



Recursos y Herramientas de Lingüística de Corpus y PLN para la Monitorización e Investigación de los Usos Académicos del Euskera



III Workshop Intele, Madrid, 13-14 Septiembre 2022

María Jesús Aranzabe, Izaskun Aldezabal, Igone Zabala
{maxux.aranzabe, izaskun.aldezabal, igone.zabala}@ehu.eus

Contexto y Motivación

Euskera: una lengua minorizada en proceso de desarrollo y fijación de la terminología

- Comunidades de discurso especializado -> “contexto natural” para el desarrollo y la autorregulación. Los miembros de dichas comunidades son a la vez: **expertos/as** (conocimiento semántico-pragmático); **autores/as** (de los textos); **usuarios/as** (de los términos)
- El predominio del inglés como *lingua franca* para la comunicación académica: tesis de máster y doctorado, trabajos fin de grado, presentaciones de seminarios...
- Gran riesgo de la desaparición de un ámbito de uso crucial para la vitalidad de una lengua
- La **Universidad**: **Agente** primordial para el desarrollo de los registros académicos

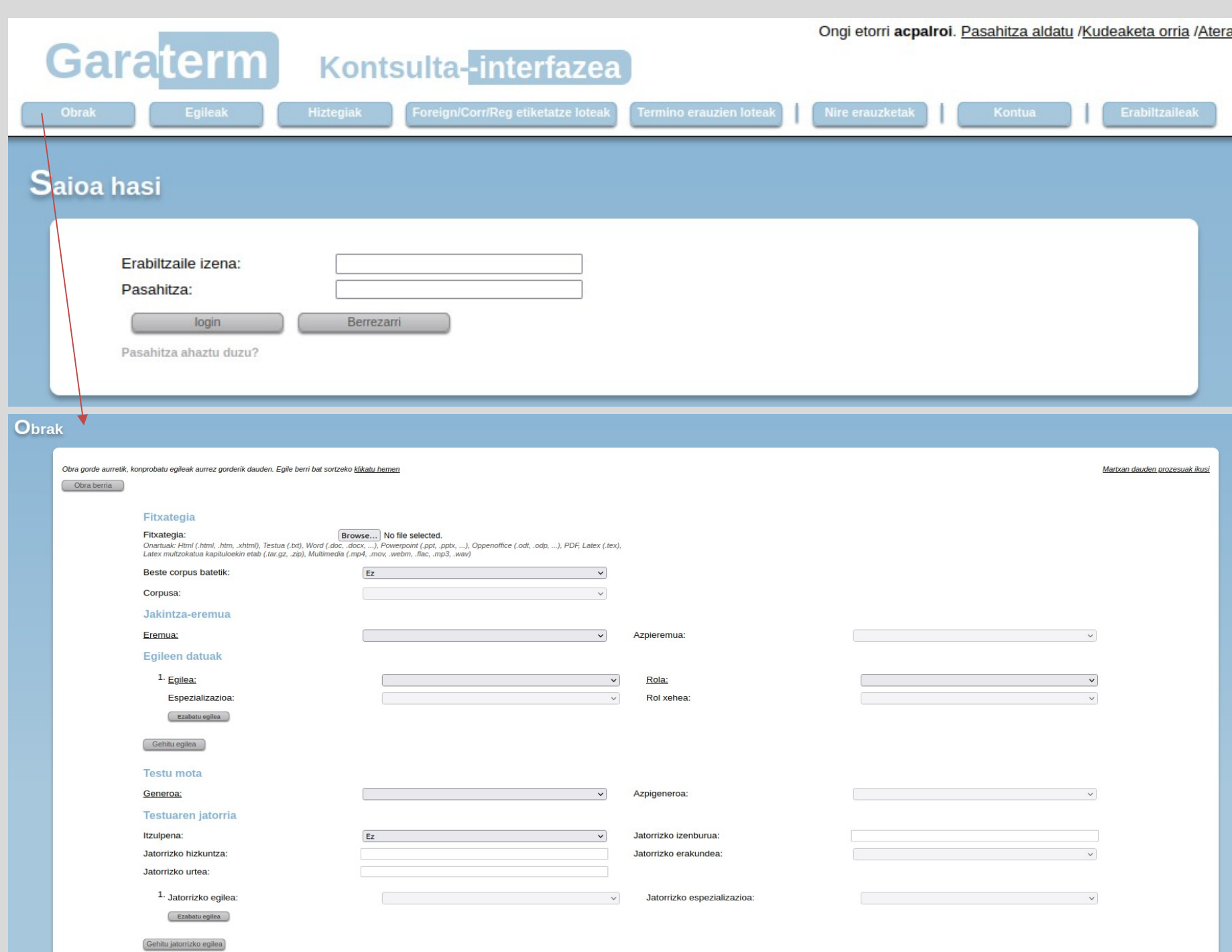
Objetivo

Diseño de un entorno estratégico de recursos para la armonización y coordinación de la didáctica, la investigación y el desarrollo de los registros académicos en las diferentes áreas de conocimiento, mediante tecnologías del lenguaje (Zabala et al., 2008)



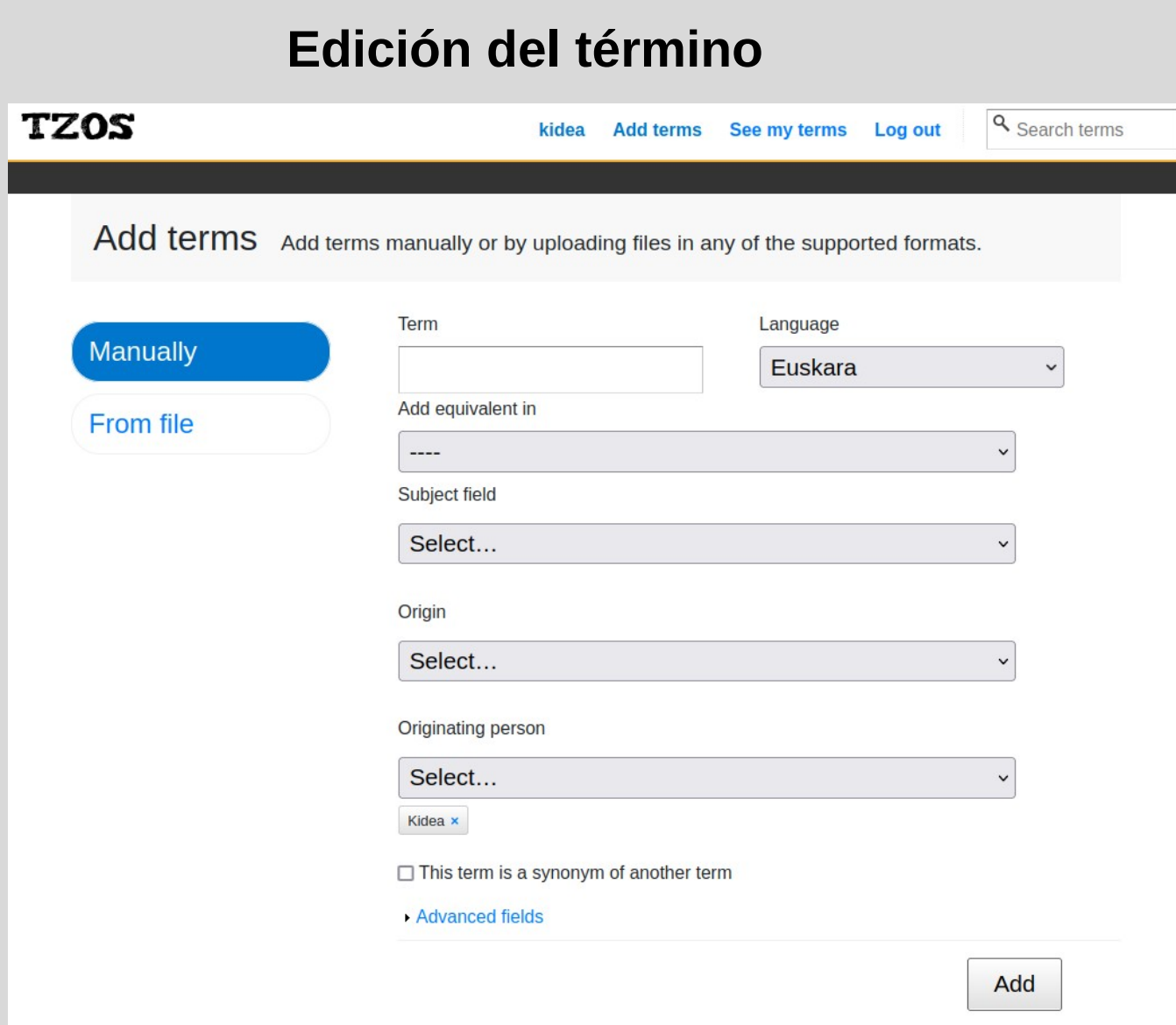
ENTORNO GARATERM para la integración de las herramientas y recursos

