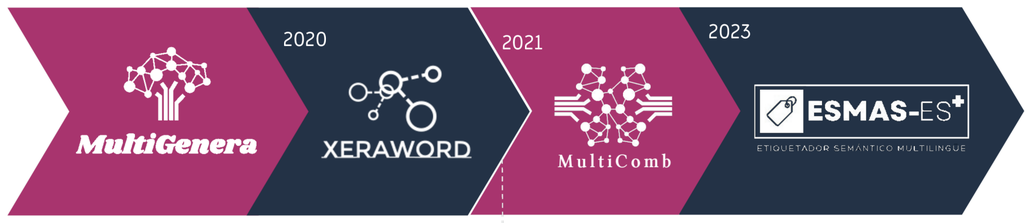
Novas aplicacións das humanidades dixitais: O etiquetador semántico multilingüe automático e sostible ESMAS-ES+

Celia Fernández Vasco / Elena Martín-Cancela UNIVERSIDADE DA CORUÑA



CONTEXTO

- 🔍 *Panorama actual*: múltiples ferramentas de PLN orientadas ao inglés, pouco integradas e con acceso limitado ao significado.
- 🌐 *Necesidade*: recursos lexicográficos multilingües para o tratamento automático da semántica nos nomes.
- 🧠 *Marco conceptua*l: semántica combinatoria + ontoloxías + PLN como resposta ás limitacións dos sistemas actuais.
- 🏛️ *Proxectos previos*: continuidade de PORTLEX, que desenvolveu o Dicionario da valencia do nome en catro linguas.

OBXECTIVOS

- Deseñar un **prototipo de etiquetador semántico multilingüe** centrado en substantivos.
- Integrar **semántica léxica, ontoloxías e técnicas de PLN** baixo un enfoque **modular e automático**.
- Crear **paquetes léxicos computacionais reutilizables** e **corpus anotados** que sirvan como base para novas aplicacións lingüísticas.
- Desenvolver un sistema **adaptable, sostible e interoperable**, válido para español, galego, francés e alemán.

NOVIDADES E IMPACTO

- **Enfoque multilingüe e sostible:** etiquetador semántico-ontolóxico automático para español, galego, francés e alemán. Un método modular único para todas as linguas.
- **Uso de recursos abertos** (WordNet, TreeTagger, Freeling, etc.) e **criterios de interoperabilidade** para garantir a sostibilidade e a reutilización de datos nunha mellora continua.
- **Paquetes léxicos computacionais reutilizables:** os paquetes léxicos son interlingüísticos e serven como recurso adestrable e ampliable para futuros proxectos.
- **Gold Standard multilingüe para a avaliación:** Desenvolvemento dun corpus Gold Standard que permite medir cuantitativamente a eficacia do etiquetador e guiar sucesivas fases de optimización.

APROXIMACIÓN METODOLÓXICA

- 🔗 **Recursos de partida**: léxicos existentes, corpus anotados, WordNet, Freeling, TreeTagger...
- 🧩 **Modelo bottom-up**: elaboración dunha ontoloxía adaptada a cada lingua a partir de datos reais.
- 🧠 **Etiquetado automático**: identificación e asignación de etiquetas semánticas mediante regras e algoritmos.
- 🔍 **Validación**: revisión manual (Gold Standard) para axustar e mellorar o rendemento do prototipo.
- 🔁 **Circularidade**: os datos alimentan e melloran sucesivamente o modelo e a anotación.

RESULTADOS

1. Compilación de **3600 paquetes léxicos multilingües** ✓

2. Creación dun **corpus Gold Standard** paralelo ✓

3. Elaboración dunha **ontoloxía bottom-up,** que mellora a adecuación cultural e lingüística en cada lingua ✓

4. Elaboración do **prototipo de etiquetador,** con resultados moi prometedores na **asignación automática de categorías semántico-ontolóxicas** ✓

FUTURAS LIÑAS DE TRABALLO E RETOS

Tradución automática de paquetes: explotar TraduWord para xerar versións automáticas de paquetes en diferentes linguas, revisando e filtrando o *output* para maximizar a interoperabilidade.

Mellora da **desambiguación polisémica** e identificación das principais **diverxencias cuantitativas** (volume de anotación) e **cualitativas** (erros ontolóxicos) entre etiquetas manuais e automáticas.

Desenvolvemento dun motor automático para o etiquetado directo de textos, automatizando desde a lematización ata a anotación semántica completa.

Colaboración: desenvolvemento de interfaces gráficas para que a comunidade lexicográfica aporte validacións.

MÁIS INFORMACIÓN
http: //portlex.usc.gal/

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES · Cofinanciado por la Unión Europea · AGENCIA ESTATAL DE INVESTIGACIÓN
ESMAS-ES+ ETIQUETADOR SEMÁNTICO MULTILINGÜE

Proxecto Etiquetador semántico multilingüe automático e sostible - ESMAS-ES+ (PID2022-137170OB-I00) financiado por MICIU/AEI/ 10.13039/501100011033 e por FEDER,UE