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: **Manuel Torres Rodríguez**

Tutora: **Rocío Díaz Bravo**

FIDUS INTERPRES

ÍNDICE DE CONTENIDOS

1. RESUMEN

Abordamos este Trabajo Fin de Grado como un estudio transversal y basado en métodos mixtos (cuantitativo y cualitativo) del procesamiento del lenguaje natural de las variedades del español de España en la aplicación Siri. Este proyecto ha consistido en la elaboración de una encuesta, a la que posteriormente han respondido voluntarios de distintas partes de España. El trabajo se ha completado con la transcripción que Siri ha aportado de las muestras, el análisis final de los resultados y nuestras propuestas de mejora.

Al iniciar la investigación, sin haber recopilado aún muestras, partíamos de varias hipótesis: nos temíamos que las variedades más estándares, como el castellano, eran mejor transcritas por los asistentes de voz virtuales y, en cambio, calculábamos que las hablas andaluzas y canarias iban a presentar mayores dificultades al ser interpretadas por Siri.

No solo despertaba nuestra atención lo inherente a las hablas andaluzas, canarias y al castellano, también nos interesa el contacto del castellano con las lenguas cooficiales de las comunidades autónomas bilingües: en primer lugar, Cataluña, las Islas Baleares y la Comunidad Valenciana; en segundo término, el País Vasco y el norte de Navarra, donde el español se encuentra en contacto permanente con el euskera; y, por último, Galicia, comunidad en la que el español comparte usuarios con el gallego.

2. MOTIVACIÓN Y OBJETIVOS

El interés compartido con la compañera Begoña Fernández Martínez en lingüística computacional, *machine learning* y procesamiento del lenguaje natural es lo que nos ha motivado a llevar a cabo esta investigación. Uno de los objetivos perseguidos es dar una explicación bien fundamentada de por qué se considera necesaria una mejora del

rendimiento de las aplicaciones de reconocimiento de habla. De igual modo, dar visibilidad a la estigmatización que sufren algunas variedades del español era otro de nuestros propósitos.

3. ESTADO DE LA CUESTIÓN

El desafío más importante al que nos hemos enfrentado ha sido el **hueco investigador** existente en la materia, hecho que ensalza la relevancia de este TFG. El estudio de los asistentes virtuales que utilizan el PLN siempre se ha abordado desde la disciplina informática, no desde la lingüística, de modo que encontrar estudios que abarquen la segunda de las disciplinas ha sido una tarea muy complicada.

Si bien es cierto que se puede saber la historia y la evolución de Siri a golpe de clic, no es tan sencillo obtener información de cómo procesa las distintas lenguas en las que está disponible y de su funcionamiento interno para un usuario lego en informática. Podemos decir, por tanto, que abrir esta nueva área de investigación y arrojar luz sobre ella han sido otras de nuestras finalidades.

4. METODOLOGÍA

Nada más comenzar el trabajo, mi compañera y yo acordamos una distribución metódica y precisa de las tareas. Para ello, establecimos las distintas partes del trabajo y cómo las íbamos a abordar: ha habido partes de las que solo se ha ocupado uno de los miembros y otras, la mayoría, que hemos trabajado en conjunto.

El estudio se ha dividido en tres fases bien diferenciadas:

a. La revisión bibliográfica de las variedades del español en el marco teórico y la investigación sobre el procesamiento del lenguaje natural y Siri. Esta primera fase, plenamente teórica, se ha caracterizado por el repaso de contenidos que ya hemos

estudiado en otras asignaturas del grado, como ‘Lengua A Nivel 3 Español’, y la búsqueda bibliográfica, principalmente artículos de profesores universitarios, acerca del procesamiento del lenguaje natural y de Siri. Las variedades incluidas en el estudio son: las hablas andaluzas, las hablas canarias, el castellano y el castellano en contacto con el catalán, el euskera y el gallego.

b. El diseño de una encuesta y la recopilación de muestras de cada una de las variedades. En primer lugar, la compañera Begoña se encargó de diseñar una encuesta en la que se recogían aspectos relevantes fónicos, gramaticales o morfosintácticos y, en última instancia, léxicos. La encuesta consta de una primera parte con preguntas sobre el lugar de origen del participante, sus estudios y la manera de contactar con su familia si alguna vez ha vivido fuera. La segunda parte de la encuesta contiene tres imágenes: una de un cazador acompañado de un perro, otra de una casa muy infantil y una última de un carro de la compra con varios productos en su interior.

   Redactamos una solicitud al Comité de Ética en Investigación de la Universidad de Granada previa a la recopilación de las muestras. 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 que decidieron aportar su granito de arena a este proyecto. Pese a que en un principio tan solo perseguíamos que este fuese un estudio cualitativo, también ha sido cuantitativo, pues hemos contado con un total de 42 voluntarios, 7 por cada una de las variedades analizadas: las hablas andaluzas, las hablas canarias, el castellano, el español en contacto con el catalán, el español en contacto con el euskera y el español en contacto con el gallego.

c. La discusión y el análisis de las muestras recibidas. Tras compendiar la transcripción realizada por Siri de todas y cada una de las muestras, llegaba el momento de analizar los aciertos y los errores. La discusión no solo se ha ceñido a esos dos aspectos, sino que también se profundiza en los rasgos más repetidos en los planos fónico y gramatical y en los vocablos usados en el plano léxico.

