>>> revista tradumàtica
Traducció i Technologies de la Informació i la Comunicació
Número 2 : La Documentació : Novembre 2003 : http://www.fti.uab.es/tradumatica/revista :
ISSN 1578-7559

És la web pública la nova biblioteca del traductor?

Pilar Sánchez-Gijón

Departament de Traducció i d'Interpretació de la UAB

Resumen:

La World Wide Web se ha convertido en una de las fuentes de información más utilizadas por los traductores. Sin embargo, la información de la WWW puede presentar ciertas carencia en lo que a la calidad se refiere que el traductor debe tener en cuenta. Por este motivo, y dado que la información ya se encuentra en formato digital, las herramientas de análisis de corpus se convierten en un instrumento de gran utilidad en la obtención de información lingüística y factual para la realización de una traducción, y en concreto de la traducción especializada.

Palabras clave:

Traducción especializada, terminología aplicada a la traducción, lingüística de corpus, extracción de conocimiento, documentación aplicada a la traducción.

Resum:

La World Wide Web s'ha convertit en una de les fonts d'informació més utilitzades pels traductors. Tanmateix, la informació que s'hi troba pot presentar certes mancances de qualitat que el traductor ha de tenir en compte. Per aquest motiu, i pel fet que la informació ja es troba en format digital, l'aplicació d'eines d'anàlisi de corpus es revela com un instrument de gran utilitat a l'hora d'obtenir informació lingüística i factual per a la realització d'una traducció, i en concret la traducció especialitzada.

Paraules clau:

Traducció especialitzada, terminologia aplicada a la traducció, lingüística de corpus, extracció de coneixement, documentació aplicada a la traducció.

Introducció

La publicació d'informació en format digital a Internet, i més concretament a la World Wide Web, s'ha convertit en una de les formes de comunicació més immediates i assequibles de la Societat de la Informació, tant per a qui emet la informació com per a qui la consulta. Aquest mitjà de comunicació permet la publicació d'informació sobre qualsevol àmbit temàtic, alhora que li dóna un tractament de molt diversa mena; en altres paraules, la informació no es veu limitada per les condicions formals o de distribució del mitjà, ja que aquest inclou qualsevol morfologia informativa i arriba virtualment arreu del planeta i, per tant, és susceptible de ser consultada per qualsevol tipus de lector. Per tot això la WWW sembla erigir-se com la font d'informació ideal del traductor.

Les necessitats informatives del traductor acostumen a ser o bé de caire lingüístic o bé de caire factual (Mayoral 1997/1998), i el procés de documentació que seguirà tindrà per objectiu cobrir una d'aquestes necessitats, si no totes dues. Per necessitat informativa de caràcter lingüístic entenem qualsevol aspecte d'aquest nivell que el traductor no pugui resoldre amb els seus coneixements, ja sigui en la llengua de partida o en la d'arribada. Aquestes mancances de coneixement poden resoldre's consultant una obra de referència lexicogràfica o terminològica,

sobretot si es tracta d'un problema a nivell microestructural (relatiu a un problema fonamentalment lèxic o d'expressió), o recorrent a textos paral·lels si es tracta d'un problema macroestructural (relatiu al text com a unitat). Per necessitat informativa de caràcter factual entenem qualsevol mancança de coneixement relativa al tema del qual tracta el text original que el traductor ha de traslladar a la llengua d'arribada. En qualsevol dels dos casos, i si les obres de consulta tradicionals no li proporcionen la informació que necessita, el traductor recorre a la consulta de textos per tal d'obtenir-la.

En aquest context, quin paper poden jugar els textos publicats a la WWW?

A l'hora de documentar-se, els textos publicats en forma de document web d'accés públic suposen un recurs que els traductors tenen molt en compte avui dia. En el moment de cercar fonts d'informació, lingüística o textual, les pàgines web, però, presenten pràcticament tants avantatges com inconvenients respecte de la documentació a partir de textos en format paper.

