Torres, S., García, C., Cabeza, C. y Docío, L. (2020). "LSE_Lex40_UVIGO: Una base de datos específicamente diseñada para el desarrollo de tecnología de reconocimiento automático de LSE". *Revista de Estudios de Lenguas de Signos REVLES*, 2: 151-172.

# LSE_Lex40_UVIGO: Una base de datos específicamente diseñada para el desarrollo de tecnología de reconocimiento automático de LSE

Soledad Torres Guijarro[a], Carmen García Mateo[a]

Carmen Cabeza Pereiro[b] y Laura Docío Fernández[a]

*Grupo de Tecnologías Multimedia – AtlanTTic Research Center – Universidade de Vigo*[a]

*Grupo GRADES – Universidade de Vigo*[b]

**RESUMEN**


Resumen en lengua de signos española.

El reconocimiento automático de habla puede considerarse una tecnología viable y madura, sobre la que se basan numerosas aplicaciones que facilitan la comunicación entre personas, y entre persona y máquina. Sin embargo, el reconocimiento automático de lenguas de signos no está tan avanzado como el de lenguas habladas; si lo estuviera, una persona sorda podría comunicarse con alguien que desconozca la lengua de signos sin necesidad de recurrir a un intérprete, ganando en privacidad e independencia. Y podría hacer uso de sistemas automáticos controlados por vídeo, que reconocieran sus instrucciones, de forma inclusiva, rápida y móvil, similarmente a como las personas oyentes acceden a sistemas controlados por voz.

Los grupos de investigación GRADES y GTM de la Universidad de Vigo nos proponemos avanzar en el desarrollo de un reconocedor automático de lengua de signos española (LSE) basada en el reconocimiento de imágenes. De la revisión del estado del arte se concluye la necesidad de desarrollar una base de datos de LSE diseñada específicamente para este fin. La complejidad de esta tarea aconseja abordarla de forma incremental, para lo cual nos proponemos como objetivo desarrollar una metodología de grabación que permita que la base de datos crezca a lo largo del tiempo en tamaño y complejidad. Esta metodología comprende la selección del léxico, el diseño del puesto de grabación, la estructura de almacenamiento de los datos, los programas informáticos de gestión de la base de datos de vídeos y los metadatos asociados, y la protección de los datos personales de las personas informantes.

Una versión inicial de la base de datos, denominada LSE_Lex40_UVIGO, está formada por múltiples repeticiones de 40 signos aislados en LSE, realizadas por distintos informantes. Esta primera versión de la base de datos nos será útil para desarrollar un reconocedor de signos aislados en entornos diversos y con independencia de la persona usuaria, y para demostrar la utilidad de la metodología de adquisición que se describe en esta contribución.

**Palabras clave**: reconocimiento automático de lengua de signos; RALS; ASLR; base de datos de LSE.

## 1. Introducción

Actualmente, el reconocimiento automático de habla puede considerarse una tecnología viable y madura, sobre la que se basan numerosas aplicaciones que facilitan la comunicación entre personas, y entre persona y máquina. Ejemplos de ello son los sistemas de conversión de voz a texto que se emplean como herramientas de dictado, y están disponibles en ordenadores personales e incluso en teléfonos inteligentes. El reconocimiento de habla puede ser también un primer paso para la traducción automática a otra lengua; o para la comunicación con una máquina, con ejemplos que abarcan desde sistemas de respuesta telefónica automática hasta la comunicación con asistentes virtuales. Así mismo, el reconocimiento automático de habla puede aplicarse al discurso presente en un material audiovisual, con el fin de subtitularlo o anotarlo, posibilitando así realizar análisis lingüísticos o búsquedas en su contenido.

El reconocimiento automático de lenguas de signos (RALS) no está tan avanzado como el de lenguas habladas; si lo estuviera, una persona sorda podría comunicarse con alguien que desconozca la lengua de signos sin necesidad de recurrir a un intérprete, ganando en privacidad e independencia. Podría hacer uso de sistemas automáticos controlados por vídeo, que reconocieran sus instrucciones, de forma inclusiva, rápida y móvil, similarmente a como las personas oyentes acceden a

sistemas controlados por voz. Además, si existieran reconocedores automáticos de lenguas de signos, sería mucho más sencilla la anotación de estas lenguas, de tal manera que se facilitaría el análisis lingüístico y se contribuiría al estudio de las mismas. Ayudándose de un RALS, una persona que estuviera aprendiendo lengua de signos podría practicarla con un ayudante virtual que le indicaría si está realizando los signos correctamente.

Para entender por qué no existe aún un reconocedor maduro de LSE, empezamos por analizar, en el apartado 2 y en el Anexo I, qué técnicas están siendo aplicadas al problema del reconocimiento de las lenguas de signos del mundo y qué elementos son necesarios para desarrollarlas. De la revisión del estado del arte se concluye la necesidad de desarrollar una base de datos de LSE diseñada específicamente para el reconocimiento de signos. Por este motivo, nos planteamos como objetivo de esta fase del proyecto desarrollar una metodología de adquisición de esta base de datos que nos permita abordar en un futuro el reconocimiento automático de LSE.

Para adquirir una base de datos para RALSE primero hay que seleccionar un léxico, diseñar el puesto de grabación (equipos de grabación, sistema de adquisición, iluminación), diseñar una estructura de almacenamiento de la base de datos, preparar un programa informático para gestionar la base de datos y ayudar a introducir la información relativa a cada grabación (metadatos), y proteger los datos personales garantizando los derechos digitales de las personas informantes. La forma en que se han abordado todas estas tareas se describe en el apartado 4.

