HACIA UNA PROPUESTA DE CLASIFICACIÓN DE UNIDADES DE TRADUCCIÓN

Patricia Fernández Carrelo
Universidad de Deusto

***Resumen**: El presente artículo aborda la problemática del análisis y tratamiento de las unidades plurilexemáticas desde una perspectiva multilingüe. Dichas unidades se interpretan como unidades de traducción clasificables en dos grupos: unidades abiertas o infinitas, sujetas a la creatividad de la lengua; y unidades cerradas o finitas, que no admiten flexión morfológica y pueden ser recogidas en repertorios léxicos con mayor facilidad.*

Esta propuesta tiene como objetivo optimizar los resultados obtenidos de las herramientas de Traducción Automática a través del uso de diccionarios o memorias de traducción que previamente hayan integrado la clasificación presentada.

***Palabras clave**: Unidades plurilexemáticas – Unidades de traducción – Traducción automática – Euskara*

***Abstract**: This article approaches the analisys and treatment of multiword expressions from a multilingual perspective. These are defined as translation units and are classified into two groups: opened or infinite units, in the sense that they are subject to language creativity; and closed or finite units, that have not morphological inflection and are better grouped as lexical collections.*

The porpouse of this proposal is to optimise Machine Translation performance, by means of dictionaries and translation memories that integrate our taxonomy of translation units.

***Keywords**: Multiword expressions – Translation units – Machine Translation - Euskara*

1. INTRODUCCIÓN

Al hablar de traducción, una de las cuestiones que más polémica suscita tanto a nivel práctico como teórico es el de la segmentación textual. Es una consideración por todos aceptada que a la hora de traducir un texto resulta humanamente inviable proceder de forma global, por ello, el traductor necesita de la división del texto en segmentos menores para acotar unidades de traducción y poder así proceder por partes (Abaitua 1997).

Si llevamos esta reflexión al plano computacional, esta tarea adquiere una complejidad aún mayor. Mientras que el traductor humano pone en juego sus saberes lingüísticos y conocimientos del mundo para llevar a cabo el proceso de segmentación, las herramientas automáticas se basan en formalizaciones lingüísticas que les permitan discriminar unidades de traducción de manera objetiva. Este hecho exige, por parte de la máquina, poseer criterios neutros y válidos para el establecimiento de estas unidades, lo que requiere a su vez de la lingüística computacional una formulación eficaz de dichos criterios.

Si bien la traducción de determinados elementos lingüísticos entraña una dificultad de grado bajo o medio, como ocurre con los términos de especialidad o las entradas léxicas monosémicas, gran cantidad de piezas léxicas de la lengua se insertan en expresiones mayores y establecen relaciones de dependencia lingüística que condicionan tanto su

Recibido el 30-11-06

RæL 5 (2006): 141-154
ISSN 1885-9089

procesamiento y análisis como su traducción. De hecho, tal y como afirma Jakendoff (1997), el número de este tipo de expresiones en el caudal léxico de los hablantes "is of the same order of magnitude as the number of single words", y así lo demuestran datos constatables como el número de entradas poliléxicas de WordNet 1.7 (Fellbaum 1999), que alcanza el 41% del total.

Así pues, a lo largo de este artículo, se llevará a cabo una propuesta de clasificación de unidades de traducción plurimembres con vistas a su aplicación en los procesos automáticos de traducción.

2. PROPUESTA TAXONÓMICA DE UNIDADES DE TRADUCCIÓN PLURIMEMBRES

Siguiendo el criterio de Newmark (1988) de considerar una escala móvil de unidades de traducción frente al concepto de unidad con forma fija[1], presentamos una taxonomía de estas unidades formadas por dos o más piezas léxicas simples.

Esta clasificación viene condicionada por la perspectiva de las lenguas con las que se pretende validar: el castellano, tomado inicialmente como lengua de partida, y el euskara, que ha motivado la inclusión en nuestra escala de determinadas estructuras específicas de esta lengua.

Además, se incluye una primera descripción/caracterización lingüística de las estructuras que conforman cada una de las unidades, acompañada de algunos ejemplos (tomados, cuando ha sido posible, de los corpus empleados para la realización de este artículo[2], o directamente de diccionarios[3]), y de una propuesta de análisis, recogida o procesamiento de cada unidad.

2.1. Las lexías complejas[4]

Denominamos de este modo aquellas unidades léxicas que no pueden ser traducidas de forma aislada, sino relacionadas con otra u otras piezas léxicas –entre las que se establecen relaciones de dependencia en función de la composicionalidad semántica, del grado de lexicalización y de la flexibilidad sintáctica- para cobrar el sentido adecuado en el traslado de una lengua a otra. Se corresponden con la denominación de "unidades fraseológicas" (UFS) y se caracterizan, en mayor o menor medida, por el rasgo de la idiomaticidad.