Els inconvenients de recórrer a la WWW durant la fase de documentació del traductor són fonamentalment els següents: el caràcter desorganitzat de la WWW, la manca de fiabilitat dels recursos que s'hi extrauen i les diferències pragmàtiques o comunicatives.

La WWW és un entorn desorganitzat. Per accedir a un web cal conèixer la seva adreça URL, però si el traductor no sap de l'existència d'un web determinat, o no en coneix l'adreça, ha de recórrer a un cercador i confiar que aquest tindrà a la seva base de dades la informació que necessita. L'èxit d'una cerca, si per èxit entenem l'obtenció del document precís que cobrirà la necessitat informativa del traductor, dependrà fonamentalment dels següents factors: a) que el traductor formuli correctament la consulta; b) que el cercador li pugui proporcionar l'accés al document que necessita; i c) que el document que pot resoldre la necessitat informativa del traductor estigui publicat a la WWW. Val a dir, però, que generalment són els dos primers factors els que acostumen a dificultar la cerca i obtenció d'informació en forma textual a la WWW.

La WWW és un entorn poc fiable. Sovint, els textos que s'hi publiquen no passen per cap mena de procés de revisió, per la qual cosa no és difícil trobar-hi errors o imprecisions. En concret, pel que fa al traductor, aquesta manca de fiabilitat repercuteix sobretot en dos aspectes: a) la qualitat lingüística del document, i b) la qualitat de la informació que conté. Si el traductor es documenta per resoldre un problema de caràcter lingüístic, com pot ser l'adequació d'un terme o d'una expressió en un context determinat, necessita saber si el text que consulta és un exemple fiable de l'ús d'aquella llengua. D'altra banda, si el seu problema és de caràcter factual, haurà de tenir en compte si l'autor del recurs està autoritzat per tractar aquell tema. Per tal d'evitar problemes consultant un document poc fiable, el traductor intenta identificar els recursos que li ofereixin les màximes garanties de qualitat (com ara que hi consti l'autor o que el responsable secundari sigui una institució reconeguda); aquest fet encara dificulta més la fase d'identificació de recursos textuals mitjançant cercadors.

Tanmateix, si bé els dos factors anteriors, relatius al desordre de la WWW i a la manca de fiabilitat dels seus recursos, es poden eludir mitjançant un procés de documentació escrupulós, el tercer dels inconvenients, el relatiu a la situació pragmàtica dels recursos, resulta més difícil d'evitar. A l'hora de documentar-se a partir de textos paral·lels, i per tal que la informació que proporcionin pugui ser directament aplicable al text final que el traductor està elaborant, cal que les seves condicions comunicatives i pragmàtiques, tant les del text que es consulta com les del text de partida i d'arribada de la traducció, siguin equiparables. En cas que no ho siguin, el traductor pot adoptar solucions errònies degut a que corresponen a registres diferents o a que no respecten els convencionalismes del gènere al qual pertany el text que està traduint. Malauradament, però, el mitjà i les condicions de difusió d'un web no permeten observar amb claredat els factors relatius a la seva situació comunicativa, o almenys no de la mateixa manera que el traductor ho pot fer en el cas dels textos publicats en format paper (articles de diari,

revistes especialitzades, manuals, etc.), ja que aquests darrers es poden englobar en gèneres o categories de textos, la situació comunicativa dels quals queda implícita. El fet que el text d'un web sigui susceptible de ser consultat per tota mena de lectors, i no per un lector prototípic (tal com passa amb els textos tradicionals), pot provocar una alteració en les decisions que l'autor pren durant la seva fase d'elaboració. A més, el mitjà també provoca canvis que el traductor ha de tenir en compte, tant sobre el procés d'elaboració com sobre el text resultant. Per tot això, el text digital publicat al web, tret d'excepcions, deixa de tenir una situació comunicativa i pragmàtica totalment equiparable a la d'un text en format paper.

