueva línea de investigación. *Pragmalingüística*, 18: 74-107.

Maas, H.-D. (1987). The MT system Susy. En M. King (ed.) Machine translation today: the state of the art, Edinburgh: Edinburgh University Press.

MacCartney, B., Grenager, T., Marneffe, M.-C., Cer, D. y Manning, C. D. (2006). Learning to recognize features of valid textual entailments. En *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Macken, L. (2007). Analysis of Translational Correspondence in View of Subsentential Alignment. En *Proceedings of the METIS-II Workshop on New Approaches to Machine Translation*.

Mann, W. C. (2005). RST Web Site. Disponible en: www.sfu.ca/rst.

Mann, W. C. y Thompson, S. A. (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *Text,* 8(3): 243-281.

Marcu, D. (1998). A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. En *Proceedings of the Joint International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Lingiustics: Workshop on Discourse Relations and Discourse Markers*, 1-7.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: A surface based approach. *Computational Linguistics*, 26 (3): 395-448.

Marcu, D., Carlson, L. y Watanabe, M. (2000). The automatic translation of discourse structures. En *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics,* 9-17.

Marcu, D., Wang, W., Echihabi, A. y Knight, K. (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Maynard D. (1999). Term recognition using combined knowledge sources. Tesis doctoral. Manchester Metropolitan University.

Maxwell, K. G. (1992). Automatic Translation of English Compounds: Problems and Prospects. En *Studies in MT and NLP*, volume 8, 37-56.

Mcenery, T. y Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Melamed, I. D. (2001). *Empirical methods for exploiting parallel texts*. Massachusetts Institute of Techonology.

Melamed, I. D. (2004). Statistical Machine Translation by Parsing. En *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Melamed, I. D., Green, R. y Turian, J. P. (2003). Precision and Recall of Machine Translation. En *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics*.

Munday, J. (1998). A Computer-Assisted Approach to the Analysis of Translation Shifts. *Meta* 43 (4): 142-156.

Myers, G. (1991). Lexical Cohesion and Specialized Knowledge in Science and Popular Texts, *Discourse Processes*, 14: 1-26.

Nazar, R., Vivaldi, J. y Cabré, T. (2008). A Suite to Compile and Analyze an LSP Corpus. En *Proceedings of the 6th Language Resources and Evaluation Conference.*

Neubert, A. y Shreve, G. (1992). *Translation as Text*, Kent State University Press.

Newmark, P. (1991). *About Translation*. Clevedon: Multilingual Matters.

Nida, E. A. (1964). *Towards a Science of Translation*. Leiden: Brill.

Nilsson, P. (2004). Translation-specific lexicogrammar? En A. Mauranen y P. Kujamäki (eds.), *Translation Universals. Do they exist?* Amsterdam: Benjamins, 129-141.

Nirenburg, S. (1989). Knowledge-based machine translation, Machine Translation, 4: 5-24.

Nyberg, E. H., Mitamura, T. y Carbonnell, J. G. (1994). Evaluation Metrics for Knowledge-Based Machine Translation. En *Proceedings of the 15th International Conference on Computational Linguistics*, 95-99.

Och, F. J., Tillmann, C. y Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London: Routledge.

Owczarzak, K., Groves, D., Genabith, J. V. y Way, A. (2006). Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. En *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 148–155.

Padó, S., Galley, M., Jurafsky, D. y Manning, C. (2009). Robust Machine Translation Evaluation with Entailment Features. En *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Padó, S., Galley, M., Manning, C. y Jurafsky, D. (2008). Evaluating MT output with entailment technology. En *Proceedings of the Association for Machine Translation in the Americas Metrics MATR Workshop: Metrics for Machine Translation Challenge*.

Papegaaij, B. y Schubert, K. (1988). *Text Coherence in Translation*. Utrecht: Foris.

Papineni, K., Roukos, S., Ward, T. y Zhu, W.-J. (2001). *Bleu: a method for automatic evaluation of machine translation*. RC22176 (Technical Report), IBM T.J. Watson Research Center.

Pardo, T. A. S., Nunes, M. G. V. y Rino L. H. M. (2004). DiZer: An Automatic Discourse Analyzer for Brasilian Portuguese. *Lecture Notes in Artificial Intelligence*, 224-234.

Portolés, J. (1998). *Marcadores del discurso.* Barcelona: Ariel.

Popescu-Belis, A., M. King, y H. Benantar. (2002). Towards a corpus of corrected human translations. En *Proceedings of the Third Language Resources and*

Evaluation Conference: Workshop MT evaluation, human evaluators meet automated metrics, 17–21.

Puurtinen, T. (2003). Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Children's Literature. *Literary and Linguistic Computing*, 18(4): 389–406.

Reitter, D. y Stede, M. (2003). Step by step: Underspecified markup in incremental rhetorical analysis. En *Proceedings of the European Association for Machine Translation 4th International Workshop on Interpreted Corpora*.

Rodríguez Medina, M. J. (2003). *La traducción de la morfosintaxis (inglés - español). Teoría y prácticas.* Universidad de las Palmas de Gran Canaria.

Roukos, S., Graff, D. y Melamed, D. (1995). *Hansard French/English*. Linguistic Data Consortium, Philadelphia.

Russo-Lassner, G., Lin, J. y Resnik, P. (2005). *A Paraphrase-Based Approach to Machine Translation Evaluation* (LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57) (Technical Report). University of Maryland, College Park.

Sager, J. C. (1993). Language Engineering and Translation. Consequences of Automation. Amsterdam: Benjamins.

Santini, M. (2004). *State-of-the-art on Automatic Genre Identification*. Technical report, ITRI, University of Brighton.

Schauer, H. y Hahn, U. (2001). Anaphoric cues for coherence relations. En *Proceedings of International Euroconference "Recent Advances in Natural Language Processing"*, 228-234.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27: 379–423, 623–656.

Shlesinger, M. (1991). Interpreter Latitude vs. Due Process: Simultaneous and Consecutive Interpretation in Multilingual Trial, en Tirkkonen-Condit (ed.), *Empirical Research in Tranlation and Intercultural Studies: Selected Papers of the TRANS-SIF Seminar, Savonlinna 1988*, Tübingen: Gunter Narr, 147-155.

Shlesinger, M. (1995). Shifts in Cohesion in Simultaneous Interpreting, *The Translator*, 1 (2): 193-214.

Sierra, G. (2006). Diseño de corpus textuales para fines lingüísticos. En *Actas del IX Encuentro Internacional de Lingüística de Noroeste, II, Sonora: UNISON*, 445-462.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Somers, H. L. (2000). Example-based Machine Translation. En D. Robert, M. Hermann y H. L. Somers (eds), *Handbook of Natural Language Processing*, New York: M. Dekker, 611-627.

Steiner, E. (2002). Grammatical Metaphor in Translation - Some Methods for Corpus-based Investigations. En H. Hasselgard, S. Johansson, B. Behrens y C. Fabricius-Hansen (eds.), *Information Structure in a Cross-linguistic Perspective*, Amsterdam: Rodopi, 213-228.

Sumita, K., Ono, K., Chino, T., Ukita, T. y Amano, S. (1992). A discourse structure analyzer for Japanese text. En *Proceedings of the International Conference on the Fifth Generation Computer Systems*, 1133-1140.

Szymańska, I. (2011). *Mosaics. A Construction-Grammar-based approach to translation.* Warszawa: Semper.

Tofiloski, M., Brooke, J. y Taboada, M. (2009). A Syntactic and Lexical-Based Discourse Segmenter. En *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

Toury, G. (1980). *In Search of a Theory of Translation*. Tel Aviv: Porter Institute.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Toury, G. (2004). Probabilistic Explanations in Translation Studies: Welcome as they are, would they qualify as Universals? En A. Mauranen y P. Kujamäki (eds.), *Translation Universals. Do they Exist?* Amsterdam: Benjamins, 15–32.

Trujillo, I. A. (1995). Lexicalist Machine Translation of Spatial Prepositions. Tesis doctoral. Trinity Hall, University of Cambridge.

Turian, J. P., Shen, L. y Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. En *Proceedings of the Machine Translation Summit IX*.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind 49*: 433-460.

Ulrych, M. y Murphy A. (2008). Descriptive Translation Studies and the Use of Corpora: Investigating Mediation Universals. En C. T. Torsello, K. Acherley y E. Castello (eds.), *Corpora for University language teachers*, Berlin: Peter Lang, 141-166.

van Halteren, H. (2008). Source language markers in europarl translations. En *Proceedings of the International Conference on Computational Linguistics*, 937–944.

van Leuven-Zwart, K. M. (1989). Translation and original: Similarities and dissimilarities. *Target*, 1(2): 151-181.

Vanderauwera, R. (1985). *Dutch Novels Translated into English: The Transformation of a "Minorirty" Literature*, Amsterdam: Rodopi.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation. En *IFIP Congress*, Edinburgh, 254-260.

Vinay, J.-P. y Darbelnet, J. (1958). *Stylistique Comparée du Français et de l'Anglais: Méthode de Traduction*. Paris: Didier.

Vivaldi J. (2001). *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*. Tesis doctoral. Universidad Politécnica de Catalunya, Departament de Llenguatges i Sistemes Informàtics.

Vivaldi, J. y Rodriguez, H. (2011). Extracting terminology from Wikipedia. *Procesamiento del lenguaje natural*, 47: 65-73.

Webber, B. (2004). D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, 28(5): 751–779.

White, J. S., O'Connell, T. y O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. En *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, 193–205.

Widdowson, H. G. (1978). *Teaching Language as Communication*. Oxford: Oxford University Press.

Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. New York: Springer.

Diccionarios

Hornby, A. S. (2011). *Oxford Advanced Learner's Dictionary*. Disponible en: http://oald8.oxfordlearnersdictionaries.com

Lackie, J. M. y Dow, J. A. T. (eds.) (1995). *The Dictionary of cell biology*. London: Academic.

Real Academia Española. (2001). *Diccionario de la lengua española*, 22ª edición. Disponible en: http://lema.rae.es/drae/

Widmer, F. y Beffa, R. (2000). *Diccionario de bioquímica y biología molecular*. Traducido por M. J. Arrizubieta Balerdi. Zaragoza: Acribia.

