UNIVERSIDAD DE GRANADA

FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Memoria individual del Trabajo Fin de Grado

SESGOS EN EL *SOFTWARE* DE PROCESAMIENTO DEL LENGUAJE NATURAL DE LA APLICACIÓN SIRI EN DISTINTAS VARIEDADES DEL ESPAÑOL DE ESPAÑA

Curso 2020/2021

Autor: **Begoña Fernández Martínez**

Tutora: **Rocío Díaz Bravo**

FIDUS INTERPRES

ÍNDICE DE CONTENIDOS

1. RESUMEN

Entre mi compañero, Manuel Torres Rodríguez y yo, planteamos nuestro Trabajo Fin de Grado como una investigación transversal y basada en métodos mixtos (cuantitativo y cualitativo) del *software* de procesamiento del lenguaje natural de la aplicación Siri en las distintas variedades del español de España.

El estudio comenzó con la elección de las variedades que queríamos analizar. Tras una revisión bibliográfica exhaustiva decidimos contar con 6 variedades: castellano, hablas andaluzas, hablas canarias, español en contacto con el catalán, el español en contacto con el euskera y el español en contacto con el gallego. A continuación, diseñamos una encuesta para que la respondiesen los voluntarios de las distintas zonas de España. Una vez recibimos las muestras de los participantes, se las reprodujimos directamente a Siri para comprobar la transcripción que la aplicación generaba. Por último, analizamos los resultados de las transcripciones y ofrecimos propuestas de mejora.

Antes de comenzar el estudio ya presagiábamos los resultados. Desde el principio, creímos firmemente que las variedades como el castellano, consideradas más estándares, proporcionarían mejores resultados que otras variedades como las hablas andaluzas o las canarias.

2. MOTIVACIÓN Y OBJETIVOS

Nuestro interés en lingüística computacional, *machine learning* y procesamiento computacional de variedades lingüísticas y dialectos es lo que nos ha motivado a llevar a cabo

esta investigación. Son temas de gran importancia en la actualidad, pero encontramos que aún no se han llevado a cabo estudios sobre el español y sus variedades dentro de los *software* de procesamiento natural del lenguaje. Los objetivos que perseguimos son aportar nuevos focos de investigación y dar visibilidad a la estigmatización que sufren muchas variedades del español, en este y en el resto de ámbitos, y explicar por qué es tan necesaria la mejora del rendimiento de aplicaciones de reconocimiento del habla.

3. ESTADO DE LA CUESTIÓN

Al comenzar nuestra revisión bibliográfica nos encontramos con un enorme **hueco investigador** en este campo, algo que hace este trabajo aún más relevante. El procesamiento del lenguaje natural no solo es una disciplina dentro de la inteligencia artificial, sino que también es un campo de estudio dentro de la lingüística, la informática o la psicología cognitiva, sin embargo, encontrar estudios que versasen sobre estas últimas resultó una tarea realmente complicada.

Otro reto imposible fue encontrar información sobre el funcionamiento interno de Siri. Aunque sí conseguimos encontrar datos sobre su historia y desarrollo, la programación de este *software* es toda una incógnita. Por este motivo, esta investigación ha tratado de arrojar luz sobre esta nueva área de investigación.

4. METODOLOGÍA

Para simplificar el trabajo tan desafiante que nos acontecía, decidimos acordar una distribución metódica de las tareas. En algunas ocasiones y por fuerza mayor, hemos trabajado por separado, aunque la mayoría de las veces hemos trabajado de manera conjunta.

En el desarrollo del estudio hemos tenido fases perfectamente diferenciadas. En primer lugar llevamos a cabo la revisión bibliográfica de las variedades del español en el marco teórico y la investigación sobre el procesamiento del lenguaje natural y Siri. Después diseñamos nuestra encuesta y sus instrucciones en un documento PDF, el cual enviamos a los participantes para ir recolectando las muestras. En primer lugar informábamos a los participantes sobre nuestra identidad y el tema principal de nuestra investigación. Decidimos no dar demasiados detalles sobre los objetivos que perseguíamos porque pensamos que podría influir en las muestras. A continuación les explicamos que debían grabar un audio y enviárnoslo por WhatsApp o por correo electrónico a una dirección que habíamos creado con este propósito. También les informamos de que al enviar el audio estaban dando su consentimiento para que utilizásemos el mismo con fines científicos.

En el audio debían responder a unas preguntas y describir unas imágenes. Las preguntas fueron diseñadas con el objetivo último de encontrar una respuesta en la que hallar rasgos lingüísticos característicos de cada variedad. Antes de la recopilación de las muestras, tuvimos que redactar una solicitud al Comité de Ética en Investigación de la Universidad de Granada. Una vez que esta Comisión emitió un informe favorable a nuestra investigación (Herrera-Viedma y O'Valle-Ravassa, 2021), comenzamos a recibir las pruebas de los voluntarios. Pese a que en un principio tan solo perseguíamos que este fuese un estudio cualitativo, también es cuantitativo, pues hemos contado con un total de 42 voluntarios, 7 por cada una de las variedades analizadas. La mayoría de los voluntarios que se han prestado a participar en nuestro Trabajo Fin de Grado son jóvenes, cursan aún estudios universitarios o los han terminado recientemente. De los 42 participantes, 25 son mujeres y 17 son hombres. Su

procedencia geográfica y el número exacto de la muestra de cada uno de los participantes está disponible en el punto I de los anexos.

