GOBIERNO DE ESPAÑA
MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL
SECRETARÍA DE ESTADO PARA LA SOCIEDAD DE LA INFORMACIÓN Y LA AGENDA DIGITAL

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

INVENTARIO DE RECURSOS LINGÜÍSTICOS DE LA ADMINISTRACIÓN PÚBLICA PARA TRADUCCIÓN AUTOMÁTICA

Plan de Impulso de las Tecnologías del Lenguaje

Autores de UPM [en orden alfabético]:

Guadalupe Aguado de Cea
Mª Socorro Bernardos Galindo
Asunción Gómez Pérez
Jorge Gracia del Río
Elena Montiel Ponsoda
Silvia Sebastián

Otros autores de ReTeLe:

Núria Bel (Universitat Pompeu Fabra)
Montserrat Marimón (Universitat Pompeu Fabra)
Mikel Forcada (Universidad de Alicante)

Octubre 2016

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para la Sociedad de la Información y la Agenda Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

Índice

INVENTARIO DE RECURSOS LINGÜÍSTICOS DE LA ADMINISTRACIÓN PÚBLICA PARA TRADUCCIÓN AUTOMÁTICA

1 INTRODUCCIÓN

El presente informe resume las actividades llevadas a cabo en el contexto de la red de excelencia ReTeLe[1] para elaborar un **inventario de recursos lingüísticos** que se hayan producido y se estén produciendo por la Administración del Estado y en las Administraciones de las diferentes Comunidades Autónomas, y que sean susceptibles de ser utilizados para alimentar sistemas de **traducción automática**.

Con tal fin se han llevado a cabo una serie de entrevistas y reuniones bilaterales con miembros de la administración pública, tomando como punto de partida la lista de participantes en el taller del European Language Resource Coordination (ELRC)[2] celebrado en Madrid el 26 de enero de 2016. En dicho taller, el consorcio ELRC convocó a miembros de la administración pública española con una doble finalidad. Primero, para presentar las políticas europeas en el área de la traducción automática, así como herramientas de traducción automática y sus aplicaciones prácticas. Y en segundo lugar, para invitarles a compartir con el resto de asistentes los procedimientos que se siguen en los diferentes organismos de la administración pública en España en cuanto a la traducción de documentos, las herramientas con las que se cuenta, casos de éxito y necesidades desatendidas. Los participantes del taller ELRC conformaban una lista representativa de los organismos de la administración productores de traducciones y por tanto como potenciales proveedores de recursos de traducción para traducción automática, y por ese motivo, se decidió tomarla como punto de partida.

Previo a la realización de las entrevistas se envió a todos un cuestionario cuyo resultado sirvió de base para las mismas, en el que se recopiló cierta información básica sobre el tipo de recursos de traducción utilizados por las distintas administraciones.

También se ha desarrollado una plataforma *online* para catalogar los recursos identificados en este proyecto por alguna de las vías mencionadas anteriormente (entrevistas y cuestionarios). El objetivo de dicha plataforma (que en adelante llamaremos “catálogo ReTeLe” o simplemente “catálogo”) es

[1] http://retele.linkeddata.es
[2] http://lr-coordination.eu/es/spain

recopilar información y datos descriptivos sobre recursos lingüísticos en general, lo cual incluye los recursos para la traducción con los que cuenta la administración pública. Dicho catálogo está preparado para publicar los datos en un formato interoperable y compatible con otras iniciativas como datos.gob.es y con el esquema Meta-Share. Para ello se ha creado una ontología, ReTeLe-share, que da soporte a los datos del catálogo.

2 LISTADO DE RECURSOS

2.1 DISEÑO DE LAS ENTREVISTAS Y REALIZACIÓN DE LAS MISMAS.

Previo a la realización de las entrevistas con los representantes de las distintas administraciones públicas, se envió a todos un cuestionario cuyo resultado sirvió de base para las mismas. La carta en la que se pedía colaboración para realizar dicha encuesta la remitió la SETSI y se adjunta como Anexo 1 a este documento. Más adelante se envió una nueva carta de recordatorio vía email, esta vez remitida por la red ReTeLe, que se adjunta en el Anexo 2. El cuestionario se adjunta como Anexo 3.

Se enviaron cartas a 25 personas representativas de diferentes departamentos de la Administración Pública, con el fin de que completaran el cuestionario. Para la elaboración de dicha lista de destinatarios se tomó como punto de partida la relación de participantes en el taller ELRC. La relación de destinatarios con los datos de las instituciones a las que pertenecen y los teléfonos de contacto figuran en el Anexo 4. En una primera fase (mayo de 2016) se obtuvieron solamente tres respuestas, que fueron completadas con siete más en la segunda fase, tras el envío de los recordatorios (junio/julio de 2016). Se obtuvo, pues, un 40% de respuestas.

Tras la recepción de los cuestionarios se creyó conveniente contactar por correo o por teléfono con las personas que habían respondido (la relación de encuestas recibidas se adjunta en el Anexo 5). El objetivo era doble; por un lado, se pretendía ampliar la información que habían incluido en la encuesta y, por otro, se trataba de mostrarles la plataforma (el portal web del catálogo descrito en el apartado 4) y explicarles cómo proceder para incluir los datos. Es preciso tener en cuenta que no todos los funcionarios están familiarizados con los metadatos de la plataforma. No obstante, no se pudieron hacer todas las entrevistas por no conseguir contactar con los interesados telefónicamente, o bien porque afirmaban no contar con ningún recurso que fuera reutilizable.

Como resultado, las entrevistas realizadas y sus fechas fueron:

1. 06/07/2016 con D. Salvador Estevan Martínez, Dirección de Tecnologías de la Información y las Comunicaciones (DTIC) del Ministerio de Hacienda y AAPP.

2. 13/07/2016 con D. Leandro Valencia, de la OIL del Ministerio de Asuntos Exteriores
3. 30/09/2016 con D. Javier Samper y Dña. Elena Navascués Guillot, del Servicio de Traducción del Ministerio de Justicia.
4. 30/09/2016 con Dña. Noemí Lera de Red.es
5. 06/10/2016 con Dña. Lucía Escapa y sus colaboradores de la Subdirección General de Tecnologías y Servicios de Información. Subsecretaría. Ministerio de la Presidencia

Nótese que la disponibilidad de los entrevistados condicionó la selección de las fechas, algunas de las cuales exceden la fecha de finalización que se había previsto para este proyecto en un primer momento.

En todas las entrevistas estuvieron presentes dos personas de la red ReTeLe participantes en el proyecto: Dña Guadalupe Aguado y Dña Elena Montiel-Ponsoda en las cuatro primeras, y Dña Guadalupe Aguado y Dña Nuria Bel en la última.

Todas las entrevistas se plantearon en tres partes: en primer lugar, se explicaba nuevamente el proyecto, los objetivos y las ventajas del mismo, ya que en algún caso el destinatario de la carta inicial nos había remitido a una segunda persona, pues dicho destinatario inicial no era la persona encargada de proporcionar los datos o la persona capacitada para tomar la decisión final. En segundo lugar, se hacían preguntas al interesado relacionadas con las respuestas que hubieran dado en la encuesta. En tercer lugar, se mostraba la plataforma para incluir los datos o se les indicaba cómo entrar. Este tercer paso no siempre fue posible, bien porque en aquel momento no supieran los recursos exactos con los que contaban o bien porque consideraban que tenían que pedir permiso a su superior, contactar con otros departamentos, o bien porque no tenían asignado más tiempo para la entrevista. En cualquier caso, para mayor facilidad, se les registraba en el portal del catálogo o se les facilitaba nuevamente la dirección web del catálogo para que introdujeran los datos en cuanto tuvieran ocasión.

Dada la limitada respuesta obtenida de la lista de participantes en el taller ELRC, se pensó en otras alternativas para identificar posibles fuentes de recursos de traducción. En particular se recurrió a las web de contratación pública de las CCAA, España y la UE para extraer los organismos públicos que hubieran contratado servicios de traducción. Este estudio pretendía, por un lado, identificar grandes productores de traducciones, y por otro, localizar contenidos ya traducidos que pudieran ser accesibles, por ejemplo, si eran traducciones de páginas web. Esta selección se podía considerar ya

como datos re-utilizables, ya que estaban publicados y la traducción no había sido mediante traducción automática, o en su caso, había sido corregida para alcanzar niveles de calidad estándar. La tarea consistió básicamente en la búsqueda para todas las fechas disponibles de todos los contratos con código CVP 79530000, correspondiente a servicios de traducción[3]. También se realizaron búsquedas para servicios de programación de software y de auditoría (CPV 72200000), pero no se encontró ningún contrato relacionado con traducción. Una vez estudiados los resultados se procedió a elaborar una tabla con los contratantes más frecuentes y su dirección y teléfono de contacto si dicha información estaba disponible. Dicha tabla se encuentra en el Anexo 6.

Finalmente, como otro medio para la identificación de recursos de traducción, se organizó el "taller ReTeLe" en septiembre de 2016 durante la conferencia SEPLN. Si bien el alcance de dicho taller era más amplio y se pretendía documentar cualquier tipo de recurso lingüístico producido en España, también cubría los recursos de la administración pública. Véase el apartado 6 para más detalles.

2.2 LISTADO DE RECURSOS PARA LA TRADUCCIÓN AUTOMÁTICA

Tras las respuestas recibidas en las encuestas y las entrevistas realizadas se identificaron los siguientes tipos de recursos:

1. En la mayoría de casos, los diferentes organismos de la administración cuentan con corpus monolingües, relativos a convocatorias, becas o legislación.

2. La sección encargada del servicio Atención al Ciudadano (dependiente de DTIC) dispone de bases de datos en Excel, que relacionan palabras clave con la respuesta que posiblemente necesita el usuario (áreas y/o documentos). También es monolingüe.

3. Asimismo, DTIC cuenta con un listado de las posibles preguntas que hacen los ciudadanos a dicho servicio de atención ciudadana. Esta lista está en castellano.

4. En la DTIC disponen también de ARCHIVE, una aplicación web de archivo definitivo de expedientes y documentos electrónicos[4]

5. La DTIC, como responsable de la plataforma de traducción automática PLATA, dispone de las memorias de traducción que se han utilizado para alimentar dicha plataforma.

[3] En el caso de la web de contratación de la UE además se restringió la búsqueda a contratos cuyo origen fuera España.
[4] http://administracionelectronica.gob.es/ctt/archive#.WBoVOyTCfo1

6. La Oficina de Interpretación de Lenguas del Ministerio de Asuntos Exteriores cuenta con memorias de traducción.

7. En el Ministerio de Justicia no se utilizan memorias de traducción. El Servicio de Traducción está compuesto por cuatro personas que se encargan de las traducciones del inglés y francés. El resto de las lenguas se externalizan. Cuentan con un archivo en la red interna para las dos lenguas principales, pero no siempre tienen versiones en ambas lenguas en formato electrónico, ya que a veces los textos originales les llegan en papel.

8. En muchas de las entrevistas, los entrevistados sugieren que contactemos con otros departamentos de las mismas instituciones, en donde también se realizan traducciones. Ellos mismos se comprometen a enviarnos los datos de contacto, pero aun no se han recibido. A modo de ejemplo, cabe mencionar la Subdirección general de cooperación jurídica internacional, quienes previsiblemente disponen de documentos en la lengua origen (inglés, francés) y la lengua meta, el castellano.

No obstante dichos recursos no se han documentado todavía siguiendo el modelo de metadatos desarrollado en este proyecto (ver apartado 3) bien por la falta de tiempo o de información por parte de los responsables entrevistados. Por el momento la lista de recursos lingüísticos disponibles en el catálogo proviene principalmente de los asistentes al taller ReTeLe y, si bien son todos ellos útiles para alimentar sistemas de traducción automática, provienen principalmente de entornos académicos y no de las administraciones públicas.