ANEXOS

Anexo A. Candidatos a términos extraídos de la traducción humana y de las traducciones automáticas

Traducción humana

CT (lema)	CD	Patrón	Frecuencia
adaptación	1.00	N	3
adenovirus	0.63	N	2
aislamiento	0.64	N	1
análisis	0.85	N	2
anestesia	0.50	N	1
angustia	0.50	N	1
animal	1.00	N	25
antígeno	0.53	N	14
apoptosis	0.71	N	1
asma	0.60	N	1
ataque	0.85	N	4
aterosclerosis	0.61	N	2
bacteria	0.89	N	34
biología	1.00	N	11
biomedicina	1.00	N	1
brazo	1.00	N	3
caenorhabditis elegans	0.50	N	1
cáncer	0.67	N	30
cara	1.00	N	1
cardiopatía	0.67	N	1
célula	1.00	N	210
cepa	1.00	N	3
cerdo	0.63	N	3
cerebelo	0.59	N	3
cerebro	0.59	N	33
cicatrización	0.80	N	1
cirugía	0.83	N	2
clon	0.75	N	2
código	0.85	N	1
convulsión	1.00	N	6
corazón	0.63	N	2
cráneo	0.54	N	2
cromosoma	1.00	N	14
cuerpo	0.85	N	10
dedo	1.00	N	3
deglución	0.54	N	1
desarrollo	1.00	N	25
diabetes	0.65	N	2
diagnóstico	1.00	N	5
diana	0.57	N	4

división	1.00	N	1
dominio	1.00	N	23
dosis	0.50	N	5
droga	0.50	N	1
drosophila melanogaster	0.86	N	1
eritrocito	0.61	N	1
estímulo	1.00	N	1
excitabilidad	0.52	N	1
expresión	0.76	N	8
extensión	0.85	N	1
fagocito	0.58	N	2
familia	1.00	N	4
fenilcetonuria	0.68	N	2
fisiología	1.00	N	1
garganta	1.00	N	3
gemelo	0.64	N	5
gen	0.75	N	209
gene targeting	0.50	N	1
genética	1.00	N	1
gonorrea	0.53	N	1
hemofilia	1.00	N	1
herida	0.67	N	1
herpesvirus	0.63	N	3
hígado	0.65	N	1
hipercolesterolemia	0.56	N	1
hipersensibilidad	0.80	N	1
huésped	0.52	N	10
ictus	0.62	N	1
implante	0.75	N	2
infección	0.78	N	8
inflamación	0.57	N	1
inmunidad	0.75	N	3
inmunoterapia	0.75	N	1
interleuquina	0.83	N	1
leucemia	0.75	N	5
leucocito	0.58	N	2
linfocito	0.53	N	7
liposoma	0.60	N	1
macrófago	0.58	N	1
malformación	0.95	N	6
mamífero	0.95	N	12
medicamento	0.50	N	1
medicina	1.00	N	4
médico	0.95	N	6
melanoma	0.50	N	1
membrana	0.85	N	15
meningitis	1.00	N	2
metabolismo	0.60	N	3

metilación	0.73	N	1
método	0.85	N	16
microinyección	0.50	N	2
microorganismo	1.00	N	3
mitocondria	0.71	N	1
mosca	0.80	N	2
múrido	0.63	N	1
mutación	0.50	N	25
mutagénesis	1.00	N	4
neisseria gonorrhoeae	1.00	N	2
neurobiología	0.53	N	1
neurocirugía	0.80	N	1
neurología	1.00	N	1
neurona	0.61	N	48
núcleo	0.64	N	21
oncólogo	0.63	N	1
operación	0.71	N	6
organismo	0.85	N	30
orgánulo	1.00	N	2
páncreas	0.60	N	1
patogénesis	0.67	N	1
patógeno	0.67	N	4
patología	0.75	N	6
personalidad	0.50	N	1
pierna	1.00	N	3
plaqueta	0.61	N	1
polio	0.58	N	1
prevención	0.85	N	2
procariota	1.00	N	2
progenie	1.00	N	1
pronóstico	1.00	N	1
pulgar	1.00	N	2
pulmón	0.55	N	2
quimiotaxis	0.61	N	1
quimioterapia	0.63	N	2
rana	0.95	N	2
reacción	0.85	N	10
receptor	0.64	N	45
represión	0.50	N	1
respuesta	0.85	N	22
retrovirus	0.63	N	14
roedor	0.63	N	1
sangre	0.50	N	7
secuencia	0.85	N	59
sinapsis	0.53	N	3
síndrome	1.00	N	1
síntoma	1.00	N	13
soma	0.85	N	3

terapia	0.85	N	47
tétanos	0.86	N	1
transcripción	1.00	N	32
transfusión	1.00	N	1
transplante	0.71	N	2
trasplante	0.75	N	8
tratamiento	1.00	N	29
traumatismo	1.00	N	1
tumor	0.67	N	19
ubre	0.54	N	1
vida	1.00	N	8
virulencia	0.80	N	1
accidente cerebrovascular	0.65	N+Adj	7
agente etiológico	0.63	N+Adj	1
anemia falciforme	0.63	N+Adj	2
célula dendrítico	0.51	N+Adj	2
cerebro humano	0.56	N+Adj	1
código genético	1.00	N+Adj	5
comunicación celular	1.00	N+Adj	1
corteza cerebral	0.59	N+Adj	1
cuerpo humano	1.00	N+Adj	2
diagnóstico precoz	0.95	N+Adj	1
división celular	1.00	N+Adj	4
enfermedad degenerativo	0.95	N+Adj	1
enfermedad genético	0.95	N+Adj	3
enfermedad hereditario	0.95	N+Adj	1
enfermedad infeccioso	0.71	N+Adj	1
enfermedad neurodegenerativo	0.95	N+Adj	1
enfermedad neurológico	0.95	N+Adj	4
esclerosis tuberoso	0.95	N+Adj	1
fibrosis quístico	0.60	N+Adj	4
función vital	0.95	N+Adj	1
genoma humano	1.00	N+Adj	3
grupo sanguíneo	1.00	N+Adj	1
líquido tisular	0.53	N+Adj	1
marcador genético	0.60	N+Adj	2
material genético	1.00	N+Adj	2
médula espinal	0.51	N+Adj	3
membrana celular	0.85	N+Adj	3
membrana nuclear	0.95	N+Adj	1
mutación génico	0.95	N+Adj	1
nervio craneal	0.52	N+Adj	5
organismo pluricelular	1.00	N+Adj	5
proceso evolutivo	0.85	N+Adj	2
producto génico	1.00	N+Adj	1
receptor celular	1.00	N+Adj	1
reproducción sexual	1.00	N+Adj	1
sistema inmunitario	0.53	N+Adj	22

sustancia negro	0.50	N+Adj	7
tejido conjuntivo	1.00	N+Adj	1
tejido nervioso	0.56	N+Adj	1
torrente sanguíneo	0.51	N+Adj	2
tracto respiratorio	0.50	N+Adj	4
trastorno genético	0.95	N+Adj	1
tronco cerebral	0.71	N+Adj	7
tronco encefálico	0.71	N+Adj	4
vaso sanguíneo	0.54	N+Adj	1
vector génico	0.82	N+Adj	1
vía intravenoso	0.95	N+Adj	1
sistema nervioso central	0.53	N+Adj+Adj	3
cáncer de piel	0.50	N+Prep+N	1
enfermedad de parkinson	0.50	N+Prep+N	11
factor de transcripción	1.00	N+Prep+N	9
transmisión de señal	0.95	N+Prep+N	2

CT	CD	Patrón	Frecuencia
adenovirus	0.63	N	2
análisis	0.85	N	5
anestesia	0.50	N	1
animal	1.00	N	31
antígeno	0.53	N	14
apoptosis	0.71	N	1
asma	0.60	N	1
ataque	0.85	N	7
aterosclerosis	0.61	N	2
bacteria	0.89	N	41
biología	1.00	N	9
brazo	1.00	N	3
cáncer	0.67	N	48
capilar	0.54	N	1
cara	1.00	N	3
célula	1.00	N	235
cepa	1.00	N	4
cerdo	0.63	N	3
cerebelo	0.59	N	3
cerebro	0.59	N	39
choque	1.00	N	1
clon	0.75	N	7
código	0.85	N	3
convulsión	1.00	N	2
corazón	0.63	N	2
cráneo	0.54	N	2
cromosoma	1.00	N	17
cuadro	0.85	N	2
cuerpo	0.85	N	26
desarrollo	1.00	N	25
diabetes	0.65	N	2
diagnóstico	1.00	N	3
dominio	0.85	N	21
dosis	0.50	N	6
droga	0.50	N	4
encéfalo	0.59	N	3
espacio	0.85	N	1
estatura	1.00	N	1
evolución	0.85	N	5
excitabilidad	0.52	N	1
expresión	0.76	N	6
fagocito	0.58	N	1
familia	1.00	N	4
fenilcetonuria	0.68	N	2
fijación	0.50	N	1

fisiología	1.00	N	2
garganta	1.00	N	3
gemelo	0.64	N	3
gen	0.75	N	227
genetista	0.95	N	8
gonorrea	0.53	N	1
hemofilia	1.00	N	1
herpesvirus	0.63	N	2
hígado	0.65	N	3
hipersensibilidad	0.80	N	1
huésped	0.52	N	13
implante	0.75	N	1
infección	0.78	N	9
inflamación	0.57	N	1
inmunidad	0.75	N	3
inmunoterapia	0.75	N	2
interleucina	0.88	N	1
inyección	0.85	N	5
leucemia	0.75	N	6
limitación	0.95	N	1
linfocito	0.53	N	4
linfoma	0.54	N	1
liposoma	0.60	N	2
macrófago	0.58	N	1
malformación	0.95	N	6
mamífero	0.95	N	12
medicamento	0.50	N	2
medicina	1.00	N	4
médico	0.95	N	12
membrana	0.85	N	9
meningitis	1.00	N	1
metabolismo	0.60	N	2
método	0.85	N	13
microbio	0.95	N	3
microinyección	0.50	N	2
mitocondria	0.71	N	1
mosca	0.80	N	2
mutación	0.50	N	25
nervio	0.55	N	2
neurobiología	0.53	N	1
neurobiólogo	0.51	N	1
neurocirugía	0.80	N	1
neurología	1.00	N	1
neurona	0.61	N	48
núcleo	0.64	N	28
oncólogo	0.63	N	2
operación	0.71	N	5
organismo	0.85	N	22

