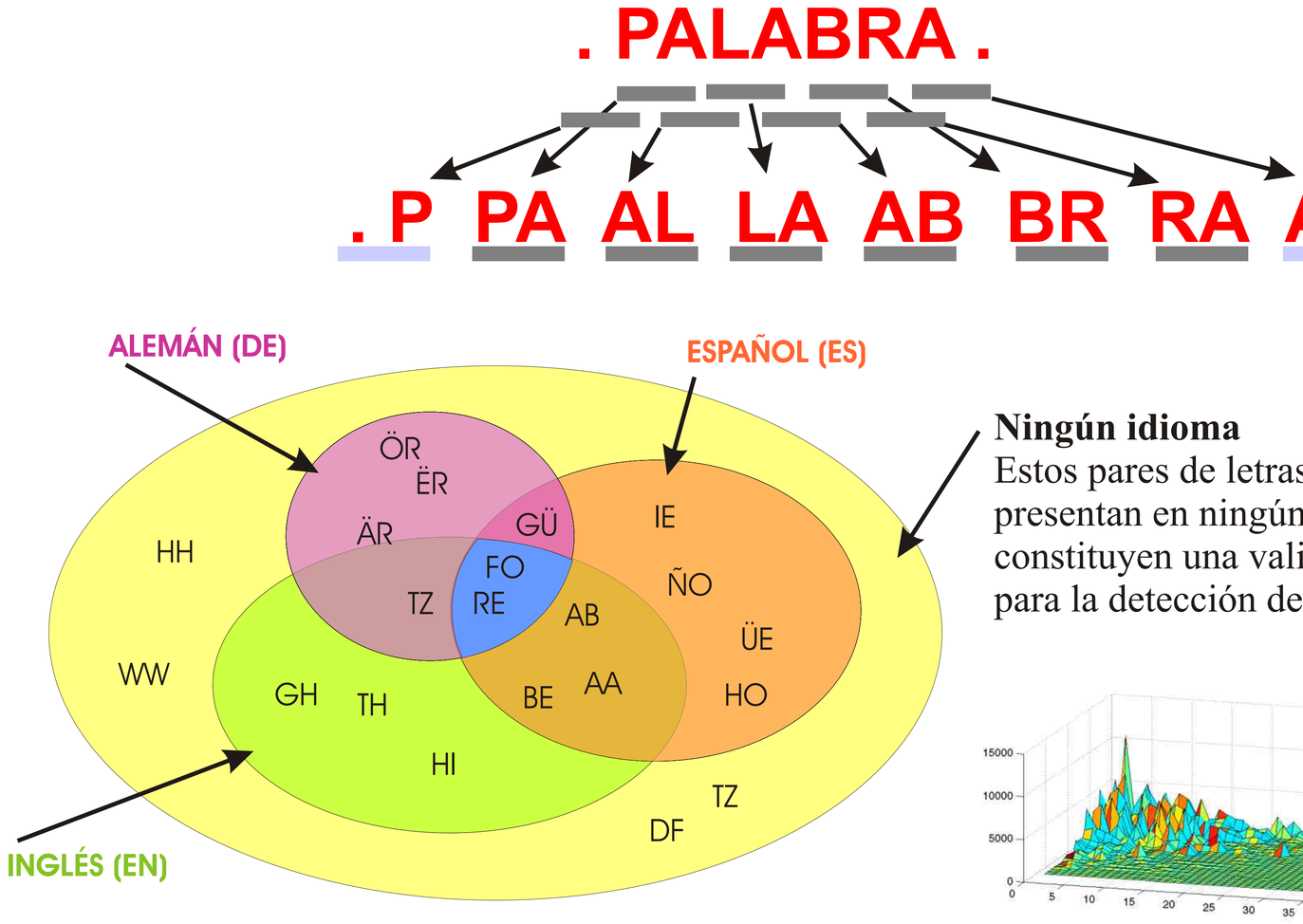
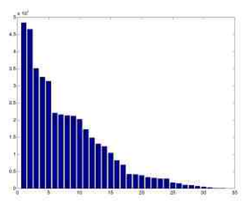
Algoritmos eficientes para detección temprana de errores y clasificación idiomática

Inteligencia Artificial
Area: Procesamiento de lenguaje natural (NLP)
Aplicación: procesadores/correctores de texto



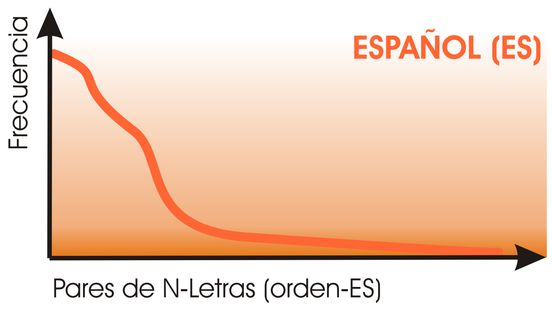
Ningún idioma
Estos pares de letras no se presentan en ningún idioma y constituyen una valiosa fuente para la detección de errores.