Malgrat aquests inconvenients, uns més fàcils d'eludir que altres, el gran avantatge del procés de documentació a partir de textos extrets de la WWW és fonamentalment la immediatesa del mitjà i la possibilitat de trobar-hi virtualment qualsevol contingut temàtic. Aquests dos aspectes fan del WWW una de les principals fonts d'obtenció d'informació en forma textual per als traductors, ja que hi poden accedir des del seu lloc de treball amb molta facilitat i, per regla general, hi trobaran informació més actualitzada que no pas a les obres de consulta tradicionals.

Com eludir els inconvenients que els textos de la WWW presenten durant la fase de documentació del traductor?

Llevat pel que fa al seu marc comunicatiu, aspecte que els traductors han de tenir en compte a l'hora de resoldre certs problemes de caràcter principalment lingüístic, les mancances més rellevants dels textos digitals publicats a la WWW presentades com a inconvenients en l'apartat anterior concerneixen a cada text per separat, de manera individual. Si, pel contrari, el traductor realitza un apropament a un grup de textos alhora, aspectes com la manca de fiabilitat de cada text es dilueixen en el conjunt. A més, la possibilitat de consultar un gran conjunt de textos simultàniament no obliga el traductor a identificar el recurs que s'adeqüi perfectament a les seves necessitats informatives, per la qual cosa la fase d'identificació i obtenció de textos es simplifica.

De quina manera, doncs, pot accedir el traductor a la informació d'un conjunt de textos alhora? La resposta a aquesta pregunta s'ha d'anar a buscar a una branca de la lingüística aplicada, en concret a la lingüística de corpus, que proposa una metodologia d'anàlisi lingüística de caràcter quantitativa. Bàsicament, la lingüística de corpus permet extraure i observar els contextos d'una paraula o expressió determinada al llarg de tots els textos que componen el corpus mitjançant eines informàtiques dissenyades per aquest propòsit. D'aquesta manera, cada consulta rep com a resposta la informació obtinguda del conjunt de textos, i no d'un text en concret, per la qual cosa la informació queda validada de manera estadística.

Una de las propostes de la lingüística de corpus que amb més facilitat es pot adoptar en la pràctica de la traducció professional, i també en la didàctica de la traducció, és la dels corpus *ad hoc* o corpus especials. Un corpus és un conjunt de textos, en format digital per tal de poder ser processat amb eines informàtiques, representatius de l'estat d'una llengua o d'una varietat de llengua. S'anomena corpus *ad hoc* aquell conjunt de textos, generalment de dimensions reduïdes, compilat amb l'objectiu de recollir informació molt concreta, com ara informació lingüística i/o factual per a la presa de decisions en el marc d'una traducció.

Figura 1: Interfície de consulta de WebCorp

Algunes eines informàtiques, com ara WebCorp, prenen la WWW pública com a corpus susceptible de ser analitzat en línia, sense que calgui compilar un corpus. WebCorp cerca una paraula o expressió mitjançant un cercador i presenta els resultats en forma de llistat de concordances o *Key Word in Context* (KWiC).

Figura 2: Resultats de la cerca de contextos de meteor storm.

Amb aquesta mena d'eines, el resultat s'obté de manera immediata. Això no obstant, no sempre resulten ser l'instrument adequat per al traductor, ja que aquest desconeix els tipus i la quantitat de textos que està consultant. A més, les consultes són poc flexibles i difícils de manipular.

Els instruments necessaris per compilar un corpus *ad hoc* a partir del WWW estan a l'abast de qualsevol traductor. Tot el que necessita és un instrument que li permeti trobar les pàgines web que formaran part del seu corpus i descarregar-les al seu disc dur. Val la pena assenyalar en aquest punt que aquests textos no seran manipulats ni es tornaran a publicar amb una altra forma, sinó que únicament es consultaran de manera local, per la qual cosa no es transgredeix cap dret de propietat intel·lectual.

