Oharrak

Notas

Remarques

Notes

Euskera Ikerketa Aldizkaria. 2026, 71 (2). Bilbo
ISSN 0210-1564

Euskorpora Summit 2026: inteligencia artificial, datos lingüísticos y soberanía digital para las lenguas pequeñas

ARRANZ, Victoria
Euskorpora
victoria.arranz@euskorpora.eus

ARREGI, Sara
Euskorpora
sarregi@euskorpora.eus

BARAÑANO, Leire
Euskorpora
lbaranano@euskorpora.eus
ORCID: https://orcid.org/0009-0007-7908-4612

GARCÍA-PABLOS, Aitor
Euskorpora
agarcia@euskorpora.eus
ORCID: https://orcid.org/0000-0001-9882-7521

DOI: https://doi.org/10.59866/eia.v2i71.355

1 Introducción
2 La brecha lingüística en la IA: diagnosis compartida
3 La producción de datos lingüísticos: complejidad técnica, legal y económica
4 Casos nacionales: estrategias comparadas para lenguas pequeñas
4.1 **El modelo finlandés**
4.2 **El caso islandés**
4.3 **El caso letón: Tilde como empresa privada europea de referencia**
5 El ecosistema vasco: fortalezas, brechas y retos
6 El European Language Data Space y la gobernanza europea
7 Impacto del evento
8 Líneas de investigación emergentes
9 Conclusiones
Referencias y recursos
Anexos

Crónica académica de la jornada celebrada el 11 de junio de 2026 en Bilbao

Nota editorial

Este documento recoge las principales contribuciones y debates del Euskorpora Summit 2026 en un formato orientado a su publicación como crónica académica. Las anotaciones reconstruyen el contenido de cada intervención a partir de las presentaciones originales, la documentación del evento, las métricas de impacto y notas de seguimiento. La ponencia de Andrejs Vasiljevs (Tilde) no contó con presentación digital; su contenido se ha completado con información pública disponible en tilde.ai. De las 108 personas inscritas, finalmente, asistieron 92.

1 Introducción

El Euskorpora Summit 2026 reunió en la sede de EITB en Bilbao a investigadores, responsables de política lingüística, empresas tecnológicas y representantes institucionales para examinar la situación de las lenguas europeas de menor extensión ante la revolución de los grandes modelos de lenguaje (LLM) y la inteligencia artificial generativa. La jornada combinó una perspectiva comparada internacional —con ponentes de Alemania, Francia, Letonia, Finlandia e Islandia— con un análisis aplicado del ecosistema vasco a través de una mesa redonda con organizaciones socias de Euskorpora.

La jornada congregó 92 asistentes de más de 40 organizaciones distintas entre instituciones públicas, empresas tecnológicas, centros de investigación, medios de comunicación y entidades culturales, y generó una cobertura mediática en 5 medios y un impacto de 1142 impresiones en LinkedIn con una tasa de interacción del 78,7 %, muy superior a la media del sector.

Este texto sintetiza las principales contribuciones organizadas temáticamente, e identifica los consensos emergentes, los debates abiertos y las líneas de investigación que se desprenden del conjunto de intervenciones.

2 La brecha lingüística en la IA: diagnosis compartida

Existe un diagnóstico convergente entre todos los ponentes internacionales: los sistemas de IA actuales refuerzan sistemáticamente las desigualdades lingüísticas preexistentes. Más del 90 % de los datos de entrenamiento de los grandes modelos de lenguaje está en inglés (Vasiljevs 2026), generando un efecto de retroalimentación donde las lenguas con mayor representación digital obtienen mejor rendimiento en IA, ampliando progresivamente la brecha.

Georg Rehm presentó el Digital Language Equality Metric (proyecto ELE), que cuantifica el soporte tecnológico de más de 80 variedades lingüísticas europeas. Los datos de 2022-2023 muestran que la distancia entre el inglés y el resto de las lenguas europeas aumentó durante ese periodo, incluidas lenguas con millones de hablantes y estatus oficial en la UE como el polaco, el húngaro o el finlandés.

Este fenómeno tiene consecuencias estructurales para lenguas como el euskera: incluso con avances significativos en normalización social y educativa durante décadas, la ausencia de representación en los corpus de entrenamiento puede traducirse en invisibilización tecnológica que cuestione la viabilidad del uso de la lengua en los entornos digitales del futuro.

> *Except for English and, to a certain extent, German, Spanish and French, all European languages are massively under-resourced in terms of data, tools and technologies.* — *Georg Rehm*

3 La producción de datos lingüísticos: complejidad técnica, legal y económica

La ponencia de Khalid Choukri (ELDA/ELRA) aportó una de las perspectivas más originales al cuantificar la complejidad real de producir recursos lingüísticos de calidad. La ratio de trabajo humano necesario para transcribir audio varía enormemente según la lengua: entre 12 y 20 horas por hora de audio para lenguas bien documentadas, y entre 60 y 70 horas para lenguas orales sin sistema de escritura estandarizado. Esta asimetría explica estructuralmente por qué el mercado libre no puede resolver el problema de las lenguas pequeñas.