orgánulo	1.00	N	2
páncreas	0.60	N	1
personalidad	0.50	N	1
pierna	1.00	N	3
planta	1.00	N	1
plaqueta	0.61	N	1
polio	0.58	N	1
prevención	0.85	N	3
pronóstico	0.85	N	2
pulgar	1.00	N	2
pulmón	0.55	N	2
quimiotaxis	0.61	N	1
quimioterapia	0.63	N	2
rana	0.95	N	2
reacción	0.85	N	12
receptor	0.64	N	43
respuesta	0.85	N	22
retrovirus	0.63	N	17
sangre	0.50	N	10
secuencia	0.85	N	55
sinapsis	0.53	N	3
síntoma	1.00	N	15
terapia	0.85	N	57
tirosina hidroxilasa	1.00	N	1
transcripción	1.00	N	40
transfusión	1.00	N	1
trasplante	0.75	N	11
tratamiento	1.00	N	26
traumatismo	1.00	N	1
tronco	0.85	N	3
tumor	0.67	N	12
ubre	0.54	N	1
variación	0.51	N	3
vida	1.00	N	11
virulencia	0.80	N	1
virus	0.67	N	24
accidente cerebrovascular	0.65	N+Adj	3
biología molecular	0.50	N+Adj	3
célula dendrítico	0.51	N+Adj	2
cerebro humano	0.56	N+Adj	1
ciencia médico	0.95	N+Adj	1
código genético	1.00	N+Adj	5
columna vertebral	0.50	N+Adj	2
comunicación celular	1.00	N+Adj	1
corteza cerebral	0.59	N+Adj	1
cuerpo humano	1.00	N+Adj	2
derrame cerebral	0.62	N+Adj	4
división celular	1.00	N+Adj	3

enfermedad cardiaco	0.63	N+Adj	1
enfermedad degenerativo	0.95	N+Adj	3
enfermedad genético	0.95	N+Adj	4
enfermedad infeccioso	0.71	N+Adj	1
enfermedad neurológico	0.95	N+Adj	7
enfermedad raro	0.95	N+Adj	2
envoltura nuclear	1.00	N+Adj	1
envoltura viral	0.63	N+Adj	1
esclerosis tuberoso	0.95	N+Adj	1
fibrosis quístico	0.60	N+Adj	4
función vital	0.95	N+Adj	1
genoma humano	1.00	N+Adj	2
glóbulo blanco	0.59	N+Adj	1
glóbulo rojo	0.58	N+Adj	1
material genético	1.00	N+Adj	3
médula espinal	0.51	N+Adj	3
membrana celular	0.85	N+Adj	4
nervio craneal	0.52	N+Adj	4
núcleo celular	1.00	N+Adj	1
organismo unicelular	0.95	N+Adj	3
proceso evolutivo	0.85	N+Adj	1
receptor celular	1.00	N+Adj	1
reproducción sexual	1.00	N+Adj	1
sistema inmunitario	0.53	N+Adj	2
sustancia negro	0.50	N+Adj	5
tallo cerebral	0.71	N+Adj	1
tejido conectivo	0.95	N+Adj	1
tejido nervioso	0.56	N+Adj	2
torrente sanguíneo	0.51	N+Adj	2
tracto respiratorio	0.50	N+Adj	3
trastorno genético	0.95	N+Adj	1
tronco cerebral	0.71	N+Adj	7
vaso sanguíneo	0.54	N+Adj	2
vía intravenoso	0.95	N+Adj	1
vía respiratorio	0.50	N+Adj	1
sistema nervioso central	0.53	N+Adj+Adj	3
cáncer de piel	0.50	N+Prep+N	1
factor de transcripción	1.00	N+Prep+N	11
medio de cultivo	1.00	N+Prep+N	1
transmisión de señal	0.95	N+Prep+N	1
enfermedad de parkinson	0.50	N+Prep+N	9
anemia de célula falciforme	0.60	N+Prep+N+Adj	2

CT	CD	Patrón	Frecuencia
adenovirus	0.63	N	1
altura	0.85	N	6
análisis	0.85	N	3
anestesia	0.5	N	1
animal	1	N	31
antígeno	0.53	N	13
apoptosis	0.71	N	1
asma	0.6	N	1
ataque	0.85	N	4
bacteria	0.89	N	36
biología	1	N	12
brazo	1	N	3
cáncer	0.67	N	58
célula	1	N	237
cerdo	0.63	N	3
cerebelo	0.59	N	3
cerebro	0.59	N	57
cirugía	0.83	N	1
clon	0.75	N	7
código	0.85	N	4
cráneo	0.54	N	2
cromosoma	1	N	17
cuerpo	0.85	N	26
desarrollo	1	N	21
diabetes	0.65	N	2
diagnóstico	1	N	2
dominio	0.85	N	2
dosis	0.5	N	6
droga	0.5	N	3
estatura	1	N	1
estímulo	1	N	1
expresión	0.76	N	6
extensión	0.85	N	3
fagocito	0.58	N	1
familia	1	N	4
fibrosis	1	N	4
fisiología	1	N	2
garganta	1	N	3
gemelo	0.64	N	5
gen	0.75	N	259
genetista	0.95	N	8
hemofilia	1	N	1
herido	0.63	N	0
hígado	0.65	N	3
hipersensibilidad	0.8	N	1

implante	0.75	N	1
infección	0.78	N	9
inflamación	0.57	N	1
inmunidad	0.75	N	4
leucemia	0.75	N	6
limitación	0.95	N	1
linfocito	0.53	N	2
linfoma	0.54	N	1
liposoma	0.6	N	2
malformación	0.95	N	6
mamífero	0.95	N	12
medicamento	0.5	N	3
medicina	1	N	6
médico	0.95	N	10
melanoma	0.5	N	1
membrana	0.85	N	13
metabolismo	0.6	N	2
método	0.85	N	11
microbio	0.95	N	3
microinyección	0.5	N	1
mitocondria	0.71	N	1
mutación	0.5	N	24
neurona	0.61	N	48
núcleo	0.64	N	32
operación	0.71	N	7
organismo	0.85	N	24
orgánulo	1	N	2
personalidad	0.5	N	1
pierna	1	N	3
planta	1	N	1
plaqueta	0.61	N	1
polio	0.58	N	1
pronóstico	0.85	N	2
pulgar	1	N	2
pulmón	0.55	N	2
quimiotaxis	0.61	N	1
quimioterapia	0.63	N	2
rabia	0.62	N	1
rana	0.95	N	2
receptor	0.64	N	45
respuesta	0.85	N	24
retrovirus	0.63	N	17
riñón	0.57	N	1
sangre	0.5	N	9
secuencia	0.85	N	55
sinapsis	0.53	N	4
síntoma	1	N	15
substantia nigra	0.5	N	6

terapia	0.85	N	59
tétanos	0.86	N	1
transcripción	1	N	40
transfusión	1	N	1
transplante	0.71	N	7
tratamiento	1	N	21
trauma	0.85	N	1
tumor	0.67	N	21
ubre	0.54	N	1
variación	0.51	N	3
vida	1	N	11
virulencia	0.8	N	1
virus	0.67	N	26
aparato respiratorio	0.53	N+Adj	4
célula dendrítico	0.51	N+Adj	2
célula eucariótico	0.95	N+Adj	1
cerebro humano	0.56	N+Adj	1
ciencia médico	0.95	N+Adj	1
código genético	1	N+Adj	5
corteza cerebral	0.59	N+Adj	1
cuerpo humano	1	N+Adj	2
desorden genético	0.95	N+Adj	1
división celular	1	N+Adj	3
enfermedad genético	0.95	N+Adj	4
enfermedad neurológico	0.95	N+Adj	8
genoma humano	1	N+Adj	2
glóbulo blanco	0.55	N+Adj	1
material genético	1	N+Adj	3
médula espinal	0.51	N+Adj	3
membrana celular	0.85	N+Adj	4
mutación génico	0.95	N+Adj	1
nervio craneal	0.52	N+Adj	5
receptor celular	1	N+Adj	1
reproducción sexual	1	N+Adj	1
sistema inmunitario	0.75	N+Adj	20
tejido conectivo	0.95	N+Adj	1
tejido nervioso	0.56	N+Adj	2
vaso sanguíneo	0.54	N+Adj	2
cáncer de piel	0.5	N+Prep+N	1
factor de transcripción	1	N+Prep+N	11
medio de cultivo	1	N+Prep+N	1
tipo de sangre	0.95	N+Prep+N	1
enfermedad de parkinson	0.50	N+Prep+N	9

Anexo B. Clasificación de las diferencias entre la traducción humana y las traducciones automáticas a nivel léxico-terminológico

Oración	TO	TH	*Shift* TO-TH	*Google*	*Shift* TO-*Google*	*Lucy*	*Shift* TO-*Lucy*
383	0	patogénesis	adición (obligatoria)	0	alteración	0	alteración
837	AIDS-causing agent	agente etiológico de el sida	explicitación (opcional)	agente causante de el sida	-	agente que causar el sida	-
159	animal	ratón mutante	explicitación (opcional)	animal	-	animal	-
898	animal	roedor	explicitación (opcional)	animal	-	animal	-
441	animal	individuo manipulado	modificación (opcional)	animal	-	animal	-
394	biomedical researcher	biomedicina	modificación (opcional)	investigador biomédico	-	investigador biomédico	-
1060	blood type	grupo sanguíneo	variación	sangre de tipo	alteración	tipo de sangre	variación
267	body	cuerpo humano	explicitación (opcional)	cuerpo	-	cuerpo	-
2	body	organismo	modificación (opcional)	cuerpo	-	cuerpo	-
1005	brain stem	tronco cerebral	variación	tronco cerebral	-	tallo de cerebro	variación
307	cancer	patología	implicitación (opcional)	cáncer	-	cáncer	-
1005	cell body	soma	implicitación (opcional)	cuerpo celular	-	cuerpo celular	-
361	defense	ataque	modificación (opcional)	defensa	-	defensa	-
303	desease-causing bacterium	bacteria patógeno	implicitación (opcional)	bacteria causante de enfermedad	-	bacteria que causar enfermedad	-
1013	diagnose	diagnóstico	modificación (opcional)	diagnosticar	-	diagnosticar	-

307	early detection	diagnóstico precoz	explicitación (opcional)	detección temprano	-	primero detección	alteración
1023	eat	deglución	modificación (opcional)	comer	-	comiendo	alteración
662	enhancer	intensificador (enhancer)	adición (opcional)	potenciador	-	potenciador	-
662	enhancer	intensificador (enhancer)	-	potenciador	alteración	potenciador	alteración
316	evolution	proceso evolutivo	explicitación (opcional)	evolución	-	evolución	-
111	gene targeting	sustitución dirigida de genes (gene targeting)	adición	orientación de genes	-	gen que apunta	-
111	gene targeting	sustitución dirigida de genes (gene targeting)	modificación (obligatoria)	orientación de genes	alteración	gen que apunta	alteración
128	geneticist	0	omisión (opcional)	genetista	-	genetista	-
89	healing	cicatrización	explicitación (opcional)	cicatrización	explicitación (opcional)	que curar	alteración
622	heart disease	cardiopatía	variación	enfermedad cardiaco	variación	enfermedad de corazón	variación
340	host	huésped	variación	huésped	-	anfitrión	variación
756	human	genoma humano	explicitación (opcional)	humano	-	humano	-
808	infectious desease	enfermedad infeccioso	variación	enfermedad infeccioso	-	enfermedad contagioso	variación
227	inherited ills	enfermedad hereditario	-	mal heredado	alteración	ills heredado	alteración
782	interleukin	interleuquina	variación	interleucina	variación	interleukin	alteración
847	life	biología	modificación (opcional)	vida	-	vida	-
763	marker genes	marcador genético	modificación (opcional)	gen marcador	-	gen de marcador	alteración