Los rasgos principales de las lexías complejas son su carácter polilexemático (terminología propuesta por las investigadoras de la Pompeu Fabra Cabré/Estopà/Lorente, 1996); la fijación o estabilidad de sus componentes y la imposibilidad de ser sustituidos y/o, en algunos casos, flexionados; y la idiomaticidad, esto es, su significado translaticio idiomático.

Tipológicamente pueden incluir desde el sintagma formado por dos piezas léxicas, siempre y cuando no se trate de combinaciones discursivas libres, hasta la oración compuesta.

Con vistas a su aplicación en repertorios léxicos y memorias de traducción, nuestra taxonomía distingue dos grandes grupos de lexías complejas: aquellas que son abiertas o infinitas, es decir, que se encuentran sujetas a la capacidad creativa de la lengua; y las finitas, que pueden ser recogidas en repertorios léxicos cerrados y, además, son inalterables

Recibido el 30-11-06

RæL 5 (2006): 141-154
ISSN 1885-9089

gramaticalmente.

2.1.1. Lexías complejas abiertas (no finitas, creativas)

Se incluyen en este grupo aquellas expresiones de cuyo uso no disponemos un listado completo cerrado y limitado, como las siguientes:

– *Compuestos nominales:*

Sintagmas nominales lexicalizados. Inalterables sintácticamente pero con flexión morfológica. En su mayoría admiten variación de número en su primer elemento, no obstante, algunas expresiones sólo aparecen en singular o en plural, como por ejemplo, "vacas flacas". Muchas de ellas se recogen ya en los diccionarios tradicionales, normalmente dentro de la acepción de la primera palabra del compuesto.

Ejs: viaje de bodas - eztei-bidaia
golpe de Estado - Eztatu-kolpe
retrato robot - erretratu robot

Categorialmente se componen de un núcleo sustantivo más un adyacente, que puede ser un sintagma preposicional (en castellano), un adjetivo u otro sustantivo en aposición.

La traducción del adyacente se puede realizar en euskara a través de alguna de las marcas del genitivo o mediante la aposición misma, en ocasiones marcada con el uso del guión.

No realizamos aquí un análisis pormenorizado de los tipos de compuestos nominales en castellano y vasco, para ello *vid*. Pérez *et al*. (eds.)(2004: 117-139).

Ejs: flora intestinal - *hesteetako flora* (recogido en el diccionario Elhuyar y clasificable también como término de especialidad)
salud intestinal - *hesteen osasuna* (no aparece en los diccionarios)

-Otros sintagmas nominales:
cultivos de bacterias - *bakterio-hazkuntzak*
raíz de achicoria – *txikoria-sustrai*
hábitos de alimentación - *elikabide-azturak*

– *Perífrasis verbales:*

Expresiones equivalentes a una forma verbal simple expresada a través de varias palabras. Las perífrasis, tanto en la lengua origen como en la meta, pueden ser traducidas por un único verbo o mediante otra perífrasis.

Los ejemplos que se presentan a continuación han sido extraídos de noticias de la Revista *Consumer*, es decir, no son entradas léxicas tomadas del diccionario, sino ejemplos de uso en los que resultan equivalentes estas expresiones verbales bitextuales:

Ejs: preocupar - *kezkagarri eduki* (lit. 'tener como preocupante')

crecer - *haziz etorri* (lit. 'venir creciendo')
favorecer - *-aren mesedetan (ihardun)* (lit. 'estar a favor de alguien')
relajarse - *utzi gerorako* (lit. 'dejar para después')
transmitir - *-en eroale izan* (lit. 'ser transmisor de')

Recibido el 30-11-06
RæL 5 (2006): 141-154
ISSN 1885-9089

(no) afectar - *eraginik (ez) eduki* (lit. '(no) tener influencia')

tienen la propiedad potencial – *dezaketen* (lit. 'pueden')
ser conscientes de - *barneratu* (lit. 'interiorizar')
realizar la selección de - *sailkatu* (lit. 'clasificar')

soler jugar - *ibili jolas horretan* (lit. 'andar en ese juego')

La mayor parte de estas perífrasis se componen de verbo + sintagma nominal (objeto), y están formadas por verbos delexicalizados -que serán analizados más adelante-; no obstante, conviene llevar a cabo un estudio más preciso de sus componentes gramaticales desde la perspectiva interlingüística.

– *Colocaciones:*

Expresiones que se caracterizan por estar sometidas a restricciones combinatorias léxicas determinadas por el uso. Responden al fenómeno de selección léxica y morfosintácticamente se pueden definir como sintagmas cuyos componentes han desarrollado ciertas preferencias de combinación, aunque mantienen gran libertad gramatical y sintáctica. Por lo general, no están recogidas en los diccionarios y poseen varias posibilidades de traducción.

A la hora de caracterizar estas unidades, es posible tener en cuenta distintos criterios.