2.3 CONJUNTO DE REQUISITOS FUNCIONALES Y NO FUNCIONALES QUE SE UTILIZARON PARA DESCRIBIR EL ESTADO Y LA MADUREZ DE LOS RECURSOS DE TRADUCCIÓN AUTOMÁTICA IDENTIFICADOS Y MANUAL PARA EJERCICIOS POSTERIORES.

Para la elaboración del conjunto de requisitos funcionales y no funcionales se elaboró un modelo de madurez de los recursos (se incluye como anexo al final de este documento[5]). En este informe proponemos criterios para evaluar la efectividad del suministro de los recursos de traducción de acuerdo con unos requisitos basados en la disponibilidad de los recursos y en sus características (modelo de madurez de recursos en Tabla 1 y metadatos Tabla 2) y se dan directrices para adecuar los protocolos que se deben adaptar para, efectivamente, hacer que la administración pública pueda

[5] También se encuentra disponible en https://arxiv.org/abs/1607.01990

ser considerada proveedora sistemática de estos datos. Dicho modelo de madurez fue presentado en la conferencia Metaforum 2016[6].

Los modelos de madurez[7] permiten trazar estos criterios para evaluar a diferentes organismos de la administración pública como proveedores regulares de datos para entrenar sistemas de traducción automática estadística, a partir de un modelo de madurez de los recursos mismos que producen.

La administración, en tanto que productora de traducciones (de forma interna o por subcontratación) es a su vez productora de recursos lingüísticos: datos en forma de documentos y sus traducciones, memorias de traducción, terminologías y glosarios bilingües o multilingües. La gestión por parte de la organización de los procesos de traducción, las condiciones de archivo de los recursos generados y de la infraestructura disponible para darle soporte determinan la eficiencia y eficacia con la que los materiales producidos pueden convertirse en datos reutilizables.

Por este motivo, enfocamos en este modelo como objeto a evaluar la organización misma, que se puede dotar de protocolos para archivar documentos (y datos) de forma organizada y con unos requisitos en cuanto al formato, a la información asociada o metadatos, a las licencias posibles y al respeto a la confidencialidad y privacidad de datos. La existencia de estos protocolos, la gestión del proceso de traducción y la infraestructura disponible, son la base para elaborar criterios objetivos que permitan evaluar qué organizaciones pueden ser consideradas proveedoras. Este modelo ofrece también, en consecuencia, una guía de los elementos a mejorar por la organización para conseguir que se convierta en proveedora de recursos para su reutilización efectiva.

[6] http://www.meta-net.eu/events/meta-forum-2016/
[7] Por ejemplo: http://cmmiinstitute.com/ y Test Maturity Model TTMI, http://www.tmmi.org/pdf/TMMi.Framework.pdf

Los criterios identificados se describen en la siguiente tabla:

Madurez de los recursos lingüísticos para su re-utilización por sistemas de traducción automática estadística									
NIVEL	**Archivo**	**documento × fichero × lengua**	**PDF**	**Texto llano** (.txt), **ODF** (.odt), **HTML , OOXML** (.docx), .doc, **XML**	**documentos alineados**	**Memorias × documentos: alineación de oraciones**	TMX	**Memorias x dominios / áreas**	**metadatos estándares**
0		✓	✓						
1		✓		✓					
2	✓	✓		✓	✓				
3	✓	✓		✓	✓	✓			
4	✓	✓		✓	✓	✓	✓	✓	
5	✓	✓		✓	✓	✓	✓	✓	✓

Tabla 1: Madurez de los recursos lingüísticos para su reutilización por sistemas de traducción automática

La siguiente tabla describe la información que se puede obtener del proveedor y que se convierte en metadatos (estándares) especializados:

NIVEL	lenguas	respon-sable	tamaño en seg-men-tos / pala-bras	fecha de creación	dominio	codi-fica-ción de carac-teres	recursos aso-cia-dos	docu-men-tación asociada	priva-cidad	confi-den-cia-lidad	licencia
0											
1											
2	✓	✓									
3	✓	✓	✓	✓							
4	✓	✓	✓	✓	✓	✓	✓	✓			
5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Tabla 2: Principales metadatos por nivel de madurez

2.4 RECOPILACIÓN DE INFORMACIÓN PARA EVALUAR EL POTENCIAL DE AUTOMATIZACIÓN DE LA TRADUCCIÓN EN LOS DIFERENTES ORGANISMOS ENTREVISTADOS. BASE DE DATOS DE CONTACTOS DE RESPONSABLES.

Tras analizar las respuestas recibidas y las entrevistas realizadas, se ha observado que los diferentes departamentos de la administración son pequeñas islas dentro de un conjunto amplio en los diferentes ministerios e instituciones. Las carencias más importantes son las siguientes:

- Solamente en la Oficina de Interpretación de Lenguas del Ministerio de Asuntos Exteriores (OIL-MAEC) y en la Dirección de Tecnologías de la Información y las Comunicaciones (DTIC) se utilizan memorias de traducción. Sin embargo, en la OIL-MAEC no mantienen un registro de todas las traducciones clasificadas por áreas o dominios, por confidencialidad, etc., estando todos los documentos juntos. También se detectó la falta de sistematicidad a la hora de utilizar la memoria de traducción; incluso hay traductores que no la usan.

- Cuentan con sus propios glosarios, más específicamente, glosarios de los términos propios de cada uno de los ministerios o de las áreas de trabajo de cada ministerio, pero no se

encuentran en bases de datos compartidas, sino que cada traductor mantiene sus glosarios en formato de texto plano word.

- Los traductores no tienen los documentos archivados en un único repositorio, ni existe un registro único, sino que cada uno almacena los que traduce, lo cual dificulta igualmente la recuperación masiva de documentos paralelos. Por otra parte, el contenido de los documentos que almacenaban no suele repetirse (discursos, cartas), es decir, son documentos para una única ocasión y por lo tanto consideraban que no merecía la pena incluirlos en una futura memoria de traducción.

- En algunos departamentos de otros ministerios (Mº de Justicia), salvo los documentos que están en inglés, francés y español, las traducciones de otras lenguas se externalizan, y la memoria de traducción, en caso de haberla, la tiene la empresa. En esos casos la empresa entrega la traducción, con lo que el organismo en cuestión es dueño de esa traducción pero, de nuevo, el esfuerzo de recopilar los originales y sus traducciones sería considerable, y en ningún caso dichos documentos estarían alineados (como es el caso de las memorias de traducción), con lo que su reutilización no sería inmediata.

El listado con los datos de contacto de los responsables entrevistados se adjunta en el Anexo 7 al final de este documento.

3 MODELO DE DESCRIPCIÓN DE METADATOS

3.1 MODELO PARA LA DESCRIPCIÓN DE METADATOS DE RECURSOS DE TRADUCCIÓN AUTOMÁTICA.

Como se ha mencionado en la introducción, el catálogo ReTeLe es un catálogo online diseñado para describir y catalogar los recursos lingüísticos y de traducción identificados en este proyecto. La forma habitual de describir recursos es mediante los llamados metadatos, es decir, etiquetas descriptivas que dan lugar a campos que se tienen que rellenar con la información que describe al recurso (por ejemplo, autor, fecha de creación, número de palabras, etc.). Para la selección de los metadatos (o descriptores) que describieran los recursos identificados se siguieron tres criterios fundamentales: 1) concisión y adecuación al tipo de recursos en cuestión, 2) interoperabilidad con otros catálogos de metadatos que describen recursos similares, e 3) implementación de acuerdo con formalismos de la Web Semántica que permitan consultas eficientes (tanto por humanos como por máquinas), reutilización y publicación (llegado el caso) en la Web. Teniendo en cuenta dichos criterios, se creó

un modelo de datos basado en una ontología, de forma que los descriptores puedan ser consultados por humanos y por máquinas (agentes software), sean interoperables con otros catálogos y se puedan utilizar en otras plataformas, llegado el caso, y permitirán describir los recursos de forma abierta, si así se desea.

Para la descripción semántica de los recursos se han analizado los siguientes modelos de metadatos con el fin de garantizar la interoperabilidad con los mismos:

- Metashare-owl[8]. Esta ontología constituye la versión en OWL del modelo de recursos lingüísticos utilizado por la iniciativa Meta-Share. Ha sido desarrollado en el contexto de la comunidad *Linked Data for Language Technologies*[9] (LD4LT) del *World Wide Web Consortium* (W3C). Dicho modelo puede encontrarse en http://purl.org/net/def/metashare

- ELRC-Share. Es el modelo utilizado por la iniciativa ELRC para dar soporte a su catálogo de recursos lingüísticos: http://lr-coordination.eu/resources. El modelo no es una ontología propiamente dicha, sino un modelo basado en XML, que a su vez está basado en el modelo de Meta-Share. No es de carácter público (por el momento) por lo que el código fue solicitado y obtenido de sus propios desarrolladores.

- DCAT. Es un vocabulario en RDF diseñado para facilitar la interoperabilidad entre catálogos de datos publicados en la Web, disponible en http://www.w3.org/TR/vocab-dcat/. Dicho vocabulario tiene el estatus de "recomendación" por el W3C y es de uso extendido en administraciones públicas (por ejemplo en http://datos.gob.es/), especialmente mediante su perfil denominado DCAT-AP[10].

El desarrollo de la ontología ha seguido los siguientes pasos:

1. Se ha tomado Metashare-owl como punto de partida, dado que es una ontología ya definida en OWL y refleja un amplio consenso de las comunidades Meta-Share y LD4LT.

2. Dicha ontología, Metashare-owl, fue diseñada con la compatibilidad con DCAT en mente, reutilizando elementos de la misma o proponiendo equivalencias entre ambos esquemas

[8] J. P. McCrae, P. Labropoulou, J. Gracia, M. Villegas, V. R. Doncel, and P. Cimiano, "One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the web," in Proc. of 12th Extended Semantic Web Conference (ESWC 2015) Satellite Events, Portorož, Slovenia, vol. 9341, Jun. 2015, pp. 271-282.
[9] https://www.w3.org/community/ld4lt/
[10] https://joinup.ec.europa.eu/catalogue/distribution/dcat-ap-version-11

cuando se consideró oportuno. Por tanto la reutilización de Metashare-owl garantiza la compatibilidad de nuestro modelo con DCAT.

3. Nuestra ontología inicial, basada en Metashare.owl, se ha reducido para incluir sólo las entidades (y sus axiomas asociados) que se usan en ELRC-Share y que son obligatorios en dicho catálogo. De esta manera manera se ha construido un modelo a la vez sencillo y manejable pero conteniendo la información esencial necesaria para modelar recursos lingüísticos.

4. Se han añadido a nuestro modelo algunas entidades que corresponden a campos que se documentan en el portal de ELRC-Share pero que no están en el modelo inicial Metashare.owl, como es el campo "ISLRN" (*International Standard Language Resource Number*). También se han incluido algunos campos no obligatorios en ELRC pero que, por su interés, conviene documentar también en nuestro modelo, como por ejemplo el campo "conforme a los estándares".

5. Finalmente, la ontología ha sido documentada y publicada utilizando la herramienta Ontoology[11].

Los principales criterios de diseño para nuestro propio modelo de metadatos de recursos han sido la interoperabilidad y la simplicidad. Nuestra ontología es, en la práctica, un subconjunto de ELRC-Share, que a su vez es un subconjunto de Meta-Share. Entendemos nuestro modelo y nuestro sistema como un "punto de entrada" para identificar y documentar recursos lingüísticos con cierta información básica utilizable por personas no especialistas. Si se desea documentar un recurso con un mayor nivel de detalle, los metadatos asociados al mismo obtenidos por nuestra plataforma serán exportables a otros modelos como Meta-Share, en cuya plataforma podría continuarse un trabajo de documentación más exhaustivo. Este aspecto, que queda fuera del alcance de este proyecto, será objeto de estudio por parte de la red ReTeLe.

Así mismo, hemos tratado de mantener el modelo lo más general posible para dar soporte no sólo a recursos de traducción sino a cualquier tipo de recurso lingüístico, con el fin de mantener nuestro catálogo dentro de los objetivos más generales de la red ReTeLe.