Entorno de trabajo GARATERM (Zabala et al., 2013)



- Procesamiento y etiquetado de textos, y gestión de la extracción de terminología y fraseología

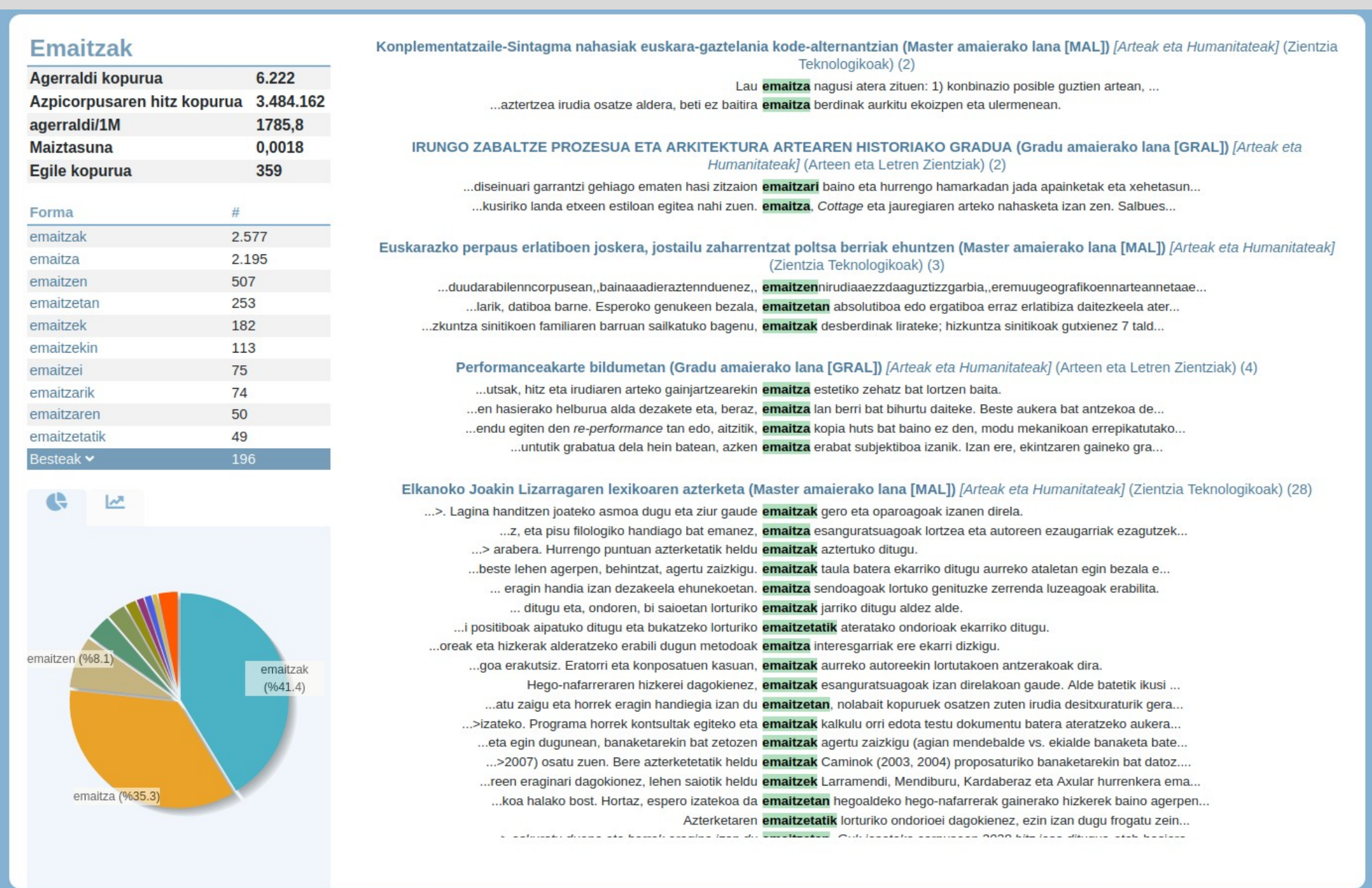
Base de datos terminológica TZOS (Arregi et al., 2013)



