

---

## **DISEÑO E IMPLEMENTACIÓN DE MÉTRICAS DE SIMILITUD ENTRE FRASES**

**VALERIA ALEJANDRA ARAYA MORA  
INGENIERÍA CIVIL EN COMPUTACIÓN**

### **RESUMEN**

Los servicios de traducción automática son sin duda una herramienta útil en la actualidad, puesto que ayudan en la comunicación facilitando la comprensión de ideas expresadas entre dos idiomas. Sin embargo, aún necesitan de mucho refinamiento porque en general cometen bastantes errores. En el caso particular entre Inglés -Español dado que ambos idiomas se basan en reglas gramaticales tan diferentes, las traducciones de un idioma al otro y viceversa se tornan complejas. Los errores no sólo son gramaticales sino que además las traducciones en ocasiones suelen ser literales perdiendo por completo el sentido de la idea que el texto original desea expresar. Esta memoria pretende colaborar con las mejoras a las traducciones automáticas. Específicamente entre los idiomas Inglés y Español, basándose en la idea de que se puede implementar una buena medida de cálculo de similitud entre frases. El objetivo de una medida de este tipo es identificar y calificar frases candidatas a ser la mejor traducción para una frase que se traducen desde el Inglés al Español. Con esto, se pueden ofrecer más opciones de traducción para una idea y por ende, al ampliar la gama de opciones es posible capturar dentro de los resultados el real sentido que a una frase se le quiere impregnar. Para cumplir con lo anterior, en este trabajo se realiza una investigación sobre métodos existentes para medir similitud entre palabras, de los cuales se eligieron tres para su implementación. Luego del estudio de dichas medidas se realizan propuestas originales para medir similitud exclusivamente entre frases y que se complementen con las métricas para palabras.

De esto surgen tres medidas bases: Distancia de Edición adaptada, Medida secuencial y Métrica básica, de las cuales sólo las dos primeras se refinan con implementaciones que incorporen funciones para medir similitud entre palabras.

Finalmente, se testean las propuestas comprobando que en su conjunto son capaces de recuperar y posicionar bastante bien frases potenciales a ser mejores traducciones que una frase recibida en Español y que supone ha sido traducida desde el Inglés. Palabras claves: Traducción automática, similitud entre frases, noción básica de similitud, Distancia de Edición, Damerau-Levenshtein, WordNet.

## ABSTRACT

The machine translation services are certainly a useful tool at present, since they help in facilitating communication of ideas expressed understanding between two languages. However, much still needs refinement because generally commit numerous errors. In the particular case among English - Spanish since both languages are based on grammar rules so different translations from one language to another and vice versa become complex. Mistakes are not only grammatical but also sometimes translations are usually literal losing all sense of the idea that the original text wishes. This report aims to contribute to improvements specifically to automatic translations between English and Spanish, based on the idea that you can implement a good measure for calculating similarity between sentences. The aim of a measure of this type is to identify and qualify candidates for phrases best translation for a phrase translated from English to Spanish. With this, they can offer more translation options for an idea and thus to extend the range of options you can capture the results in the real sense as a phrase you want to impregnate.

To accomplish this, in this paper an investigation of existing methods is performed to measure similarity between words, of which three were chosen for implementation. After studying original proposals such measures are performed to measure similarity between sentences only and that complement the metrics for words. Edit Distance adapted, sequential measurement and basic metrics, of which only the first two are refined implementations that incorporate functions to measure similarity between words this basis three measures arise.

Finally, check that the proposals as a whole are able to recover well enough potential and position statements to be better received than a phrase translations in Spanish and has been translated means from English are tested.

Keywords: Machine Translation, similarity between sentences, basic notion of similarity, Edit Distance, Damerau-Levenshtein, WordNet.