Uno de ellos es el de la distancia colocacional. No existe una única teoría sobre este aspecto. Mientras que Jones & Sinclair (1974), como pioneros, y Sinclair (1991) independientemente en posteriores estudios barajan la cifra de 4 posiciones a la derecha o a la izquierda del núcleo como distancia máxima entre los colocados, o el proyecto COBUILD considera el tope en 5 unidades, Greenbaum (1988: 114) señala que los colocados no tienen por qué aparecer siquiera en una misma frase, pueden aparecer, incluso, en frases dichas por diferentes hablantes[5].

Otro aspecto es el grado de restricción de los colocados, que permite distinguir las colocaciones de las locuciones, pues según Cowie (1981, siguiendo a Mitchell 1971), la colocación permite al menos la sustitución de uno de sus componentes, sin que la expresión sufra alteración semántica.

En cuanto al nivel semántico, podemos caracterizar las colocaciones considerando que en ellas:

- se produce una especialización semántica que restringe las posibilidades de conmutación
- la expresión cobra un significado abstracto o figurativo
- su significado está casi gramaticalizado, como ocurre en las colocaciones de verbo delexicalizado

Además, como señala Corpas Pastor (1997: 83), "las bases suelen seleccionar acepciones secundarias, abstractas o figurativas de sus colocados. Por tanto, podemos decir que el significado de una colocación es parcialmente composicional" (también en Alonso Ramos 1993: 162, Heid 1994: 232).

Por último, en función de su composición morfológica podemos distinguir los siguientes tipos de colocaciones:

1-Sustantivo (sujeto) + verbo (que en castellano puede ser pronominal):
Ejs: estallar una guerra – *gerra hasi* (lit. 'empezar una guerra')
desatarse una tormenta – *tormenta hasi* (lit. 'empezar una tormenta')

2-Verbo + sustantivo (objeto):
Ejs: citar ejemplos – *adibideak eman* (lit. 'dar ejemplos')

3-Sustantivo + adjetivo (también en gradación). El adjetivo intensifica su base:
Ejs: lucrativo negocio - *negozio oparo* (lit. 'negocio abundante')
fotos paradisíacas - *argazki ederrak* (lit. 'fotos hermosas')
precio mayor - *prezio biziagoa* (lit. 'precio más vivo')
craso error - *akats nabarmena* (lit. 'error notable')

4-Sustantivo + preposición + sustantivo (en castellano) / sustantivo [guión] sustantivo (en euskara):
Ejs: diente de ajo – *baratxuri-ale*
rebanada de pan – *ogi xerra*
banco de peces – *arrain-sarda*

5-Verbo + adverbio:
Ejs: negar(se) rotundamente – *bipilki ezetz esan*

6-Adjetivo + adverbio:
Ejs: altamente cualificados – *biziki kualifikatuak*

Algunas de estas expresiones, como diente de ajo – *baratxuri ale* o negar(se) rotundamente – *bipilki ezetz esan* también aparecen en los diccionarios, por lo que podríamos considerarlas ya como lexicalizadas o gramaticalizadas y clasificarlas, por tanto, dentro del siguiente grupo.

Cualquiera de las combinaciones mencionadas puede pertenecer a la categoría de colocación en una de las lenguas pero no en la otra. En todo caso, siempre que se dé este fenómeno, la expresión debe ser clasificada como colocación dentro del repertorio léxico pertinente para poder disponer de la expresión adecuada a la hora de su traducción.

Por último, cabe destacar que las colocaciones admiten modificación, movimiento y extracción de sus elementos, siempre en función de la categoría morfológica de los mismos.

– *Locuciones:*

En la mayor parte de los casos, su traducción es única y de conjunto, pues difiere de la traducción de cada uno de los componentes de la combinación por separado debido a su significado no composicional. Se caracterizan por no respetar las reglas gramaticales generales: en ellas se producen restricciones morfosintácticas, imposibilidad de alterar el orden de los componentes o de insertar otros nuevos, etc. Se corresponden con los *idioms* de la tradición lingüística inglesa, que Collins (2000)

Recibido el 30-11-06
RæL 5 (2006): 141-154
ISSN 1885-9089

define como “a group of words which have a different meaning when used together from the one it would have if the meaning of each word were taken individually”.

Las clasificaciones más precisas realizadas para el castellano de estas unidades (v.g. Casares 1992; Corpas Pastor 1997, entre otros) se han basado tradicionalmente en la función oracional que desempeñan las locuciones en la oración, independientemente de que sean conmutables por palabras simples o por sintagmas.

Las locuciones, de indudable interés para la lexicografía, se incluyen en los diccionarios dentro de la acepción de alguno de sus componentes a modo de subentrada, y reciben, por tanto, un tratamiento similar al de las entradas propiamente dichas. No obstante, su inclusión depende del concepto de “lexicalización” adoptado por los responsables de la obra lexicográfica.