[11] http://ontoology.linkeddata.es/

Nuestra ontología de recursos lingüísticos, resultante del proceso anteriormente descrito, se ha denominado "**Retele-Share**". A dicha ontología se le ha asignado un identificador permanente mediante el servicio W3ID[12] respaldado por el W3C:

https://w3id.org/def/retele-share

La ontología y su documentación asociada han sido alojadas en un servidor de la UPM, si bien el servicio W3ID permite su redirección a cualquier otra ubicación en la Web si se estima oportuno. El identificador o URI anterior es "dereferenciable" y se ha implementado la negociación de contenidos, de modo que si se accede a dicha URI desde un navegador web se muestra la documentación asociada a la ontología para consumo humano, mientras que si es un agente software el que accede a dicha URI, se le entrega la información en RDF (*Resource Description Format*). El siguiente enlace contiene la versión RDF de la ontología:

http://ontoology.linkeddata.es/publish/retele-share/ontology.ttl

3.2 MANUAL DE DESCRIPCIÓN DE RECURSOS SEGÚN EL MODELO

Una descripción detallada de todas las entidades de la ontología Retele-share está disponible en: http://ontoology.linkeddata.es/publish/retele-share/index-en.html.

Los recursos se han clasificado en tres grandes bloques: corpus, gramática o modelo, y recurso léxico-conceptual.

a. De manera amplia, se entiende por **corpus**, un conjunto de documentos o textos en una o varias lenguas. Algunos ejemplos son:

   - Conjunto de documentos de la administración oficial (por ejemplo, decisiones de ministerio, actos legales, etc.).
   - Conjunto de documentos de un periódico, revista, boletín, blog de artículos,etc.
   - Corpus paralelo (por ejemplo, documentos de traducción con el documento original, sin procesar).
   - Memorias de traducción (por ejemplo, documentos de traducción alineados con sus documentos originales).

[12] https://w3id.org/

b. En segundo lugar están las **Gramáticas o modelos,** que incluyen aquellos recursos que describen la lengua, es decir,

- Gramáticas (por ejemplo, un conjunto de reglas que formalizan una lengua).
- Modelos de lenguas y de traducción que contengan información estadística que asigne una probabilidad a una parte de un texto no visto (basado en un entrenamiento de datos).

c. En tercer lugar están los **Recursos léxico conceptuales,** que engloban recursos diversos:

- Léxico terminológico, glosarios, etc., incluyendo listas de términos con alguna otra información (definiciones, ejemplos, traducciones equivalentes, información lingüística...) o sin ella.

4 METADATOS EN RDF EN FORMATOS COMPATIBLES CON DATOS.GOB.ES Y CON CEF.AT

4.1 SELECCIÓN DE RECURSOS DISPONIBLES PARA SU REUTILIZACIÓN EN SISTEMAS DE TRADUCCIÓN AUTOMÁTICA Y DESCRIPCIÓN FORMAL DE LOS MISMOS (EN FORMATO COMPATIBLE META-SHARE Y ELRC-CEF .AT).

Como se ha discutido en el apartado 2, la información recopilada hasta el momento sobre recursos disponibles para traducción automática es escasa, y en el momento de escribir este documento ninguna de las instituciones y responsables entrevistados ha documentado sus recursos en el catálogo. Aun habiendo concluido este proyecto, está en la intención de los miembros de ReTeLe el mantener abierto y disponible el catálogo de recursos al menos durante la duración de la propia red, y comunicar a la SETSI una versión actualizada del listado de recursos siempre que se le solicite.

Por el momento la lista de recursos lingüísticos disponibles en el catálogo provienen en su mayoría de los asistentes al taller ReTele, si bien tienen su origen principalmente en entornos académicos y no en las administraciones públicas.

4.2 SOFTWARE QUE GENERA RDF A PARTIR DEL CATÁLOGO

Se han definido un conjunto de *mappings* y *scripts* para la conversión de los datos del catálogo desde su formato inicial, basado en el modelo de base de datos relacional que da soporte al portal web (ver apartado 5), al formato RDF siguiendo el modelo Retele-Share.owl (https://w3id.org/def/retele-

share). Dicha conversión se lleva a cabo mediante el sistema Morph-RDB[13] desarrollado por el *Ontology Engineering Group*. El formato resultante es compatible con datos basados en otros modelos (ver apartado 3) con el fin de garantizar la interoperabilidad entre ellos. No obstante la conversión de un formato a otro (mediante la definición de nuevas reglas de *mapping* y sus correspondientes scripts de conversión) queda fuera del alcance de este proyecto.

4.3 *END-POINT* EN RDF CON LOS METADATOS DE RECURSOS PARA TRADUCCIÓN AUTOMÁTICA.

Si bien la lista de recursos disponibles en nuestro catálogo es reducida por el momento (por las razones descritas en apartados anteriores) la infraestructura semántica se ha montado igualmente con el fin de dar soporte a los metadatos que pudieran llegar en el futuro.

El portal semántico y su SPARQL *endpoint* está basado en la plataforma Virtuoso y será accesible tanto desde el portal del catálogo http://catalogo.retele.linkeddata.es/en como en la siguiente dirección: http://linguistic.linkeddata.es/retele-share

4.4 VEINTE CONSULTAS IMPLEMENTADAS EN SPARQL PARA RECUPERAR DATOS DEL *END-POINT*

Tanto el conjunto de consultas predefinidas al catálogo como el acceso al propio SPARQL *endpoint* para llevar a cabo cualquier consulta libre estarán disponibles en la dirección http://linguistic.linkeddata.es/retele-share/sparql-editor/

5 PORTAL WEB

5.1 DISEÑO DE UNA INTERFAZ WEB DE RECOGIDA DE INFORMACIÓN DE RECURSOS LINGÜÍSTICOS PARA LA TRADUCCIÓN AUTOMÁTICA

Se ha diseñado un portal web para la recogida de metadatos de recursos lingüísticos y su posterior consulta. Está accesible tanto por medio del portal de ReTeLe (http://retele.linkeddata.es/): como directamente en la dirección http://catalogo.retele.linkeddata.es/

Para poder operar con el sistema y dar de alta un recurso, el usuario se ha de haber registrado en el sistema. Una vez accede a la opción de "documentar" un recurso, debe seleccionar el tipo del mismo,

[13] https://github.com/oeg-upm/morph-rdb/

de entre los tres grandes grupos referidos en el apartado 3.2: corpus, gramática o modelo y recurso léxico-conceptual. Una vez hecho eso, se abre un formulario (ver figura 1) para introducir toda la información que describe al recurso según la ontología ReTeLe-share (ver el apartado 4). Dicha información queda finalmente grabada en una base de datos mySQL al finalizar el proceso.

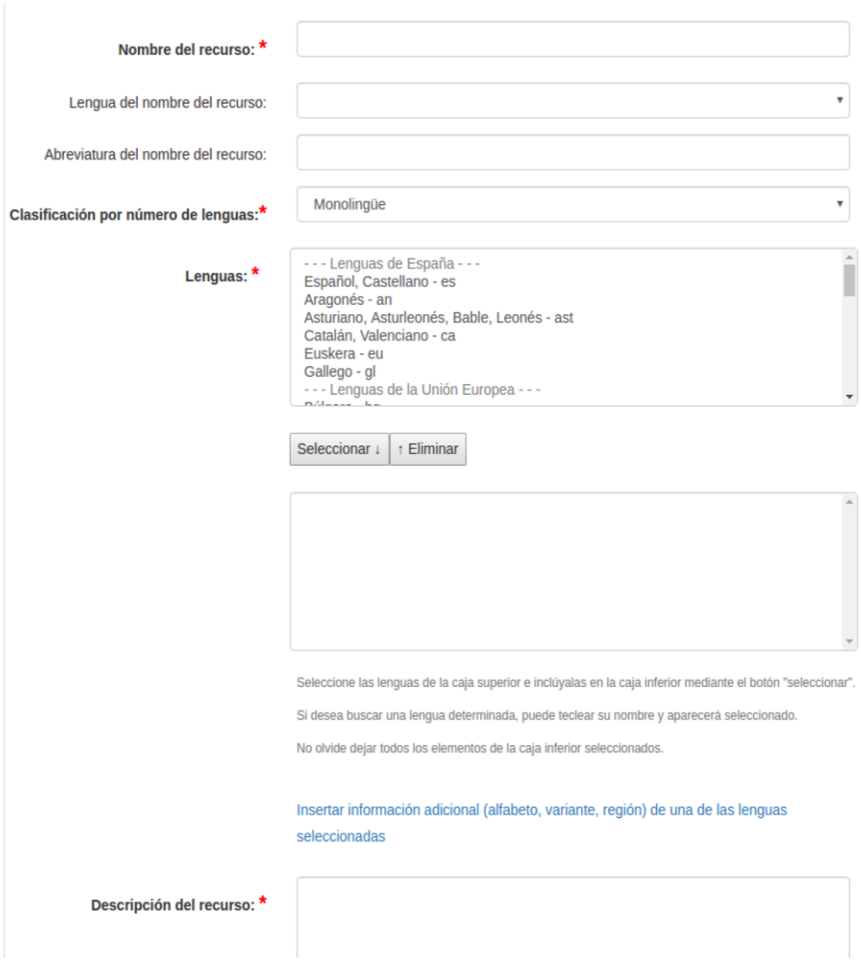


Figura 1: captura de pantalla de la interfaz web para la introducción de datos del recurso

5.2 DISEÑO DE UNA INTERFAZ WEB DE CONSULTA DE INFORMACIÓN DE RECURSOS LINGÜÍSTICOS PARA LA TRADUCCIÓN AUTOMÁTICA

El portal web del catálogo (http://catalogo.retele.linkeddata.es/) permite también la consulta y la modificación de los datos del mismo. La figura 2 muestra una captura de pantalla de una consulta que enumera la lista recursos disponibles. La figura 3 muestra un ejemplo de pantalla en la que se lleva a cabo la modificación de los datos de uno de ellos. No todos los usuarios pueden modificar los metadatos de todos los recursos sino sólo de aquellos de fueron creados por ellos mismos.

Nombre del recurso	Descripción	Estado		
CLUVI Parallel Corpus	Text type: literary, subtitling and specialized (legal, tourism, computing, science divulgation, economy) Register: formal	Publicable		Ver
Apertium RDF	Apertium RDF is the linked data version of the Apertium family of bilingual dictionaries. It covers 22 dictionaries currently.	Publicable	Modificar	Ver
Terminoteca RDF	This dataset makes available a set of multilingual terminologies as Linked Data on the Web. Terminoteca RDF aims to integrate different terminologies into a single unified graph and constitute a single entry point to them. In that way information coming from different sources and developed in isolation can now be traversed and searched in an easy way by following Semantic Web standards. Terminoteca RDF aggregates the following sources at this moment: 1. Terminesp is a terminological database in Spanish created by AETER (Asociación Española de Terminología) by extracting the terminological data from the UNE documents produced by AENOR (Asociación Española de Normalización y Certificación). It contains the terms and definitions used in the UNE Spanish norms and amounts to more than thirty thousand terms with equivalences in other languages whenever they are available. 2. Termcat is the centre for terminology in the Catalan language. It was established in 1985 by the Government of	Publicable		Ver

