Desarrollo de la aplicación Post-editing Calculeffort para la estimación del esfuerzo en posedición

Miguel Ángel Candel-Mora

Carla Borja-Tormo

revista.tradumàtica
tecnologies de la traducció

Miguel Ángel Candel-Mora
Universitat Politècnica de València
mcandel@upv.es;
ORCID:
0000-0001-8754-6046

Carla Borja-Tormo
Universitat Politècnica de València
mcandel@upv.es;
ORCID:
0000-0003-4815-3455

Resumen

En la actualidad, después de la posedición, la calidad de la traducción automática puede ser equiparable a la de la traducción humana. Este trabajo describe el diseño de la aplicación *Post-editing Calculeffort*, que estima esfuerzo y tiempo necesarios de posedición, y determina si es factible procesar un texto con TA.

Palabras clave: traducción automática; posedición; esfuerzo; calidad

Resum

Actualment, després de la post-edició, la qualitat de la traducció automática pot ser equiparable a la de la traducció humana. Aquest article descriu el disseny de l'aplicació *Post-editing Calculeffort*, que calcula l'esforç i el temps necessaris de post-edició, i determina si és factible processar un text amb TA.

Paraules clau: traducción automática; posedición; esfuerzo; calidad

Abstract

Nowadays, after post-editing, the level of quality of machine translation output can be compared to human translation. This paper describes the design of the application *Post-editing Calculeffort* that estimates the effort and time required for post-editing, and determine if it is feasible to process a text with MT.

Keywords: machine translation; post-editing; effort; quality

Recepció: 24 de març de 2017 / Acceptació: 31 d'agost de 2017 / Publicació avançada: 20 de novembre de 2017

Revista Tradumàtica. Número 15 (desembre 2017).
Cita recomanada: CANDEL-MORA, M.A.; BORJA-TORMO, C. (2017). « Desarrollo de la aplicación Post-editing Calculeffort para la estimación del esfuerzo en posedición». Revista Tradumàtica. Tecnologies de la Traducció, 15, 1-9. <https://doi.org/10.5565/rev/tradumatica.190 >

Agradecimientos: Los autores desean agradecer la colaboración de Jesús Hernández en la programación de Post-editing Calculeffort.

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

1. Introducción

La posedición de textos previamente procesados con motores de traducción automática (TA) no es un fenómeno nuevo (Allen, 2003), sin embargo, el desarrollo de sistemas de traducción automática estadística basados en corpus ha dado lugar a un aumento en la calidad de la producción de TA, por lo que cada vez más traductores e instituciones han comenzado a incluir TA en su flujo de trabajo de traducción profesional (Torres Hostench et al., 2016).

Dentro de este nuevo contexto, existe un escaso número de metodologías y criterios sobre cómo realizar trabajos de posedición disponibles, ya que la mayoría de criterios o pautas no son accesibles por razones de confidencialidad de las empresas que los han diseñado, lo que dificulta la posibilidad de proporcionar una visión más general sobre prácticas de posedición.

Según la literatura consultada (Allen, 2003; Guzmán, 2007; Mitchell et al., 2014; SAE International, 2001; TAUS, 2010), entre las categorías más comunes de errores se encuentran los errores terminológicos, de ambigüedad léxica, de sintaxis, de concordancia, por omisión y repetición, o errores de puntuación; y dependiendo del tipo de indicadores utilizados, con diferentes pesos para cada error dependiendo del nivel de gravedad asignado. La literatura también enfatiza que además de criterios específicos de gramática o léxicos existen criterios generales como la legibilidad y aceptabilidad del texto procesado con TA, y de funcionalidad según el tipo de texto (Görög, 2014).

Hasta el momento, la investigación sobre posedición se ha abordado desde diferentes puntos de vista: calidad (Aramberri, 2014, Koby et al., 2014, Specia et al., 2010), pautas de evaluación (Babych, 2014), esfuerzo cognitivo (O'Brien, 2005), aceptabilidad de la producción de TA (Görög, 2014), la combinación de estrategias como preedición y uso de lenguajes controlados para mejorar la traducibilidad (Temnikova, 2010), o desde un contexto comercial para comprobar la productividad en conjunción con herramientas de traducción asistida (Parra y Arcedillo, 2015), mientras que el enfoque para predecir el esfuerzo requerido para la posedición dentro de una perspectiva de gestión de proyectos no ha sido estudiado en profundidad.