Les estratègies per compilar un corpus *ad hoc* dependran de les necessitats informatives del traductor, o en altres paraules, de la informació que en vulgui obtenir. Si la informació que cerca és de caràcter factual (relativa al tema que es tracta en el text que està traduint), la compilació del corpus *ad hoc* es basarà en una cerca a partir de les paraules clau del tema, i es podrà dur a terme en qualsevol de les llengües que el traductor utilitzi (la de partida, la d'arribada o fins i tot qualsevol altra que li permeti obtenir el coneixement factual necessari per dur a terme la seva traducció). Tanmateix, sempre li resultarà més pràctic documentar-se en les llengües implicades en la traducció, ja que d'aquesta manera adquirirà informació lingüística lligada a la informació factual que consulta o validarà la que ja té. Per tal de treure tot el fruit de les anàlisis que posteriorment en farà, els corpus *ad hoc* recollits amb aquest propòsit seran monolingües; no obstant això, si el traductor es documenta en més d'una llengua, per exemple en la llengua de partida i en la llengua d'arribada, recolliria dos corpus *ad hoc* monolingües i comparables entre ells pel que fa al tema. En qualsevol cas, aquesta mena de corpus estaran formats per text pur, sense etiquetar sintàcticament o morfològicament, ja que el procés de marcatge requereix una gran inversió de temps i faria aquest recurs poc rentable per al traductor. No hem d'oblidar que el traductor ha de realitzar la seva tasca en les millors condicions possibles dins d'un marge de temps limitat.

Si, pel contrari, la necessitat informativa del traductor és de caràcter lingüístic, haurà de prendre més precaucions a l'hora de compilar el seu corpus. En aquest cas, el corpus haurà d'estar format per textos digitals que coincideixin amb les característiques, formals o pragmàtiques, de la informació que el traductor està buscant. Així, si necessita conèixer amb més profunditat aspectes macroestructurals relatius a un gènere o a un tipus de text, la consulta de textos que pertanyen a altres categories podria fer-lo prendre decisions errònies. D'altra banda, però, si el que busca és resoldre possibles problemes de traducció a nivell microestructural (com ara problemes de caire fraseològic), la documentació a partir de textos de diferents gèneres no resultarà tan perjudicial com en el cas anterior. Per tot això, cal que el traductor aprengui a reconèixer els diferents tipus de llocs web i el tractament que donen a la informació que contenen.

Com s'utilitza un corpus *ad hoc*?

L'explotació del corpus compilat es durà a terme amb l'ajuda d'un programa d'anàlisi de corpus1 i obeirà a les necessitats informatives del traductor. Així, si el traductor ha de resoldre un problema de traducció determinat, farà una consulta puntual sobre el corpus per tal d'obtenir la informació que li permeti resoldre'l. Aquesta mena de problemes acostumen a ser de caire lingüístic en primer o darrer terme. D'altra banda, si el traductor s'ha de documentar perquè desconeix el tema de la traducció que ha de fer, el tipus de consulta que efectuarà serà de caràcter sistemàtic per tal d'obtenir així informació sobre aquell tema de manera estructurada.

Les consultes que es facin sobre el corpus amb la finalitat d'obtenir informació relativa a una paraula o expressió determinada es poden concebre de manera oberta o tancada. Les consultes obertes són aquelles en què s'observen tots els contextos en els quals apareix la paraula o expressió que es consulta. Les consultes tancades o limitades són aquelles que es

basen en l'extracció de certs contextos de l'expressió que es consulta en funció de criteris addicionals.

A l'hora, per exemple, d'observar l'ús que es fa d'una paraula en context, si aquesta acostuma a formar part d'una col·locació determinada o si coapareix habitualment amb altres unitats lingüístiques, realitzarem una cerca oberta que proporcionarà com a primer resultat una llista amb els contextos de la paraula que hem cercat.(2)

Figura 3. Concordances o KWiC de meteor storm.

La reordenació alfabètica d'aquests contextos en funció de les paraules que estan a dreta i esquerre de la paraula o expressió que hem cercat ens permetrà comprovar visualment qualsevol patró lèxic repetitiu del qual pugui formar part.