Tras compendiar la transcripción realizada por Siri de todas y cada una de las muestras, llegó el momento de analizar los aciertos y los errores. La discusión no solo se ciñe a esos dos aspectos, sino que además profundizamos en los rasgos más repetidos en los planos fónico y gramatical y en el vocabulario utilizado en el plano léxico.

5. RESULTADOS Y CONCLUSIONES

Tras completar el proceso descrito anteriormente, extrajimos las siguientes conclusiones con respecto a los planos fónico, gramatical y léxico.

En el plano fónico, el rasgo que más errores genera en la transcripción del habla natural es la pérdida o la aspiración de la *-s* final de palabra. En cambio, ante la pérdida de otros sonidos finales como *-d*, la efectividad de Siri es casi del 100 %.

En relación con el plano gramatical, Siri se ha visto en apuros para identificar y procesar correctamente las sinalefas, hasta un total de 21 errores repartidos en todas las variedades. Esto nos llamó especialmente la atención ya que las sinalefas forman parte del lenguaje común y, aun así, el índice de acierto ha sido menor que con, por ejemplo, la pluralización del verbo *haber* impersonal, un rasgo muchos más estigmatizado y peor considerado.

Las imágenes que propusimos a los participantes han dado un gran juego para comprobar la riqueza léxica dentro de las distintas variedades del español de España. Aun existiendo gran variedad entre el vocabulario utilizado, Siri ha sabido procesar las distintas formas de referirse

a un mismo concepto como es el caso de *papas* y *patatas* o *carro de la compra* y *carrito de la compra*.

Tras este análisis y como se muestra en la ilustración 1, las variedades que se ven más perjudicadas durante el procesamiento y la transcripción de Siri son, con diferencia, las hablas andaluzas y canarias, como ya presagiábamos al comenzar la investigación. En cambio, el castellano y el español en las zonas bilingües muestra una eficacia casi perfecta, cometiendo Siri errores sobre todo por sinalefa o taquilalia.

Esto deja claro que, aunque la herramienta es bastante precisa, aún se necesita un conjunto de datos de entrenamiento o *input* más representativo de las distintas variedades del español para no contribuir a la estigmatización de las variedades del español que ya son discriminadas tanto lingüísticamente como en otros muchos aspectos sociales.

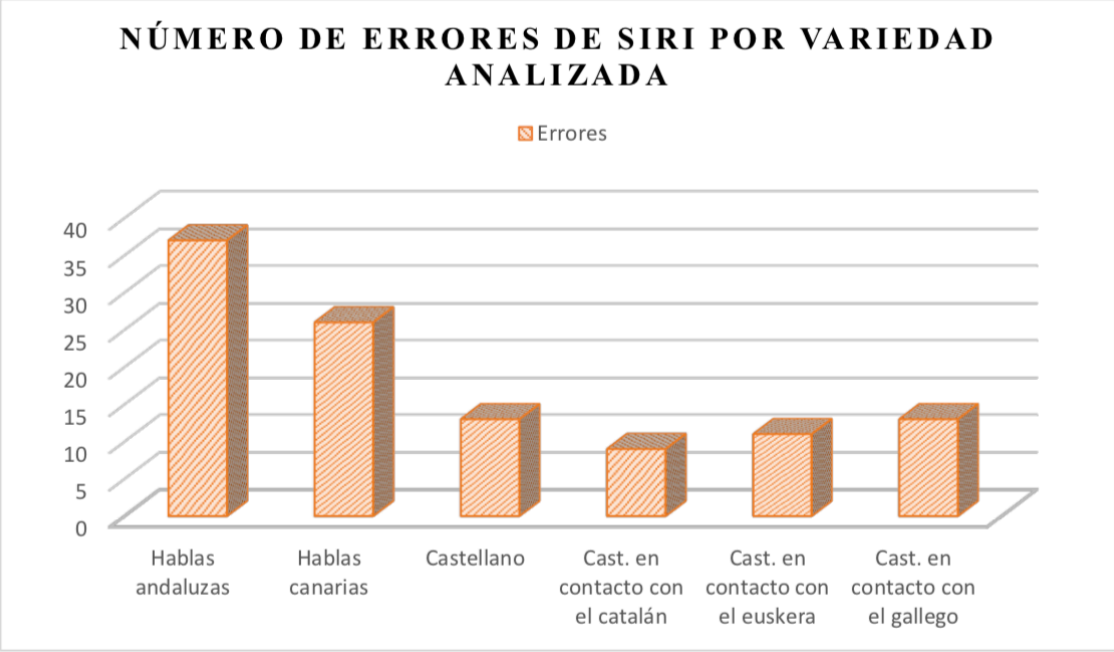

Ilustración 1. Número de errores de Siri por variedad analizada.

6. REFERENCIAS BIBLIOGRÁFICAS

Herrera-Viedma, E. & O'Valle-Ravassa, F. J. (2021). *Comité de Ética en Investigación de la Universidad de Granada*. Granada, España: Vicerrectorado de Investigación y Transferencia.