A continuación, presentamos una serie de ejemplos -extraídos directamente de los corpus ya citados- clasificados en función de criterios como la categoría gramatical a la que equivale la expresión o la tipología morfológica de los componentes de la misma. Como en este caso, su clasificación se ha realizado tomando el castellano como lengua de partida:

1-Locuciones verbales: verbo + complementos

Ejs: no dejar nada a la improvisación - *inprobisazioaren esku ezer ez uztea* (lit. ‘no dejar nada en mano de la improvisación’)
tener claro - *argi eta garbi deliberatu* (lit. ‘deliberar claro y limpio’)
tener poco que ver - *ia zerikusirik ez eduki* (lit. ‘no tener casi nada qué ver’)
quedarse en tierra - *herrian geratu* (lit. ‘quedarse en la tierra’)
hacer de algo su mejor aliado - *ez apartatu -tik* (lit. ‘no apartarse de’)
pasar factura (pagar caro) - *larrutik ordaindu behar izan* (lit. ‘tener que pagar desde la piel’)
vivir en las propias carnes - *larrutik ordaindu* (lit. ‘pagar desde la piel’)
dejarse conquistar por el estómago - *mahaira esertzea zure kirola izan* (lit. ‘ser tu deporte el sentarse a la mesa’)

2-Locuciones clausales: oración con todos sus constituyentes (sujeto + verbo + complementos)

Ejs: otros aprovechan para hacer su particular agosto - *besteak pagotxaren peskizan* (lit. ‘otros con esperanza de la ganga’)
el sol aprieta - *eguzkiak majo estutzen gaitu* (lit. ‘el sol nos aprieta estupendamente’)
las prisas y los agobios no son las mejores consejeras - *Presaka eta korrika ibiltzea ez da biderik hoberena* (lit. ‘andar de prisa y corriendo no es el mejor camino’)

3-Expresiones equivalentes a un sustantivo o a un adjetivo:

Ejs: lo que no pasa de ... – ... *hutsa baizik ez dena* (lit. ‘lo que no es sino un(a) simple...’)

un descalabro para su bolsillo - *sekulako hondamendia (...) zure poltsikoari* (lit. 'menudo daño (...) a tu bolsillo')

Como ocurre con las colocaciones, estas expresiones pueden ser locuciones en una lengua pero no en la otra; a pesar de ello, dentro de los repertorios léxicos deben ser clasificadas como locuciones para obtener una sistematización apropiada de su traducción.

Existen, además de los mencionados, otros criterios que sirven para la caracterización lingüística de estas expresiones. Entre ellos se encuentran el grado de institucionalización, la estabilidad sintáctico-semántica que se establece entre sus componentes (cohesión semántica y morfosintáctica), su significado composicional (como en *ir de mal en peor*) o no composicional (como en los ejemplos citados), la imposibilidad de sustitución o eliminación de un elemento por otro sin pérdida o distorsión del significado de la expresión, deficiencias transformativas en el plano gramatical (por ejemplo, la imposibilidad de pasivización) o determinados aspectos formales entre los que destacan peculiaridades fónicas tales como la presencia de aliteraciones, similicadencias o disposiciones rítmicas.

-*Verbos delexicalizados (verbos ligeros, verbos soporte, cópulas y auxiliares):*

Verbos sin carga semántica que participan o equivalen a expresiones más complejas en las que el aporte semántico corre a cargo de un sustantivo u otro componente de la expresión. Se utilizan también como verbos comodín en sustitución de un verbo especializado en determinados contextos. Es posible establecer grados dentro de la indefinición semántica de este tipo de verbos, no obstante, en esta ocasión los presentamos todos como un gran grupo en el que se establecen equivalencias interlingüísticas con verbos con significado pleno.

En castellano son buenos ejemplos de estos verbos: *ser, dar, tomar, hacer* o *poner*, en euskara, *izan* o *eduki*, y en inglés, *make, give, do* y *have.*

Son muy frecuentes en la traducción de determinados verbos poco comunes en la lengua meta.

Ejs: representan – *dira* (lit. 'son')
se hace – *da* (lit. 'es')
convertirse – *izan* (lit. 'ser')

Una base de datos de equivalencias entre verbos plenos y verbos delexicalizados podría resultar también de gran interés para el ejercicio automático de la traducción.

-*Verba dicendi* y otras partículas discursivas*:*

Estos verbos, que sirven para la introducción del estilo directo en el discurso indirecto, son frecuentes en noticias o textos científicos en los que se citan teorías de otros. Muchas veces adquieren un significado equivalente y resultan intercambiables. Así ocurre en el siguiente ejemplo:

Ejs: aseguran – *diote* (lit. 'dicen')

También se produce este fenómeno de equivalencia con las partículas discursivas del euskara que sirven para introducir palabras ajenas con mayor o menor participación

Recibido el 30-11-06
RæL 5 (2006): 141-154
ISSN 1885-9089

del narrador en la presentación de dichas palabras. Estas partículas son *ei* y *omen.*

-*Rección preposicional* nominal y verbal.