Figura 4: Reordenació de les KWiC de meteor storm a partir de la primera paraula a l'esquerre (marcada en vermell).

A més d'utilitzar les KWiC d'una paraula o expressió per observar els seus contextos, o si aquests no proporcionen prou informació, una altra manera d'aprofundir en les característiques del cotext de l'expressió que cerquem consisteix a extraure un llistat de cosituats, és a dir, de les paraules que més habitualment es troben en el cotext de la paraula cercada i la posició que hi ocupen.

Figura 5: Llistat amb els primers 40 cosituats de meteor storm.

Aquest llistat ens permetrà intuir repeticions lèxiques que puguin conduir a col·locacions o termes compostos, a més de comprovar si en el cotext sobresurt cap tret semàntic significatiu.

Si l'objectiu, però, és el d'obtenir informació de manera estructurada, i generalment de caire factual, les consultes intentaran extraure únicament aquells contextos susceptibles de contenir la dada que es cerca. És en aquests casos quan el traductor efectuarà consultes tancades o limitades. Donat que el corpus que es consulta no està etiquetat, el filtratge dels contextos s'haurà de fer en funció de la presència o absència d'una paraula o expressió determinades en el cotext de la paraula sobre la qual es fa la consulta. Aquestes expressions permetran identificar contextos que continguin una informació determinada sobre la paraula que es consulta. Així doncs, per exemple, si el que es cerca és una definició en context d'un terme determinat, la consulta es farà a partir del propi terme i d'expressions que explicitin un context definitori, com ara "és un" en català. La formulació d'aquestes consultes, doncs, serà similar a la d'una equació matemàtica, en la qual a partir de dos elements coneguts s'intentarà aïllar un element desconegut:

TERME + EXPRESSIÓ DE FILTRATGE = DADA CERCADA

Així doncs, per exemple, la cerca dels contextos "meteor" que incloguin l'expressió "is a" dóna com a resultat el següent conjunt de KWiC:

Figura 6: KWiC de meteor i is a.

Les expressions que vehiculen una informació determinada i expliciten una dada d'interès per al traductor acostumen a ser construccions del llenguatge natural. La intuïció lingüística del traductor li ha de permetre elaborar consultes utilitzant expressions d'aquesta mena, que li proporcionaran informació de manera estructurada en funció de les seves necessitats. Aquestes expressions, que no sempre seran de caràcter verbal, sovint estan íntimament relacionades amb les peculiaritats de la llengua en aquella àrea temàtica, és a dir, a les característiques de la llengua d'especialitat. Per aquest motiu, unes expressions determinades seran més productives en unes àrees temàtiques que en altres.

Així, es poden obtenir contextos que responguin preguntes com què, qui, com, quan, on o per què en relació a un terme determinat mitjançant expressions que expliciten les relacions conceptuals d'un terme dins de la seva àrea temàtica, així com les seves característiques més rellevants. Si bé aquesta mena de consultes no extrauran tota les dades que el corpus pot facilitar sobre un terme concret, almenys sí que permetran de sistematitzar el procés de cerca i efectuar-lo amb la major celeritat possible. A més, en consultar la informació de manera estructurada en funció de les expressions de filtratge, el traductor podrà assimilar-la amb més facilitat.

Conclusió

Donat que un dels avantatges més destacables de les pàgines web en relació amb el procés de documentació que segueix un traductor rau en la facilitat d'accedir-hi i el seu format digital, en consultar-les es pot recórrer a noves metodologies pròpies de la lingüística de corpus que, d'una banda, permetran eludir alguns dels inconvenients que aquesta mena de documents presenten en tant que font d'informació per a la traducció, i de l'altra proporcionaran la informació de manera concisa i estructurada. En aplicar aquestes metodologies, el procés de documentació passaria per la compilació de corpus *ad hoc*, l'objectiu dels quals consisteix en proporcionar la informació que el traductor necessita per dur a terme la seva tasca, ja sigui durant la fase de comprensió de l'original o la d'elaboració del text final.

