Universidad del País Vasco Euskal Herriko Unibertsitatea

SERVICIO DE ACTUALIZACIÓN Y DESARROLLO DE LOS CORPUS DEL INSTITUTO DEL EUSKERA DE LA UPV/EHU

PLIEGO DE PRESCRIPCIONES TÉCNICAS PARTICULARES

1. ANTECEDENTES

El Instituto de Euskera de la UPV/EHU, a través de la contratación de una empresa especializada en digitalización externa, ha desarrollado y publicado a lo largo de varios años diccionarios y corpus que recogen el uso del euskera en su página web. El euskera es una lengua viva, y para la descripción e investigación sincrónica y diacrónica de la lengua es imprescindible actualizar adecuadamente los recursos que la recogen (alimentarla), desarrollar los necesarios y cruzar los datos entre los recursos (para una correcta actualización).

La herramienta utilizada hasta ahora para poder indexar y consultar públicamente los textos que están detrás de los recursos del Instituto de Euskera no ha sido una herramienta homologada y pública, sino privada. Considerando que ello puede obstaculizar en el futuro el desarrollo del Instituto de Euskera, es necesario que todos los corpus se desarrollen con la herramienta pública open-source softaware, Apache Solr, a la que también se deben adaptan los corpus existentes en la actualidad, creando una nueva interfaz de consulta que recoja, como mínimo, todas las funcionalidades existentes en la actualmente en funcionamiento.

2. OBJETO

El presente pliego tienen por objeto la definición de las características técnicas que rigen la prestación consistente en el Servicio de Actualización y Desarrollo de los Corpus del Instituto del Euskera de la UPV/EHU.

3. CARACTERÍSTICAS TÉCNICAS QUE DEBE REUNIR EL SERVICIO A PRESTAR

El contrato tiene un doble finalidad:

1. Actualización de los corpus del Instituto de Euskera de la UPV/EHU en su página web: https://www.ehu.eus/es/web/eins y cruce de datos entre los recursos.

   Los corpus a actualizar son los siguientes:

   a. Ereduzko Prosa Dinamikoa (EPD)
   b. Egungo Testuen Corpusa (ETC)
   c. Hizkuntzen arteko Corpusa (HAC)
   d. 0-9 Urtekoentzako Testuen Corpusa

   Tambien se deberá realizar la inserción del corpus "Adinekoen corpusa"

2. Desarrollo de una nueva herramienta pública y open-source, software Apache Solr, de explotación de corpus y adaptar también a esta herramienta los corpus actuales, creando una nueva interfaz de consulta, que recoja como mínimo todas las funcionalidades existentes en la actualmente en funcionamiento.

3.1 Actualización de corpus

Todos los textos de los corpus deberán ser lematizados.

El suministro de los ficheros de textos será también labor de la empresa contratista.

3.1.a. Ereduzko Prosa Dinamikoa (EPD)

Este corpus se actualizará cada año incorporando los textos del año anterior.

Las actualizaciones se realizarán en el año natural.

Se añadirán 3 millones de palabras de texto al año.

La composición del corpus será:

- Diario ***Berria*** (1,3 millones)
  (el punto de partida son los archivos HTML)

- Boletín semanal ***Herria*** (0,2 millones)
  (el punto de partida son los archivos del programa InDesign)

- Libros (1.5 millones) — de los que al menos el 30% serán traducciones —
  (el punto de partida son los archivos PDF, y además de extraer el texto, habrá que etiquetar el número de página para luego poder visualizarlo en el resultado de las búsquedas)

El instrumento público de consulta deberá incorporar el mayor número posible de funcionalidades y capacidades que permitan realizar consultas en diferentes modos. El instrumento público de consulta deberá recoger, como mínimo, las funcionalidades actualmente existentes, que son:

— Búsqueda por palabra o lema
— Visualización de ejemplos, señalando el origen.
— Utilización del carácter comodín

3.1.b. Egungo Testuen Corpusa (ETC)

Habrá que incluir los textos de 2020 a 2025 y la cantidad de nuevos textos será de al menos 50 millones de palabras.

Los textos que se añadan deberán provenir, al menos, de las siguientes fuentes:

- Diario Berria (completo).
- Semanario Argia (completo).
- Semanario Goienkaria (completo).
- El semanario Herria (completo).
- Libros (mínimo 50 libros al año), de los que al menos el 20% serán traducciones.
- Wikipedia en euskera.