Por lo tanto, este trabajo propone el diseño de la herramienta Post-editing Calculeffort basada en pautas de evaluación de posedición existentes y las categorías de error más generalizadas en la literatura para proporcionar un valor específico del esfuerzo necesario para revisar un texto completo según cálculos realizados sobre un fragmento de ese texto. Tras la posedición de un fragmento de un texto procesado con TA, esta herramienta proporciona un informe que incluye las categorías de errores más recurrentes en ese texto, y el tiempo total estimado de la posedición, contribuyendo así a estimar el tiempo y el esfuerzo que conllevaría la posedición completa de todo el texto de donde se extrajo el fragmento, o simplemente determinar si es factible procesar un texto en concreto con TA y realizar su posterior posedición.

2. Traducción automática, calidad y posedición

Tradicionalmente, la calidad de la traducción automática se ha contemplado desde la finalidad del texto traducido (Allen, 2003, TAUS, 2010): para ser publicado y difundido, o si sólo tiene como objetivo orientar al lector sobre su significado general, además de diferentes factores como las especificaciones del cliente, el volumen de la documentación que se espera procesar, o las expectativas con respecto al nivel de calidad del borrador final del texto traducido, entre otros (Allen, 2003: 301).

De hecho, en los últimos tiempos ha habido un incremento en la demanda de lo que se conoce como gisting translation (Candel-Mora, 2015) o lo que es lo mismo, traducciones hechas únicamente de manera automática, sin intervención del traductor humano, que ayudan a proporcionar una aproximación al contenido del texto. Este cambio de paradigma se debe a los grandes volúmenes a traducir, lo cual requiere una automatización cada vez mayor del proceso de traducción y con ello, la calidad ha pasado a ser algo gradual, cuyo nivel varía en base a la finalidad del texto.

Allen (2003) distingue entre dos tipos de actividades de TA: los textos traducidos únicamente para que puedan ser comprendidos y aquellos que se traducen con el fin de difundirse, publicarse, etc., respectivamente. Sin embargo, hay otras maneras de emplear estos motores de TA con el fin de obtener una alta calidad en lo que se traduce: combinar la TA del texto con la posedición, que ha hecho posible que ésta pueda ser empleada a nivel profesional gracias a la mejora de calidad que supone (Aziz, De Sousa, y Specia, 2012).

Con respecto al nivel de calidad necesario o requerido de una traducción, TAUS (2010) a su vez diferencia dos niveles de posedición en función de la calidad que interese obtener: full post-editing y light post-editing. Al mismo tiempo, establece dos niveles en cuanto a la posible calidad deseada dependiendo de qué nivel de posedición se aplique: good-enough y high-quality human translation and revision. Una traducción considerada como good enough sería aquella cuya calidad sería baja, pero el texto es comprensible. Al contrario, una traducción high-quality podría considerarse equiparable a una traducción hecha por un traductor humano.

O'Brien (2011) plantea la medición de la calidad como una aportación para los poseditores para medir el esfuerzo cognitivo que podía suponer la posedición. No obstante, para valorar la calidad de una traducción realizada automáticamente se han creado métricas o escalas de valoración automáticas, ya que las evaluaciones humanas pueden llevar meses y suponen un trabajo que luego no puede ser reutilizado (Papineni et al., 2002). Según Banerjee y Lavie (2005:65), "evaluating an MT system using such automatic metrics is much faster, easier and cheaper compared to human evaluations, which require trained bilingual evaluators". IBM desarrolló el sistema BLEU, diseñado para evaluar traducciones automáticas basándose en la siguiente premisa: "The closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002).

Con sistemas de TA basados en reglas, algunos de los errores más comunes a los que se enfrenta el poseditor son impropiedades terminológicas, falsos sentidos o sinsentidos, palabras sin traducir, repeticiones, alteración del orden de las palabras, uso incorrecto de preposiciones, uso incorrecto de tiempos verbales, calcos sintácticos y léxicos de la lengua origen, extranjerismos innecesarios, errores en la traducción de metáforas lexicalizadas y alteración de la puntuación (Alarcón Navío, 2003).

Además, en la traducción inglés-español, Guzmán (2007) muestra que se producen errores típicamente de confusión entre nombres e infinitivos; no se distingue entre si un texto está escrito en estilo personal o impersonal; se emplea la voz pasiva en lugar de la activa y viceversa. Guzmán también menciona el uso incorrecto de artículos y preposiciones, así como errores en el orden de las palabras dentro de las frases.