Figura 2: captura de pantalla de la consulta de recursos disponibles

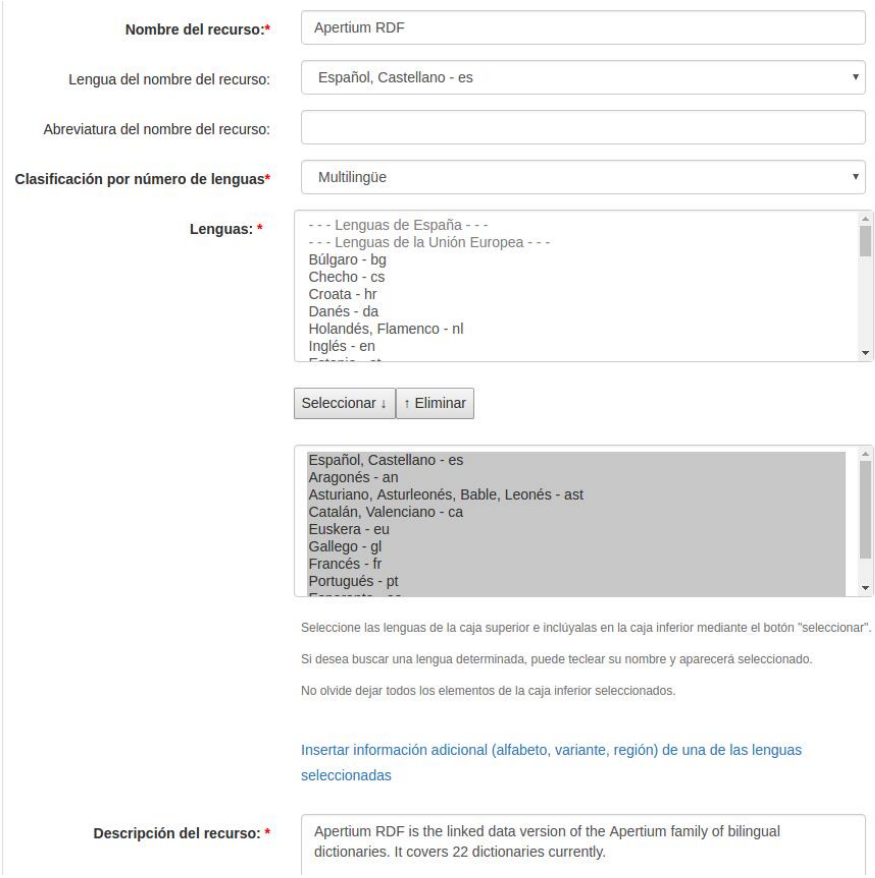


Figura 3: captura de pantalla de la funcionalidad de modificación un recurso

5.3 IDENTIFICACIÓN DE LAS TECNOLOGÍAS SUSCEPTIBLES DE FORMAR PARTE DE LA PLATAFORMA DEL PLAN EN EL EJE 3.1.

Con el objetivo de ayudar a los diferentes productores de recursos lingüísticos de la administración a generar metadatos de forma sistemática y, con el fin de garantizar la calidad de los mismos, se ha

desarrollado un portal Web (ver apartados 5.1 y 5.2) que permita su edición y modificación en base a los campos definidos en la ontología ReTeLe-share (ver apartado 3).

Mediante procesos de conversión automáticos (ver apartado 4) se puede generar la versión de dichos metadatos en RDF. A su vez el portal Web dará acceso a los mecanismos de consulta en SPARQL de dichos datos.

Todas estas herramientas serán candidatas a formar parte de la plataforma de procesamiento para la Administración Pública objeto del Eje 3.1 del Plan Nacional de Impulso.

6 TALLER DE METADATOS SOBRE RECURSOS PARA TRADUCCIÓN AUTOMÁTICA

La red ReTeLe organizó un taller que se celebró el día 13 de septiembre de 2016, como evento asociado al XXXII Congreso Internacional de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2016) (http://www.congresocedi.es/sepln). La información y el programa de dicho taller se pueden encontrar en http://retele.linkeddata.es/taller-retele-2016

El objetivo principal del taller era invitar a distintos grupos productores de recursos lingüísticos a presentar los recursos de procesamiento del lenguaje natural que han desarrollado. Este taller, además, pretendía discutir la necesidad de mantener un catálogo con metadatos para facilitar la ubicación y el uso de estos recursos y herramientas con el objetivo de incluirlos en el catálogo ReTeLe.

El taller se organizó en dos bloques: (1) breve presentación de recursos y (2) sesión práctica de metadatos estándares, en la que los participantes crearon metadatos que describen sus recursos y herramientas con el objetivo de incluirlos en el catálogo ReTeLe. Las instituciones que presentaron sus recursos fueron: el grupo *Ontology Engineering Group* de la Universidad Politécnica de Madrid, el grupo *Tecnologies dels Recursos Lingüístics* de la Universitat Pompeu Fabra, el Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante, el *Grup de Recerca Interuniversitari en Aplicacions Lingüístiques* de la Universitat de Barcelona, el grupo IXA de la Universidad del País Vasco, el Laboratorio de Humanidades Digitales de la UNED, el Departamento de Informática de la Universidad Carlos III de Madrid, Centro Singular de Investigación en Tecnoloxías da Información de la Universidade de Santiago de Compostela, el grupo SINAI de la Universidad de Jaén, el Grupo LyS de la Universidade da Coruña, el grupo TALG de la Universidade de Vigo, el grupo Transducens de la Universidad de Alicante y La Real Academia Española. Además, tuvieron lugar tres presentaciones:

"Ventajas de las licencias libres/de código fuente abierto para recursos lingüísticos" por Miquel Esplà-Gomis de Transducens-UA, "Editor de metadatos" por Elena Montiel Ponsoda del OEG de la UPM y "Estándares e interoperabilidad", por Núria Bel de TRL-UPF.

7 CONCLUSIONES

El estudio realizado confirma que hay dos tipos de posibles proveedores de recursos de traducción: organismos que externalizan la traducción y organismos que tienen servicios de traducción propios. Estos últimos, identificados por su participación en el taller del consorcio ELRC, fueron los organismos contactados, primero mediante una encuesta online que debían contestar con información sobre la existencia o no de los datos de traducción, y después, a los que contestaron se les hizo una entrevista presencial para identificar las condiciones de los datos de traducción.

De los que encargan traducciones a agencias de traducción externas pudimos observar que los departamentos que piden servicios de traducción son frecuentemente: la policía (estatal y de comunidades autónomas), la administración de justicia, también local, y los organismos que publican información turística. Frecuentemente los servicios incluyen traducción e interpretación, así como corrección lingüística.

De los que tienen servicios de traducción propios, ya conocíamos el dato publicado en el informe *Libro blanco de la traducción e interpretación institucional* en España (2011)[14]. Menos de un 8% de los traductores de la administración pública en España usan herramientas de traducción asistida por ordenador, y en general la traducción se produce de forma convencional. Estos datos se confirmaron con las entrevistas y las encuestas realizadas a las que, como ya se ha mencionado, contestaron solamente 10 de los 25 organismos contactados. Esta baja respuesta confirma que no hay una organización de los servicios de traducción, si no que los traductores trabajan de forma aislada, sin crear, en la mayoría de casos, archivos de datos de traducción que puedan ser reutilizados fácilmente.

De las encuestas respondidas se extrajeron los siguientes datos:

- 6/10 departamentos archivan los documentos y sus traducciones en formatos del procesador de textos.

[14] http://ec.europa.eu/spain/pdf/libro_blanco_traduccion_es.pdf

- 4/10 dicen tener documentos de más de 300 palabras por documento, y solamente 1 de ellos dice tener una base de más de 50.000 documentos
- 4/10 dicen que los documentos que almacenan son o contienen datos confidenciales. Nótese que son los mismos que en el caso anterior, con lo que la disponibilidad de los archivos queda comprometida.

Únicamente 3/10 departamentos trabajan con herramientas de traducción asistida, de los cuales:

- Solamente 2 centralizan los archivos de memorias de traducción
- Solamente 1 utiliza el estándar de memorias de traducción TMX
- En ningún caso se dispone de protocolos para mantener y actualizar las memorias de traducción.

Proyectada esta información en el Modelo de Madurez mencionado antes (ver Tabla 1) supondría que: de las 10 respuestas recibidas, situaríamos 4 departamentos en el nivel 1 de madurez, 3 en el nivel 2 y 3 en el nivel 3.

La administración pública produce datos de traducción y efectivamente puede ser un potencial proveedor de datos reutilizables. No obstante, es importante primero que las organizaciones mismas sean conscientes de que están produciendo datos valiosos y, segundo, que organicen sus procesos internos con protocolos que estandaricen y centralicen los datos de traducción producidos.

Es indudable que todas las traducciones que se hagan dentro de la administración pública, bien sea interna o externamente, forman un conjunto de recursos lingüísticos y terminológicos (ya sean corpus bilingües, paralelos, glosarios, diccionarios, etc.) de un valor inestimable. Sin embargo, uno de los problemas con el que nos hemos encontrado es que gran parte de los representantes de la administración española entrevistados no son realmente conscientes del valor añadido que dichos recursos aportarían. Consiguientemente, la mayor parte de dichos recursos no están clasificados de forma que se puedan reutilizar fácilmente a corto plazo, para los fines que se pretende, es decir, para alimentar sistemas de traducción automática.

Si tomamos como referencia el modelo de madurez (Bel et al. 2016) descrito en el apartado 2 de este informe, y la explicación recogida en el apartado 2.4, puede pensarse que una de las causas sea la escasa atención que tradicionalmente se ha prestado a la organización de los departamentos de traducción, y en general a los traductores, por parte de la administración pública en España, salvo en

el caso de la Oficina de Interpretación de Lenguas del Ministerio de Asuntos Exteriores. No han recibido formación ni información sobre las tecnologías que pueden hacer más eficiente el proceso de la traducción profesional. Un ejemplo más es que entre los entrevistados, algunos habían oído hablar del sistema PLATA pero nadie lo utilizaba. No así en las instituciones y organismos autonómicos que han necesitado traducir una gran cantidad de documentos a las otras lenguas del estado español y que han financiado y disponen de sistemas de traducción asistida y de traducción automática desde hace años.

De acuerdo con dicho modelo, son pocos los departamentos estudiados que pasan del nivel 1 de madurez, es decir, los documentos originales y las correspondientes traducciones no se archivan como documentos relacionados. Por consiguiente, la recuperación de dichos documentos para su reutilización en sistemas de TA representaría una tarea ingente y costosa en tiempo, esfuerzos personales y económicos.

En este sentido, ya se cuenta con un modelo a seguir, como son los recursos multilingües desarrollados por la Comisión Europea, disponibles como datos abiertos para uso público y que, en este momento, se están utilizando como base para la mayoría de sistemas estadísticos de traducción automática que son entrenados, entre otros textos, con las memorias de traducción de la DGT.

La gestión de la organización del proceso de traducción, las características de los archivos de los recursos generados y la posibilidad de ofrecerles una infraestructura para que puedan implementarlos sin costes adicionales determina la eficiencia y la eficacia con que esos materiales producidos puedan convertirse en datos reutilizables.

8 OBSERVACIONES PARA EL FUTURO

Para finalizar apuntaremos algunas observaciones de cara al futuro. Realizar un inventario de recursos lingüísticos de la administración pública española para fomentar su reutilización en los sistemas de traducción automática es un paso fundamental si se quiere mantener la presencia del español como lengua de comunicación en Europa. Si no se provee de recursos a los sistemas de TA, no se traducirá de o al español.

A la vista de lo dicho anteriormente, dos son las medidas más urgentes que cabe plantearse. Por un lado, es necesario concienciar a los responsables de los organismos de la Administración Pública española del valor añadido que tienen los documentos que producen, ya que son recursos imprescindibles para los sistemas de traducción automática. Es muy importante que comprendan

que, además, la reorganización de los procesos de traducción para no perder los recursos que se generan contribuye a mejorar los mismos servicios de traducción en términos de una mayor eficiencia y calidad. Por otro lado, es sumamente necesario proporcionar formación a los traductores de la administración y prestarles apoyo para diseñar conjuntamente el proceso de recogida, clasificación y organización de los datos minimizando su impacto para los individuos. Creemos que es prioritario convencer a los traductores de los beneficios de este sistema. El hecho de dedicar un tiempo a aprender a manejar con soltura estas herramientas de traducción revertirá indudablemente en la obtención de mejores resultados a largo plazo, y al hecho de contar finalmente en la administración pública con recursos que sean reutilizables. Por ello, consideramos que esta formación especializada para traductores o para otros profesionales de la administración que hagan estas tareas de traducción se podría incluir como módulos en los cursos que imparte la Escuela de Administración Pública.