303	microbe	microorganismo	implicitación (opcional)	microbio	-	microbio	-
443	microinjecting	microinyección	modificación (opcional)	microinyección	modificación (opcional)	microinyectar	-
106	mouse	múrido	implicitación (opcional)	ratón	-	ratón	-
153	mouse	animal	implicitación (opcional)	ratón	-	ratón	-
157	mutant animal	animal mutante	-	animal mutante	-	animal de mutante	alteración
2	neurobiologist	0	omisión (opcional)	neurobiólogo	-	neurobiologist	alteración
1014	neurobiology	neurobiología	-	neurobiología	-	neurobiology	alteración
306	neurological condition	enfermedad neurológico	modificación (obligatoria)	enfermedad neurológico	modificación (obligatoria)	condición neurológico	alteración
940	neurosurgery	neurocirugía	-	neurocirugía	-	neurosurgery	alteración
837	pathogen	patógeno	-	agente patógeno	explicitación (opcional)	pathogen	alteración
831	patient	organismo	modificación (opcional)	paciente	-	paciente	-
341	phagocytic cell	fagocito	implicitación (opcional)	célula fagocítico	-	célula fagocítico	-
808	preventing	prevención	modificación (opcional)	prevención	modificación (opcional)	prevenir	-
263	red cell	eritrocito	variación	glóbulo rojo	variación	célula blanca	variación
322	respiratory tract	tracto respiratorio	variación	tracto respiratorio	-	aparato respiratorio	variación
363	respiratory tract	tracto respiratorio	variación	vías respiratorias	variación	aparato respiratorio	variación
852	seizure	convulsión	-	convulsión	-	toma	alteración
911	seizure	convulsión	-	embargo	alteración	golpe	alteración
915	seizure	convulsión	variación	ataque	variación	golpe	alteración
316	sickle cell anemia	anemia falciforme	implicitación (opcional)	anemia de célula falciforme	-	anemia de célula de hoz	alteración

911	stroke	accidente cerebrovascular	-	accidente cerebrovascular	-	toma	alteración
926	stroke	accidente cerebrovascular	-	accidente cerebrovascular	-	golpe	alteración
909	stroke	accidente cerebrovascular	-	carrera	alteración	golpe	alteración
915	stroke	ictus	variación	derrame cerebral	variación	toma	alteración
782	T cell	linfocito	variación	célula T	variación	célula T	variación
350	tissue fluid	líquido tisular	variación	fluido tisular	variación	fluido de tejido	alteración
908	transplant	trasplante	modificación (opcional)	trasplantar	-	trasplantar	-
926	trauma	traumatismo	modificación (opcional)	traumatismo	modificación (opcional)	trauma	-
745	treating	tratamiento	modificación (opcional)	tratar	-	que invitar	alteración
729	treating	tratamiento	modificación (opcional)	tratamiento	modificación (opcional)	tratar	-
263	white cell	leucocito	variación	glóbulo blanco	variación	célula roja	variación
746	white cell	leucocito	variación	célula blanca	variación	célula blanca	variación

Anexo C. Codificación morfosintáctica en formato EAGLES para el español

	Posición	Atributo	Valor	Código
Adjetivos	1	Categoría	Adjetivo	A
	2	Tipo	Calificativo	Q
			Ordinal	O
			-	0
	3	Grado	-	0
			Aumentativo	A
			Diminutivo	C
			Superlativo	S
	4	Género	Masculino	M
			Femenino	F
			Común	C
	5	Número	Singular	S
			Plural	P
			Invariable	N
	6	Función	-	0
			Participio	P

Adverbios	1	Categoría	Adverbio	R
	2	Tipo	General	G
			Negativo	N

Determinantes	1	Categoría	Determinante	D
	2	Tipo	Demostrativo	D
			Posesivo	P
			Interrogativo	T
			Exclamativo	E
			Indefinido	I
			Artículo	A
	3	Persona	Primera	1
			Segunda	2
			Tercera	3
	4	Género	Masculino	M
			Femenino	F
			Común	C
			Neutro	N
	5	Número	Singular	S
			Plural	P
			Invariable	N
	6	Poseedor	Singular	S
			Plural	P

Nombres	1	Categoría	Nombre	N
	2	Tipo	Común	C
			Propio	P
	3	Género	Masculino	M
			Femenino	F
			Común	C
	4	Número	Singular	S
			Plural	P
			Invariable	N
	5 y 6	Clasificación semántica	Persona	SP
			Lugar	G0
			Organización	O0
			Otros	V0
	7	Grado	Aumentativo	A
			Diminutivo	D

Verbos	1	Categoría	Verbo	V
	2	Tipo	Principal	M
			Auxiliar	A
			Semiauxiliar	S
	3	Modo	Indicativo	I
			Subjuntivo	S
			Imperativo	M
			Infinitivo	N
			Gerundio	G
			Participio	P
	4	Tiempo	Presente	P
			Imperfecto	I
			Futuro	F
			Pasado	S
			Condicional	C
			-	0
	5	Persona	Primera	1
			Segunda	2
			Tercera	3
	6	Número	Singular	S
			Plural	P
	7	Género	Masculino	M
			Femenino	F

Pronombres	1	Categoría	Pronombre	P
	2	Tipo	Personal	P
			Demostrativo	D
			Posesivo	X
			Indefinido	I
			Interrogativo	T
			Relativo	R
			Exclamativo	E
	3	Persona	Primera	1
			Segunda	2
			Tercera	3
	4	Género	Masculino	M
			Femenino	F
			Común	C
			Neutro	N
	5	Número	Singular	S
			Plural	P
			ImpersonalMInvariable	N
	6	Caso	Nominativo	N
			Acusativo	A
			Dativo	D
			Oblicuo	O
	7	Poseedor	Singular	S
			Plural	P
	8	Politeness	Polite	P

Conjunciones	1	Categoría	Conjunción	C
	2	Tipo	Coordinada	C
			Subordinada	S

Interjecciones	1	Categoría	Interjección	I

Preposiciones	1	Categoría	Adposición	S
	2	Tipo	Preposición	P
	3	Forma	Simple	S
			Contraída	C
	4	Género	Masculino	M
	5	Número	Singular	S

Signos de puntuación	1	Categoría	Puntuación	F

Cifras y numerales	1	Categoría	Cifra	Z

Anexo D. Diferencias significativas en las frecuencias de aparición de n-gramas de etiquetas POS entre la traducción humana y las traducciones automáticas

Unigramas

POS	Frecuencia TH	Frecuencia *Google*	X^2 TH-*Google*	Frecuencia *Lucy*	X^2 TH-*Lucy*
ao	50	46	0.56	52	0.01
aq	2198	2101	15.17	1812	51.35
cc	686	755	0.27	708	0.02
cs	484	610	7.32	608	11.43
da0	3099	3385	0.71	2333	143.04
dd0	280	323	0.84	348	5.85
di0	1003	1131	1.47	1139	5.85
dp1	45	52	0.14	54	0.61
dp3	177	139	7.60	140	5.31
dt0	14	5	4.91	8	1.79
nc	6065	6519	0.05	6474	6.84
np	556	634	1.19	590	0.35
pd00	33	20	4.15	42	0.87
pi00	128	138	0.00	168	4.49
pp10	45	28	5.22	23	7.68
pp1n	3	10	3.31	4	0.12
pp30	594	486	19.92	451	23.75
pp3a	67	26	21.06	33	12.44
pp3d	32	10	13.11	27	0.56
pp3o	11	10	0.14	1	8.59
pr00	493	531	0.01	533	0.74
pt00	28	56	7.52	83	25.97
rg	732	1045	37.06	1168	93.07
rn	158	153	0.79	121	5.90
sps00	4256	4544	0.03	4424	0.51
vaii	44	59	1.28	74	6.92
vaip	146	226	12.26	185	3.71
van0	2	4	0.54	8	3.45
vap0	3	5	0.37	6	0.93
vasi	6	2	2.29	4	0.45
vasp	6	2	2.29	1	3.70
vmg0	71	87	0.71	165	35.33
vmic	119	96	4.31	158	4.60
vmif	47	22	10.88	36	1.75
vmii	216	52	112.64	374	38.96
vmip	1131	1139	2.25	1091	2.14
vmis	138	178	2.71	1	138.83
vmm0	29	7	15.02	25	0.40
vmn0	679	836	7.51	804	7.91

vmp0	652	738	1.09	698	0.66
vmsi	58	8	41.49	56	0.10
vmsp	121	21	77.75	84	7.65
vsg0	3	7	1.34	2	0.23
vsic	9	16	1.51	13	0.63
vsif	4	4	0.01	8	1.24
vsii	24	36	1.65	51	9.08
vsip	137	254	27.68	260	35.41
vsis	4	30	18.17	0	4.10
vsn0	21	97	44.03	58	16.46
vsp0	4	26	14.69	13	4.55
vssi	8	1	5.94	5	0.77
vssp	28	23	0.90	16	3.58
z	175	189	0.01	208	2.10