- Descripción de la terminología utilizada en la comunicación académica (descripción *activa*)
- Trabajo terminológico guiado por corpora
- Conectado con el corpus Garaterm: términos en el contexto discursivo (*in vivo*)

CORPORA: Garaterm y HARTAeus

Ejemplo de la búsqueda del lema “emaizta” (resultado)



HARTAeus (Aranzabe et al., 2022)



- Único corpus académico del euskera
- Constituido por material docente y trabajos de investigación académica
- Corpus monitor: estudios diacrónicos. En continuo incremento
- Textos espontáneos y controlados: adecuados para analizarlos y utilizarlos en tareas de enseñanza

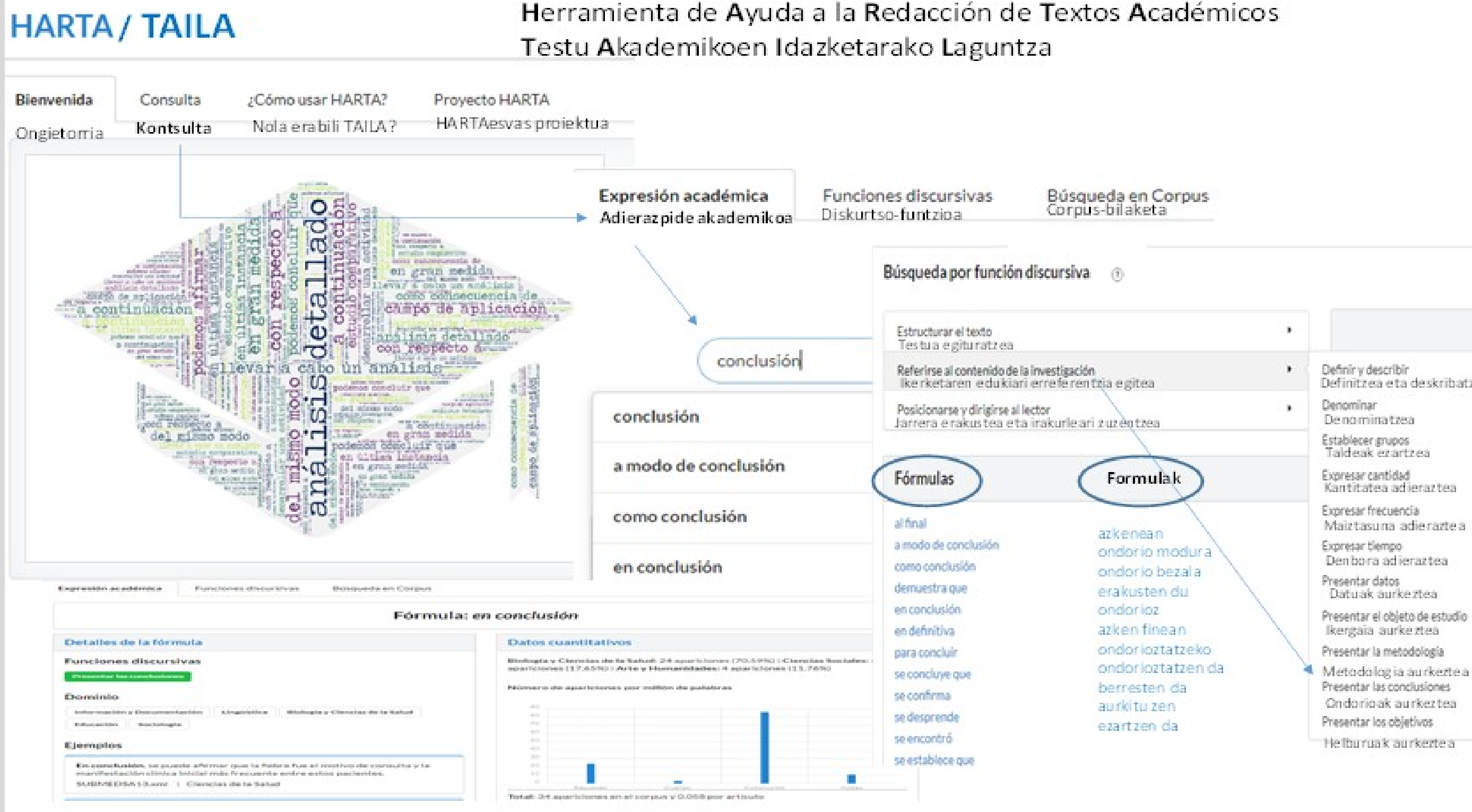
Extractor automático Erauzterm (Alegria et al., 2002)