Parece imprescindible proporcionar herramientas de asistencia a la traducción en forma de servicio, mediante una aplicación web que contribuiría a minimizar el impacto en los sistemas actuales al tiempo que solucionaría la recogida de documentos traducidos en forma de memorias de traducción (con los estándares habituales TMX, TBX) que estarían almacenadas de forma centralizada. La aplicación web permitiría también generar metadatos para una mejor gestión de las memorias de traducción. La recogida y almacenamiento centralizados permitirían organizar tareas para una mejor reutilización: la anonimización y la asignación de licencias (según la información recogida y niveles de confidencialidad).

Así pues recomendamos la creación de una infraestructura que apoyara la producción de datos de traducción y que debería tener las siguientes características:

- Herramientas de traducción asistida por ordenador como un servicio a través de una aplicación web mantenida centralmente. Esta aplicación podría dar acceso, mediante APIS, a servicios asociados como traducción automática y bases de datos terminológicas.

- Un archivo común (asociado a la herramienta anterior) donde las memorias producidas por los traductores pudieran procesarse para la creación de recursos documentados y preprocesados:

    - generación de metadatos a partir de los datos del traductor y su organización,

- anonimización de datos personales susceptibles de crear problemas de privacidad y confidencialidad,
- generación de recomendaciones sobre la distribución del recurso y licencia adecuada del recurso generado en ese caso.

- Asimismo cabe señalar la necesidad de proporcionar a los traductores una mayor preparación en la gestión de la documentación ya que eso finalmente revierte en procesos más optimizados.
- En relación con la producción de documentos, es importante, como ya se ha apuntado previamente que los traductores sean conscientes de los beneficios que tanto ellos como otros traductores de otros departamentos pueden obtener en un futuro a medio y largo plazo.

9 GLOSARIO DE SIGLAS Y ACRÓNIMOS

AAPP	Administraciones Públicas
API	Application Programming *Interface*
CCAA	Comunidades Autónomas
CVP	*Common Procurement Vocabulary*
DGT	Dirección General de Traducción
DTIC	Dirección de Tecnologías de la Información y las Comunicaciones
ELRC	European Language Resource Coordination
LD4LT	*Linked Data for Language Technologies*
LYS	*Language in the Information Society*
OEG	*Ontoly Engineering Group*
OIL	Oficina de Interpretación de Lenguas
RDF	*Resource Description Framework*
ReTeLe	Compra Pública de Innovación
SEPLN	Sociedad Española para el Procesamiento de Lenguaje Natural
SETSI	Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información
SINAI	Sistemas Inteligentes de Acceso a la Información
SPARQL	*SPARQL Protocol and RDF Query Language*
TA	Traducción Automática

TALG	Tecnologías y Aplicaciones para la Lengua Gallega
TBX	*Term-Base Exchange*
TMX	*Translation Memory Exchange*
TRL-UPF	Tecnologías de los Recursos Lingüísticos- Universitat Pompeu Fabra
UE	Unión Europea
UNED	Universidad Nacional de Educación a Distancia
UPM	Universidad Politécnica de Madrid
W3C	*World Wide Web Consortium*

10 ANEXOS

10.1 ANEXO 1: Modelo de carta de invitación a la realización del cuestionario

Desde la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información aprobamos en octubre de 2015 el Plan de Impulso de las Tecnologías del Lenguaje con el objetivo de fomentar el desarrollo de estas tecnologías en nuestro país. Su correcto desarrollo es de gran importancia puesto que suponen un habilitador para la prestación de avanzados servicios digitales.

En este sentido, disponer de fuentes de datos de calidad es un elemento básico. Es por ello que este plan de impulso incluye entre sus medidas una línea de actuación orientada a aprovechar el marco de la política Reutilización de la Información del Sector Público (RISP) para generar datos abiertos de interés para las tecnologías del lenguaje.

Las Administraciones públicas disponemos de una gran cantidad de información que tiene un enorme valor para el desarrollo de esta tecnología y sus productos y servicios, especialmente en el ámbito de la traducción automática. Muestra de este valor es que el elemento más descargado del portal de datos abiertos de la UE es el corpus paralelo de traducciones de la Dirección General de Traducción de la Comisión Europea.

Además, hay que señalar que entre los potenciales beneficiarios de estos recursos se encuentran las propias plataformas de traducción automática para las Administraciones Públicas que están implantando tanto la Unión Europea como la Administración General del Estado, entre otras administraciones.

Es por ello que, para facilitar el desarrollo de estas aplicaciones, se ha encargado a la red de excelencia ReTeLe (Red de Recursos para las Tecnologías de la Lengua) la elaboración de un inventario de potenciales recursos para traducción automática.

Por consiguiente, me dirijo a usted para solicitarle, en el marco que permiten las normas legales aplicables, dentro de los ámbitos subjetivo y objetivo de aplicación de la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público; y sin perjuicio del régimen aplicable al derecho de acceso a los documentos y a las especialidades previstas en su normativa reguladora, su colaboración en la realización de este inventario.

Agradeciendo su interés,

David Pérez Fernández.

Coordinador de área de tecnología

Recogida de información

Conscientes de la variedad de niveles de disponibilidad de los materiales traducidos que se producen en la administración, y con el objetivo de diseñar eficientemente la recogida de información, le pedimos que contesten a este breve cuestionario. En caso de que no pueda contestarlas fácilmente, por favor marque la opción "Pendiente de determinar" y podrá suministrar la información más tarde, durante una entrevista con los técnicos que considere que pueden aportarla. Para ello, le agradeceremos que nos indique una persona de contacto que pueda recibirnos y ayudarnos a recoger los datos necesarios.

Acceso a encuesta

https://docs.google.com/forms/d/1AbE42RSzRie8grqmHyN25r273EYvKoLYLhiSXoRUV2w/viewform

Contacto: montserrat.marimon@upf.edu

10.2 ANEXO 2: Carta de recordatorio para la realización del cuestionario

Estimad@ XXXX

La Secretaría de Estado de Telecomunicaciones y Sociedad de la Información (SETSI), ha lanzado una iniciativa, en el marco del "Plan de Impulso de las Tecnologías del Lenguaje", con el fin de recopilar información sobre los recursos lingüísticos ya existentes en la administración española, que pueden resultar útiles para la traducción automática (documentos y sus traducciones, diccionarios y terminologías multilingües) y facilitar esta tarea a diferentes organismos públicos. Por esta razón, nos ponemos en contacto con usted.

El objetivo es ofrecer dichos recursos como datos abiertos para promover su reutilización en el marco de la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público. La participación de su institución en esta iniciativa reflejará públicamente el compromiso de la misma con las directrices españolas (recogidas en el preámbulo de la ley 37/2007) y europeas (Directiva 2013/37/UE, que modifica la Directiva 2003/98/CE relativa a la reutilización de la información del sector público) y podrá servir como ejemplo de buenas prácticas en este ámbito.

En el taller “Coordinación de Recursos de Idiomas”, que tuvo lugar en Madrid el pasado 26 de enero, en el que usted participó, ya se puso de manifiesto, por parte de los asistentes, el gran interés de esta iniciativa para la administración pública. Para lograr este objetivo, la "Red de Excelencia de Recursos para Tecnologías de la Lengua" (ReTeLe, http://retele.linkeddata.es/) ha recibido el encargo de coordinar y ayudar a los posibles proveedores en la identificación de los materiales susceptibles de ser considerados recursos lingüísticos reutilizables. Con toda la información que se recoja, elaboraremos un informe de la situación de los recursos lingüísticos reutilizables y un análisis de la situación de los servicios de traducción de la administración como proveedores de datos.

Por todo ello, le solicitamos que, bien sea usted mismo o quien considere de su departamento, conteste a una breve encuesta online (duración aproximada 5 minutos), antes del 30 de junio de 2016.

Agradeciendo sinceramente su colaboración, quedamos a su disposición para cualquier aclaración o consulta.

Dña. Asunción Gómez Pérez,

Coordinadora de ReTeLe (TIN2015-68955-REDT)

10.3 ANEXO 3: Cuestionario sobre recursos para la traducción automática

Información sobre recursos para la traducción automática

Sección 1: Memorias de traducción

¿Utilizan los traductores programas de traducción asistida (SDL TRADOS, Deja Vu, Transit, MemoriaQ, OmegaT, etc.)?

- Sí
- No
- Pendiente de determinar

(En caso negativo: saltar a la sección 2.)
(En caso afirmativo: responder a las siguientes preguntas)

¿Centralizan la gestión de las memorias de traducción?

- No.
- Sí, usando un sistema específico.
- Sí, usando un depósito común (DropBox, Drive, disco compartido).
- Pendiente de determinar

En caso de que su respuesta sea afirmativa especifique qué sistema o depósito utilizan

¿Utilizan el formato TMX en las memorias de traducción?

- Sí
- No, utilizan otro formato
- Pendiente de determinar

Si no lo utilizan, ¿qué otro formato utilizan?

¿Utilizan algún sistema de control de versiones (CVS, Subversion, Git, Mercurial, etc.) para las memorias de traducción?

- Sí
- No
- Pendiente de determinar

En caso de que su respuesta sea afirmativa, especifique cuál

¿Existe algún protocolo de actualización de las memorias de traducción utilizadas por parte de la organización?

- Sí
- No
- Pendiente de determinar

Sección 2: Archivos de documentos y sus traducciones

¿Archivan de forma sistemática los documentos y sus traducciones?
- Sí
- No
- Pendiente de determinar

(En caso negativo, saltar a la sección 4)
(En caso afirmativo: responder a las siguientes preguntas)

¿Centralizan la gestión de los documentos y sus traducciones?
- No.
- Sí, usando un sistema específico.
- Sí, usando un depósito común (DropBox, Drive, disco compartido).
- Pendiente de determinar

En caso de que su respuesta sea afirmativa, especifique qué sistema o depósito utilizan

¿En qué formato se guardan los archivos de los documentos y sus traducciones?
- Procesador de textos: ODF (.odt), .doc, OOXML (.docx), RTF
- PDF
- texto sin formato (.txt)
- HTML
- Pendiente de determinar

¿Utilizan algún sistema de control de versiones (CVS, Subversion, Git, Mercurial, etc.) para los documentos y sus traducciones?
- Sí
- No
- Pendiente de determinar

En caso de que su respuesta sea afirmativa, especifique cuál

¿Existe algún protocolo de actualización de los documentos y sus traducciones por parte de la organización?

- Sí
- No
- Pendiente de determinar

Sección 3: Archivos de documentos

Por favor, identifique cual es el tamaño de los documentos con los que su organizaciçon trabaja

- Documentos de 300 palabras o menos (1 folio)
- Documentos de más de 300 palabras
- Pendiente de determinar

Por favor, identifique el número de documentos archivados

- más de 100
- más de 1000
- más de 50000
- Pendiente de determinar

Por favor, identifique si

- Los documentos y la traducción pertenecen al organismo o se dispone del copyright.
- Los documentos y la traducción son materiales publicados (web o papel).
- Los documentos y la traducción tienen restringido el copyright.
- Los documentos y la traducción no se pueden publicar porque contienen datos privados
- Los documentos y la traducción no se pueden publicar porque contienen información confidencial
- Pendiente de determinar

Por favor, señale los pares de lenguas de las traducciones:

- Español y catalán/valenciano
- Español y euskera
- Español y gallego
- Español e inglés
- Español y lenguas europeas
- Español y otras lenguas
- Pendiente de determinar

Sección 4: ¡Gracias por su participación!