Bigramas

POS	Frecuencia TH	Frecuencia *Google*	X^2 TH-*Google*	Frecuencia *Lucy*	X^2 TH-*Lucy*
ao nc	38	42	0.00	38	0.04
ao z0	3	3	0.01	9	2.75
aq aq	93	80	2.40	73	3.37
aq cc	119	119	0.43	106	1.43
aq cs	38	53	1.37	59	3.69
aq da	23	10	6.30	17	1.18
aq di	11	10	0.17	15	0.46
aq nc	305	318	0.29	209	22.58
aq np	9	6	0.88	8	0.11
aq pp	39	36	0.51	35	0.42
aq pr	65	54	2.17	52	2.07
aq rg	24	56	10.27	40	3.34
aq rn	10	9	0.17	8	0.32
aq sp	551	551	2.02	496	5.89
aq va	10	22	3.55	19	2.42
aq vm	166	138	5.54	158	0.70
aq vs	27	48	4.25	35	0.72
cc aq	35	34	0.22	26	1.75
cc cs	10	7	0.82	9	0.10
cc da	96	142	5.46	93	0.27
cc dd	5	6	0.03	8	0.57
cc di	47	39	1.58	42	0.54
cc dp	14	18	0.22	19	0.56
cc nc	92	73	4.11	125	3.71
cc np	31	39	0.36	31	0.03
cc pi	15	15	0.05	15	0.01
cc pp	29	26	0.52	22	1.29
cc pr	2	6	1.68	7	2.57
cc pt	1	6	3.17	8	5.15
cc rg	39	61	3.17	65	5.44
cc rn	18	17	0.18	9	3.41
cc sp	70	62	1.40	55	2.51
cc vm	129	121	1.39	140	0.10
cc vs	5	7	0.19	10	1.46
cc z0	8	6	0.48	5	0.83
cs aq	4	8	1.02	7	0.69
cs cs	4	14	4.75	14	5.14
cs da	175	252	8.23	213	2.28
cs dd	12	20	1.38	19	1.29
cs di	40	75	7.93	74	8.75
cs dp	7	13	1.33	13	1.55
cs nc	8	34	14.01	50	28.67
cs np	9	19	2.78	18	2.63

cs pd	7	2	3.22	3	1.78
cs pi	6	6	0.02	11	1.26
cs pp	43	31	3.10	31	2.50
cs rg	13	19	0.68	20	1.20
cs rn	17	12	1.34	10	2.13
cs sp	14	8	2.19	10	0.85
cs va	5	9	0.83	3	0.59
cs vm	90	60	8.85	83	0.67
cs vs	8	21	4.79	12	0.64
cs z0	0	2	1.84	6	5.75
da ao	29	25	0.73	32	0.05
da aq	115	112	0.70	100	1.80
da di	10	8	0.42	11	0.01
da nc	2667	2949	0.40	1962	155.75
da np	100	123	0.83	53	16.60
da pi	9	6	0.88	18	2.63
da pr	83	89	0.01	68	2.22
da rg	9	17	1.84	22	4.91
da sp	22	12	3.86	5	11.46
da vm	25	25	0.09	24	0.09
da z0	28	14	5.94	36	0.69
dd aq	11	4	3.89	3	4.93
dd cs	2	20	13.27	0	2.09
dd nc	256	289	0.18	329	6.31
di aq	45	75	5.20	29	4.19
di cc	5	6	0.03	3	0.59
di da	35	37	0.02	44	0.67
di dd	1	1	0.00	6	3.36
di di	11	20	1.91	22	3.21
di nc	825	924	0.37	959	5.31
di np	9	8	0.17	9	0.01
di pi	10	20	2.55	21	3.45
di sp	17	9	3.19	8	3.64
di vm	9	7	0.45	7	0.34
di z0	20	7	7.42	13	1.81
dp aq	21	14	2.06	12	2.86
dp nc	192	169	4.09	172	2.15
dp np	6	5	0.20	5	0.14
dt nc	14	5	5.07	8	1.91
nc aq	1405	1216	36.39	1096	55.76
nc cc	228	220	1.64	232	0.08
nc cs	26	39	1.62	50	6.59
nc da	42	35	1.37	40	0.17
nc dd	9	6	0.88	1	6.76
nc di	24	16	2.35	27	0.07
nc nc	76	116	5.33	73	0.26
nc np	100	95	0.90	58	13.11

nc pi	2	0	2.18	6	1.83
nc pp	108	119	0.01	120	0.22
nc pr	208	257	1.87	308	15.48
nc rg	110	224	30.13	228	36.54
nc rn	31	47	2.08	40	0.79
nc sp	1873	2249	10.84	2228	19.00
nc va	39	87	14.52	85	15.18
nc vm	606	604	2.40	690	2.49
nc vs	74	122	8.09	130	13.11
nc z0	24	11	6.01	16	1.97
np aq	38	36	0.36	24	3.80
np cc	54	77	2.34	66	0.74
np cs	3	6	0.76	6	0.88
np da	4	6	0.25	8	1.17
np nc	33	56	4.17	5	21.89
np pp	13	9	1.11	13	0.01
np pr	13	15	0.02	15	0.07
np rg	7	14	1.78	16	3.15
np rn	6	6	0.02	6	0.01
np sp	68	107	5.74	102	5.43
np va	12	15	0.13	13	0.01
np vm	79	67	2.27	83	0.00
np vs	10	20	2.55	17	1.53
pd vm	9	5	1.51	10	0.02
pd vs	7	6	0.19	13	1.55
pi cc	9	7	0.45	8	0.11
pi pp	7	3	1.96	7	0.01
pi pr	6	4	0.59	5	0.14
pi rg	7	6	0.19	10	0.41
pi sp	46	68	2.60	82	8.66
pi va	3	7	1.28	5	0.42
pi vm	11	15	0.32	25	4.86
pp aq	11	10	0.17	10	0.10
pp cc	8	6	0.48	7	0.12
pp cs	12	5	3.51	2	7.59
pp da	15	5	5.89	3	8.53
pp nc	11	1	9.22	8	0.61
pp pp	11	3	5.28	11	0.01
pp rg	11	4	3.89	2	6.63
pp sp	68	42	8.58	37	10.57
pp va	80	60	4.82	64	2.55
pp vm	467	407	10.98	367	17.02
pr da	19	34	3.08	32	2.78
pr di	5	9	0.83	8	0.57
pr nc	15	1	13.49	2	10.52
pr pp	99	93	1.04	85	1.76
pr rg	14	18	0.22	19	0.56

pr rn	13	15	0.02	10	0.53
pr va	13	32	6.51	27	4.32
pr vm	277	281	0.70	305	0.41
pr vs	14	32	5.62	23	1.82
pt da	5	15	4.20	2	1.42
pt pp	1	3	0.84	7	4.25
pt va	1	2	0.25	7	4.25
pt vm	16	13	0.62	52	17.57
pt vs	2	2	0.01	6	1.83
rg aq	112	186	12.75	194	18.70
rg cc	12	12	0.04	22	2.53
rg cs	35	43	0.28	50	2.04
rg da	41	46	0.02	44	0.02
rg dd	5	2	1.55	10	1.46
rg di	28	35	0.30	46	3.64
rg nc	32	31	0.21	56	5.56
rg pi	2	5	1.05	10	5.00
rg pp	28	42	1.75	23	0.73
rg pt	1	2	0.25	7	4.25
rg rg	26	51	6.16	89	31.94
rg rn	9	14	0.71	7	0.34
rg sp	98	142	4.81	173	17.75
rg va	3	17	8.67	2	0.25
rg vm	102	167	10.77	187	21.61
rg vs	11	12	0.00	7	1.07
rg z0	10	22	3.55	27	7.10
rn aq	8	9	0.00	2	3.87
rn pp	32	18	5.21	22	2.31
rn rg	7	4	1.09	9	0.17
rn sp	8	4	1.70	4	1.51
rn va	12	13	0.00	4	4.36
rn vm	69	62	1.21	46	5.67
rn vs	17	34	4.33	28	2.24
sp ao	5	9	0.83	6	0.05
sp aq	83	87	0.06	57	6.04
sp cs	68	66	0.44	59	1.09
sp da	1640	1755	0.26	1003	193.50
sp dd	127	134	0.06	146	0.63
sp di	416	454	0.00	422	0.18
sp dp	113	90	4.94	79	7.62
sp nc	1008	1038	1.63	1784	195.37
sp np	196	215	0.01	264	7.40
sp pd	11	1	9.22	5	2.52
sp pi	38	21	6.46	29	1.63
sp pp	44	35	1.93	28	4.29
sp pr	30	14	7.27	12	8.52
sp pt	9	15	1.04	19	3.16

sp rg	20	45	7.64	51	12.25
sp sp	9	8	0.17	14	0.88
sp vm	365	426	0.98	341	2.22
sp vs	4	21	10.19	12	3.66
sp z0	59	101	7.80	63	0.02
va aq	1	11	7.52	6	3.36
va rg	4	11	2.71	12	3.66
va sp	14	9	1.55	13	0.09
va va	3	5	0.35	7	1.43
va vm	180	228	2.34	226	3.44
va vs	4	27	15.21	12	3.66
vm aq	84	53	9.93	52	9.01
vm cc	46	64	1.63	73	5.03
vm cs	143	204	6.24	202	7.76
vm da	477	456	4.00	338	30.65
vm dd	41	36	0.89	57	1.97
vm di	255	261	0.49	286	0.70
vm dp	52	41	2.41	49	0.27
vm nc	289	273	2.85	464	34.03
vm np	16	23	0.74	30	3.68
vm pd	1	2	0.25	8	5.15
vm pi	21	22	0.02	17	0.61
vm pp	167	78	40.58	83	32.16
vm pr	6	11	1.08	5	0.14
vm pt	7	9	0.11	18	4.38
vm rg	188	168	3.48	261	9.01
vm rn	4	4	0.01	8	1.17
vm sp	885	888	3.04	832	4.92
vm va	3	5	0.35	12	5.02
vm vm	288	277	2.18	372	7.48
vm vs	21	83	32.03	58	15.80
vm z0	13	12	0.17	16	0.19
vs aq	73	138	15.00	97	2.44
vs cs	7	18	3.96	13	1.55
vs da	51	48	0.52	28	7.74
vs di	20	42	6.08	44	8.01
vs nc	19	12	2.23	24	0.39
vs rg	22	93	38.20	87	36.10
vs sp	13	22	1.62	17	0.37
vs vm	12	92	55.18	83	50.16
z0 cc	14	15	0.00	15	0.00
z0 nc	74	82	0.01	106	4.41
z0 np	3	6	0.76	6	0.88
z0 sp	22	31	0.86	31	1.17
z0 vm	4	5	0.04	8	1.17