El traductor ha de compilar el corpus i explotar-lo en funció de les seves necessitats informatives. Durant la fase d'explotació del corpus, els seus coneixements lingüístics li permetran d'identificar expressions per filtrar contextos que continguin la informació que necessita, sobretot si es tracta d'una necessitat informativa de tipus factual. D'altra banda, les eines d'anàlisi de corpus proporcionen instruments de consulta directa que el traductor pot utilitzar per tal d'obtenir informació que li permeti resoldre problemes de traducció. D'aquesta manera, el traductor arriba a les millors solucions consultant molts textos simultàniament; així doncs, les seves decisions no es basen en un sol text, cosa que pot resultar arriscada.

Bibliografia

AGUILAR-AMAT, Anna; PIQUÉ, Ramon (1999). "La informática y el proceso de la traducción". A: Antonio Argüeso (ed.) *Traducción, interpretación, lenguaje*. TIL 2. Madrid: Fundación Actilibre. Pàgines 141-148.

ALEXANDER, Jan; TATE, Marsha Ann (2001). "Evaluating web resources". [en línia] Wolfgram Memorial Library. 8 d'agost de 1996, actualitzat el 25 de juliol de 2001. <http://www2.widener.edu/Wolfgram-Memorial-Library/webevaluation/webeval.htm> [Consulta: 4 d'agost de 2003]

BOWKER, Lynne (2000). "Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources." A: *International Journal of Corpus Linguistics*, Vol. 5 (1). John Benjamins Publishing. Pàgines 17-52.

CODINA, Lluís (2002). "Fonaments de teoria de recuperació d'informació". A Cristòfol Rovira i Lluís Codina (eds.): *Organització i recuperació de la informació*. Barcelona: Universitat Oberta de Catalunya, pàgines 5-32.

CODINA, Lluís et al. (2001). "Dificultades de representación del conocimiento especializado y propuesta teórica de solución". A: M. Teresa Cabré Castellví, Judit Feliu (eds.) *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica* (DGES PB96-0293). Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. Pàgines 187-191.

CONDAMINES, Anne (1995). "Analyse de textes spécialisés pour le ecueil de données terminologiques". A: *Terminologies Nouvelle. Terminologie et diversité culturelle*, vol. 14. Pàgines 35-42.

CORPAS, Gloria (2001). "Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada". A *Trans*. Revista de traductología, 5. Universidad de Málaga. Pàgines 155-184.

MAYORAL, Roberto (1997/1998). "La traducción especializada como operación de documentación". A: *Sendebar*. Boletín de la Facultad de Traducción e Interpretación. Núms. 8/9. Universidad de Granada. Pàgines 137-153.

PALOMARES, Rocío (2000): *Recursos documentales para el estudio de la traducción*. Málaga: Servicio de Publicaciones Univ. de Málaga.

PINTO, María; CORDÓN, José Antonio (eds.) (1999). *Técnicas documentales aplicadas a la traducción*. Madrid: Síntesis.

SÁNCHEZ-GIJÓN, Pilar (2002). "Aplicaciones de la lingüística de corpus a la práctica de la traducción – Complemento de la traducción asistida por ordenador". A: *Terminologie et Traduction*, 2.2002, Commission des Communautés euorpéennes. Pàgines 84-106.

WebCorp [en línia]. Research and Development Unit for English Studies, University of Liverpool, 2001. <http://www.webcorp.org.uk>[Consulta: 4 d'agost de 2003].

WordSmith Tools [programa informàtic]. Mike Scott, Oxford University Press.

Notes

1 Els exemples que s'inclouen a continuació han estat elaborats amb l'ajuda del programa WordSmith Tools.

2 El corpus utilitzat per il·lustrar l'explotació d'un corpus *ad hoc* va ser compilat com a part de la tesi doctoral de l'autora i consisteix en pàgines web obtingudes cercant les paraules clau *Leonids* i *astronomy* en els cercadors d'Internet més habituals.