Basque Research
Sitio web de la I+D+i vasca

NOTICIAS

Contacto:
Prentsa Bulegoa
UPV/EHU

Datos de contacto:
komunikazioa@ehu.es
(+34) 946012065

15/3/2010

Analizan los errores que se cometen en euskera para aplicarlos en correctores automáticos y programas de aprendizaje

El grupo IXA de la Facultad de Informática de la UPV/EHU lleva años investigando el desarrollo de sistemas (semi)automáticos beneficiosos para el euskera. Entre estos sistemas, se encontrarían el tratamiento automático de los errores en euskera y las herramientas que permiten el aprendizaje de la lengua con medios informáticos. Larraitz Uria, miembro del grupo IXA, ha fijado en su tesis doctoral presentada en la UPV/EHU las bases para el desarrollo de estos dos sistemas, mediante el establecimiento de varios criterios de análisis de errores y desviaciones.

La tesis doctoral de Uria se titula *Euskarazko erroreen eta desbideratzeen analisirako lan-ingurunea. Determinatzaile-erroreen azterketa eta prozesamendua* (Entorno de trabajo para el análisis de errores y desviaciones en euskera. Evaluación y procesamiento de errores con determinantes). En primer lugar, se han diferenciado los errores y las desviaciones, y ésta es una de las aportaciones de la investigación. Los errores son fallos en la ortografía o la gramática. Las desviaciones son palabras gramaticalmente correctas pero inapropiadas para un contexto determinado; están relacionadas con el registro o el dialecto. El objetivo es que los sistemas automáticos del futuro diferencien los dos conceptos, por lo que la distinción es relevante.

Uria informa sobre dos bases de datos en las que ya se han comenzado a recopilar ejemplos sobre errores y desviaciones. Han sido puestas en marcha por el grupo IXA, y están adaptadas a dos aplicaciones. En la primera se almacena la información necesaria para desarrollar los tratamientos automáticos de los errores en euskera (correctores, marcadores de variaciones dialécticas, etc.). En la segunda, se recopilan los datos que faciliten la creación de herramientas para el aprendizaje de la lengua con medios informáticos. Es totalmente inusual fusionar estas dos líneas, pero muchos de los datos para el tratamiento automático de errores son útiles para el aprendizaje con medios informáticos, y viceversa. Ésta es una de las aportaciones de este trabajo.

Imprescindible para desarrollar un detector de errores

Otra de las aportaciones de la tesis es el corpus, el cual está ya en funcionamiento y es el principal soporte en el que se apoyan las bases de datos. De ahí se están empezando a extraer los primeros ejemplos de errores y desviaciones, los cuales son imprescindibles para desarrollar un sistema que sea capaz de detectarlos. Se ha formado ya un corpus de 113.290 palabras, derivadas de la recopilación de textos de estudiantes de euskera de varios niveles. De la misma manera, se han incluido algunos textos de estudiantes de euskera técnico y de hablantes comunes. En este primer paso, se ha establecido una cantidad de información importante para comenzar el análisis, y se han definido los criterios para crear el corpus.

El próximo paso a seguir es el etiquetado. En esta tesis doctoral, y como punto de partida de la investigación, se han etiquetado mayoritariamente los errores cometidos con determinantes. Como los fallos con determinantes en euskera son poco comunes, pero a su vez son muy graves cuando se cometen, Uria ha considerado que es un ejemplo adecuado para realizar una primera prueba. De todas maneras, su intención en un futuro es desarrollar la detección de todo tipo de errores y desviaciones. Para el proceso de etiquetado se ha valido de EtikErro, un editor creado por el grupo IXA. Además de etiquetar errores, exporta a las bases de datos los ejemplos etiquetados, incluyendo también la

información lingüística necesaria para el análisis.

En cuanto a la fase de clasificación -justo después del etiquetado- se ha hecho una gran aportación. Se ha definido la estructura principal de la clasificación, desarrollando especialmente la categoría referente a los errores con determinante. Finalmente, y después de cumplir las fases ya mencionadas, se ha procedido a la creación de las dos bases de datos. Ambas almacenan los mismos ejemplos e información lingüística, pero también tienen diferencias. La base de datos para el tratamiento automático de errores en euskera incluye información técnica. En cambio, la base de datos para el aprendizaje del idioma con medios informáticos almacena información psicolingüística.

Primeros resultados del tratamiento automático

Uria ya ha realizado, junto al grupo IXA, las primeras pruebas para comprobar los resultados que da el tratamiento automático de errores basado en los instrumentos mencionados. Mediante una técnica y una serie de reglas adecuadas para los errores cometidos con determinantes, ha medido la precisión del tratamiento. Es decir, ha comprobado la eficacia del tratamiento con un programa informático. En un principio, la precisión fue sólo de un 45,5 %. Sin embargo, si previamente se eliminan los errores que no están etiquetados, el "ruido" desaparece y la precisión se eleva al 80 %. Uria ha concluido también que cuanto más extenso sea el corpus, mayor será su eficacia. La aportación que supone su tesis no es más que un primer paso para un reto de futuro.

Sobre la autora

Larraitz Uria Garín (Hernani, 1977) es licenciada en Filología Inglesa y Magisterio de Educación Primaria. Ha redactado la tesis bajo la dirección de Igone Zabala Unzalu y Montse Maritxalar Anglada, profesoras del Departamento de Filología Vasca y de la Facultad de Informática, respectivamente. Actualmente es investigadora en el grupo IXA de la UPV/EHU y en el grupo IKER de la Universidad de Baiona.

Dirección de Internet
www.ehu.es

Información adicional

Imágenes