

Plan TL

Plan de Impulso de las
Tecnologías del Lenguaje



GOBIERNO
DE ESPAÑA

MINISTERIO
DE ECONOMÍA
Y EMPRESA

PLAN DE IMPULSO DE LAS TECNOLOGÍAS DEL LENGUAJE

SEPLN 2019

- Eje 0: Gobernanza
- Eje I: Apoyo al desarrollo de infraestructuras lingüísticas
- Eje II: Impulso de la Industria TL
- Eje III: La Administración como impulsor de la Industria del Lenguaje
- Eje IV: Proyectos Faro
- Plan TL y la Estrategia IA

- OTG Sanidad BSC:
 - Alcance: 6M € / 4 años
 - OTG Sanidad / Plataforma HPC PlanTL
 - Comienzo del encargo SEAD-BSC (enero 2019)
- OTG:
 - Sistemas conversacionales: David Griol
 - Inteligencia Artificial: Jerónimo Arenas
- Informes:
 - Sistemas conversacionales
 - Datos Abiertos en AAPP
 - Análisis de campañas evaluación español y lenguas cooficiales



PRESENTE

- RAE Corpus Capitel (Corpus Anotado del Plan de Impulso de las Tecnologías del Lenguaje): Textos periodísticos de distintas fuentes y dominios (anotación morfosintáctica, entidades nombradas, anotación sintáctica)
- EFE (Convenio en trámite, se espera firma 2019)
- Vanguardia (Convenio en trámite, se espera firma 2019)
- IEC (Convenio en trámite, se espera firma 2019)
- BNE Crawling web española 10 años
- Corpus Legal Jurix2019



FUTURO

- Terminesp (Instituto Cervantes + CSIC)
- RAE anotación de más niveles semánticos CAPITEL
- Convenio RTVE-SEAD: más campañas diarización y subtitulado (Iberspeech 2020)
- Olimpiadas TL
- Infraestructuras lingüísticas Iberoamericanas

- Ayudas SEAD, SGFSI TDH 2019 (9M €), incluyendo TL (Abril 2019)
- INAP Formación funcionariado AGE + CCAA + EELL (MOOCs + capsulas formativas)
- Comunicación: Página web. LinkedIn
- Infodays:
 - Vigilancia Sectorial Mayo
 - Infoday TL Salud (BSC, Barcelona), diciembre 2019.
 - Infoday Justicia (Jurix2019, Madrid), diciembre 2019
- Catálogo productos/svc. Censo empresas y grupos investigación
- Estrategia formativa (UPV, SEPLN)
- Congresos. Participación en congresos y encuentros nacionales e internacionales sobre TL e Inteligencia Artificial

Eje II: Impulso de la Industria TL

- Ayudas Red.es 2019-2020, incluyen las TL entre las tecnologías digitales habilitadoras
- Puesta en marcha plan formación INAP
- Informe Argentina, estudio sector TL similar al estudio realizado para México

- Plataforma Traducción Automática
 - Proyectos EU CEF: MTHub / ELRI / NTEU / NecData
 - Taller ELRI en España (marzo 2019) en la sede de la Comisión Europea en Madrid: Infraestructura de Recursos y Servicios de Traducción Automática para las Instituciones Públicas
- Plataforma Sistemas Conversacionales
 - Informe Sistemas Conversacionales
 - Incorporación David Griol OTG
- Plataforma Procesamiento de Lenguaje Natural
 - Resultados Proyecto eInfra OpeMinTed y adaptación al entorno HPC (ej. procesado crawling BNE)

- Plataforma Traducción Automática
 - CEF AT : 2 proyectos, recolección ELRI
 - BOE, CGPJ, BNE crawling
- Plataforma Sistemas Conversacionales
 - SGAD: Proyecto 060. Arquitectura de referencia
 - RTVE: Subtitulado, diarización, campañas. Recursos
 - Ministerio de la Presidencia: Corpus Parlamento
- Plataforma Procesamiento de Lenguaje Natural
 - Independencia entorno de ejecución, entorno de despliegue, mantener componentes NLP, pipelines de lógica
 - Uso de GPUs (tanto modelos de neuronales como tareas masivas NLP), BSC

- CNIO: publicación de desarrollos en Github y Zenodo
 - Recursos y componentes sw
- Corpus de casos clínicos en español: Anotación morfosintáctica
- Base de datos de abreviaturas médicas en español
- Glosarios médicos bilingües
- Archivos para traducción automática neuronal
- Crawling webs medicina (600Gb texto)
- Campañas de evaluación en Sanidad
 - Reconocimiento de sustancias en textos médicos en español
 - Anonimización
 - Indización semántica médica en español

Eje IV: Proyectos Faro: Salud

- Convenios inmediatos: IACS
- Convenios pendientes de tramitación
 - SAS
 - AEMPS
 - Clinic
 - H12O-ISCIH
 - Son Espases
 - La Paz
- Infoday Salud BSC 4 Diciembre, Barcelona

Eje IV: Proyectos Faro: Vigilancia sectorial

- Convenio SEUIDI + FECYT
 - Dirección de políticas públicas I+D
 - Análisis sector TL, IA
- Incibe: Análisis contratación y empleo sector ciberseguridad
- Apertura código CorpusViewer 1.0
- Integración desarrollos experimentales (perfilado, evolución temporal ... CorpusViewer 1.5)
- Infoday: Jornada sobre Inteligencia Artificial aplicada a las políticas públicas de ciencia, innovación y emprendimiento
- Participación en las Jornadas Técnicas RedIRIS (Abril)
- Corpus Viewer finalista en los premios CNIS 2019 (marzo 2019)

Eje IV: Proyectos Faro: Vigilancia sectorial

- Continuación Convenio SEUIDI+FECYT
- CorpusViewer 1.5 y 2.0
- Colaboración a entidades EU (JRC, INEA, EM) y organismos internacionales (OCDE, WIPO)
- Otros organismos I+D nacionales (CDTI, CSIC, ...). Usuarios CorpusViewer.

Eje IV: Proyectos Faro: Otros

- Legislación y Justicia
 - Jurix2019
 - Corpus Legal 1.250M palabras. Modelos embeddings, Transformers, tópicos, grafos relaciones interdocumentales
- Cultura
 - BSC-BNE: Procesado crawling web españolas 10 años, 5Tb
 - EFE: Anotación corpus noticias 20M

- Justicia
 - Anotación corpus Justicia: legislación, sentencias, ...
 - Recurso terminológico
 - Convenios pendientes: CGPJ, MJusticia, CAM, MPR

- Cultura
 - Convenio BNE: crawling, BDH, HD, epubS
 - EFE

- Presidenta Comisión anunció prioridades Digital Single Market + Energía
- 2500M € / año, 8% ES
- IA europea -> IA ciudadano
 - Orientación servicio público (sectores estratégicos: salud, justiciar, educación, I+D, ...)
 - XAI: orientación a evaluación, generación de evidencias
 - Cuestiones éticas: evaluación regulador
- Sector TL dentro de IA: 45% sector IA nacional es TL
- Colaboración Planes CCAA: Agenda Digital Vasca, PlanIA Cat

Gracias

www.plantl.es

plantecnologiaslenguaje@mineco.es