En la definición de esta metodología de creación de la base de datos para RALSE se ha tenido presente en todo momento una premisa fundamental: que la base de datos fuera incrementable. De esta forma, su tamaño y complejidad podrá ir creciendo a medida que seamos capaces de abordar problemas de RALSE más complejos, contemos con la participación de más informantes, y grabemos en entornos diferentes.

La primera base de datos adquirida con esta metodología es una base de datos de signos aislados, que denominamos LSE_Lex40_UVIGO. Esta base de datos, y la metodología empleada para su grabación y sus futuras ampliaciones, son la contribución fundamental de este trabajo. El progreso actual de las grabaciones, junto con las estrategias que emplearemos para ampliarlas, se exponen para finalizar en el apartado 5.

## 2. El aprendizaje automático aplicado al reconocimiento de lengua de signos

Las técnicas que nos proponemos aplicar al RALS se basan en el aprendizaje automático, o aprendizaje de máquinas (del inglés machine learning). Es una rama de la inteligencia artificial que enseña a las máquinas a aprender a partir de la experiencia. Esto se logra desarrollando algoritmos matemáticos que, en una primera fase, se "entrenan" para aprender a distinguir unos elementos de otros, y en una segunda fase, se emplean para reconocer una nueva representación de algún elemento previamente aprendido.

Los elementos que una máquina puede aprender a reconocer son de muy diversa índole: la matrícula de un coche a partir de una fotografía, una huella digital escaneada, el mensaje contenido en un vídeo de una persona signando, o en una grabación de voz. En cualquier caso, el primer elemento necesario para enseñar a la máquina es una base de datos suficientemente grande para abarcar toda la variabilidad presente en aquello que debe aprender. Esta base de datos debe estar además anotada, es decir, cada elemento incluido en la base de datos necesita estar identificado mediante una etiqueta que le indica a la máquina qué información contiene. Por ejemplo, la base de datos de fotos de matrículas debe contener fotos de matrículas con todos los posibles formatos, encuadres, e iluminaciones, y cada foto de la base de datos debe estar etiquetada con el conjunto de cifras y letras que la componen.

Aplicando estas ideas a nuestro problema, y tras un análisis detallado del estado del arte (ver Anexo I), llegamos a la conclusión de que para entrenar un reconocedor automático de lengua de signos hace falta una base de datos que cumpla los siguientes requisitos: debe estar anotada, haber sido adquirida con la tecnología adecuada, y contar con un número de repeticiones y de informantes suficientemente alto para que el proceso de aprendizaje de la máquina sea robusto frente a variaciones en la realización de los signos o en el entorno donde se graba el vídeo.

Desafortunadamente, solo algunas lenguas de signos ofrecen bases de datos lingüísticas con material suficiente para permitir el entrenamiento de reconocedores complejos (Tilves-Santiago *et al.*, 2018). Una de ellas es la RWTH-PHOENIX-Weather, que consta de 5.356 oraciones y 1.200 signos, aproximadamente 600.000 fotogramas de lengua de signos alemana, procedentes de la información meteorológica emitida por televisión. La lengua de signos que tiene más bases de datos disponibles es la estadounidense, por ejemplo la conocida ASL Lexicon Video Dataset (Athitsos *et al.*, 2008); aunque estas bases de datos son más pequeñas que la RWTH-PHOENIX-Weather, son lo suficientemente grandes como para usarlas en reconocimiento automático. En UK hay una interesante base de datos de lengua de

signos británica (BSL) (ESRC, 2018), que sirve de base a un proyecto de reconocimiento y traducción al inglés escrito (ExTOL, 2018).

El estudio del estado del arte resumido en los siguientes párrafos y en el Anexo I demuestra que la LSE está absolutamente infrarrepresentada en la literatura científica dedicada a reconocimiento automático de lenguas de signos. Se han realizado algunos esfuerzos para recopilar la variedad de signos de LSE, pero para fines muy diferentes al reconocimiento automático. El Centro de Normalización Lingüística de la lengua de signos española (CNLSE) lleva años desarrollando un corpus en colaboración con numerosas asociaciones y centros de investigación del estado. Se trata de grabaciones anotadas de discurso espontáneo muy útiles para recoger la variación geográfica, generacional, de género y de tipo de discurso de la LSE, pero poco apropiadas para el entrenamiento del RALS en una primera fase, que requeriría de una base de datos con un elevado número de repeticiones por signo y una clara segmentación temporal de los signos recogidos en las grabaciones.

El Basque Center on Cognition Brain and Language (BCBL), con la colaboración de la Confederación Estatal de Personas Sordas (CNSE), ha recopilado el corpus LSE-Sign (Gutierrez-Sigut *et al.*, 2016). Contiene 2.400 signos y 2.700 no-signos, anotados gramaticalmente a partir del diccionario estandarizado de LSE de 2008 (CNSE, 2008). A pesar de que este corpus controlado es muy útil para estudiar la variabilidad de los signos en LSE y se puede utilizar para probar algoritmos de extracción de características tanto para signos estáticos como dinámicos, la escasa variabilidad de las personas informantes (un hombre y una mujer que signan medio diccionario cada uno) y la pequeña resolución de la imagen corporal la excluye de su uso para el entrenamiento de modelos de aprendizaje automático que permitan el reconocimiento de signos continuos independiente de quién signe.

Martínez-Hinarejos, del Centro de Investigación de Tecnología de Reconocimiento de Patrones y Lenguaje Humano de la Universidad Politécnica de Valencia (Martínez-Hinarejos y Parcheta, 2017), adquirió una base de datos completamente diferente: empleó el sensor de infrarrojos Leap Motion que captura, a corta distancia, la posición de las manos y los dedos, de manera similar a un guante electrónico pero sin contacto. La base de datos, disponible para el público, está compuesta por un conjunto principal de 91 signos repetidos 40 veces por 4 personas (3.640 adquisiciones) y un subconjunto de 274 oraciones formado por 68 palabras del conjunto principal. Por limitaciones tecnológicas, Leap Motion solo puede captar el movimiento de las manos si están cerca del dispositivo y no se tapan la una a la otra; no captura los brazos, los movimientos corporales ni las expresiones faciales. En cualquier caso, este sistema es bastante útil para probar la precisión de los modelos

dinámicos en 3D para detectar signos basados en las manos y los dedos y, por supuesto, el deletreo manual.

