Ficha resumen

Estudio sobre datos reutilizables como recursos lingüísticos

Plan de impulso de las Tecnologías del Lenguaje

Antonio Moreno, Doroteo Torre, Ana Valverde, Leonardo Campillos (UAM)

Resumen

Los objetivos centrales de estudio, financiado por la SEAD y Red.es, son 1) realizar un censado de recursos de las diferentes administraciones públicas que puedan ser convertidos en recursos lingüísticos (RL), y 2) proponer un plan de acción para abordar su conversión en RL. Se ha elaborado una metodología específica para el censado y evaluación de la madurez de los datos. Tras el análisis preliminar de 101 recursos, se seleccionaron 24 para su análisis detallado y evaluación. El informe incluye también un repaso de estudios similares en otros países. Concluye con unas recomendaciones genéricas y estrategias concretas para algunos recursos.

Para la elaboración del censo y su evaluación, se creó una ficha técnica para la recogida de información, compuesta por diferentes campos (Tabla 1):

1. **Identificación del recurso:** nombre, url, clasificación, lenguas, licencia, etc.
2. **Persona de contacto u organización responsable**
3. **Creación del recurso:** proveedor, proyecto financiador.
4. **Descripción del recurso:** variedad de lengua, niveles de anotación, estándares, tamaño, unidad, formato, dominio, etc.
5. **Otros recursos relacionados.**
6. **Grado de madurez de datos conforme al modelo de la tabla 2.**
7. **Posibles aplicaciones del futuro recurso lingüístico.**

Tabla 1: Estructura y principales contenidos de la ficha técnica

Adicionalmente, se elaboró una ficha específica para la evaluación de la madurez (Tabla 2), teniendo en cuenta factores técnicos requeridos por los RL usados en PLN y aspectos legales para su (re)utilización.

De este primer censo, se seleccionaron 24 recursos para su análisis exhaustivo, según los criterios de interés (calidad, cantidad y disponibilidad de los datos), plurilingüismo, estado de la propiedad intelectual, variedad temática, grado de madurez y tipología del RL. La mayoría de los conjuntos analizados en este informe presentan una madurez baja o media. La explicación se debe a que los requisitos para ser considerado un recurso maduro son muy estrictos: solo los ya procesados y en formatos directamente usables por los investigadores de PLN pueden ser considerados propiamente RL.

Puntos que considerar para evaluar el grado de madurez
Aspectos técnicos:
1. Digitalización (conversión a formato procesable).
2. Transcripción (ortográfica, fonológica, …).
3. Alineación vídeo/sonido y texto.
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1).
5. Anotación morfológica y/o sintáctica.
6. Anotación de entidades nombradas.
7. Otros tipos de anotación (semántica, pragmática, palabras clave, …).
8. Revisión de aspectos formales (ortografía, formato de etiquetado, …).
9. Revisión de contenido (incoherencias, redundancia de datos, …).
10. Anotación conforme a estándares PLN.
11. Presencia de metadatos.
Aspectos legales:
12. Necesidad de anonimización de datos personales.
13. Necesidad de solicitud de permiso de uso.

Tabla 2: Plantilla para la evaluación de la madurez como RL de un recurso

En el plano internacional, se han dado iniciativas para la creación de RL a partir de la reutilización de datos o contenidos de las administraciones públicas. Destaca el portal de la *European Language Resource Coordination*, que proporciona una lista de recursos existentes en Europa de carácter abierto. España se encuentra entre los países de cabeza en cuanto a RL en formato libre y disponible para TL. A nivel europeo, Francia y Reino Unido son los referentes, si bien la Comisión Europea ha advertido del retraso con respecto al inglés y a las compañías americanas, que han *capitalizado* el uso de los *Big Data* lingüísticos, permitiendo que las grandes multinacionales tecnológicas como Google, Microsoft, Amazon o IBM ofrezcan servicios lingüísticos en muy diferentes dominios y lenguas.