Trigramas

POS	Frecuencia TH	Frecuencia *Google*	X^2 TH-*Google*	Frecuencia *Lucy*	X^2 TH-*Lucy*
ao nc aq	6	7	0.01	7	0.04
ao nc sp	11	14	0.12	13	0.10
aq aq cc	8	10	0.06	4	1.49
aq aq sp	25	18	1.98	22	0.32
aq aq vm	7	4	1.16	6	0.12
aq cc aq	28	25	0.62	18	2.58
aq cc da	16	14	0.42	9	2.24
aq cc di	6	4	0.63	2	2.16
aq cc nc	10	12	0.03	18	1.99
aq cc pp	7	5	0.57	5	0.42
aq cc rg	8	6	0.53	6	0.37
aq cc sp	14	17	0.06	8	1.88
aq cc vm	23	26	0.00	29	0.48
aq cs da	19	23	0.08	22	0.12
aq cs di	2	9	3.78	5	1.17
aq cs nc	1	3	0.81	6	3.38
aq cs pp	4	2	0.89	6	0.33
aq cs vm	2	4	0.48	7	2.59
aq da nc	18	8	4.96	12	1.44
aq di nc	10	8	0.48	11	0.02
aq nc aq	45	44	0.35	20	10.62
aq nc cc	9	9	0.05	9	0.01
aq nc nc	1	6	3.09	2	0.30
aq nc pp	3	6	0.72	2	0.24
aq nc pr	9	11	0.05	12	0.32
aq nc rg	3	8	1.79	4	0.11
aq nc sp	120	149	0.87	92	4.88
aq nc vm	21	19	0.41	17	0.59
aq pp va	5	4	0.24	7	0.26
aq pp vm	33	31	0.44	28	0.63
aq pr pp	12	8	1.27	8	0.96
aq pr vm	46	38	1.81	38	1.10
aq rg aq	6	9	0.33	10	0.85
aq rg sp	1	8	4.76	5	2.52
aq rg vm	8	14	1.08	8	0.01
aq sp aq	9	7	0.50	3	3.24
aq sp cs	10	5	2.23	5	1.87
aq sp da	251	245	2.02	149	30.39
aq sp dd	9	15	0.95	16	1.70
aq sp di	45	59	0.73	50	0.11
aq sp dp	18	12	1.90	10	2.61
aq sp nc	116	90	6.54	146	2.39
aq sp np	14	13	0.21	24	2.26

aq sp rg	2	5	1.00	6	1.85
aq sp vm	69	81	0.12	71	0.00
aq va vm	7	15	2.15	14	2.07
aq vm cs	9	9	0.05	11	0.13
aq vm da	21	12	3.48	11	3.53
aq vm di	13	18	0.38	18	0.63
aq vm nc	17	15	0.42	27	1.91
aq vm rg	11	11	0.06	10	0.09
aq vm sp	48	28	7.55	36	2.21
aq vm vm	20	18	0.41	14	1.30
aq vm vs	3	8	1.79	5	0.43
aq vs aq	3	10	3.09	3	0.00
aq vs da	8	6	0.53	4	1.49
aq vs rg	4	11	2.60	9	1.74
aq vs vm	0	5	4.51	9	8.66
cc aq sp	6	7	0.01	6	0.00
cc da nc	79	131	8.15	79	0.06
cc da sp	6	4	0.63	3	1.12
cc dd nc	5	6	0.02	8	0.58
cc di nc	37	33	0.83	36	0.08
cc dp nc	13	18	0.38	18	0.63
cc nc aq	20	12	2.92	18	0.20
cc nc np	5	4	0.24	6	0.06
cc nc sp	28	21	1.85	39	1.41
cc nc vm	8	6	0.53	12	0.65
cc np sp	3	9	2.42	5	0.43
cc np vm	5	9	0.77	8	0.58
cc pi sp	5	6	0.02	6	0.06
cc pp va	7	2	3.32	3	1.76
cc pp vm	22	23	0.04	18	0.57
cc rg da	2	10	4.56	5	1.17
cc rg nc	7	11	0.53	12	1.13
cc rg rg	2	2	0.01	8	3.37
cc rg sp	8	7	0.21	9	0.03
cc rg vm	6	15	3.00	7	0.04
cc rn sp	6	3	1.34	1	3.77
cc rn vm	5	9	0.77	2	1.40
cc sp da	27	19	2.34	19	1.72
cc sp dd	1	1	0.01	7	4.27
cc sp di	8	11	0.22	8	0.01
cc sp nc	12	9	0.79	13	0.01
cc sp vm	7	11	0.53	5	0.42
cc vm aq	6	1	4.11	0	6.24
cc vm cs	3	4	0.06	6	0.89
cc vm da	20	22	0.00	13	1.77
cc vm di	13	9	1.20	14	0.01
cc vm dp	10	3	4.54	3	4.05

cc vm nc	15	10	1.58	23	1.39
cc vm pp	9	5	1.59	11	0.13
cc vm rg	7	8	0.00	7	0.01
cc vm sp	28	31	0.00	30	0.01
cc vm vm	7	13	1.24	17	3.79
cc z0 nc	6	3	1.34	0	6.24
cs da aq	4	5	0.03	8	1.18
cs da nc	155	227	7.27	191	2.51
cs da np	6	8	0.12	3	1.12
cs dd nc	12	20	1.27	17	0.68
cs di nc	36	61	4.15	62	5.95
cs dp nc	6	13	1.92	13	2.32
cs nc aq	2	4	0.48	7	2.59
cs nc cc	1	2	0.24	6	3.38
cs nc sp	2	17	10.38	19	13.13
cs np cc	2	7	2.29	6	1.85
cs pp vm	42	27	5.00	26	4.41
cs rg vm	4	5	0.03	7	0.71
cs rn vm	8	7	0.21	3	2.47
cs sp da	7	6	0.21	4	0.94
cs vm cs	2	6	1.62	7	2.59
cs vm da	18	9	4.01	12	1.44
cs vm di	17	7	5.27	19	0.05
cs vm nc	6	4	0.63	4	0.48
cs vm rg	5	4	0.24	6	0.06
cs vm sp	17	17	0.09	11	1.53
cs vm vm	10	8	0.48	7	0.65
cs vs aq	4	6	0.22	5	0.08
da ao nc	22	22	0.12	23	0.00
da aq nc	82	90	0.00	82	0.06
da aq sp	12	9	0.79	10	0.27
da di nc	10	7	0.88	10	0.01
da nc aq	605	584	5.83	335	90.39
da nc cc	103	108	0.16	59	13.76
da nc cs	8	9	0.00	11	0.37
da nc da	15	15	0.08	7	3.23
da nc di	11	7	1.35	5	2.49
da nc nc	36	49	0.88	23	3.39
da nc np	55	50	1.03	15	24.47
da nc pp	48	54	0.01	44	0.36
da nc pr	99	112	0.02	94	0.40
da nc rg	28	72	15.18	45	3.34
da nc rn	14	21	0.78	17	0.19
da nc sp	918	1123	5.19	805	12.95
da nc va	17	45	9.98	38	7.24
da nc vm	274	249	5.22	222	7.73
da nc vs	35	61	4.65	49	1.83

da nc z0	17	7	5.27	0	17.67
da np cc	6	9	0.33	4	0.48
da np nc	17	26	1.08	0	17.67
da np sp	18	28	1.27	24	0.64
da np vm	11	10	0.21	6	1.67
da pi sp	2	3	0.11	14	8.55
da pr da	12	18	0.67	17	0.68
da pr di	3	6	0.72	5	0.43
da pr pp	18	13	1.41	20	0.04
da pr vm	33	37	0.00	13	9.50
da rg aq	5	9	0.77	6	0.06
da sp da	10	7	0.88	2	5.65
da vm da	7	2	3.32	4	0.94
da vm nc	2	7	2.29	4	0.59
da vm sp	11	11	0.06	4	3.54
da vm vm	1	0	1.11	6	3.38
da z0 nc	11	6	2.03	0	11.43
dd aq nc	11	4	4.04	3	4.89
dd cs da	2	15	8.68	0	2.08
dd nc aq	24	27	0.00	30	0.46
dd nc cc	3	4	0.06	6	0.89
dd nc pp	16	22	0.43	24	1.31
dd nc rg	4	13	3.90	15	5.96
dd nc rn	5	8	0.42	4	0.15
dd nc sp	54	42	3.00	58	0.03
dd nc va	5	9	0.77	9	0.99
dd nc vm	65	61	0.87	79	0.88
dd nc vs	6	24	9.06	18	5.55
di aq nc	39	67	4.83	23	4.78
di da nc	31	32	0.08	35	0.11
di dd nc	1	1	0.01	6	3.38
di di nc	9	19	2.63	19	3.20
di nc aq	269	223	10.47	236	3.66
di nc cc	24	18	1.59	21	0.33
di nc cs	5	7	0.16	9	0.99
di nc nc	4	9	1.45	8	1.18
di nc np	10	7	0.88	10	0.01
di nc pp	8	9	0.00	10	0.15
di nc pr	27	55	6.94	53	7.50
di nc rg	32	63	7.22	50	3.30
di nc sp	238	300	2.23	313	7.63
di nc va	5	7	0.16	7	0.26
di nc vm	81	98	0.34	102	1.68
di nc vs	2	6	1.62	6	1.85
di pi sp	7	11	0.53	11	0.74
di sp da	9	4	2.48	3	3.24
di vm nc	7	5	0.57	0	7.28

di z0 nc	7	3	2.04	0	7.28
di z0 sp	8	2	4.25	0	8.32
dp aq nc	16	12	1.06	9	2.24
dp nc aq	37	23	4.88	21	5.06
dp nc cc	6	8	0.12	8	0.21
dp nc rg	0	3	2.71	7	6.74
dp nc sp	54	59	0.01	61	0.20
dp nc va	4	6	0.22	4	0.00
dp nc vm	17	21	0.11	24	0.94
dt nc sp	1	3	0.81	6	3.38
nc aq aq	85	72	2.84	70	2.09
nc aq cc	80	80	0.42	74	0.52
nc aq cs	13	13	0.07	18	0.63
nc aq da	12	6	2.67	6	2.24
nc aq di	6	6	0.03	6	0.00
nc aq nc	17	28	1.68	18	0.00
nc aq pp	35	29	1.35	30	0.60
nc aq pr	56	41	4.13	38	4.19
nc aq rg	19	45	8.10	35	4.15
nc aq rn	7	8	0.00	4	0.94
nc aq sp	408	353	11.87	326	12.82
nc aq va	7	17	3.21	14	2.07
nc aq vm	131	99	8.41	114	1.94
nc aq vs	19	43	7.02	30	2.07
nc cc aq	1	6	3.09	4	1.69
nc cc da	47	72	3.02	32	3.46
nc cc di	20	15	1.32	15	0.92
nc cc dp	7	2	3.32	1	4.74
nc cc nc	65	42	7.63	82	1.37
nc cc pi	6	6	0.03	3	1.12
nc cc pp	6	10	0.63	7	0.04
nc cc rg	3	8	1.79	15	7.55
nc cc sp	24	19	1.21	16	1.93
nc cc vm	37	27	2.77	37	0.03
nc cs da	6	12	1.44	12	1.78
nc cs nc	2	8	3.02	10	5.03
nc cs np	3	3	0.02	7	1.45
nc cs vm	4	4	0.02	6	0.33
nc da nc	36	30	1.34	35	0.08
nc dd nc	6	4	0.63	1	3.77
nc di nc	20	11	3.63	22	0.03
nc nc aq	10	15	0.55	8	0.31
nc nc rg	2	7	2.29	5	1.17
nc nc sp	14	28	3.35	20	0.84
nc nc vm	9	12	0.18	10	0.02
nc np aq	9	7	0.50	3	3.24
nc np cc	11	9	0.46	6	1.67