En este mismo sentido el grupo Graphics, Interaction & Learning Technologies de la Universidad de Porto está desarrollando un proyecto de comunicación bidireccional para la lengua de signos portuguesa, VirtualSign, que ahora se está ampliando a otras lenguas europeas. La parte de generación de signos a través de un avatar está muy avanzada, siendo relativamente fácil, según sus creadores, añadir una lengua nueva una vez que los expertos lingüistas y los conocedores de la lengua de signos especifican todos los movimientos que debe hacer el avatar para cada signo. Sin embargo la parte de reconocimiento de signos visuales para convertir el mensaje signado a texto no está tan avanzada, debido de nuevo a la escasez de bases de datos anotadas, y además está restringida a utilizar guantes sensorizados.

Del estudio del estado del arte que se resume en el Anexo I se concluye que contamos con diversos elementos que facilitan el desarrollo de un RALS: parece viable emplear sólo cámaras de vídeo y sensores de profundidad para grabar los signos, prescindiendo de elementos hápticos, como los guantes sensorizados, que encarecen el sistema y limitan su uso a situaciones específicas. También existen técnicas prometedoras de segmentación y reconocimiento de gestos (movimiento de manos, tronco y expresiones faciales), basadas en procesado de imagen y aprendizaje automático, que pueden aplicarse a reconocimiento de signos.

Por otro lado, existen técnicas maduras de reconocimiento de voz, basadas en procesado de señal, que podrían ser aplicables a RALS. Pero hay que tener en cuenta que las reglas de articulación, secuenciación de palabras y gramática en LSE dependen mucho de la comunidad e incluso la localidad. La lengua de signos también varía mucho con cada persona y su forma de expresar la misma idea[1].

Una diferencia importante entre una lengua hablada y una signada, desde el punto de vista del reconocimiento automático, es el número de fonemas o primitivos estructurales para construir los mensajes. Frente a los 22 ó 24 fonemas del español, la LSE tiene, según Herrero (2009), 42 configuraciones, 24 orientaciones (6 de dedos por 4 de palma), 44 lugares (16 en la cabeza, 12 en el tronco, 6 en la mano/brazo dominado y 10 en el espacio), 4 movimientos direccionales y 10 formas de movimiento; si bien no existe unanimidad en esta clasificación (ver, por ejemplo, CNSE, 2008).

[1] La variabilidad de una lengua con la comunidad o la localidad no tienen que ver con su modalidad, hablada o signada, sino con el estado de estandarización y, secundariamente, con la existencia o inexistencia de una tradición escrita.

Otra diferencia clave entre la lengua hablada y la lengua de signos es que la última depende mucho más de la expresión facial y los movimientos corporales para generar el mensaje. Por ejemplo, para hacer una pregunta polarizada (es decir, de las que esperan una respuesta sí o no) en lengua de signos se usan otros canales además del movimiento de la mano: levantar las cejas, abrir los ojos ampliamente e inclinar el cuerpo hacia adelante. Otros tipos de pregunta suelen ir asociadas a otra configuración de las cejas.

Debido a esta complejidad, y a la necesidad de adquirir nuestra propia base de datos, parece razonable abordar primero el reconocimiento de signos aislados, lo que nos permite soslayar los problemas que conlleva la coarticulación, es decir, la mezcla del final de un signo con el inicio del siguiente. Además, comenzaremos por un número pequeño de signos aislados, y haremos crecer el léxico progresivamente en versiones sucesivas de la base de datos.

## 3. Reconocimiento de signos aislados

Las tareas involucradas en el reconocimiento de signos aislados se representan en la Figura 1 en forma de diagrama de bloques. En la primera fase, correspondiente al entrenamiento, la base de datos de vídeos anotados se procesa para extraer las características relevantes; a continuación, la base se divide en tres partes: una parte se emplea como conjunto de entrenamiento, a partir del cual se definirán los modelos matemáticos que representarán a los signos; otra parte sirve para ajustar los parámetros de los modelos; y una tercera se utiliza como conjunto de prueba para evaluar los modelos. Una vez realizado este proceso de entrenamiento, el sistema está preparado para pasar a la fase de reconocimiento, en la que se emplean los modelos previamente entrenados para identificar signos presente en un nuevo vídeo.