Estas construcciones, equivalentes en muchos casos a las posposiciones del euskara, determinan el uso de determinadas estructuras que son regidas por el sustantivo o adjetivo núcleo del sintagma en el que se insertan.

Ejs: *estudio* anual *sobre* la Seguridad - *Segurtasunaren gaineko urteko azterketa*
más de medio millar - *bostehunetik gora*
expertos en + sustantivo – *-t(z)en adituak*
encontrarse con – -(ar)ekin (aurrez aurre) topatu
englobados dentro de - tzat sailkaturiko
fermentada con - en bidez hartzituriko
acción sobre - -n daukaten eragina

– *Otras fórmulas y expresiones idiomáticas*:

Destacan también aquellas expresiones en las que la distancia lingüística entre el castellano y el euskara conlleva la necesidad de transformación de una estructura.

A continuación mostramos un esquema de estas fórmulas: expresiones formadas por más de una palabra gráfica, que participan del rasgo de la idiomaticidad y tienen una estructura lingüística diferente en las dos lenguas que analizamos.

Distinguimos, entre otros, los siguientes grupos, que presentamos ordenados por campos semánticos:

– *Números*

-*Horas*:

Ejs: *a eso de* las diez – *hamaikak aldera* (lit. ‘cerca de las diez’)

-*Fechas*:

Ejs: *Desde* 2000 – *2000tik hona* (lit. ‘de 2000 aquí’)
con dos semanas *de antelación* - *data baino bi astebete lehenago* (lit. ‘dos semanas antes de la fecha’)

-*Cantidades* (aproximaciones, conjeturas, etc.):

Ejs: De entre cinco o seis - *bospasei*

– *Tiempo atmosférico*

Ejs: va a llover – *euria dakar* (lit. ‘trae lluvia’)
había viento – *haizea zebilen* (lit. ‘andaba el viento’)

– *Sentimientos*

Ejs: Tengo frío - *Hotzak nago* (lit. ‘estoy el frío’)

Como ya se ha mencionado anteriormente, para lograr un tratamiento computacional óptimo de todas las unidades descritas, es necesaria la creación progresiva de repertorios

basados en corpus que respondan, por una parte, a las distintas tipologías textuales, y por otra, a los diferentes ámbitos de conocimiento. De este modo, se podrá recopilar un gran número de casos de uso de estas expresiones tan sujetas a la capacidad creativa del lenguaje que permita realizar con mayor precisión su traslado interlingüístico por medio de herramientas automáticas.

2.1.2. Lexías complejas cerradas (de repertorio finito)

Denominamos así a aquellas expresiones que pueden recogerse con mayor precisión que las anteriores en repertorios léxicos. Además, se caracterizan por carecer de flexión morfosintáctica.

En los procesos automáticos, la traducción de estas expresiones no tiene por qué estar sometida a las reglas gramaticales (flexión morfológica, estructura sintáctica), sino que es posible llevarla a cabo como una mera sustitución de elementos.

Se incluyen aquí:

- *Conectores/marcadores de discurso:*

Partículas que se insertan en el ámbito de la oración pero que apuntan más allá de ella, pues actúan como elementos que dan cohesión y coherencia al discurso.

En cuanto a su categoría estos conectores abarcan estructuras muy distintas: locuciones adverbiales, conjuntivas y preposicionales, sintagmas nominales o preposicionales, e incluso cláusulas enteras. Estas estructuras pueden ser además equivalentes a adverbios, a otras locuciones, a otros sintagmas o a las posposiciones del euskara. No obstante, todas ellas coinciden en su carácter invariable y en su casi total lexicalización. Es frecuente, asimismo, la acumulación de estas partículas: (en castellano) *pues bien*, *ni aun siquiera*, *o sea que*, etc.

La subclasificación que presentamos responde al aporte de contenido que ofrece cada conector (criterio semántico). Así, encontramos marcadores que sitúan el discurso en un tiempo o un lugar, otros que hablan del modo en que se realiza la acción del verbo, los que relacionan oraciones o cláusulas indicando que una es la causa, consecuencia o el impedimento para que se lleve a cabo la otra, y aquellos que, en un contexto menos definido dan información temática, introducen ejemplos, ratifican, enfatizan, añaden, etc., es decir, aportan cohesión textual dentro del contexto. Algunos ejemplos de los grupos mencionados son:

-Tiempo / lugar: en algún momento – *inoiz*
el año pasado – *iaz*
en la actualidad, - *egun*
cada año – *urterik urte*
a muchos kilómetros de distancia - *kilometro asko eta askotara*
de punta a punta - *alderik alde*