Por otro lado, Vilar et al. (2006), proponen una estructura jerarquizada de errores en la que sitúan en el primer nivel los siguientes: omisión de palabras, palabras incorrectas, orden de las palabras, palabras desconocidas y errores de puntuación.

Asimismo, TAUS coincide con algunos de los errores destacados por los anteriores autores (Alarcón Navío, 2003; Guzmán, 2007): terminología, repeticiones, orden de las palabras, ambigüedad, puntuación, etc. (TAUS, 2010).

3. Diseño de la aplicación

Para este trabajo ha sido necesario el diseño de una escala propia de valoración de errores para la posedición que recogiera los errores más comunes y relevantes y sirviera para elaborar un informe que ayudara a la estimación del tiempo invertido y trabajo de posedición e identificación de patrones de errores más frecuentes. Para ello, se han consultado aportaciones a la literatura académica al respecto que, aunque proceden de su aplicación a diferentes sistemas de TA como los basados en reglas o estadísticos, proporcionan un completo panorama de las categorías más comunes en escalas de valoración de TA (Alarcón Navío, 2003; Guzmán, 2007; TAUS, 2010; Vilar et al., 2006). Con el fin de simplificar el uso de la aplicación, tras recopilar toda la información de escalas de valoración de errores anteriores, los errores se han agrupado en tres categorías: estilo, terminología y gramática. En dicha escala se han asignado valores para cada tipo de error en función de su gravedad, igualmente siguiendo las escalas consultadas.

3.2 Diseño de la aplicación Post-editing Calculeffort

Una vez analizadas las escalas anteriores, se ha hecho un inventario de las coincidencias encontradas en las distintas escalas. Uno de los errores más frecuentemente citados en la literatura es la alteración del orden de las palabras, al que le sigue la terminología y después, el uso incorrecto de preposiciones y tiempos verbales. La alteración de la puntuación y la ambigüedad son algunos de los errores que, aunque menos comunes, son relevantes para este estudio.

Los errores de la categoría "Estilo", como la alteración del orden de las palabras o la alteración de la puntuación son los más destacados por la literatura consultada, de manera que este tipo de errores se situó en primer lugar en el orden de importancia de la escala para la aplicación Post-editing Calculeffort y se le asignó una puntuación de 3. Las escalas estudiadas otorgan importancia en segundo lugar a los errores de terminología y vocabulario, por lo que en este trabajo se asignó a esa categoría una puntuación de 2. Por último, los errores gramaticales y sintácticos, entre los que se encuentran el uso incorrecto de preposiciones, el uso incorrecto de artículos, o el uso incorrecto de tiempos verbales no aparecen con tanta frecuencia en las escalas estudiadas y se les ha puntuado con un 1.

La interpretación de la puntuación asignada es la siguiente: se considerará que cada error de cada categoría equivale a tantas palabras incorrectas como puntos asignados tenga, es decir, si hay un error de estilo, dada su gravedad, equivaldrá a 3 errores. De esta forma, la puntuación total acumulada de estilo a lo largo de todo el texto será: Puntuación total estilo = n.º errores de estilo x 3. Por lo tanto, cuantos más puntos acumulados, más imperfecto será el texto que se ha poseditado. Análogamente, la puntuación de terminología acumulada al finalizar la posedición será: Puntuación total terminología = n.º errores de terminología x 2. Finalmente, y de la misma forma que los dos anteriores, la puntuación total de errores de gramática que se tendrá una vez finalizada la revisión será: Puntuación total gramática = n.º errores de gramática x 1.

Durante la fase inicial de pruebas de la aplicación se priorizó la identificación del mínimo número de palabras que era necesario poseditar para poder estimar el tiempo y el esfuerzo de posedición del texto completo. Para ello se comparó los resultados obtenidos al poseditar los textos completos y los resultados de poseditar sus fragmentos. La intención es comprobar con qué número de palabras se puede extrapolar el resultado obtenido (número de errores y tiempo necesario) al número total de palabras del texto. La metodología utilizada consistió en registrar el tiempo, el número y el tipo de errores que identificaron cuatro traductores humanos que poseditaron el mismo texto: primero el fragmento y después el texto completo, y de ahí se obtuvieron los promedios que ayudaron a extrapolar los datos de la posedición del fragmento a la posedición del texto completo. Las pruebas se realizaron en varias fases, y con varios textos del mismo género, pero de diferente extensión y con el mismo motor de traducción, Traductor de Google. Estos resultados iniciales motivaron la segunda fase de la investigación, actualmente en curso, consistente en realizar las mismas pruebas con otros sistemas de traducción automática y otros géneros textuales. Con ello, se podría desarrollar una tarea de consultoría lingüística rápida y eficaz en la que, mediante la revisión de un fragmento del texto traducido automáticamente, se podría aproximar el esfuerzo y tiempo que la posedición completa supondría al poseditor y ayudaría a ofrecer un presupuesto aproximado al cliente en un entorno profesional.