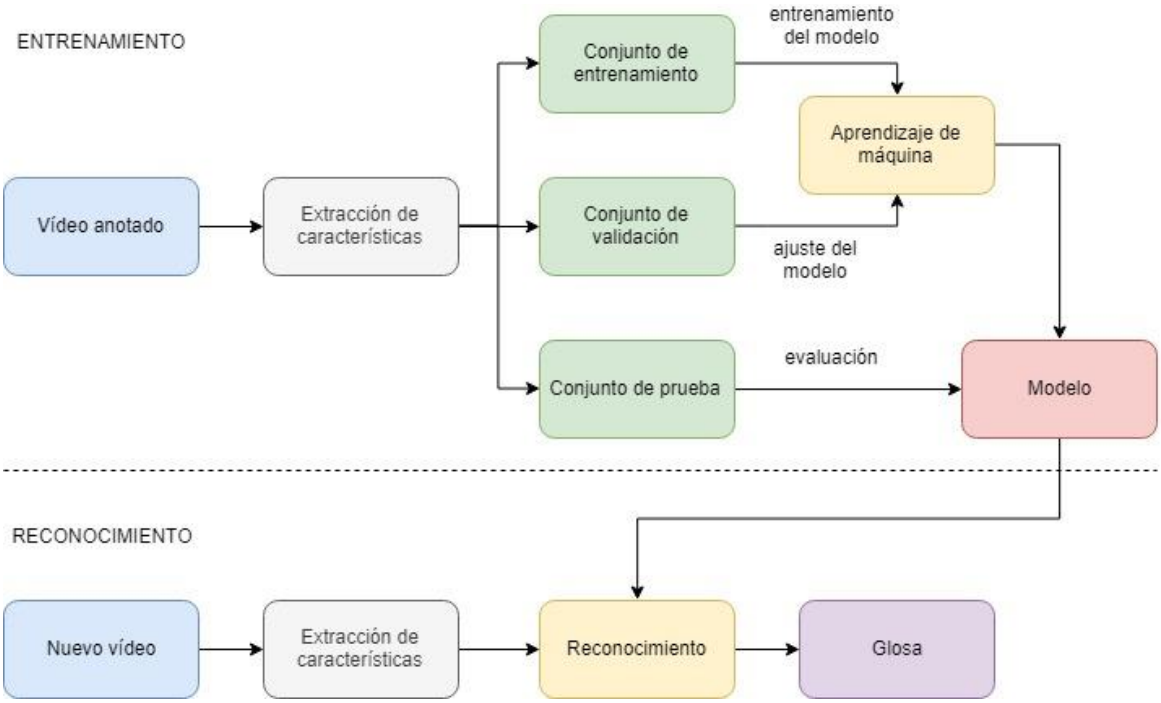


Figura 1: tareas involucradas en el reconocimiento automático de signos aislados.

La primera tarea a abordar es generar una base de datos de LSE que contenga vídeos de signos aislados, anotados y segmentados temporalmente, es decir, con marcas temporales de comienzo y final de signo; con un léxico inicialmente reducido pero que irá creciendo progresivamente; y con numerosas repeticiones de cada signo, realizadas por distintos informantes.

## 4. Desarrollo de LSE_Lex40_UVIGO

Se describe a continuación la metodología de desarrollo de una base de datos incrementable de LSE para reconocimiento automático, y su aplicación a la creación de la primera versión de la base de datos, LSE_Lex40_UVIGO, compuesta por múltiples repeticiones de 40 signos aislados.

La metodología de desarrollo de la base de datos incluye las siguientes tareas: selección del léxico, diseño del puesto de grabación (equipos de grabación, sistema de adquisición, iluminación), diseño de la estructura de almacenamiento de la base de datos, preparación del programa informático para gestionar la base de datos y la información relativa a cada grabación (metadatos), y protección de los datos personales garantizando los derechos digitales de las personas informantes.

### *4.1. Selección del léxico de la base de datos*

La utilidad de la base de datos LSE_Lex40_UVIGO es entrenar un reconocedor de signos aislados, que pueda funcionar con cualquier signante. Por ese motivo el número de signos se limita inicialmente a 40, de manera que la mayoría de las personas informantes puedan signarlos en su totalidad en una sola sesión de grabación, lo que nos ayuda a obtener múltiples repeticiones de cada signo realizadas por personas diferentes.

Los 40 signos se han seleccionado atendiendo a criterios lingüístico-morfológicos de forma que queden reflejadas distintas modalidades de articulación que puedan afectar a la complejidad del reconocimiento automático del gesto. Así, la base de datos contiene signos estáticos y dinámicos, con intervención de una o las dos manos, con movimiento simétrico o asimétrico de las dos manos, con distintas configuraciones de la(s) mano(s), orientaciones de la(s) palma(s), localización espacio-temporal, contacto con diferentes partes del cuerpo, modificadores de velocidad, oclusiones entre mano y mano o mano(s) y cara, etc.

Por otro lado, se han escogido signos de uso común, y que se articulan habitualmente de forma parecida, intentando evitar así que las realizaciones de un mismo signo por parte de informantes diferentes fueran muy distintas entre sí, lo que dificulta su aprendizaje automático. Por el mismo motivo, durante el proceso de grabación se solicita a la persona informante el signo que queremos, y no una equivalencia en español, para evitar que produzca un sinónimo o una variante. Algunos de los signos escogidos, en concreto los numerales, suelen presentar de hecho mucha variación en las lenguas signadas (Schembri y Johnston, 2012).

Los 40 signos seleccionados, recogidos en la tabla 1 junto con sus identificadores, están agrupados en cuatro bloques de 10 signos cada uno: estáticos (con sólo un pequeño movimiento local), monomanuales, bimanuales simétricos y bimanuales asimétricos. La estructuración en bloques facilita el proceso de grabación, permitiendo pausas entre bloques para que la persona informante descanse si lo necesita. En la página (GTM, 2019) se enlazan vídeos de referencia para cada signo.

| Signos estáticos | Signos monomanuales | Signos bimanuales simétricos | Signos bimanuales asimétricos |
|---|---|---|---|
| w0001: Uno | w0011: Bien | w0021: Bicicleta | w0031:Intérprete |
| w0002: Dos | w0012: Otro | w0022: Ganas | w0032: Carácter |
| w0003: Tres | w0013: Barato | w0023: Asociación | w0033: Deporte |
| w0004:Cuatro | w0014: Colegio | w0024: Trabajar | w0034: Arroz |
| w0005: Cinco | w0015: Contento | w0025:Integración | w0035: Lunes |

| Signos estáticos | Signos monomanuales | Signos bimanuales simétricos | Signos bimanuales asimétricos |
|---|---|---|---|
| w0006: Seis | w0016: Mujer | w0026: Sufrir | w0036: Viernes |
| w0007: Siete | w0017: Hombre | w0027: Abril | w0037: Árbol |
| w0008: Ocho | w0018: Gallego | w0028: Oscuro | w0038: Hasta |
| w0009:Nueve | w0019:Identidad | w0029: Ayudar | w0039: Ascensor |
| w0010: Diez | w020:Sentir | w0030: Ensalada | w0040: Calle |