-Modales: a contrarreloj - *korrika bizian*

-Causales: Debido a que ... - *...nez,*

por lo que... - ... *nez,*
Dada esta situación, - *Gauzak horrela*
Esta es la razón que justifica que – *Horregatik*
Es por eso que – *Horrenbestez,*
Así las cosas, - *Horrenbestez,*

-Concesivos: cuando... - *...arren*
A pesar de los datos - *datuak datu*

-"Contextuales": con este término se designa a - hartzen dira halakotzat
En opinión de ... - *...-(ar)en aburuz,*
En materia de ... - *... kontuetan*
En este contexto - *Eszenatoki horretantxe*
*e*n casos concretos - *kasu zehatz-zehatzetan*

-Otros:

(*ejemplifican*) como... – *hala nola...* (+ enumeración)
por citar algunos ejemplos – *adibide pare bat ematearren*
(enumeran) etc. - *eta tankerako(ekin) / eta beste hamaikatxo(tan)*
(enfatizan) fundamentalmente – *batik bat*
precisamente, durante el verano - *udan bertan*
sobre todo – *batez ere*

especialmente – *batik bat*
(añaden) Pero hay más*: - Areago,*
a partir de - *kontuan hartuta*
(refuerzan) *t*ambién - *horrekin batera*
menos aún - *are gutxiago*

Un último consejo*:- Amaitzeko, beste bat*
ni siquiera así - *hala eta guztiz,*
*...*tan habituales en estas fechas, - *Data hauetan ohi-ohikoak diren ...*

Como en casos anteriores, un detenido análisis de estos elementos permitirá establecer subdivisiones más claras -que atiendan a la función o funciones que desempeña cada partícula en el texto- junto a equivalencias interlingüísticas estables y sistemáticas que faciliten la tarea traductológica computacional.

– *Fórmulas comunicativas:*

Propias de contextos conversacionales cuya aparición viene determinada en mayor o menor medida por situaciones comunicativas precisas. Son expresiones convencionales, ritualizadas y rutinarias según el contexto comunicativo y la lengua. Se pueden establecer también repertorios o glosarios de equivalencias entre ambas lenguas. Su número es limitado.

Las fórmulas de este tipo más comunes son: saludos, expresiones de agradecimiento y de petición de perdón, y sus respectivas respuestas.

Ejs: Bienvenido – *Ongi etorri*
A seguir bien – *Ondo izan*
Muchas gracias – *Eskerrik asko*
Perdón – Barkatu / No te preocupes, no importa – *Ez kezkatu, ez dio axola*

-*Paremias: proverbios, refranes...*

Enunciados completos en sí mismos. Se caracterizan por su carácter de inmutabilidad morfosintáctica.

Las *paremias*, siguiendo la terminología de Zuluaga (1980: 192), pueden también denominarse *enunciados fraseológicos*, y se caracterizan por funcionar "como secuencias autónomas del habla, su enunciación se lleva a cabo en unidades de entonación distintas; en otras palabras, son unidades de comunicación mínimas".

Se caracterizan lingüísticamente por los siguientes rasgos (Corpas Pastor 1997: 136):

-Lexicalización
-Autonomía sintáctica
-Autonomía textual
-Valor de verdad general
-Carácter anónimo

Esta autora distingue, además, entre enunciados de valor específico, citas y refranes. No entramos en la caracterización de cada una de estas subdivisiones por la escasez de ejemplos de este tipo encontrados en los corpus analizados.

Estas expresiones forman parte del grupo cerrado de lexías complejas ya que carecen de flexión morfológica, variación sintáctica, posibilidad de modificación interna o elisión de sus componentes, y pueden, como las anteriores, ser recogidas en repertorios, glosarios o bases de datos especializadas como formas inmutables (ya sea junto a la expresión idiomática equivalente en la lengua meta o junto a la paráfrasis creada por un determinado traductor en un determinado texto que sirva como equivalente para una determinada paremia).

-*Textos estereotipados o canónicos: títulos de películas, de obras literarias, etc.*

Tanto en el caso de los *títulos de libros o de películas* como en el de las *fórmulas de especialidad*, que mencionaremos a continuación, podemos hablar de la existencia de traducciones canónicas o institucionalizadas, una convención interlingüística que condiciona la traducción de la expresión y limita sus posibilidades.

En cuanto a los "títulos" institucionalizados pueden resultar esclarecedores los ejemplos cinematográficos que ofrece Abaitua (1997) o el mismo autor en su *Testuteka*[6]*:*

Ejs: *Monthy Pyton and the Holy Grail* - Los caballeros de la mesa cuadrada
Love and death - La última noche de Boris Grushenko
Play it again, Sam - Sueños de un seductor

Recibido el 30-11-06
RæL 5 (2006): 141-154
ISSN 1885-9089

-*Fórmulas rutinarias de especialidad:*

Del mismo modo, en el corpus de especialidad, corpus de Boletines Oficiales del País Vasco, observamos una serie de fórmulas que pueden ser consideradas como unidades de traducción por su repetida aparición en los documentos. Estas fórmulas abarcan desde sintagmas nominales simples (clasificables también como términos) hasta oraciones completas. En ocasiones, nos encontramos además con varias posibilidades de traducción, igualmente válidas para su traslado interlingüístico automático.