nc np sp	8	15	1.48	6	0.37
nc np vm	16	13	0.70	4	7.68
nc pp va	14	20	0.53	19	0.58
nc pp vm	92	96	0.17	96	0.00
nc pr da	6	9	0.33	11	1.29
nc pr pp	46	44	0.49	48	0.00
nc pr rg	3	8	1.79	8	2.09
nc pr rn	6	11	1.01	8	0.21
nc pr va	8	21	4.59	19	4.07
nc pr vm	121	140	0.12	191	13.23
nc pr vs	3	16	7.63	12	5.06
nc rg aq	56	80	2.14	75	2.08
nc rg cc	2	2	0.01	7	2.59
nc rg cs	2	6	1.62	5	1.17
nc rg pp	2	11	5.36	2	0.00
nc rg rg	6	9	0.33	16	4.17
nc rg sp	6	14	2.44	17	4.85
nc rg va	0	9	8.12	1	0.96
nc rg vm	25	58	10.00	68	18.31
nc rn aq	6	8	0.12	1	3.77
nc rn pp	5	9	0.77	7	0.26
nc rn va	3	6	0.72	3	0.00
nc rn vm	11	12	0.00	14	0.25
nc rn vs	3	12	4.53	11	4.27
nc sp aq	29	37	0.32	29	0.02
nc sp cs	27	25	0.42	21	1.00
nc sp da	703	809	0.56	411	90.68
nc sp dd	47	49	0.09	42	0.51
nc sp di	145	183	1.38	159	0.22
nc sp dp	32	40	0.26	25	1.15
nc sp nc	579	710	3.37	1160	180.09
nc sp np	137	150	0.01	179	4.12
nc sp pd	6	0	6.65	2	2.16
nc sp pi	22	11	4.90	11	4.11
nc sp pp	4	7	0.54	8	1.18
nc sp pr	15	7	3.80	5	5.40
nc sp pt	1	10	6.49	9	6.10
nc sp rg	4	14	4.59	17	7.56
nc sp sp	3	4	0.06	8	2.09
nc sp vm	102	141	2.92	115	0.36
nc sp z0	17	44	9.38	0	17.67
nc va aq	1	6	3.09	4	1.69
nc va rg	2	6	1.62	4	0.59
nc va sp	6	3	1.34	4	0.48
nc va vm	29	58	6.95	65	12.47
nc va vs	0	10	9.03	5	4.81
nc vm aq	11	6	2.03	8	0.60

nc vm cc	12	8	1.27	9	0.55
nc vm cs	18	43	7.87	46	11.22
nc vm da	57	46	2.59	47	1.39
nc vm di	33	42	0.36	51	3.20
nc vm dp	6	4	0.63	8	0.21
nc vm nc	30	40	0.59	65	11.61
nc vm np	4	8	0.96	8	1.18
nc vm pp	6	8	0.12	11	1.29
nc vm rg	28	31	0.00	24	0.48
nc vm sp	214	179	7.84	193	2.06
nc vm vm	82	78	0.94	112	3.57
nc vm vs	6	29	12.89	20	7.02
nc vm z0	3	7	1.22	0	3.12
nc vs aq	23	29	0.21	31	0.90
nc vs cs	2	7	2.29	5	1.17
nc vs da	13	10	0.76	8	1.39
nc vs di	5	11	1.68	12	2.62
nc vs nc	4	5	0.03	10	2.35
nc vs rg	5	29	14.62	31	17.81
nc vs sp	7	8	0.00	6	0.12
nc vs vm	5	16	4.70	20	8.44
np aq sp	7	4	1.16	3	1.76
np aq vm	4	6	0.22	4	0.00
np cc da	5	6	0.02	7	0.26
np cc dp	2	13	7.00	14	8.55
np cc nc	5	6	0.02	4	0.15
np cc np	25	34	0.61	24	0.08
np da nc	4	5	0.03	7	0.71
np nc sp	6	14	2.44	0	6.24
np nc vm	10	12	0.03	2	5.65
np pp vm	10	9	0.21	14	0.52
np pr vm	7	10	0.27	11	0.74
np rg vm	1	3	0.81	8	5.18
np sp da	22	39	3.17	18	0.57
np sp di	14	13	0.21	14	0.01
np sp dp	6	3	1.34	3	1.12
np sp nc	3	7	1.22	32	22.95
np sp np	13	25	2.67	22	1.98
np sp vm	2	8	3.02	6	1.85
np va vm	12	15	0.10	13	0.01
np vm cs	5	5	0.03	8	0.58
np vm da	7	7	0.04	2	2.98
np vm di	13	10	0.76	13	0.01
np vm nc	5	6	0.02	10	1.48
np vm sp	21	13	2.80	18	0.36
np vm vm	10	8	0.48	7	0.65
pi pp vm	7	2	3.32	5	0.42

pi sp da	33	39	0.07	42	0.76
pi sp dd	6	5	0.22	7	0.04
pi sp nc	0	3	2.71	12	11.55
pi sp pp	2	9	3.78	6	1.85
pi va vm	3	7	1.22	5	0.43
pi vm sp	3	6	0.72	5	0.43
pi vm vm	0	3	2.71	9	8.66
pp cs da	6	1	4.11	0	6.24
pp da nc	14	4	6.65	2	9.47
pp nc sp	6	0	6.65	3	1.12
pp pp vm	7	3	2.04	9	0.18
pp sp da	23	18	1.24	13	3.18
pp sp di	9	6	0.95	6	0.72
pp sp nc	17	8	4.24	10	2.10
pp va vm	79	58	5.77	62	2.77
pp vm aq	17	5	7.86	8	3.60
pp vm cc	11	12	0.00	8	0.60
pp vm cs	21	22	0.03	13	2.21
pp vm da	29	22	1.82	14	5.83
pp vm dd	7	2	3.32	2	2.98
pp vm di	29	17	4.50	2	24.59
pp vm dp	2	6	1.62	1	0.37
pp vm nc	24	12	5.34	18	1.11
pp vm rg	41	36	1.05	50	0.58
pp vm sp	198	173	5.29	118	23.64
pp vm vm	29	55	5.63	74	18.02
pr da nc	18	33	3.02	30	2.56
pr di nc	4	6	0.22	6	0.33
pr pp va	6	7	0.01	14	2.90
pr pp vm	90	83	1.47	68	3.99
pr rg vm	8	8	0.04	9	0.03
pr rn vm	3	6	0.72	3	0.00
pr va vm	10	20	2.39	19	2.46
pr va vs	1	7	3.92	3	0.92
pr vm cc	0	1	0.90	9	8.66
pr vm cs	5	15	4.04	10	1.48
pr vm da	83	74	1.86	40	16.79
pr vm dd	4	4	0.02	9	1.74
pr vm di	24	23	0.25	24	0.02
pr vm dp	6	4	0.63	6	0.00
pr vm nc	25	29	0.03	53	9.02
pr vm rg	10	10	0.05	16	1.16
pr vm sp	75	56	5.08	70	0.42
pr vm vm	27	34	0.25	34	0.56
pr vm vs	0	12	10.83	6	5.77
pr vs aq	5	14	3.40	8	0.58
pr vs rg	1	6	3.09	5	2.52

pr vs vm	0	6	5.42	4	3.85
pt da nc	5	13	2.79	2	1.40
pt vm da	7	1	5.15	19	5.09
pt vm sp	0	4	3.61	9	8.66
rg aq cc	2	7	2.29	10	5.03
rg aq cs	9	13	0.38	13	0.58
rg aq da	2	0	2.22	6	1.85
rg aq nc	5	10	1.20	7	0.26
rg aq sp	34	63	5.98	68	10.09
rg aq vm	8	10	0.06	14	1.41
rg cc rg	2	2	0.01	12	6.77
rg cc vm	7	5	0.57	6	0.12
rg cs da	12	17	0.43	20	1.71
rg cs vm	8	6	0.53	8	0.01
rg da nc	35	34	0.30	36	0.00
rg dd nc	5	2	1.62	8	0.58
rg di da	3	5	0.32	6	0.89
rg di nc	21	20	0.24	34	2.60
rg nc aq	10	4	3.23	11	0.02
rg nc sp	11	15	0.27	25	4.93
rg pi sp	0	3	2.71	6	5.77
rg pp va	4	8	0.96	2	0.75
rg pp vm	23	34	1.15	21	0.18
rg rg aq	6	8	0.12	11	1.29
rg rg cs	0	4	3.61	9	8.66
rg rg rg	0	1	0.90	8	7.70
rg rg sp	7	11	0.53	15	2.61
rg rg vm	2	5	1.00	13	7.65
rg rn vs	0	6	5.42	0	0.00
rg sp aq	2	6	1.62	2	0.00
rg sp cs	4	0	4.43	6	0.33
rg sp da	37	71	7.54	56	3.19
rg sp dd	3	5	0.32	8	2.09
rg sp di	16	15	0.22	15	0.08
rg sp nc	13	16	0.08	45	16.47
rg sp np	3	4	0.06	6	0.89
rg sp vm	7	8	0.00	14	2.07
rg sp z0	6	12	1.44	0	6.24
rg va vm	3	14	6.05	1	1.08
rg vm cs	3	16	7.63	17	9.28
rg vm da	9	22	4.21	18	2.67
rg vm di	6	16	3.59	20	7.02
rg vm nc	13	15	0.01	28	4.93
rg vm pp	4	3	0.26	9	1.74
rg vm rg	2	4	0.48	8	3.37
rg vm sp	22	35	1.79	32	1.49
rg vm vm	9	21	3.66	13	0.58