Tabla 1: léxico de LSE_Lex40_UVIGO

### *4.2. Adquisición de los datos*

La captura de las imágenes de vídeo se realiza empleando dos cámaras: una cámara digital Nikon D3400 que capta vídeos RGB a 50 imágenes por segundo y con obturación de 1/240 segundos para congelar el movimiento, y un sensor Kinect 2, que proporciona los canales RGB con una resolución mínima de 640x480 y además el canal de profundidad. Este último facilita la segmentación de las manos y la cara (Tilves-Santiago *et al*., 2018). Además de las señales de vídeo y profundidad, se almacenan también las posiciones de los puntos clave del cuerpo de la persona (dedos, palma, codos, ojos, boca…) que proporciona el programa de desarrollo (SDK) de la Kinect 2.

Se ha diseñado cuidadosamente el puesto de grabación para, aun siendo fácilmente transportable, facilite las operaciones de encuadre, enfoque, iluminación y ajuste de distancia a la persona informante. Como puede verse en la figura 2, consta de un trípode, donde se instalan la cámara de vídeo y la Kinect 2, una encima de la otra; el ordenador portátil al que se conectan ambas a través de puertos USB; y dos focos.

Figura 2: puesto de grabación desplegado en un aula de la Universidad de Vigo.

Para facilitar la introducción de los metadatos de la sesión de grabación y de la persona informante en la base de datos, se ha desarrollado una plataforma de adquisición programada en MatLab®, que permite además grabar simultáneamente vídeos de las dos cámaras. En la figura 3 se muestra la ventana de introducción de información de la sesión de grabación (fecha y lugar, operador/a, cámaras) y en la figura 4, la ventana de introducción de los metadatos de la persona informante. Éstos consisten en: nombre, sexo, año de nacimiento, lengua nativa, colegio, mano dominante, lugar de residencia, colegio donde estudió, si es sorda u oyente, y a qué edad se quedó sorda, en su caso.

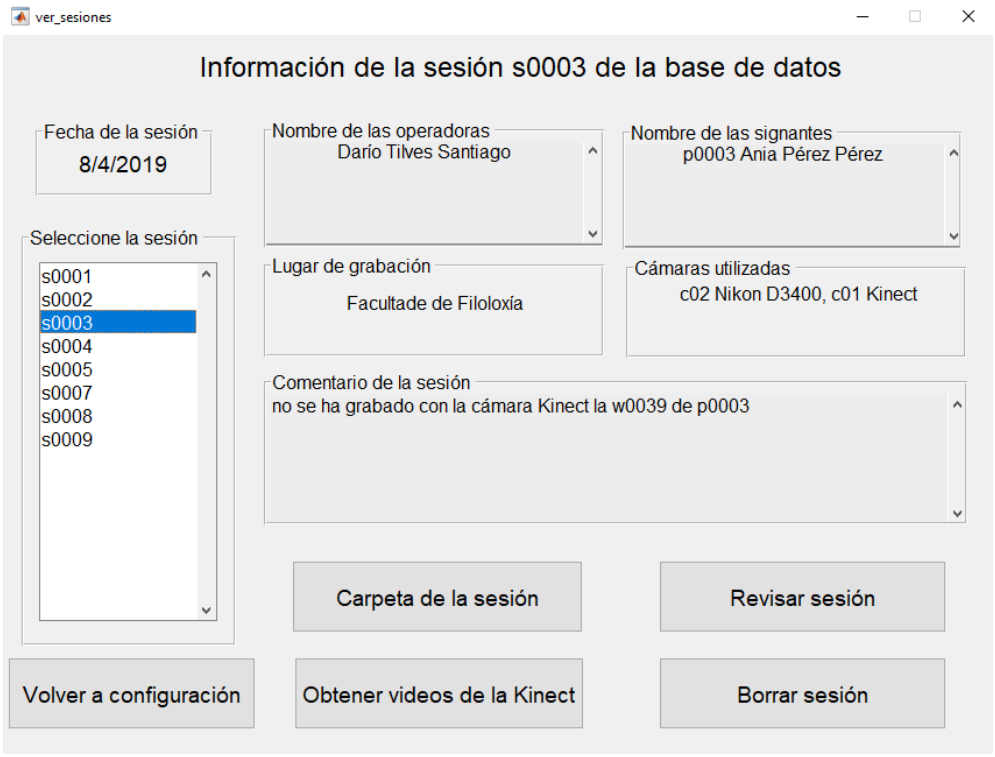


Figura 3: ventana de introducción de datos de la sesión de grabación.


Figura 4: ventana de introducción de datos de la persona informante.

### *4.3.Estructura de almacenamiento de la base de datos*

La organización de la información dentro del dispositivo donde se almacena la base de datos se ha diseñado para facilitar la identificación de cada vídeo, y el crecimiento progresivo del número de sesiones de grabación e informantes. En el Anexo II describimos la estructura de directorios donde se almacenan los vídeos y metadatos.