Ejs: Disposición Final - Azken Xedapena / Amaierako Xedapena

Criterios de selección y valoración de proyectos
Egitasmoak hautatzeko eta balioztatzeko irizpideak

El presente Decreto Foral entrará en vigor el día de su publicación en el Boletín Oficial del Territorio Histórico de Gipuzkoa.
Foru Dekretu hau Gipuzkoako Aldizkari Ofizialean argitaratzen den egunean jarriko da indarrean. / Foru Dekretu hau Gipuzkoako Aldizkari Ofizialean argitaratzen den egun berean jarriko da indarrean.

Tanto los textos canónicos como los de especialidad necesitan de grandes colecciones de equivalencias interlingüísticas institucionalizadas, pues es el único modo, tanto para humanos como para ordenadores, de tener un acceso directo a la traducción correcta de estas fórmulas tan precisas.

Muy cercanas a estos dos últimos tipos de expresiones y dentro del amplio grupo de lexías complejas, se ubican otras dos casuísticas específicas: la *onomástica*, en la que se incluyen topónimos, antropónimos, etc.; y la *terminología*, conjunto de piezas léxicas que vehiculan el conocimiento especializado.

Con las unidades terminológicas también pueden crearse repertorios en glosarios, diccionarios especializados y bases de datos terminológicas, distinguiendo entre unidades simples y compuestas; del mismo modo que con las entidades onomásticas. Estas últimas, además, carecen en ocasiones de posibilidad de traducción. No obstante, su detección resulta aún más compleja que la de otras lexías ya que, en ocasiones, ni siquiera los expertos humanos son capaces de ponerse de acuerdo en su reconocimiento y caracterización (fenómeno que se produce también con expresiones como las colocaciones o las locuciones).

En los corpus analizados, tanto terminología como onomástica son, por lo general, sintagmas nominales con patrones sencillos (núcleo + adyacente):

Ejs: Observatorio de la Seguridad - *Segurtasunaren Behatokia*
Guardia Civil - *Guardia Zibila*
Instituto de Estudios Turísticos - *Azterketa Turistikoen Institutua*
Centro de Vacunación Internacional - *Nazioarteko Txertapen Zentro*

En último lugar, cabe mencionar que no abordamos aquí otras equivalencias discursivas y lingüísticas que podrían ser consideradas como unidades de traducción. A saber, la

equivalencia entre “sustantivos” y “oraciones sustantivas” como “las discusiones - *eztabaidan aritzeak* (lit. ‘el entretenerse en la discusión’).

3. CONCLUSIONES

Como conclusión de este artículo cabe reiterar el *desideratum* metodológico vertido sobre estas páginas que tiene como objetivo optimizar la práctica de la traducción automática: la creación de recursos léxicos sistematizados que recojan todas las equivalencias descritas y, especialmente, las traducciones ya definidas o institucionalizadas de términos, locuciones, fórmulas, unidades onomásticas o incluso textos canónicos reconocidos por la tradición. De este modo, sólo sería necesario traducir lo no contenido en estos recursos, ya fuera por medio de reglas, de aproximaciones estadísticas o de métodos híbridos, y con la participación de los condicionantes gramaticales. Y es que no debemos olvidar que los ordenadores tienen más memoria que inteligencia, con lo que los resultados serían indiscutiblemente mejores.

Además, estos repertorios, ya sea en forma de bases de datos, tesauros, lexicones o diccionarios, deberán estar orientados no tanto al uso humano sino al tratamiento automático, es decir, deberán recibir un enfoque de *machine-readable technology*, con el fin de ser implementados posteriormente en herramientas de Traducción Automática.

Por ello, para lograr que este proceso se automatice y facilitar así la recopilación de este tipo de equivalencias, resulta imprescindible llevar a cabo dos tareas lingüísticas que validen y sirvan de referencia para refinar la taxonomía propuesta:

1. Encontrar más ejemplos en bitextos de referencia de cada tipo de unidad definida, para caracterizar y precisar las subclasificaciones establecidas por el momento.

2. Medir la frecuencia de aparición de cada tipo de unidad para valorar el interés/relevancia de su tratamiento a la hora de traducir los distintos géneros textuales.

1. Encontramos una completa revisión del concepto de “unidad de traducción” en Hurtado Albir (2001).

2. Incluimos ya una primera referencia a estos corpus, que responden a distintos géneros textuales:
- Jurídico: Corpus *Boletines*. Extensa colección de Boletines Oficiales del País Vasco (1994-2004).
- Técnico: Corpus *Consumer*. Noticias traducidas del castellano al euskara tomadas de la Revista Consumer (http://revista.consumer.es).
- Literario: Los dos primeros capítulos de la obra *SP-rako Tranbia / Tranvía para SP* de Unai Elorriaga.