La comunidad investigadora en PLN en España necesita RL de calidad y en cantidad suficiente para desarrollar aplicaciones competitivas en el mercado internacional. Por ello, es necesario que la conversión a RL siga **estándares técnicos internacionales** para garantizar la **interoperabilidad entre datos y procesadores**. La Tabla 3 recoge los requisitos definidos por CLARIN para conseguirlo.

- Codificación de caracteres en UTF-8.
- Segmentación del texto en unidades (palabras, sintagmas, frases o párrafos).
- Lenguajes de marcado para codificar la anotación lingüística: XML, RD, tabular vertical, JSON.
- Código estándar de lengua.
- Metadatos con la descripción de los recursos.
- Formatos estándar para codificar memorias de traducción: TMX y XLIFF.
- Formatos estándar para codificar diccionarios y recursos léxicos: LMF y TBX

Tabla 3: Recomendaciones de CLARIN para garantizar la interoperabilidad

Entre las aportaciones del estudio están una especificación sobre **formatos recomendados por tipo de recurso** (Tabla 4) y **herramientas específicas para preparar los datos como RL** (Tabla 5).

Tipo de recurso	Formato recomendado	Formatos adecuados
Corpus textuales	Anotación en XML o TXT en codificación UTF-8	JSON, CSV; no es conveniente PDF
Corpus de audios:	WAV 16 bits, 16 KHz. (voz) o 44.1 KHz (música, audio)	FLAC; MP3 (de alta calidad); otros formatos convertibles (con posible pérdida de calidad)
Corpus de vídeos	MPEG-4 (MP4) de alta calidad	H.264; cualquier otro formato de alta calidad convertible
Corpus memorias de traducción	TMX	CSV

Tabla 4: Formatos recomendados por tipo de recurso

- Herramientas de segmentación de oraciones y secciones (ej. cabecera del documento).
- Programas de segmentación de palabras (tokenización).
- Herramientas de normalización, desambiguación y expansión de abreviaturas y acrónimos (especialmente necesarios en el dominio técnico y biomédico).
- Herramientas de anonimización de datos personales.
- Anotadores morfológicos (*Pos taggers*) y sintácticos (*parsers*), y de sintagmas (*chunkers*).
- Reconocedores de entidades, especialmente, para demarcar sus límites, y con convenciones como el uso de offsets o etiquetas BIO (*Beginning, Inside, Outside*).

Tabla 5: Herramientas para preparación de RL

El estudio concluye con unas **recomendaciones de actuación** a corto y medio plazo, recogidas en la Tabla 6.

1. **Garantizar la disponibilidad y el acceso universal** a los datos abiertos para RL en todas las lenguas del Estado **a través de un portal común y único**.
2. Impulsar **la conversión de** los millones de **páginas digitalizadas en PDF o imagen en texto plano.**
3. Mejorar **la visibilidad de los conjuntos de datos** en cuanto a su **disponibilidad y madurez**.
4. Facilitar la **descarga masiva de grandes ficheros en formatos apropiados** (texto plano, XML, CSV, JSON, RDF).
5. **Creación de recursos anotados** de utilidad general y de disponibilidad abierta**.**
6. Facilitar **herramientas de conversión de datos abiertos a RL.**
7. Proporcionar la **transcripción de ficheros multimedia.**
8. Estimular la **adopción de licencias de libre uso y acceso a los datos.**
9. Estimular la reutilización de los datos mediante **la organización de competiciones tecnológicas** basadas en los mismos**.**
10. Facilitar el **acceso a capacidad de cómputo y almacenamiento** para grandes volúmenes de datos.
11. Impulsar la **publicación de conjuntos de datos anonimizados**, esencial en los documentos médicos o legales.
12. En los **recursos de traducción**, identificar las **lenguas fuente y meta**, así como el **alineamiento** de las “unidades de traducción”.

Tabla 6: Recomendaciones generales para plan de actuación