### *4.4.Protección de datos personales y garantía de los derechos digitales*

Con posterioridad a la primera sesión de grabación en la que toma parte, cada persona informante recibe un correo electrónico en el que se le informa, por escrito y en LSE, sobre los objetivos de proyecto y el equipo investigador que lo desarrolla. También se solicita su autorización para que dicho equipo pueda (i) utilizar los vídeos grabados de la persona informante, (ii) mostrarlos en publicaciones científicas, congresos y una web de la Universidad de Vigo, y (iii) cederlos a otros grupos para que los usen en sus investigaciones con las mismas condiciones y garantías. Cada informante es libre de consentir el uso para los fines que desee y, caso de no hacerlo, sus vídeos son borrados de la base de datos. Los consentimientos se almacenan de forma segura mediante el procedimiento ARCO, que se puede consultar en (GRADES y GTM, 2019).

Los grupos GRADES y GTM de la Universidad de Vigo se obligan a posibilitar a las personas informantes, en todo momento, el ejercicio de los derechos fundamentales de acceso, rectificación, cancelación, oposición, limitación del tratamiento y portabilidad sobre los datos objeto de tratamiento, tal y como regula la ley de protección de datos (España, 2018). La explicación relativa a los derechos de protección de datos se ha traducido a la LSE para facilitar la comprensión por parte de las personas signantes que colaboran.

## 5. Primeras grabaciones y líneas futuras

Por ahora, 20 informantes han contribuido a la base de datos. De estos 20 informantes, 12 son mujeres (9 sordas y 3 oyentes), y 8 hombres (7 sordos y 1 oyente); todos tienen la mano derecha dominante. La mayoría de informantes (12) son mayores de 50 años; y la mayoría (12) quedó sorda antes de los dos años. En relación con la edad a la que aprendieron la lengua de signos, dos informantes aprendieron antes de los 5 años, cinco entre los 5 y los 10, siete entre los 10 y los 20, un hombre sordo de nacimiento entre los 20 y los 30, y una mujer sorda de nacimiento aprendió con más de 50 años.

La misión de LSE_Lex40_UVIGO es servir para entrenar un reconocedor automático de signos aislados que funcione correctamente para todas las personas. Por ello, necesitamos incorporar más signantes, a ser posible de distinta edad, género, talla, vestimenta, mano dominante, procedencia, centro de escolarización, grado de destreza, etc. Nuestra estrategia en este sentido es recurrir a asociaciones de sordos de distintas localidades para incorporar el mayor número posible de informantes, y a posteriori, si fuera necesario, seleccionar el conjunto de datos de entrenamiento de forma balanceada para evitar posibles sesgos en el reconocedor.

El RALSE también debe funcionar correctamente en entornos diversos, esto es, ser robusto frente a cambios en la iluminación, el fondo, la distancia a la cámara o el encuadre. Por ahora, la mayor parte de los vídeos se han grabado en ASORVIGO, y el resto en la Escuela de Ingeniería de Telecomunicación y en la Facultad de Filología y Traducción de la Universidad de Vigo. En los tres casos la distancia a las cámaras y el encuadre fue similar, mientras que en el fondo de la imagen hay variaciones: es una pared desnuda pintada de color claro en dos de las localizaciones, y está cubierto por una tela verde para eliminar reflejos en la tercera. En futuras grabaciones incorporaremos otras localizaciones, condiciones de iluminación y tipos de fondo para mejorar la robustez del RALSE frente a este tipo de variaciones en las condiciones de grabación.

Por último, nos proponemos incrementar el léxico de la base de datos para que incorpore frases formadas por varios signos. Comenzaremos con frases sencillas de uso común, de entre dos y cinco signos concatenados, que tengan sentido completo y que se empleen bien en interacciones casuales, bien para solicitar indicaciones o informaciones sencillas. Más adelante, nos centraremos en frases relacionadas específicamente con una aplicación concreta del RALSE, aún por determinar. Para estas nuevas grabaciones emplearemos la misma metodología descrita en este artículo, que ya ha sido desarrollada previendo futuras ampliaciones de la base de datos.

## Agradecimientos

Esta investigación está financiada por el Ministerio de Ciencia, Innovación y Universidades, a través del proyecto RTI2018-101372-B-I00 *Análisis audiovisual de los canales de comunicación verbal y no verbal* (Speech&Signs); por la Xunta de Galicia y el Fondo de Desarrollo Regional Europeo a través de la Agrupación Estratégica Consolidada AtlanTTic (2016-2019); y por la Xunta de Galicia a través del Grupo de Potencial Crecimiento 2018/60.

Las autoras quieren expresar su agradecimiento a la Asociación de Personas Sordas de Vigo (ASORVIGO) y a la Federación de Asociaciones de Personas Sordas de Galicia (FAXPG) por su colaboración en la grabación de la base de datos LSE_Lex40_UVIGO. Agradecemos también los comentarios de las personas revisoras, que han contribuido notablemente a mejorar la calidad de este artículo.

## Anexo I: Estado del arte del reconocimiento automático en lengua de signos

Afortunadamente, debido a las similitudes entre la lengua signada y la hablada desde el punto de vista del procesamiento de la señal, muchos de los modelos dinámicos utilizados con éxito en reconocimiento automático de habla también se pueden aplicar a la lengua de signos. De hecho, desde mediados de los 90, los modelos ocultos de Markov (HMM) se han utilizado en los primeros sistemas para decodificar signos dinámicos, así como algunas oraciones con una estructura muy restringida (Starner *et al.*, 1998). Vogler y Metaxas pronto se dieron cuenta de que el signo es un proceso de varios canales paralelos (forma de la mano, orientación de la mano, ubicación del cuerpo y movimiento) que se influencian mutuamente, tal y como se describe en los estudios lingüísticos clásicos, desde Stokoe (1960), y más recientes, como Van der Hulst (1993) o Brentari (2012). Por ese motivo diseñaron un HMM paralelo (Vogler y Metaxas, 1999) que también resolvió parcialmente el problema de escalabilidad con el número de posibles fonemas en lengua de signos, convirtiendo la combinación multiplicativa de los elementos de los canales (unas 108 combinaciones) en una combinación aditiva (unas 200). Utilizaron guantes sensorizados (data-gloves) para extraer características y no incluyeron la expresión facial como canal complementario.