Desde el punto de vista legal, Choukri y su equipo describieron la superposición de regímenes normativos: derechos de propiedad intelectual, RGPD cuando los datos contienen información personal, el AI Act cuando se destinan al entrenamiento de modelos de propósito general, y normativa sectorial adicional en sanidad o finanzas. Esta complejidad justifica intermediarios especializados como ELDA o el LDS.

Los cuatro modelos de retorno de la inversión presentados (inversión pública, recuperación de costes, valor estratégico y ecosistema) ofrecen a iniciativas como Euskorpora un marco para articular su modelo de negocio. El modelo de ecosistema —datos compartidos libremente, monetización vía servicios y consultoría— resulta especialmente pertinente para organizaciones con misión cultural y lingüística.

4 Casos nacionales: estrategias comparadas para lenguas pequeñas

4.1 El modelo finlandés

Krister Lindén presentó el Language Bank of Finland (Kielipankki) como infraestructura nacional de referencia: 50+ gigapalabras de texto y 8000 horas de audio bajo marcos de licencia estandarizados, integrado en FIN-CLARIAH y conectado al nivel europeo a través de CLARIN/ALT-EDIC.

Esta arquitectura permitió desarrollar modelos soberanos como FinGPT y la familia Poro.

Cinco lecciones metodológicas de especial relevancia para Euskorpora: los ciudadanos donan datos cuando el proyecto lo promueven instituciones de confianza, la centralización estándar evita los silos, los marcos legales claros habilitan la innovación en IA, las lenguas pequeñas requieren intervención pública activa, y las interfaces web de acceso no técnico democratizan el uso de los corpus.

4.2 **El caso islandés**

Steinşór Steingrímsson presentó el paralelismo entre el islandés y el euskera —ambas lenguas con 400 000-800 000 hablantes, morfología compleja y escasez histórica de corpus digital— y la trayectoria del programa quinquenal 2019-2023: Icelandic Gigaword Corpus, más de 4000 horas de habla mediante *crowdsourcing* (Samrómur) y herramientas de síntesis y reconocimiento de voz.

Steingrímsson señaló una paradoja relevante para la planificación lingüística digital: el programa quinquenal financiado con fondos públicos resultó insuficiente ante la velocidad de los avances en LLMs, que han modificado radicalmente los requisitos de datos en apenas tres años. La planificación debe contemplar ciclos de revisión mucho más cortos que los habituales en política cultural.

4.3 **El caso letón: Tilde como empresa privada europea de referencia**

Andrejs Vasiljevs presentó Tilde como demostración de que una empresa privada europea puede desarrollar tecnologías avanzadas para lenguas pequeñas con sostenibilidad económica. Con 30 años de historia, 150+ empleados, 700+ clientes y certificación ISO 27001, Tilde opera en los tres países bálticos y combina traducción automática, asistentes conversacionales, tecnologías de habla y servicios de datos para IA.

TildeOpen LLM —30B parámetros, ganador del Large AI Grand Challenge, entrenado con 2M horas GPU en LUMI, disponible en código abier-

to desde septiembre de 2025— ofrece una demostración empírica de que invertir en arquitecturas diseñadas para lenguas europeas de menor recurso produce mejoras medibles: +41 % en eficiencia en letón, +37 % en lituano, +31 % en finlandés frente a LLaMA-3, y 84,7 % de precisión en el *benchmark* Belebele.

5 **El ecosistema vasco: fortalezas, brechas y retos**

La mesa redonda moderada por Leire Barañano permitió analizar las condiciones específicas del ecosistema vasco. Del conjunto de intervenciones emergen cuatro fortalezas diferenciales: la Academia normativa (Euskaltzaindia) con recursos lexicográficos de alta calidad, el mayor archivo audiovisual en euskera del mundo (EITB), industria tecnológica local con experiencia en productos conversacionales en euskera (Logikaline), y grandes empresas del sector financiero cooperativo con voluntad de invertir en IA en euskera (Laboral Kutxa).

Las brechas identificadas son igualmente claras: insuficiencia de datos de habla espontánea y dialectal, ausencia de *benchmark*s públicos estandarizados para evaluar modelos en euskera, dificultades jurídicas para reutilizar el archivo audiovisual de EITB, y escasez de modelos de base que funcionen con calidad comparable al castellano o el inglés en tareas aplicadas.

El debate sobre la Ley BAZ como *driver* regulatorio resultó especialmente revelador: la obligación de prestar servicios accesibles en euskera convierte la calidad de las tecnologías conversacionales en un imperativo de cumplimiento normativo, no solo en una opción estratégica. Esto abre una línea de investigación sobre el papel de la regulación como mecanismo de incentivo para el desarrollo tecnológico de lenguas pequeñas.