La función principal de esta aplicación es estimar el tiempo y esfuerzo dedicado a la posedición de un texto en base a los cálculos sobre un fragmento del mismo. Como se puede observar en la Figura 1, a la izquierda se ha incluido la escala de errores utilizada para el diseño de la escala utilizada en la aplicación. Además, cuenta con un

botón de "Inicio" que pone en marcha el temporizador que calculará el tiempo invertido en la posedición de cada documento. Dicho cronómetro se detendrá en el momento en el que se pulse la opción "Fin". Durante el transcurso de este tiempo, el poseditor dispone de un botón para cada tipo de error que identifique. El número de errores irá acumulándose en el contador de la parte superior. Para anotar un error, el poseditor únicamente marca con el cursor (se sitúa al inicio de la secuencia y mantiene pulsado el botón izquierdo del ratón arrastrando el cursor hasta el final de dicha secuencia) la palabra o frase en cuestión y a continuación, pulsa el botón que corresponde al tipo de error encontrado, y éste se destaca mediante un código de colores.

Post-editing Calculeffort se ha diseñado como una aplicación independiente para ser utilizada en el entorno Windows y consiste en una sencilla, práctica e intuitiva interfaz, como muestra la Figura 1, con una ventana central donde se pega el fragmento en formato texto y mediante los botones de la barra lateral se van anotando los errores sobre ese texto.

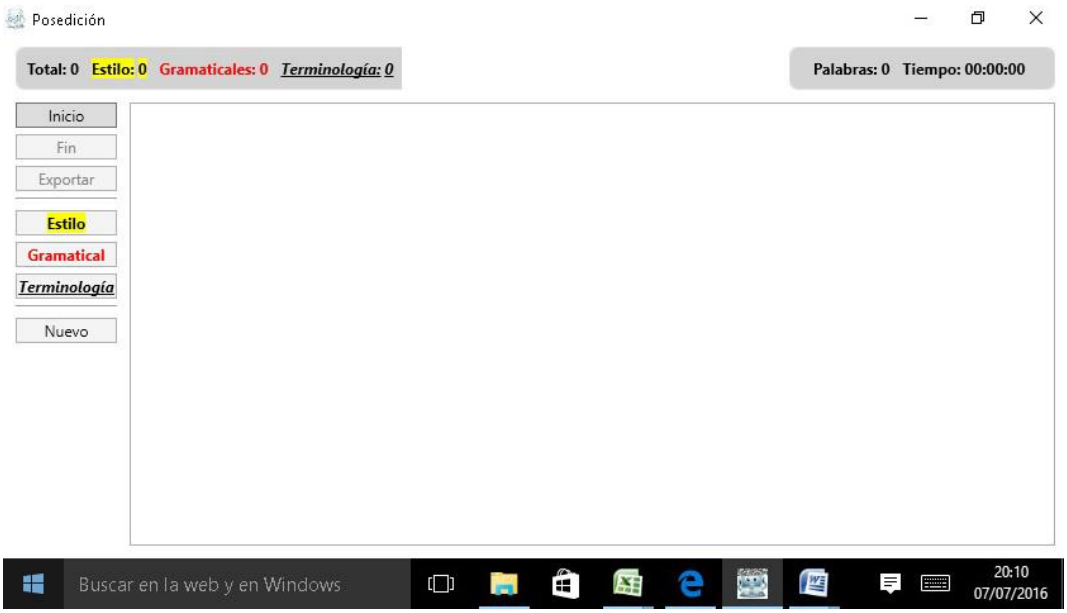

Figura 1. Interfaz de la aplicación Post-editing CalculEFFORT.