rg vm vs	2	6	1.62	2	0.00
rg vs aq	6	9	0.33	3	1.12
rg z0 nc	5	13	2.79	0	5.20
rn pp va	10	1	8.34	3	4.05
rn pp vm	22	17	1.26	17	0.85
rn va vm	11	7	1.35	3	4.89
rn vm da	3	6	0.72	4	0.11
rn vm di	4	7	0.54	9	1.74
rn vm nc	9	10	0.00	8	0.10
rn vm rg	10	1	8.34	2	5.65
rn vm sp	14	8	2.32	10	0.83
rn vm vm	14	16	0.01	7	2.61
rn vs aq	6	14	2.44	7	0.04
rn vs rg	3	7	1.22	12	5.06
sp ao nc	5	9	0.77	5	0.00
sp aq nc	67	56	2.45	40	7.92
sp cs da	25	40	2.10	29	0.16
sp cs di	5	8	0.42	10	1.48
sp cs pp	9	4	2.48	3	3.24
sp cs vm	17	2	13.47	6	5.70
sp da ao	16	12	1.06	16	0.01
sp da aq	55	48	1.47	38	3.81
sp da nc	1406	1515	0.62	799	202.20
sp da np	64	83	0.90	32	11.97
sp da pi	6	5	0.22	8	0.21
sp da pr	45	49	0.01	61	1.84
sp da rg	3	7	1.22	7	1.45
sp da vm	18	16	0.41	16	0.21
sp da z0	19	8	5.70	0	19.75
sp dd nc	119	126	0.13	140	1.00
sp di aq	17	29	2.03	12	1.07
sp di da	14	14	0.07	14	0.01
sp di di	7	10	0.27	12	1.13
sp di nc	340	376	0.00	361	0.08
sp di pi	2	6	1.62	6	1.85
sp di vm	6	2	2.44	2	2.16
sp di z0	15	5	6.10	0	15.59
sp dp aq	12	9	0.79	4	4.32
sp dp nc	98	78	4.82	71	5.44
sp nc aq	249	171	23.89	252	0.09
sp nc cc	53	49	0.84	81	4.84
sp nc cs	2	11	5.36	9	4.19
sp nc da	9	11	0.05	20	3.76
sp nc di	4	6	0.22	13	4.43
sp nc nc	16	28	2.17	22	0.73
sp nc np	10	17	1.17	15	0.82
sp nc pp	14	12	0.43	31	5.79

sp nc pr	30	54	4.64	98	33.67
sp nc rg	27	38	0.91	60	11.31
sp nc rn	4	8	0.96	12	3.70
sp nc sp	244	270	0.00	518	89.97
sp nc va	3	12	4.53	20	11.93
sp nc vm	87	103	0.21	172	24.89
sp nc vs	14	11	0.74	34	7.59
sp np aq	17	16	0.22	12	1.07
sp np cc	16	21	0.26	23	1.00
sp np nc	9	11	0.05	3	3.24
sp np pp	5	1	3.10	10	1.48
sp np pr	10	10	0.05	10	0.01
sp np rg	2	6	1.62	12	6.77
sp np sp	22	38	2.79	46	7.59
sp np va	3	6	0.72	7	1.45
sp np vm	30	25	1.12	45	2.46
sp np vs	3	7	1.22	8	2.09
sp pi sp	10	8	0.48	11	0.02
sp pp aq	8	7	0.21	3	2.47
sp pp vm	4	6	0.22	4	0.00
sp pr pp	12	4	4.88	1	9.74
sp pr vm	10	4	3.23	4	2.81
sp pt da	4	6	0.22	1	1.92
sp pt vm	4	2	0.89	12	3.70
sp rg di	2	5	1.00	7	2.59
sp rg sp	4	13	3.90	9	1.74
sp rg vm	1	2	0.24	8	5.18
sp vm aq	9	11	0.05	9	0.01
sp vm cc	7	20	5.01	12	1.13
sp vm cs	10	23	3.89	14	0.52
sp vm da	83	116	2.63	43	14.33
sp vm dd	14	13	0.21	18	0.36
sp vm di	31	48	2.13	27	0.45
sp vm dp	12	10	0.45	12	0.01
sp vm nc	58	54	0.85	85	4.13
sp vm pp	54	34	6.86	24	12.75
sp vm pt	3	4	0.06	6	0.89
sp vm rg	6	8	0.12	8	0.21
sp vm sp	49	56	0.02	48	0.08
sp vm vm	7	2	3.32	10	0.42
sp vs aq	0	8	7.22	3	2.89
sp vs vm	0	8	7.22	6	5.77
sp z0 cc	5	9	0.77	0	5.20
sp z0 nc	27	36	0.53	0	28.08
sp z0 sp	5	17	5.39	0	5.20
va aq sp	1	8	4.76	4	1.69
va rg vm	2	6	1.62	5	1.17

va sp nc	3	2	0.32	7	1.45
va sp vm	6	1	4.11	0	6.24
va vm aq	9	9	0.05	5	1.30
va vm cs	12	17	0.43	15	0.23
va vm da	22	27	0.13	23	0.00
va vm di	20	12	2.92	18	0.20
va vm nc	11	11	0.06	25	4.93
va vm rg	28	23	1.14	50	5.40
va vm sp	48	88	8.06	58	0.60
va vm vm	7	2	3.32	3	1.76
va vs rg	0	6	5.42	1	0.96
va vs vm	2	15	8.68	8	3.37
vm aq cc	7	3	2.04	3	1.76
vm aq da	7	0	7.76	2	2.98
vm aq nc	41	28	3.98	16	11.97
vm aq sp	10	10	0.05	9	0.10
vm aq vm	6	1	4.11	4	0.48
vm cc nc	3	4	0.06	7	1.45
vm cc sp	4	7	0.54	6	0.33
vm cc vm	27	34	0.25	37	1.20
vm cs cs	3	5	0.32	6	0.89
vm cs da	58	89	3.77	71	0.86
vm cs dd	5	6	0.02	6	0.06
vm cs di	15	24	1.26	23	1.39
vm cs nc	3	14	6.05	15	7.55
vm cs pp	5	3	0.73	11	2.03
vm cs rg	4	5	0.03	6	0.33
vm cs vm	24	23	0.25	31	0.64
vm cs vs	2	8	3.02	5	1.17
vm da ao	6	4	0.63	5	0.13
vm da aq	25	24	0.25	22	0.32
vm da nc	423	407	4.23	291	30.29
vm da np	13	16	0.08	6	2.86
vm dd nc	37	35	0.45	51	1.73
vm di aq	21	23	0.00	7	7.56
vm di da	5	5	0.03	7	0.26
vm di nc	213	218	0.69	249	1.61
vm dp nc	44	33	2.92	40	0.38
vm nc aq	78	67	2.36	116	6.08
vm nc cc	8	6	0.53	17	2.91
vm nc cs	4	2	0.89	9	1.74
vm nc np	8	4	1.78	4	1.49
vm nc pr	14	9	1.67	25	2.70
vm nc rg	6	10	0.63	27	12.58
vm nc sp	108	115	0.09	178	14.65
vm nc vm	10	8	0.48	19	2.46
vm nc vs	3	2	0.32	7	1.45

vm np sp	2	4	0.48	12	6.77
vm pi sp	12	15	0.10	10	0.27
vm pp cc	7	5	0.57	5	0.42
vm pp cs	12	4	4.88	1	9.74
vm pp da	11	2	7.21	0	11.43
vm pp nc	7	0	7.76	3	1.76
vm pp rg	10	3	4.54	1	7.72
vm pp sp	66	37	11.47	33	12.34
vm pp vm	17	14	0.68	22	0.46
vm pr vm	3	6	0.72	3	0.00
vm pt vm	4	2	0.89	12	3.70
vm rg aq	15	10	1.58	15	0.01
vm rg cc	5	7	0.16	6	0.06
vm rg cs	9	8	0.21	15	1.28
vm rg da	31	14	8.32	8	14.48
vm rg di	12	9	0.79	15	0.23
vm rg nc	19	9	4.69	28	1.40
vm rg rg	7	15	2.15	24	8.69
vm rg sp	50	58	0.06	90	9.98
vm rg vm	14	6	4.09	13	0.09
vm rg z0	6	6	0.03	0	6.24
vm sp aq	19	17	0.41	11	2.45
vm sp cs	12	15	0.10	16	0.43
vm sp da	362	370	1.22	195	57.50
vm sp dd	21	26	0.14	29	0.99
vm sp di	105	116	0.00	117	0.27
vm sp dp	26	20	1.53	20	1.03
vm sp nc	169	124	12.42	247	11.90
vm sp np	17	12	1.46	37	6.67
vm sp pi	3	5	0.32	10	3.51
vm sp pp	11	6	2.03	4	3.54
vm sp pt	7	1	5.15	2	2.98
vm sp rg	6	20	6.19	19	6.28
vm sp vm	108	128	0.27	103	0.39
vm sp vs	3	6	0.72	3	0.00
vm sp z0	11	13	0.02	0	11.43
vm va vm	2	4	0.48	9	4.19
vm vm aq	8	4	1.78	4	1.49
vm vm cc	5	4	0.24	8	0.58
vm vm cs	11	16	0.49	12	0.01
vm vm da	51	59	0.05	47	0.35
vm vm dd	4	4	0.02	7	0.71
vm vm di	18	26	0.75	30	2.56
vm vm dp	2	4	0.48	6	1.85
vm vm nc	35	29	1.35	72	11.44
vm vm pp	63	16	33.14	6	49.41
vm vm rg	13	8	1.76	46	17.24

vm vm sp	48	78	4.42	81	7.25
vm vm vm	11	5	2.92	19	1.84
vm vm vs	2	5	1.00	7	2.59
vm vs aq	8	18	2.90	12	0.65
vm vs di	1	7	3.92	8	5.18
vm vs rg	3	15	6.84	10	3.51
vm vs vm	0	28	25.28	16	15.40
vm z0 nc	7	10	0.27	0	7.28
vs aq cc	6	7	0.01	8	0.21
vs aq cs	11	13	0.02	13	0.10
vs aq sp	17	61	20.60	36	6.11
vs aq vm	7	6	0.21	6	0.12
vs da aq	8	5	1.04	3	2.47
vs da nc	34	37	0.01	21	3.60
vs di aq	1	7	3.92	4	1.69
vs di nc	18	33	3.02	35	4.83
vs nc aq	4	5	0.03	8	1.18
vs nc sp	8	7	0.21	11	0.37
vs rg aq	8	49	25.54	54	32.44
vs rg da	0	5	4.51	6	5.77
vs rg rg	2	6	1.62	9	4.19
vs rg vm	5	17	5.39	3	0.58
vs sp nc	5	5	0.03	10	1.48
vs vm rg	0	8	7.22	13	12.51
vs vm sp	4	57	40.94	47	34.67
z0 cc rg	4	8	0.96	0	4.16
z0 cc z0	7	4	1.16	0	7.28
z0 nc aq	10	11	0.00	0	10.39
z0 nc sp	22	23	0.04	0	22.88
z0 sp da	5	10	1.20	0	5.20
z0 sp nc	4	6	0.22	0	4.16
z0 sp z0	3	9	2.42	0	3.12