3. En concreto, los diccionarios consultados han sido los de Euskaltzaindia (monolingüe) y Elhuyar (bilingüe) en su versión disponible a través de la web.

4. Denominación acuñada modernamente por B. Pottier (1972: 55).

5. Como vemos en Corpas Pastor 1997:79: “Berry-Rogghe (1973) también aplica la distancia de 4 posiciones, aunque la restringe a 2 en el caso de los adjetivos. Haskel (1972) aplica una distancia de +/-3; Al-Madi (1986) de +/-1, pero extensible a 5 ó 6 según los casos; Smadja (1989) se decanta por 5 posiciones hacia la izquierda. Posteriormente, Miall (1992) ha estimado que la distancia de cinco unidades a partir del núcleo es insuficiente para estudiar las colocaciones de la prosa de Coleridge, por lo que la ha ampliado a 15, aunque la distancia colocacional media empleada es de 9 unidades.”

6. http://paginaspersonales.deusto.es/abaitua/deli/testuteka/index.html

Recibido el 30-11-06

RæL 5 (2006): 141-154
ISSN 1885-9089

BIBLIOGRAFÍA

Abaitua, J. 1997. “Segmentos y unidades de traducción”. [Documento de Internet disponible en http://paginaspersonales.deusto.es/abaitua/konzeptu/l2uutts.htm]

Alonso Ramos, M. (1993): *Las funciones léxicas en el modelo lexicográfico de I. Mel'čuk*, Tesis Doctoral, Universidad Nacional de Educación a Distancia, Madrid.

Cabré, M.T., Estopà, R., Lorente, M. 1996. “Terminología y fraseología”. Actas del V Simposio de Terminología Iberoamericana. México D. F.

Casares, J. 1992. Introducción a la lexicografía moderna. Madrid: CSIC.

Corpas Pastor, G. 1997. Manual de Fraseología Española. Madrid: Gredos.

Collins. 2000. Collins Cobuild Dictionary of Idioms. Harper Collins Publishers.

Cowie, A.P. (1981): “The treatment of Collocations and Idioms in Learne's Dictionaries”, *Applied Linguistics* 2 (3): 223-235.

Fellbaum C. 1999. WordNet: An electronic Lexical Database. Cambridge, Massachusetts. London: The MIT Press.

Greenbaum, S. 1988. Good English and the Grammarian. Londres: Longman.

Heid, U. 1994. “On Ways Words Work Together – Topics in Lexical Combinatorics”, Martin, W. *et al.,* eds. Euralex 1994. Proceedings. Papers submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, Amsterdam. 226-257.

Hurtado Albir, A. 2001. Traducción y traductología. Introducción a la traductología. Madrid: Cátedra.

Jakendoff, R. 1997. The Architecture of the Language Faculty. Cambridge: MA MIT Press.

Jones, S. & Sinclair, S. 1974. “English Lexical Collocations. A Study in Computational Linguistics”. *Cashiers de Lexicologie* 24: 15-61.

Mitchell, T.F. 1971. “Linguistic 'goings on': collocations and other lexical matters arising on the syntagmatic record”. *Archivuum Linguisticum* 2: 35-39.

Newmark, P. 1988. A textbook of translation. London: Prentice Hal.

Pérez Gaztelu, E., Zabala, I., Gràcia, L., eds. 2004. Las fronteras de la composición en lenguas románicas y en vasco, San Sebastián, Universidad de Deusto.

Pottier, B. 1972. Presentación de la lingüística. Madrid.

Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford-New York: Oxford University Press.

Zuluaga, A. 1980. Introducción al estudio de las expresiones fijas, “Studia Romanica et Linguistica”, 10, Francfort-Berna-Cirencester, Peter D. Lang.

Diccionarios y repositorios

- Diccionario de la RAE, *Diccionario de la Lengua Española*, 2001, 21ª edición actualizada *on-line,* http://www.rae.es
- Elhuyar, 2000. Euskal Hiztegi Modernoa, versión on-line, http://www.hiztegia.net
- Euskaltzaindia, 2000. *Hiztegi Batua*, versión on-line, http://www.hiztegia.net
- Diccionario *on-line WordReference*, http://www.wordreference.com/es/
- *Testuteka*. Grupo DELi, diciembre 2001: http://paginaspersonales.deusto.es/abaitua/deli/testuteka/index.html
- *Revista Consumer*: http://revista.consumer.es/
- *Boletines Oficiales del País Vasco*: http://www.euskadi.net
- Respositorio LegeBi de DELiWeb http://www.deli.deusto.es/Resources/BOPV

CLUVI (Corpus Lingüístico da Universidade de