Dado que un mismo error puede pertenecer a dos categorías al mismo tiempo, o pertenecer a un error dentro de una secuencia de palabras más extensa, Post-editing CalculEFFORT permite la anotación de errores de diferentes maneras y que ese error tenga un valor según las categorías a las que pertenezca, como cuando ocurre un error que es tanto gramatical, como terminológico o de estilo a la vez, por ejemplo.

Finalmente, una vez acabado el proceso de identificación de errores, se pulsa la opción "Fin", y el contador dejar de sumar y el cronómetro se detiene. Con el botón "Exportar" se pueden extraer los resultados de la revisión en un documento de texto en el que se incluye el número de errores de cada tipo y el tiempo empleado (Figura 2).

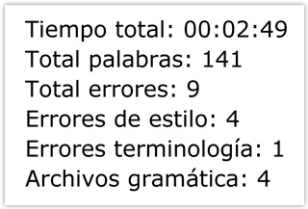

Figura 2. Datos exportables de Post-editing Calculeffort.

Estos datos de tiempo y patrones de errores (Figura 2) se utilizaron durante la fase inicial de ensayos de la aplicación para realizar diversas pruebas de posedición con fragmentos de textos de diferente longitud, y contrastar los resultados obtenidos por varios poseditores, hasta llegar a datos extrapolables al total de los textos. De esta forma, la información obtenida puede servir también para verificar, en primer lugar, el nivel de calidad en los textos traducidos con TA y si el rendimiento de los motores de TA varía en función de la temática o del género textual.

4. Conclusiones

Debido a que la traducción mediante motores de TA no alcanza el nivel de calidad de la traducción humana, se hace énfasis en el desarrollo de técnicas de posedición y estudios que contribuyan a la mejora de la TA y su aplicación en entornos profesionales. Por ello, el principal objetivo de esta investigación ha sido intentar mejorar el proceso de los proyectos de TA diseñando una aplicación que permita estimar el esfuerzo necesario en un proyecto de posedición y contabilizar los errores de cada tipo (terminología, gramática y estilo), así como el tiempo invertido en la posedición, ofreciendo la posibilidad de exportar los resultados para su posterior tratamiento.

Una vez identificados los errores más frecuentes que produce la TA y la elaboración de una escala simplificada que agrupa esos errores en tres categorías, el siguiente paso fue la creación de una aplicación informática para el entorno Windows de ayuda al cálculo del esfuerzo de posedición de proyectos de TA de grandes dimensiones. Con Post-editing CalculEFFORT es posible realizar la posedición de una muestra de texto, ya que mediante diferentes pruebas se ha obtenido una estimación muy precisa de la extensión de la muestra necesaria para que los tiempos y el número de errores identificados en esa muestra resulten representativos para el total del texto objeto de posedición. En la segunda fase de la investigación, en curso actualmente, el objetivo consiste en repetir las pruebas iniciales con diferentes textos pertenecientes a otros géneros textuales como textos divulgativos o manuales de instrucciones, junto con el uso de otros sistemas de TA extendidos. También, está prevista la reprogramación de la aplicación para hacerla disponible online y poder computar datos de otras combinaciones de idiomas y aumentar el número de usuarios con el fin de comprobar la precisión del esfuerzo en posedición.

5. Bibliografía

Alarcón Navío, E. (2003). "Traducción automática versus traducción humana: tipología de errores". En: Muñoz Martín, R. (ed.). *Actas del I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación* 2, p. 721-738, <http://www.aieti.eu/pubs/actas/I/AIETI_1_EAN_Traduccion.pdf> [Consulta: 8 de junio de 2016].

Allen, J. (2003). "Post-editing". En: Somers, H. (ed.). Computers and Translation: *A Translator's Guide. Amsterdam* [etc.]: John Benjamins, p. 297-318.

Aranberri. N. (2014). "Posedición, productividad y calidad". *Revista Tradumàtica: tecnologies de la traducció,* n. 12, p. 471-477. <https://doi.org/10.5565/rev/tradumatica.62> [Consulta: 5 de febrero de 2017].

Aziz, W.; De Sousa, S.; Specia, L. (2012). "PET: a Tool for Post-editing and Assessing Machine Translation". En: *Proceedings of LREC 2012.* <http://wilkeraziz.github.io/dcs-site/publications/2012/AZIZ+LREC2012.pdf>. [Consulta: 4 de junio de 2016].

Babych, B. (2014). "Automated MT Evaluation Metrics and their Limitations". *Revista Tradumàtica: tecnologies de la traducció*, n.12, p. 464-470. <https://doi.org/10.5565/rev/tradumatica.70>. [Consulta: 5 de febrero de 2017].