El instrumento público de consulta deberá incorporar el mayor número posible de funcionalidades y capacidades que permitan realizar consultas en diferentes modos. El instrumento público de consulta deberá recojer, como mínimo, las funcionalidades actualmente existentes, que son:

- Búsqueda por palabra o lema
- Visualización de ejemplos, señalando el origen.
- Utilización del carácter comodín

También existe en la web del Instituto de Euskera de la UPV/EHU el Egungo Euskararen Hiztegia (EEH), en el que abundan las referencias al número de apariciones en ETC. La empresa contratista deberá realizar una propuesta en la que se prevea el impacto que esta nueva actualización del ETC tendría en dicho diccionario, y se proponga la forma de plasmar y actualizar los nuevos datos (la base de datos del diccionario EEH está disponible en la web del Instituto y se puede descargar públicamente).

3.1.c. Hizkuntzen arteko Corpusa (HAC)

Durante el primer año natural se añadirán 40 nuevos libros al corpus.

El suministro de textos en castellano, francés e inglés será también obra de la empresa contratista.

Asimismo, se deberán lematizar textos de otros idiomas para poder realizar búsquedas mediante "palabra" o "lema".

El instrumento público de consulta deberá reunir la mayor cantidad posible de características y habilidades para poder realizar consultas en diferentes modos. El instrumento público de consulta deberá recojer, como mínimo, las funcionalidades actualmente existentes, que son:

- Búsqueda por palabra o lema
- Visualización de ejemplos, señalando el origen.
- Utilización del carácter comodín

El contratista deberá presentar una lista de esos 40 libros que considere aptos para enriquecer el corpus.

3.1.d. 0-9 Urtekoentzako Testuen Corpusa

En cuanto a este corpus, el objeto de este contrato es también doble:

a) Dotar al corpus de nuevos textos. Se aumentarán al menos 400 libros. El conseguir los textos será labor de la empresa contratista.

b) Añadir nuevas funcionalidades. Teniendo en cuenta que el corpus está pensado principalmente para investigadores y profesionales de la logopedia, la empresa contratista deberá realizar una propuesta especificando cuáles pueden ser estas nuevas funcionalidades más allá de las actuales.

El instrumento público de consulta deberá reunir las características de la versión actual:

- Poder buscar la "palabra" o el "lema".
- «Poder realizar una búsqueda fonémica en función del tipo de vocal o consonante y la combinación entre ellos».

3.1.e. Inserción del "Adinekoen corpusa"

En 2024 el Instituto de Euskera comenzó a crear un corpus específico con producciones orales de personas mayores. La tarea consistirá en poner en la red las transcripciones de las grabaciones que realizará el Instituto.

3.2. Desarrollo de una nueva herramienta de explotación de corpus.

Desarrollo de una nueva herramienta pública y open-source, software Apache Solr, de explotación de corpus y adaptar también a esta herramienta los corpus actuales, creando una nueva interfaz de consulta (que recoja todas las funcionalidades existentes).

"Explotación de corpus" hace referencia a la herramienta necesaria para indexar y luego consultar públicamente los textos existentes en dichos corpus.

Todos los corpus deberán desarrollarse con la herramienta pública Apache Solr (https://solr.apache.org/), software open-source.

Por tanto, a la finalización del presente contrato de servicios se deberán cumplir tres objetivos:

- Adaptar a este instrumento los cuatro corpus mencionados inicialmente:
    - Ereduzko Prosa Dinamikoa (EPD)
    - Egungo Testuen Corpusa (ETC)
    - Hizkuntzen arteko Corpusa (HAC)
    - 0-9 Urtekoentzako Testuen Corpusa

    Asi como realizar la inserción del "Adinekoen corpusa"

- Crear una interfaz pública en función de esta herramienta. Esta nueva interfaz deberá incluir todas las funcionalidades existentes hasta la fecha.

- Garantizar la utilización de la herramienta Apache Sorl para la futura alimentación de estos corpus, de forma que en el futuro pueda ser actualizada por terceros.

Además de la herramienta de explotación, se definirá el procedimiento de formación del corpus. La alimentación del Corpus es continua. A medida que se vayan obteniendo nuevos textos, se ofrecerán procedimientos y programas para poder integrarlos en la herramienta de consulta del corpus, siendo parte de este proyecto estos trabajos de alimentación.

4. LUGAR DE TRABAJO

Los trabajos se llevarán a cabo en las dependencias de la contratista, con sus propios recursos físicos y logísticos.

Las reuniones para el seguimiento de resultados se realizarán preferiblemente en la sede del Instituto de Euskera de la UPV/EHU, o en caso de que no fuera posible, se harán de forma telemática.

ICIAR SAN MARTIN EGUIA - 15389693W

Firmado digitalmente por ICIAR SAN MARTIN EGUIA - 15389693W
Fecha: 2026.02.24 09:48:59 +01'00'