Con el nuevo siglo la cámara Microsoft Kinect, y posteriormente la Kinect2, permitieron extraer fácilmente la información 3D, ya que estas cámaras de vídeo proporcionan una señal de profundidad (depth, D) adicional a los tradicionales rojo-verde-azul (RGB). Usando su paquete de desarrollo software (SDK), que combina los canales RGBD, visión por computador y técnicas de aprendizaje automático, es posible ubicar en la señal de vídeo las posiciones de las articulaciones, las manos y los dedos, en tiempo real y a casi 30 imágenes por segundo. Esta tecnología atrajo a nuevos grupos de investigación durante la última década (Keskin *et al.*, 2011; Agarwal y Thakur, 2013; Yang 2015) y permitió desarrollar algoritmos que localizan las manos y cuerpo en la imagen general (algoritmos de segmentación), que son el primer paso para el reconocimiento de gestos, aunque muy pocos de ellos se diseñaron específicamente para la lengua de signos. Si bien muchos de los estudios de lengua de signos utilizan también elementos hápticos, por ejemplo guantes electrónicos, para capturar la forma y el movimiento de las manos, los últimos avances en algoritmos de visión por computador para la detección de objetos, estimación de movimiento y seguimiento visual, permitieron simplificar la adquisición y utilizar solo información RGB. Wong y Cipolla (2005) se centraron en 10 primitivas de movimiento e implementaron un sistema en tiempo real usando solamente una cámara a 25 imágenes por segundo. Aunque es un esquema bastante simple, su naturaleza probabilística permite escalarlo para hacer un análisis de movimiento más complejo con múltiples hipótesis. Cooper *et al.* (2012) entrenaron

subunidades a partir de datos RGB estáticos y datos de seguimiento dinámicos en 2D y 3D de una Kinect, y combinaron los flujos utilizando modelos de Markov o realce de patrones secuenciales (SP-boosting). Lograron un 73% de precisión en una gran base de datos de 984 signos de lengua de signos inglesa (BSL) utilizando datos 2D y modelos de Markov, y un 85,1% en una base de datos más pequeña de 40 signos utilizando información 3D y SP-boosting. El empleo de subunidades parece ser el enfoque más prometedor para el reconocimiento continuo de signos independiente del signante, pero los investigadores están de acuerdo en que se requieren más bases de datos anotadas lingüísticamente, y de múltiples informantes, para que los algoritmos de aprendizaje automático que se entrenen con ellos sean capaces de distinguir entre las características específicas del signante y las características independientes del signante.

El grupo de investigación que ha trabajado más intensamente en el reconocimiento de signos continuos con entradas multicanal y usando solo una cámara es el Grupo de Tecnología de Lenguaje Humano y Reconocimiento de Patrones de la Universidad RWTH de Aachen (Alemania). Durante los últimos 15 años, han ido incrementando la complejidad y la precisión de su reconocedor de signos continuos, con un vocabulario extenso, utilizando modelos de articulación y de lenguaje (Dreuw *et al*., 2007), y el procesamiento de varios canales, incluidas las expresiones faciales (Koller *et al.*, 2015).

En los últimos años también hemos visto cómo las redes neuronales profundas (DNN) han mejorado el rendimiento de muchas tareas de visión artificial. El reconocimiento de lengua de signos también está empezando a beneficiarse de esta tecnología. Las DNN son capaces, por un lado, de aprender las características óptimas de una lengua de signos dada y, por otro, de modelar dinámicas temporales. Nuevamente, el Grupo RWTH, junto con personas de CVSSP, (U. Surrey, Reino Unido), hizo algunas contribuciones en este campo al ser el primero en utilizar las salidas de una red neuronal convolucional (CNN) como verdaderas probabilidades a posteriori para entrenar un reconocedor híbrido CNN-HMM, de extremo a extremo, solo con información de manos (Koller *et al.*, 2016). Declararon conseguir con esta técnica una clara mejora en tres bases de datos diferentes (SIGNUM, RWTH-PHOENIX-Weather, 2012 y 2014). También de una colaboración entre CVSSP y RWTH surgió un modelo DNN extremo a extremo que incluye redes CNN y memoria a corto y largo plazo (LSTM) que se publicó en 2017 (Cihan *et al.*, 2017). En él se utiliza la información contextual de la parte superior del cuerpo para mejorar la subred dedicada al análisis de la forma de la mano, logrando un rendimiento competitivo en reconocimiento continuo sin segmentación explícita de signos. Esta línea de investigación es bastante prometedora y se espera que su rendimiento aumente a medida que se agreguen más canales como subredes diferentes. Muy

recientemente, Huang *et al.* (2018) diseñaron una arquitectura DNN bastante efectiva, basada en una extensión de LSTM, llamada Red de atención jerárquica con espacio latente, que no necesita segmentación temporal de signos, y utiliza CNN 3D para procesar bloques de 16 imágenes de vídeo, que extraen características espacio-temporales de cada mano y del tronco. Entrenaron y probaron el modelo con dos grandes bases de datos: la mencionada RWTH-PHOENIX-Weather (2014) y una base de datos de lengua de signos china más grande, adquirida por el propio grupo de investigación, que incluye las señales RGB, profundidad y las posiciones de las articulaciones extraídas por la Kinect. Declararon mejorar la precisión de otros métodos reconociendo signos continuos, en ambas bases de datos, utilizando tanto RGBD como sólo RGB, y utilizando solo los canales de las manos o también el canal del tronco.