Lema	Bal.	Alorra	Hiztegiak	Ereduak	Maiz.	Kop.
1 baldintzen plegu	...	...	...	...	...	...
2 aurrekontu	...	...	...	...	...	...
3 baldintza	...	...	...	...	...	...
4 hilaibete	...	...	...	...	...	...
5 bideragarritasun	...	...	...	...	...	...
6 epe	...	...	...	...	...	...
7 plegu	...	...	...	...	...	...
8 kalkulu	...	...	...	...	...	...

- Extracción automática y validación manual de los candidatos a término, del léxico académico, de colocaciones y de fórmulas discursivas

HARTA/TAILA: Herramienta bilingüe de Ayuda a la Redacción de Textos Académicos

- Conectado con el corpus HARTAeus
- Descripción del uso de las combinaciones léxicas académicas (CLA): léxico académico, colocaciones y fórmulas discursivas



Referencias bibliográficas

- Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. (2002) “Robustness and customisation in an analyser/lemmatiser for Basque”. En Third International Conference on Language Resources and Evaluation (LREC): Customizing Knowledge in NLP Applications-Strategies, Issues and Evaluation Workshop. European Language Resources Association (ELRA): 1-6.
- Aranzabe M.J., Gurrutxaga A., Zabala I. (2022) “Compilación del corpus académico de novelas en euskera HARTAeus y su explotación para el estudio de la fraseología académica”. XXXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2022), 20-23 de septiembre, A Coruña (Spain).
- Arregi X., Arriarte A., Arriola X., Zabala I., Lersundi M. (2013) “TZOS: An On-Line System for Terminology Service”. Actualizaciones en Comunicación Social. Actas XIII Simposio Int. de Comunicación Social. Centro de Lingüística Aplicada. Santiago de Cuba: 400-404.
- Villayandre, M. (2018) “HARTA de novelas: un corpus de español académico”. CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos, 5(1): 131–140.
- Zabala I., Aierbe A., Aldezabal I., Aranzabe M., Arregi X., Arriola J.M., Elordui A., Elozegi A., Elozegi K., Ezeiza J., García, J., García I., Lersundi M., San Martín I., Ugarteburu I. (2009) “GARATERM: Diskurtso akademikoko profesionalaren didaktika eta garapena uzartzeko tresna informatikoen diseinu eta integrazioa helburu duen proiektua”. Ugarteburu I. eta P. Salaburu (arg.) *Espazialitate hizkerak eta terminologia Jardunaldiak III*, Bilbo: Euskara Institutua—EHU: 211-219.
- Zabala I., Lersundi M., Leturia I., Manterola I., Santander, G. (2013) “GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea”. Xabier Alberdi eta Pello Salaburu (arg.) *Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira*. UPV/EHUko Argitalpen Zerbitzua: 98-114.

Con la ayuda de:



Grupo Ixa. Grupo de investigación consolidado del tipo A (IT-1570-22).

Proyecto HARTAvas (PID2019-109683GB-C22), financiado por el Ministerio de Ciencia e Innovación.