5. RESULTADOS Y CONCLUSIONES

Tras concluir el proceso descrito anteriormente, hemos extraído conclusiones reveladoras sobre el procesamiento de las variedades del español por parte de Siri.

En el plano fónico, los rasgos en los que más errores hemos detectado son la pérdida o la aspiración de la *-s* final de palabra y la pérdida de la *-d-* intervocálica. En cambio, ante la pérdida de *-d* final de palabra, la efectividad de Siri es plena y la procesa correctamente.

En relación con el plano gramatical, también existen datos que hablan por sí solos. Mientras que la mayoría de los voluntarios hacen un uso etimológico de los pronombres clíticos, Siri ha reconocido con acierto el leísmo en las cuatro muestras donde aparecía. No ha sido así con las sinalefas, rasgos fonético-sintácticos para los que Siri se ha visto en apuros a la hora de identificarlos.

Las imágenes han sido determinantes para analizar con detalle los planos morfosintáctico y léxico. Siri sabe distinguir a la perfección las distintas variantes que se han aportado para hacer referencia a un mismo concepto, como en el caso de *carro/carrito*, interesante desde el punto de vista morfosintáctico, y *papa/patata*, reseñable a la hora de analizar los rasgos léxicos. Para las legumbres, Siri ha identificado con acierto la mayoría de los términos empleados por los participantes: *habichuelas, judías*, *guisantes*, *garbanzos* y, en menor medida, *edamame*, palabra que no ha reconocido en una de las tres ocasiones que se ha empleado.

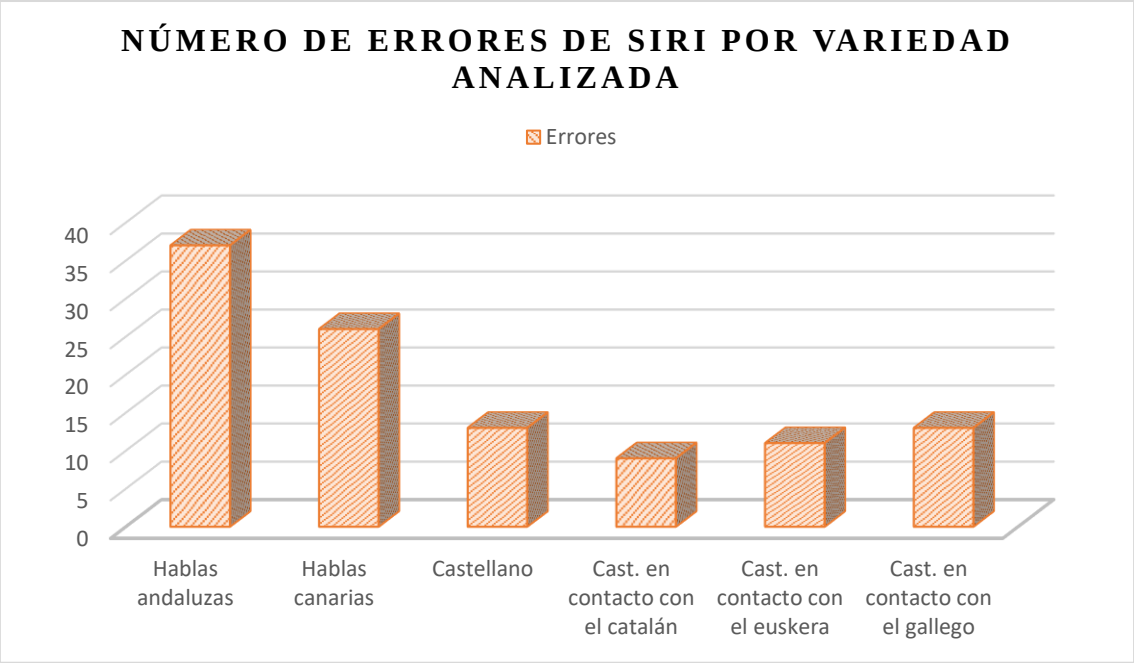

Ilustración 1. Número de errores de Siri por variedad analizada.

En definitiva, las hablas andaluzas y canarias son las que salen peor paradas al dictar a este asistente de voz virtual. Su eficacia es aceptable, pero mejorable. Las variedades castellanas en las comunidades autónomas bilingües cuentan, por su parte, con una eficiencia más que evidente. Por lo tanto, como imaginábamos, es necesaria una mejora de Siri para acabar con la discriminación relacionada con la variación diatópica, una forma más en la que se manifiesta la «brecha digital».

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.