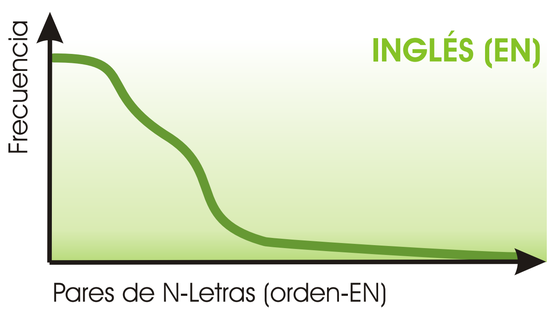
ES- Distribución de diletras
por: inicial, segunda (48424words)

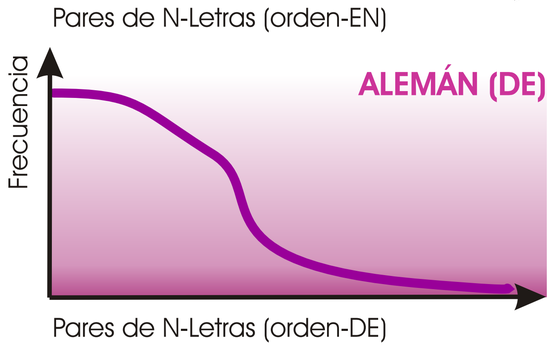


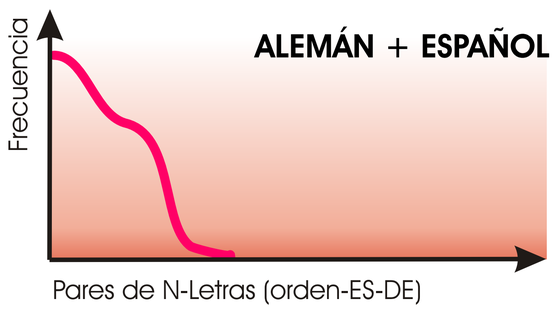
ES - Distribución Totalizada
(pares vs. letras intervinientes)

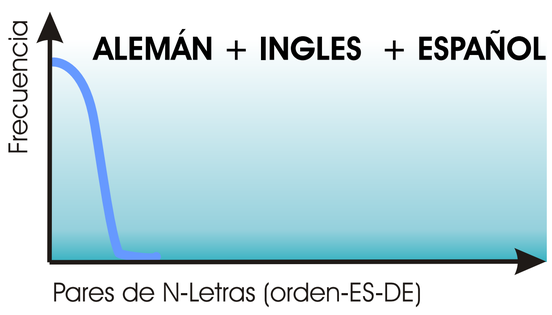
Detección de Idiomas











La intersección de **diletras** (N-gramas, N=2) entre idiomas, resulta en un área notáblemente pequeña, por lo que es apta para la ponderación probabilística.

Fundamentos
El área de pares de letras (**diletras**) inexistentes en cada idioma responde a que en cada lengua, hay un conjunto muy característico de transiciones fonéticas representadas, la cual se condice con la experiencia de "como suena" al oido.
Las intersecciones entre frecuencias son importantes (del orden del 30-40%) y en consecuencia, hay una alta probabilidad de determinar correctemente un idioma de un escrito sin recurrir más que a una tabla de frecuencias.

Palabras Mal Escritas

En cada idioma, aprox. entre el 40 y el 60% del universo de "**diletras**" son inexistentes (*)
Un filtrado con este criterio puede detectar en el orden del >50% de errores, alivianando cualquier procesamiento posterior.
En caso de rechazo, éste será rotundo y en caso de aceptación, ésta será en el idioma mas probable, generando un ránking de los idiomas candidatos, ordenados probabilísticamente.

Ventajas del Método
En un mismo proceso se evalúan simultáneamente los errores y la pertenencia a un idioma particular, posibilitando una posterior búsqueda óptima y más específica en los diccionarios.

(*) Están bajo de un umbral de significancia estadística.

Andres T. Hohendahl[1] *José F. Zelasco*[1,2]
1 Laboratorio de Estereología y Mecánica Inteligente. Dto. de Ing. Mecánica Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, (1065), Buenos Aires, Argentina.
2 INTIA, Facultad de Ciencias Exactas, UNCPBA, Universidad del Centro de la Provincia de Buenos Aires, Campus Universitario, Paraje Arroyo Seco (B7000) Tandil, Prov. de Buenos Aires, Argentina.