Banerjee, S.; Lavie, A. (2005)."METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". En: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* 29, p. 65-72. <http://www.aclweb.org>. [Consulta: 6 de Julio de 2016].

Candel-Mora, M. A. (2015). "Evaluation of English to Spanish MT Output of Tourism 2.0 Consumer-Generated Reviews with Post-Editing Purposes". En: *Proceedings of the 37th Conference Translating and the Computer*, p. 37-47.

De Almeida, G.; O'Brien, S. (2010). "Analysing Post-editing Performance: Correlations with Years of Translation Experience". En: *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. <http://www.mt-archive.info/EAMT-2010-Almeida.pdf> [Consulta: 9 de junio de 2016].

Doddington, G. (2002). "Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics". En: *HLT '02 Proceedings of the second international conference on Human Language Technology Research*. San Diego, California: Morgan Kaufmann, p.138-145. <http://dl.acm.org>. [Consulta: 29 de mayo de 2016].

Görög, A. (2014). "Traducción y calidad". *Revista Tradumàtica: tecnologies de la traducció*, n. 12, p. 388-391. <https://doi.org/10.5565/rev/tradumatica.80> [Consulta: 29 de mayo de 2016].

Guzmán, R. (2007). "Manual MT Post-editing: If it's not broken, don't fix it". *Translation Journal*, v. 11, n. 4 (October). <http://translationjournal.net/journal/42mt.htm> [Consulta: 29 de mayo de 2016].

Koby, G.; Fields, P.; Hague, D.; Lommel, A.; Melby, A. (2014). "Defining Translation Quality*". Revista Tradumàtica: tecnologies de la traducció*, n. 12, p. 413-420. <https://doi.org/10.5565/rev/tradumatica.76> [Consulta: 14 de agosto de 2017].

Mitchell, L.; O'Brien, S.; Roturier, J. (2014). "Quality Evaluation in Community Post-editing". *Machine translatio*n, v. 28, n.3-4 (December), p.237-262. <https://doi.org/10.1007/s10590-014-9160-1> [Consulta: 5 de julio de 2016].

O'Brien, S. (2011). "Towards Predicting Post-editing Productivity". *Machine translation,* v. 25, n. 3 (September), p. 197-215. <https://doi.org/10.1007/s10590-011-9096-7> [Consulta: 5 de julio de 2016].

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation". En: *Proceedings of the 40th Annual meeting on Association for Computational Linguistics*. Philadelphia, p. 311-318. <http://dl.acm.org/citation.cfm?id=1073135>. [Consulta: 5 de julio de 2016].

Parra Escartín, C.; Arcedillo, M. (2015). "Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings". En: *Proceedings of MT Summit XV*, v.1, p.131-144.

SAE International (2001). *SAE J2450: Translation Quality Metric Metric Task Force*. Warrendale, USA: Society of Automotive Engineers. <http://www.sae.org/standardsdev/j2450p1.htm> [Consulta: 14 de agosto de 2017].

Specia, L.; Raj, D.; Turchi, M. (2010). "Machine Translation Evaluation Versus Quality Estimation". *Machine Translation*, v. 24, n.1 (march), p. 39–50. <https://doi.org/10.1007/s10590-010-9077-2>. [Consulta: 18 de septiembre de 2017].

TAUS (2010). *MT Post-editing Guidelines*. <https://www.taus.net/academy/best-practices/evaluatebest-practices/adequacy-fluency-guidelines>. [Consulta: 5 de julio de 2016].

Temnikova, I. (2010). "Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment". En: *Proceedings of the LREC 2010 Conference*. Valletta: European Language Resources Association, p.17-23. <http://www.lrec-conf.org/proceedings/lrec2010/pdf/437_Paper.pdf> [Consulta: 4 de junio de 2016].

Torres Hostench, O.; Cid-Leal, Pilar ; Presas, Marisa (coord.) (2016). *El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: informe de investigación ProjecTA 2015*. Bellaterra. <https://ddd.uab.cat/record/148361>. [Consulta: 23 de noviembre de 2016].

Vilar, D.; Xu, J.; d'Haro, L. F.; Ney, H. (2006). "Error Analysis of Statistical Machine Translation Output". En: *Proceedings of LREC 2006*, p. 697-702. <http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf>. [Consulta: 6 de julio de 2016].