Por favor, introduzca su nombre y la organización para la que trabaja:

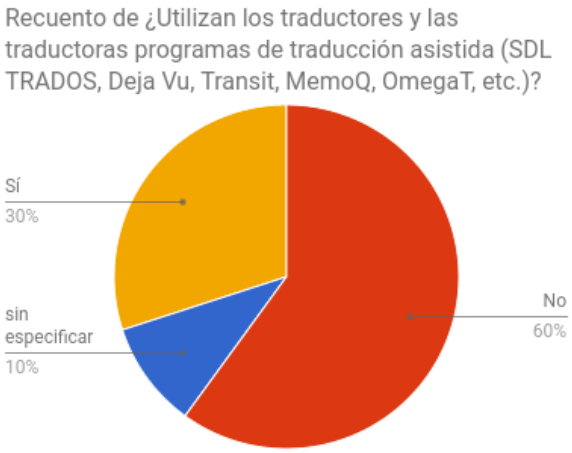
10.4 ANEXO 4: Relación de organismos a los que se envió carta

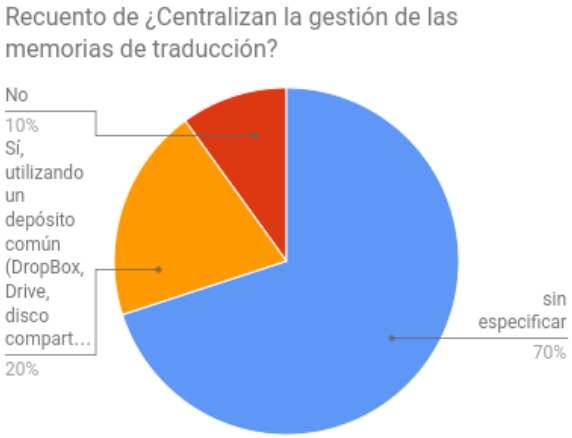
	Organismo
1	Agencia Estatal de Administración Tributaria
2	Biblioteca Nacional de España
3	Centro de Documentación Judicial - CGPJ
4	DG de Cooperación Jurídica Internacional y Relaciones con las Confesiones, Ministerio de Justicia
5	Dirección de Tecnologías de la Información y las Comunicaciones Administración General del Estado
6	Dirección General de Tráfico
7	FEMP
8	Generalitat de Catalunya
9	Gerencia de Informática de la Seguridad Social
10	Instituto Cervantes
11	Instituto Nacional de Estadística
12	IVAP. Gobierno Vasco
13	MINECO
14	Ministerio de Agricultura, Alimentación y Medio Ambiente
15	Ministerio de Defensa. Área de Patrimonio. Unidad de Normalización
16	Ministerio de Hacienda y AAPP DTIC
17	Ministerio de Industria, Energía y Turismo
18	Ministerio de la Presidencia
19	Ministerio del Interior
20	OIL - Ministerio de Asuntos Exteriores y de Cooperación
21	RED
22	Servicio Público de Empleo Estatal
23	Viceconsejería de Política Lingüística, Gobierno Vasco
24	Xunta de Galicia-Secretaría General de Política Lingüística
25	FECYT

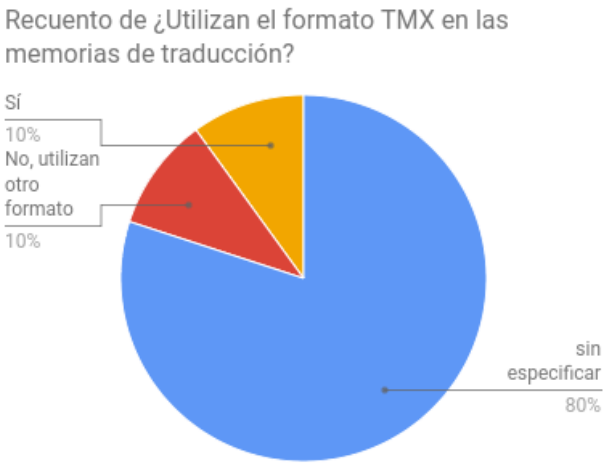
10.5 ANEXO 5: Relación de encuestas recibidas

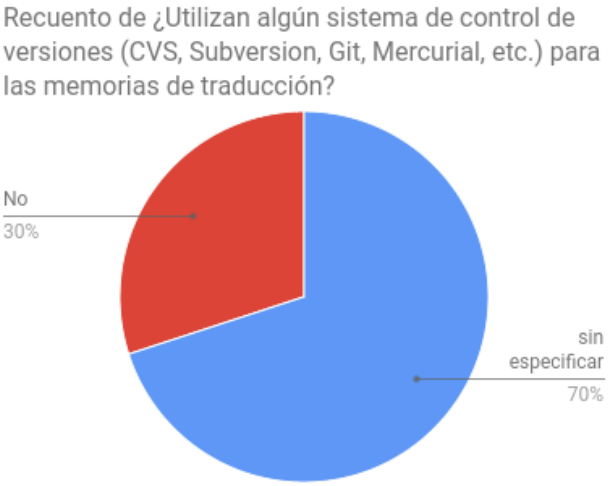
En el siguiente enlace se puede encontrar el detalle de los resultados de las encuestas realizadas: https://drive.google.com/open?id=0B6L1GIK9dcoCLWxKUklSZGlWVEk

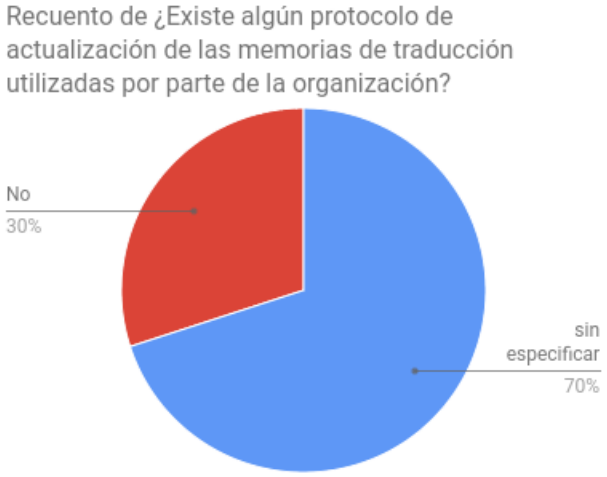
A continuación se resumen gráficamente algunos de los resultados obtenidos:

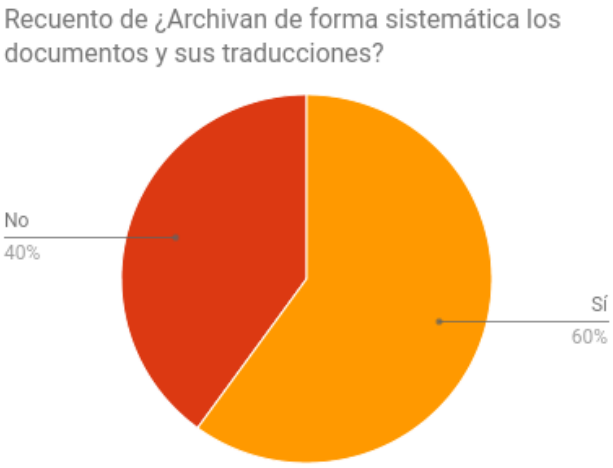




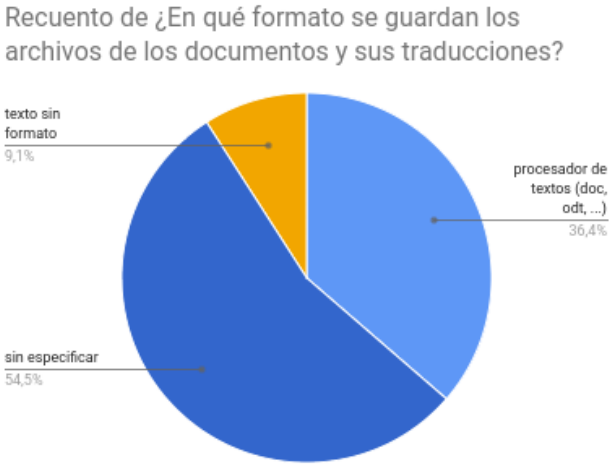

Recuento de ¿Utilizan el formato TMX en las memorias de traducción?
Sí
10%
No, utilizan otro formato
10%
sin especificar
80%

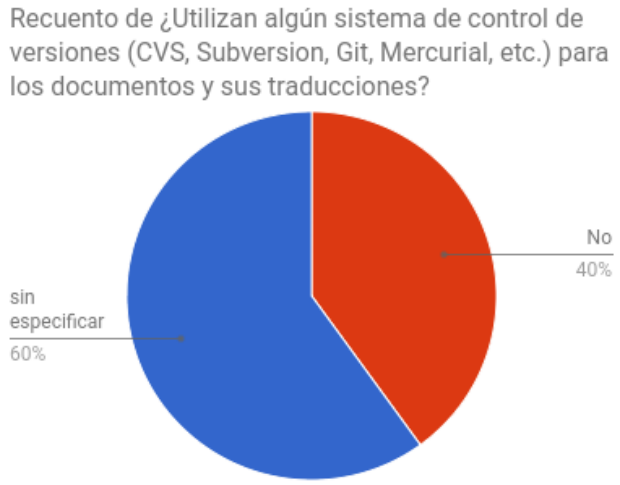

Recuento de ¿Utilizan algún sistema de control de versiones (CVS, Subversion, Git, Mercurial, etc.) para las memorias de traducción?
No
30%
sin especificar
70%

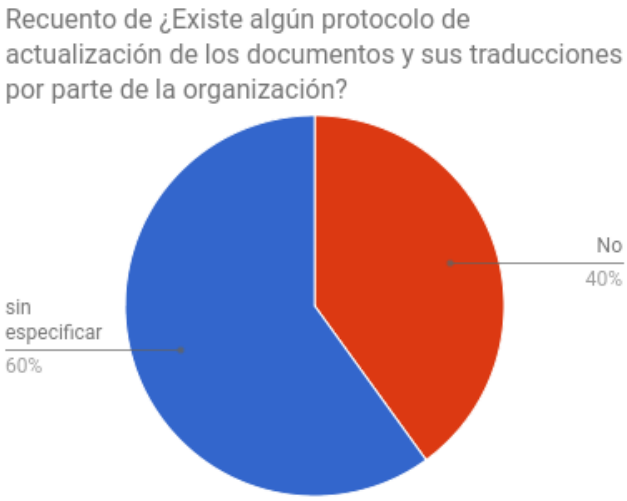

Recuento de ¿Existe algún protocolo de actualización de las memorias de traducción utilizadas por parte de la organización?
No
30%
sin especificar
70%

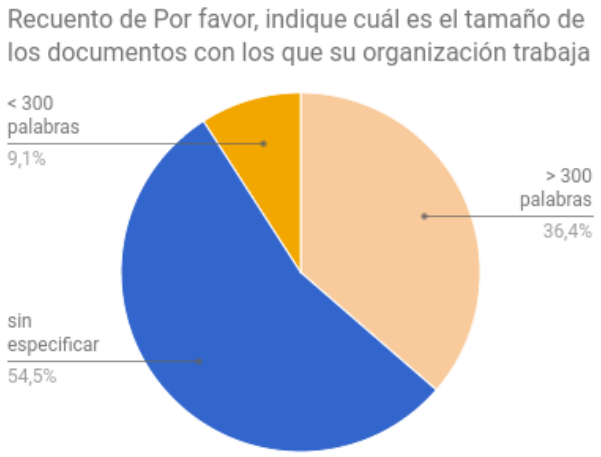

Recuento de ¿Archivan de forma sistemática los documentos y sus traducciones?
No
40%
Sí
60%

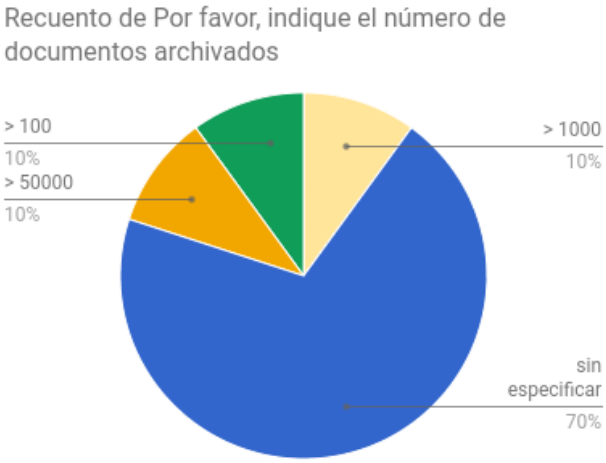

Recuento de ¿Centralizan la gestión de los documentos y sus traducciones?
No
Sí (sistema específico)
Si (servidor común)
sin especificar
Sí (disco compartido)
0
1
2
3
4
5


