

## **TRATAMIENTO AUTOMÁTICO DE LA MORFOLOGÍA VASCA**

**Diseño y construcción de un procesador morfológico  
robusto para el euskara y del  
corrector ortográfico realizado con esa base.**

**Iñaki Alegria Loinaz**

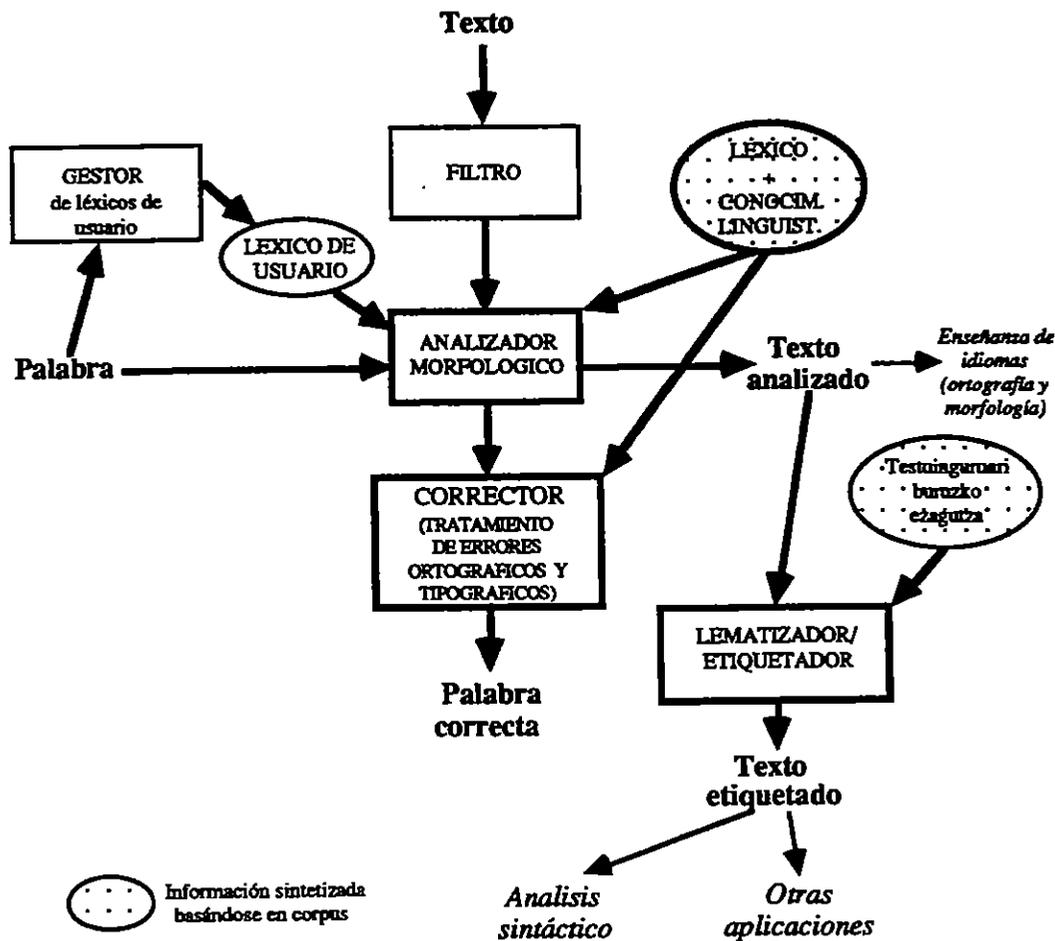
**Directores: Xabier Artola Zubillaga  
Kepa Sarasola Gabiola**

**Donostia, Junio de 1995.**

*Euskal morfologiaren tratamendu automatiko tresnak (Herramientas para el tratamiento automático de la morfología del euskara)* quiere ser un primer paso en el procesamiento automático del euskara.

Este trabajo debe ubicarse dentro de un amplio proyecto a largo plazo para impulsar y desarrollar el procesamiento automático del euskara, y para ello estamos trabajando en equipo un grupo de lingüistas e informáticos y recoge el diseño y ejecución de dos herramientas básicas: un procesador morfológico y un corrector ortográfico, siendo ambos de escala real y reusables, basados en normas y conocimiento morfológico pero verificados con corpus reales. Para el trabajo en la morfología ha sido fundamental la preparación de la base de datos léxica (EDBL) y el manejo de un conjunto de textos que han servido como fuente de conocimiento y como banco de pruebas para la evaluación de resultados.

El esquema general del trabajo realizado se refleja en la figura siguiente:



### La herramienta básica: el procesador morfológico automático.

El procesador morfológico se ha desarrollado en dos etapas, por un lado la construcción de un procesador morfológico estándar y posteriormente su complementación para conseguir un procesador robusto que reconoce un mayor número de palabras —mayor *coverage* o cobertura.

Las técnicas utilizadas en las dos fases se basan en la morfología de dos niveles evitando las soluciones particulares y consiguiendo un sistema totalmente homogéneo. Sobre la base del formalismo de dos niveles se han realizado tres extensiones: por un lado se le ha añadido la posibilidad de léxico(s) particular(es), por otro lado, se le ha dado una “nueva” aplicación en el tratamiento de variantes lingüísticas no estándares y por último se ha robustecido el sistema con el “análisis sin léxico” que anteriormente se había utilizado en fonología pero no en morfología.

Se han propuesto diversas mejoras del modelo de morfología de dos niveles durante el desarrollo de este proyecto.

### **El producto comercial: el corrector ortográfico Xuxen.**

Desde un principio la realización de un corrector ortográfico para el euskara era uno de nuestros objetivos centrales ya que coincidía plenamente con los objetivos de aplicación y escala real que nos habíamos marcado y es una herramienta de gran importancia para el euskara debido al proceso de normalización que se está desarrollando.

Teniendo en cuenta esos objetivos se ha dado prioridad al tratamiento de errores que surgen del incipiente proceso de normalización —los llamaremos variaciones lingüísticas y errores de competencia— dejando el tratamiento del resto de errores —los llamados errores tipográficos— en un segundo plano.

El proceso de corrección se compone de dos módulos complementarios: el de los errores de competencia y el de los errores tipográficos. Mientras que para el tratamiento de los primeros se ha llevado a cabo un tratamiento innovador basado en la morfología de dos niveles el tratamiento de los errores tipográficos se hace de forma bastante convencional centrándose el trabajo en cuestiones de eficiencia.