6 **El European Language Data Space y la gobernanza europea**

El LDS representa un cambio de paradigma relevante para iniciativas como Euskorpora: permite monetizar datos propios sin perder el control, acceder a datos de terceros bajo condiciones acordadas y participar en pro-

yectos europeos de LLMs con datos propios como contribución reconocida. A junio de 2026 cuenta con 79 participantes y 18 conectores operativos; en 2027 el ALT-EDIC asumirá su operación.

La articulación entre el LDS, el ALT-EDIC (21 países miembros) y proyectos como OpenEuroLLM o LLMs4EU configura una arquitectura europea de soberanía lingüística en IA que ofrece a lenguas como el euskera un punto de entrada al nivel continental que no existía hace tres años. La posible conexión del corpus vasco al LDS es una línea de trabajo que los debates del Summit dejaron como agenda pendiente.

7 **Impacto del evento**

La jornada generó una cobertura en 5 medios —*Diario Vasco*, Gobierno Vasco (dos notas), *Euskadi Irratia* («Baipasa»), SPRI Indeusberri y Euskaltzaindia (*Plazaberri*)— y una publicación en LinkedIn con 1142 impresiones, 557 miembros alcanzados, tasa de interacción del 78,7 % (CTR del 74,1 %) y amplio número de menciones orgánicas de ponentes, socios y asistentes. Estas métricas confirman la relevancia del evento para el ecosistema profesional del sector y su potencial como plataforma de visibilidad para Euskorpora a nivel europeo.

8 **Líneas de investigación emergentes**

Del conjunto de debates se desprenden seis líneas de interés para la comunidad académica:

1. *Benchmark*s de evaluación para el euskera en IA: conjuntos de prueba estandarizados para medir el rendimiento de modelos en tareas de comprensión, generación, reconocimiento de voz y traducción.
2. Estrategias de *crowdsourcing* para habla espontánea y dialectal en euskera: qué factores determinan el éxito de las campañas ciudadanas de donación de voz en lenguas pequeñas.

3. Marcos jurídicos para la reutilización de archivos audiovisuales de medios públicos: análisis comparado de soluciones adoptadas por EITB, YLE finlandesa, RÚV islandesa y otras radiotelevisiones europeas.
4. Modelos de sostenibilidad económica para iniciativas de datos lingüísticos sin ánimo de lucro: análisis comparado de modelos de financiación y retorno adoptados por Euskorpora, Kielipankki, Árni Magnússon Institute y ELDA/ELRA.
5. Regulación como incentivo para el desarrollo tecnológico de lenguas pequeñas: estudio del impacto de la Ley BAZ, el AI Act y normativas equivalentes en la inversión privada en IA para lenguas minoritarias.
6. Interoperabilidad entre corpus nacionales y el European Language Data Space: barreras técnicas, jurídicas e institucionales para la integración del corpus vasco en el LDS y en proyectos europeos de LLMs.

9 **Conclusiones**

El Euskorpora Summit 2026 confirmó que el debate sobre las lenguas pequeñas en la era de la IA ha alcanzado madurez suficiente para articular diagnósticos compartidos, comparar estrategias nacionales y formular agendas de investigación colectivas. La convergencia entre las experiencias de Finlandia, Islandia, Letonia y Euskadi no es casual: todas estas comunidades afrontan los mismos retos estructurales y han llegado a conclusiones similares sobre las respuestas posibles.

Las claves del debate se pueden sintetizar en tres tensiones productivas: entre la velocidad del cambio tecnológico y los ciclos de la planificación lingüística pública, entre la necesidad de datos abiertos para la investigación y los derechos de los productores de contenido, y entre el carácter de bien público del patrimonio lingüístico y los incentivos económicos necesarios para sostener su desarrollo tecnológico.

Euskorpora se presenta como un experimento de gobernanza lingüística digital de primer orden: un consorcio público-privado que intenta resolver simultáneamente los problemas de producción de datos, coordinación institucional, sostenibilidad económica y articulación con el ecosistema europeo para una lengua de menos de un millón de hablantes. Su seguimiento constituye, en sí mismo, un objeto de investigación de relevancia para los estudios de política lingüística, tecnologías del lenguaje y gobernanza de datos.

Referencias y recursos

European Language Data Space: https://language-data-space.ec.europa.eu

Tilde / TildeOpen LLM: https://tilde.ai · https://huggingface.co/TildeAI/TildeOpen-30b

Rehm, G. & Way, A. (eds.) (2023). European Language Equality: A Strategic Agenda for Digital Language Equality. Springer, Cognitive Technologies (Open Access).

Language Bank of Finland (Kielipankki): https://www.kielipankki.fi

ALT-EDIC (Alliance for Language Technologies EDIC): https://alt-edic.eu

Language Technology Programme for Icelandic 2019-2023. Árni Magnússon Institute for Icelandic Studies.

ELDA/ELRA: https://www.elda.org

Euskorpora: https://www.euskorpora.eus

Anexos