Recuento de ¿En qué formato se guardan los archivos de los documentos y sus traducciones?
texto sin formato
9,1%
procesador de textos (doc, odt, ...)
36,4%
sin especificar
54,5%


Recuento de ¿Utilizan algún sistema de control de versiones (CVS, Subversion, Git, Mercurial, etc.) para los documentos y sus traducciones?
No
40%
sin especificar
60%


Recuento de ¿Existe algún protocolo de actualización de los documentos y sus traducciones por parte de la organización?
No
40%
sin especificar
60%


Recuento de Por favor, indique cuál es el tamaño de los documentos con los que su organización trabaja
< 300 palabras
9,1%
> 300 palabras
36,4%
sin especificar
54,5%


Recuento de Por favor, indique el número de documentos archivados
> 100
10%
> 50000
10%
> 1000
10%
sin especificar
70%

10.6 ANEXO 6: Selección de AAPP contratantes de servicios de traducción

	Organismo
1	Dirección Ejecutiva de la Agencia Española de Consumo, Seguridad Alimentaria y Nutrición
2	Dirección del Servicio de Gestión Económica de la Agencia Estatal de la Administración Tributaria
3	Tesorería General de la Seguridad Social. Gerencia de Informática de la Seguridad Social
4	Dirección del Consorcio de la Comunidad de Trabajo de los Pirineos
5	Consejo de Administración de la Sociedad Estatal para la Gestión de la Innovación y las Tecnologías Turísticas, S.A. (SEGITTUR)
6	Paradores de Turismo de España, S.A
7	Consejería de Administraciones Públicas y Portavoz del Gobierno
8	Departamento de Empleo y Asuntos Sociales
9	Instituto Vasco de Administración Pública
10	EUSTAT - Instituto Vasco de Estadística
11	Consell Comarcal de la Conca de Barberà
12	Vicepresidència Econòmica, de Promoció Empresarial i d'Ocupació
13	Conselleria d'Hisenda i Pressuposts
14	Institut Balear de la Natura (IBANAT)

10.7 ANEXO 7: Listado organismos entrevistados

Organismo
Dirección de Tecnologías de la Información y las Comunicaciones (DTIC) del Ministerio de Hacienda y AAPP
OIL del Ministerio de Asuntos Exteriores
Servicio de Traducción del Ministerio de Justicia.
Red.es
Subdirección General de Tecnologías y Servicios de Información. Subsecretaría. Ministerio de la Presidencia

10.8 ANEXO 8: Modelo de madurez

En las páginas siguientes se adjunta el modelo de madurez de los recursos de traducción en las administraciones públicas, en la versión que fue publicada en la conferencia Meta-Forum 2016 (en inglés)

A Maturity Model for Public Administration

Open Translation Data Providers

Núria Bel (Universitat Pompeu Fabra)

Mikel L. Forcada (Universitat d'Alacant)

Asunción Gómez-Pérez (Universidad Politécnica de Madrid)

Abstract

Any public administration that produces translation data can be a provider of useful reusable data to meet its own translation needs and the ones of other public organizations and private companies that work with texts of the same domain. These data can also be crucial to produce domain-tuned Machine Translation systems.

The organization's management of the translation process, the characteristics of the archives of the generated resources and of the infrastructure available to support them determine the efficiency and the effectiveness with which the materials produced can be converted into reusable data.

However, it is of utmost importance that the organizations themselves first become aware of the goods they are producing and, second, adapt their internal processes to become optimal providers. In this article, we propose a Maturity Model to help these organizations to achieve it by identifying the different stages of the management of translation data that determine the path to the aforementioned goal.

1. Introduction

Translated documents and their original source texts are becoming precious resources whose availability can bring about significant translation cost savings when large quantities of texts have to be translated. Existing translations are the raw material to feed *computer-assisted translation* (CAT) tools and *machine translation* (MT) systems to boost professional translation processes. An increasing demand for such resources has already been observed. However, the last survey by LT-Observatory[1] reports that while 83% of language service providers in Europe expects more language resources, in particular translation data, to become available in the next years, only a 15% is ready to pay for them.

Most likely, these expectations are based on two facts. On the one hand, the web contains large quantities of data that can be harvested and reused. The indexed web contains about 4,590 million pages,[2] with many being in multilingual sites, therefore constituting potential translation

---

1 http://ltinnovate.blogspot.com.es/2016/04/the-future-of-language-resources-for.html

2 According to www.worldwidewebsize.com at 15-05-16.

data. On the other hand, public administration organisms, such as the translation services of the European Commission, are producing translation resources that are made available for public reuse. In 2005, for instance, the EU institutions translated 2,861 million pages, according to a special report[3] written to assess the cost of translation in the EU. It represents about 70 % of the total EU translation volume.[4] Note that the expenditure for translation in 2005 amounted to 511 M€. The situation has surely changed because from 2005 to now, as Europe has moved from 15 to 24 languages.

For the first source of translations, the web, a well founded-study conducted for the project QTLaunchPad[5] (Tsiavos et al. 2015) notices that although technically is not only possible, but even easy thanks to the availability of specialized crawlers for finding multilingual webs already available, reusing unauthorized crawled resources implies legal risks, mostly related with the copyrights of web published texts. The report recommends as mitigation strategy to ask for permission to every crawled web that has not published the license conditions about its contents. Arranz and Hamon (2012), in the framework of the PANACEA project, described, however, how costly it can be to contact every provider to ask a license: on average it took 176 days to negotiate the permission for reusing web contents with actual copyright holders. Other proposals for avoiding legal issues are presented in Forcada et al. (2016).

None of these problems arise for the resources provided by public administrations, if, as in the case the of the translation memories of the European Commission Directorate General for Translation (DGT), they are published as open data. Note that this resource is the most downloaded resource of the European Open Data Portal (which contains more than 8000 different datasets). Open data, by definition, guarantees availability, access, copyright and reuse permissions, in all conditions, also with commercial purposes. Moreover, Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the reuse of information of the public sector ensures that this sharing policy extends to all public administrations of the EU Member States. In fact, the European Language Resource Coordination initiative is already working for Member States to realize this possibility and thus to contribute to the creation of *automated translation* services in the framework of the Connecting Europe Facility programs (CEF.AT). The proposal is to mine selected organizations to extract language resources from them.

The question that this invitation to mine language resources in public administration organizations raises is: what would be the cost of the exercise of identifying, gathering and making available these resources? Obviously the cost can be minimal if we talk about translation memories which are already compiled and formatted. However, the use of CAT tools in the administration might be not as spread as it might seem. For instance, the case in Spain is that according to the *Libro Blanco de la traducción y la interpretación institucional*[6] ('White Paper on Institutional Translation in Spain') only less than the 8% of the translators in these organizations works with CAT tools. What is, then, the cost of identifying, collecting and preparing translation data from the rest of potential translation data providers?

---

[3] Special Report No 9/2006 of the European Court of Auditors concerning translation expenditure incurred by the Commission, the Parliament and the Council (2007/2077(INI))

[4] Official Journal of the European Union, C 284/1 (http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52006SA0009)

[5] http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-4_5_1_0.pdf

[6] http://ec.europa.eu/spain/pdf/libro_blanco_traduccion_es.pdf

In the framework of the recently approved Spanish *Plan de Impulso de las Tecnologías del Lenguaje*[7] ('Language Technologies Support Program'), the *Recursos para las Tecnologías del Lenguaje* ('Resources for Language Technology') network of excellence[8] has been commissioned to conduct an study to identify Spanish public administration organizations that produce translation data to produce a catalogue of potential open data providers. An extensive survey is currently being carried out. In this paper we present the criteria to determine both what are the characteristics that actual translation related resources have and to what extent they can hamper its identification, collection and preparation, in order to identify the organizations that might become providers of a continuous supply of data. We suggest that these criteria can be structured as a path to follow for organizations to become regular translation data providers.

In what follows we address, in section 2, some technicalities about translation resources, in section 3, we propose a maturity model for organizations with a primary focus on Public Administration bodies. Some conclusions are provided in section 4.

2. Technicalities about language resources for CAT and SMT

Non-experts might have difficulties to realize the characteristics of the translation resources required to feed CAT tools or to build MT systems in particular. In what follow, we review most relevant aspects about these data that need to be taken into account when assessing the task of deciding about the benefit/cost balance of collecting resources.

The following translation data are of interest, but in what follows we mainly concentrate on documents either in the form of translation memories or corpora.

- Translation memories: linguistic databases that capture translations made by humans. They can be used to facilitate future translations tasks but also for training automated translation systems.
- Corpora: monolingual and multilingual corpora, comparable, aligned, parallel documents, etc.
- Lexica: monolingual and multilingual lists of words, multi-words, sentences, etc. in general or specific subject fields.
- Terminological resources: structured sets of concepts, with associated linguistic information in a specific subject field.

One of the musts for reusability is that data collections can be downloaded as a whole and freely processed, as stated by the Full Open Data definition,[9] including commercial purposes. There are "accessible" materials that are useless.[10]

Other required technical details to be taken into account are the following.

---

[7] http://www.agendadigital.gob.es/planes-actuaciones/Paginas/plan-impulso-tecnologias-lenguaje.aspx

[8] http://retele.linkeddata.es/ (founded by the Spanish Ministerio de Economía y Competitividad, TIN2015-68955-REDT)

[9] http://opendefinition.org/od/2.1/en/

[10] http://index.okfn.org/dataset/legislation/

2.1. Size

The magnitude of the resources required for building MT systems that translate with a quality that makes it worth their use for professional translation is a common topic in technical discussions. The easy answer is "the more, the best", which is initially the truth, but a threshold should be pointed out when assessing a potential repository.

Most standard MT datasets contain tens or hundreds of millions of words (Irvine & Callison-Burch, 2013). There are techniques to mitigate the need of large quantities of parallel text, but most often at the expense of resulting translation quality. As a reference of the magnitude we can take as a standard corpus the Common Crawl corpus (Smith et al. 2013) that in actual experiments like the ones by Song et al. (2014) contained 161 million words, 3,158,523 sentence pairs for the French–English language pair. Besides, again to guarantee quality, with a base of hundreds of millions, and very good data of a particular domain, about several hundreds of parallel sentences words (Pecina et al. 2014) can be enough to tune a system to achieve a good quality in a particular domain.

In the specialized literature, low-resource settings are considered to be those with parallel datasets of fewer than one million words. One million words are about 3,000 pages of parallel data: original and translated documents.

2.2. File format

Despite of the benefits that formats such as PDF might have for archiving and printing documents, documents in PDF cannot generally be directly processed. Note that in order to use them for CAT or MT system training, segmentation into sentences —or segments— is required for alignment. Plain, editable formats are thus required, and although conversion between formats is technically possible, there are still many problems when the source is a PDF file. Because the original source texts of PDF files were produced in other editable formats, it is highly recommended to use these sources for building a translation data repository.

2.3 Alignment.

Traditional document archiving protocols do not always require encoding different language versions. In the case they do, in order to process documents, the best practice is when the file name or path clearly identifies the relation between source and target documents and the language of the named document.

The effort for collecting documents and translations can significantly vary depending on the alignment information: a very disperse and undocumented or unlabelled collection of documents can demand a big effort, and therefore be completely prohibitive, in terms of effort to make it reusable.

The best option, however, is storing translation memories, as they constitute an already compiled collection of segments of the same topic related explicitly with their translations, and one that is aligned, in a particular, preferably, standard format.

2.4. Translation memories

A translation memory is a file where source texts and its translations are stored broken down into segments, which are aligned with their translations to form “translation units”. Segments are mostly, but not only, sentences. Currently, CAT tools include the programs that produce translation memories when a human translator types the translation and when two documents

are provided as source and target texts. These tools can also export in-built translation memories into a standard format, most commonly TMX: the *translation memory exchange* format.[11]

What are the contents in a translation memory file is decided by the user. A translation memory can store document-by-document translation units, or many documents belonging to a particular subject or domain can be stored into one single TMX file.

2.5. Metadata

Translation data packages need to be documented in order to be searchable by means of metadata. A standard metadata schema is already available[12] as suggested by the community and for these metadata to be assigned information has to be foreseen since the initial archiving step: languages, size, domain, character encodings as well as license of the resource or whether texts include contain private or confidential data are required as well as more clerical information: creation date, contact data of the responsible person, and whether there are associated documentation and resources.

3. Public Administration Open Translation Data Providers

Any public administration that produces translation data to meet its own translation needs is a potential provider of data for other organizations including private companies that work with texts of the same domain. The problem is that they are completely ignorant of the potentiality of the data they produce.

By producing translations (internally or by outsourcing), the administration produces language resources: data in the form of documents and their translations, translation memories, terminologies and bilingual or multilingual glossaries. The organization's management of the translation process, the characteristics of the archives of the generated resources and of the infrastructure available to support them determine the efficiency and the effectiveness with which the materials produced can be converted into reusable data. For this reason, here we focus on the organization itself, which can provide protocols for archiving documents (therefore, data) and requirements regarding the format, the associated information or metadata, the licenses and distribution restrictions because of confidentiality and privacy data. The existence of these protocols for the management of the translation process and the available infrastructure are the basis for developing objective criteria for assessing which organizations can be considered suppliers.

Here we propose a set of criteria for evaluating the potential cost of collecting and preparing these data. Criteria are provided as an organization maturity model to suggest guidelines to adapt the operations of administration offices for them to become translation data providers. This model offers, therefore, a clue about the elements that need to be improved by the organization for it to become a supplier of resources.

We take into account requirements based on the availability and readiness of the resources and their characteristics, as described in section 2. Note that, while any resource is potentially useful, the costs to effectively enable their reuse can be high, mostly because of the quantity of texts required, and, therefore, hinder the objective. Also it is worth to keep in mind that the

---

11 https://www.gala-global.org/tmx-14b

12 https://elrc-share.ilsp.gr/ELRC-SHARE_SCHEMA/v1.0/

investment in a common infrastructure (particular tools like anonymization tools, computer-assisted translation programs, assistance for licensing evaluation, treatment of private confidential data, etc.) may be a necessary requirement for an organization to effectively reach maturity stage 5, the highest in the scale that we explain below.

For our *translation data provider maturity model* we have taken into account the *capability maturity model integration* by Sally Godfrey (2008), which is shown in Figure 1 and Aymerich and Carmelo (2009) report on translation services at the PanAmerican Health Organization.

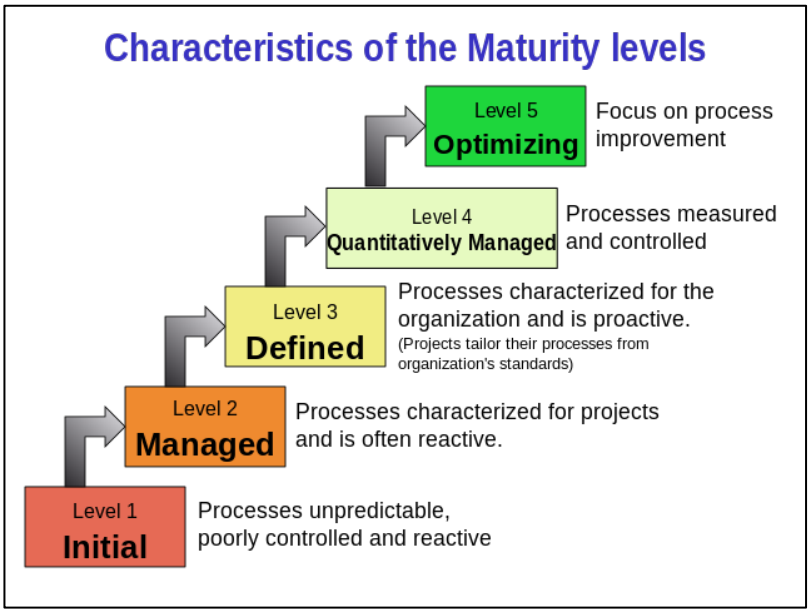


Figure 1: Capability Maturity Model Integration by Sally Godfrey (2008)

Level 1: Initial

At Level 1 or Initial ("Processes unpredictable, poorly controlled and reactive") one would find the organizations in which documents and their translations are not archived as a defined particular process. Translation data preservation is part of the general process of archiving documents, perhaps on paper or as a scanned document in PDF format, so that the relation between a document and its translation is not systematically registered and the registered copy is not processable. The organization has not been provided with tools that support the particular archiving of documents related to its translations.

At this level, linguistic resources could be reused, but only thanks to the competence and will of the people in the organization, but not as the result of to a proven process. The risk of considering these organizations as providers is that delivered materials might be inconsistent: it may be that not all translations are associated with the original document causing significant failures. In addition, organizations in maturity level 1 can commit themselves beyond their capabilities, or deliver resources of quality lower than expected, and therefore cannot guarantee the continuity of the process of compiling resources.

Level 2: Managed

At Level 2 ("Processes characterized for projects and often reactive"), the translation process has been given a particular space for archiving and it is managed by a protocol. Original texts and their translations are produced in editable formats: (.doc) Microsoft Office, OOXML (.docx), HTML, ODF (.odt), or plain text (.txt).

The *ad hoc* protocol only requires the identification (traceability) of documents and translations (e.g. there may have guidelines on how to name the folders and file documents with names where the only variation is an indication of the language of the document), but quality controls are not provided and there are no specific objectives as regards the quality of archived materials or file system.

The protocol manages the file that is in a space (common folder, Drive space, etc.) separately from other activity documentation file; however, the reuse of materials is not yet considered the ultimate aim of the process. It could be the case that translators use CAT, but each translator manages their own memories in their personal workstation, and even if they could share them, they would share them in an informal and unsystematic way.

Level 3: Defined Objective

At level 3 ("Processes characterized by the organization and is proactive: projects adapt their processes to organizational standards"), compiling and archiving of translation materials are integrated into the translation process, since the reuse of materials by using CAT programs is integrated into the translation process.

The document and translation archive in the form of a translation memory is planned from an early stage of the translation project and it is documented in the translation plan. Translation is planned relying on the availability of existing resources (reuse of translation memories of previous projects). The archiving protocol is reviewed from the experience that builds up in the organization.

A critical distinction between maturity level 2 and level 3 is the scope of the descriptions of processes, procedures and standards. At level 2 they may still vary for each instance or project. At level 3, the standards, processes and procedures are those used by the entire organization and expressed in a protocol that is tuned for each project. Another critical distinction is that at level 3 processes are described rigorously and are part of the content to be taken into account for staff training.

Level 4: Quantitatively Managed

After level 2 and 3, the organization has a technical and human infrastructure where the reuse of translations and related materials is considered an element of the translation process optimization.

In organizations in Level 4 ("Quantitative management: measured and controlled processes") compilation and archiving of translations for reuse in the form of translation memories are defined in detail as a process that is documented and measured, since it is perceived as a

fundamental basis of the translation process; the resources generated are used to evaluate the costs of translation, to evaluate the productivity of the organization and to monitor improvements.

At Level 4 translation memories are stored and managed centrally, possibly in a version control system, which is updated manually when translators (or those responsible for managing subcontract translations) choose to do so. Memories, and their versions, are documented specifically with information concerning language, format, person responsible for the translation, segment size, creation date, domain and document type, character encoding and associated resources (terminologies, glossaries). Translation data are considered as an internal good, produced and controlled by the organization itself. However, in the protocol, the active search for possible external sources that can be incorporated to optimize the system is not contemplated.

Level 5: Optimal

An organization that has managed to reach this level has a support infrastructure for compiling and archiving translation memories and related materials. At level 5 ("Emphasis on process improvement"), the organization includes actions to improve its processes based on a quantitative perspective while technological means to achieve it are provided. The organization also foresees active tasks to search external resources (translation memories, terminologies, etc.) that may contribute to the efficiency of the translation process.

As at level 4, level 5 translation memories are stored and managed centrally with an automatic control system. The archive is automatically updated when translators (or those responsible for managing translation subcontracts) send translations.

Following a defined protocol, translation memories, and their versions, are documented with information on language, format, person responsible for the translation, size segments, creation date, domain and type of document, character encoding and resources (terminologies, glossaries) associated, but in level 5 this is done using specialized metadata.

Translation resources are considered as a good, produced and used by the organization itself but the organization is also aware of the possibility of third parties to reuse its memories in an example of reuse of public sector information. Therefore, the documentation of translation memories includes specific information about the possibility that they are published: content of private information (in some cases, to be able to publish the materials, they have been anonymized), confidential information. In addition, the organization has decided on a distribution method (open data, published in the corporate web with a restricted license or other possible solutions). This information is added to the documentation and metadata.

4. Conclusions

The reuse of translation resources has become popular with CAT tools that use translation memories: the compilation of translated material that is used to find —and to reuse— previously translated text segments. These translation memories are also the best material for training statistical machine translation systems.

Any public administration that produces translation data can be a potential provider of useful reusable data to meet its own translation needs and the ones of other public organizations and private companies that work with texts of the same domain.

The organization's management of the translation process, the characteristics of the archives of the generated resources and of the infrastructure available to support them determine the efficiency and the effectiveness with which the materials produced can be converted into reusable data.

However, it is of utmost importance that the organizations themselves first become aware of the goods they are producing and, second, adapt their internal processes to become optimal providers. In this article, we propose a maturity model to help these organizations to achieve optimal organizations by identifying the different stages of the management of translation data that determine the path to the aforementioned goal. The maturity model presented is still ongoing work; it is going to be validated in an extensive study to identify Spanish public administration organizations that produce translation data in order to produce a catalogue of potential open translation data providers.

References

Victoria Arranz and Olivier Hamon (2012). On the Way to a Legal Sharing of Web Applications in NLP. In Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC 2012), May 2012, Istanbul, Turkey.

Mikel L. Forcada, Miquel Espla-Gomis, Juan Antonio Perez-Ortiz. Stand-off Annotation of Web Content as a Legally Safer Alternative to Bitext Crawling for Distribution. Proceedings of the EAMT 2016. Baltic Journal of Modern Computing.

Julia Aymerich and Hermes Carmelo (2009), The MT Maturity Model at PAHO. MT Summit XII: Proceedings of the twelfth Machine Translation Summit.

Sally Godfrey (2008) What is CMMI ?. NASA presentation.
https://ses.gsfc.nasa.gov/ses_data_2004/040601_Godfrey.ppt

Ann Irvine and Chris Callison-Burch, Combining Bilingual and Comparable Corpora for Low Resource Machine Translation, Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 262–270, Sofia, Bulgaria, August 8-9, 2013 c2013 Association for Computational Linguistics

Pavel Pecina,, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Aleš Tamchyna, Andy Way and Josef van Genabith, Domain adaptation of statistical machine translation with domain-focused web crawling. Language Resources and Evaluation 49-1, 147-193.

Xingyi Song, Lucia Specia, and Trevor Cohn. Data Selection for Discriminative Training in Statistical Machine Translation. In 17th Annual Conference of the European Association for Machine Translation, pp. 45–53, EAMT, Dubrovnik, Croatia, 2014.

Prodromos Tsiavos, Stelios Piperidis, Maria Gavrilidou, Penny Labropoulou, and Tasos Patrikakos. Qtlaunchpad public deliverable d4.5.1: Legal framework.