El proyecto DictaSign (DictaSign, 2012), financiado en la convocatoria europea FP7-ICT, se planteó como objetivo desarrollar las tecnologías que permitan interaccionar con la Web 2.0 mediante LS. Las personas usuarias signan delante de una kinect, el ordenador reconoce las frases, las convierte a una representación interna de la LS, y hace que un avatar las reproduzca en la misma LS u otra. De esta forma se anonimiza la contribución, se posibilita que las contribuciones sean modificadas por cualquier persona, y se puedan traducir. El proyecto incorpora 4 lenguas de signos: británica (BSL), alemana (DGS), griega (GSL) y francesa (LSF). En su transcurso grabaron y anotaron en HamNoSys un corpus paralelo semi-espontáneo en las 4 lenguas sobre viajes y un diccionario multilingüe de más de 1.000 signos. El producto final son 3 demostradores: traductor LS a LS en el contexto de los viajes por Europa; búsqueda por ejemplo, usando signos aislados como entrada; y Sign Language Wiki, un sistema que combina RALS, presentación con avatar y edición en LS.

El proyecto SignSpeak (SignSpeak, 2012), financiado en la misma convocatoria, se propuso como objetivo desarrollar tecnologías de traducción de LS continua a texto basadas en procesado de imagen. El principal resultado es un traductor de lengua de signos alemana a texto en un contexto de meteorología, entrenado con imágenes de televisión correspondientes a la interpretación del boletín meteorológico de las noticias. La base de datos empleada para este fin, la RWTH-PHOENIX-Weather, se ha mencionado en el apartado 2. Otras lenguas de signos trabajadas en el proyecto son la holandesa (NGT), la británica (BSL), la americana (ASL) y la irlandesa (ISL). El Grupo de Tecnología de Lenguaje Humano y Reconocimiento de Patrones de la Universidad RWTH de Aachen (Alemania) forma parte del consorcio. Sus aportaciones en el campo de RALSE se comentan unos párrafos más arriba en este mismo anexo.

En relación con la LSE, el proyecto Consignos (López-Ludueña *et al*., 2014ab) desarrolla con financiación estatal un conversor y reproductor automático de LSE. Reconoce el mensaje oral, traduce de texto en español a LSE y representa los signos mediante un agente animado. Dio como resultado demostradores para el sector hotelero y para la Empresa Municipal de Transportes de Madrid.

## Anexo II: Estructura de almacenamiento de la base de datos

La estructura se ha diseñado para facilitar la identificación de cada vídeo, y el crecimiento progresivo del número de sesiones de grabación e informantes. Como puede verse en la figura 5, parte de un directorio raíz con el nombre de la base de datos. Dentro de este directorio están cada una de las sesiones grabadas en forma de carpetas independientes. Dentro de cada sesión hay una carpeta para cada informante, y dentro de la carpeta de cada informante hay una carpeta para cada signo grabado. Dentro de la carpeta de cada signo habrá tres archivos: el vídeo grabado con la cámara Nikon, el vídeo grabado con la Kinect y los metadatos de la Kinect.

Cada carpeta de sesión se nombra con un identificador formado por la letra s seguida de un número de cuatro dígitos que se incrementa automáticamente al crear una nueva sesión en el programa de adquisición de la base de datos. De forma similar, al incorporar a una nueva persona informante se crea un identificador formado por una p y un número de cuatro dígitos con el siguiente código disponible. Los signos a grabar se identifican mediante un código de cuatro dígitos precedidos por la letra w (ver tabla 1). La base de datos LSE_Lex40_UVIGO cuenta con 40 signos, por lo que los códigos varían de w0001 a w0040. El código w0000 contiene el vídeo de la cámara Nikon de la entrevista que se realiza a cada nueva informante para preguntarle sus datos personales.

Por ejemplo, la ruta LSE_Lex40_UVIGO/s0001/p0002/w0040/ contendrá los archivos relativos al signo calle (w0040) articulado por la segunda informante (p0002) grabada en la primera sesión (s0001) de la base de datos LSE_Lex40_UVIGO.

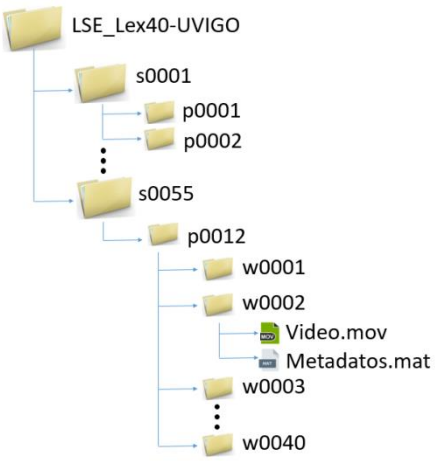


Figura 5: estructura de almacenamiento de la base de datos LSE_Lex40_UVIGO.

## Anexo III: Abreviaturas

ARCO: Acceso, Rectificación, Cancelación y Oposición
ASLR: Automatic Sign Language Recognition (reconocimiento automático de lengua de signos)
ASL: American Sign Language (lengua de signos americana)
ASORVIGO: Asociación de Personas Sordas de Vigo
BCBL: Basque Center on Cognition, Brain and Language (centro vasco de cognición, cerebro y lenguaje)
BSL: British Sign Language (lengua de signos británica)
CNLSE: Centro de Normalización Lingüística de la Lengua de Signos Española
CNN: Convolutional Neural Network (red neuronal convolucional)
CVSSP: Centre for Vision, Speech and Signal Processing (centro de visión, habla y tratamiento de señal)
DNN: Deep Neural Networks (redes neuronales profundas)
FAXPG: Federación De Asociacions De Persoas Xordas De Galicia (Federación de Asociaciones de Personas Sordas del Galicia)
GRADES: Gramática, Discurso e Sociedade
GTM: Grupo de Tecnoloxías Multimedia
HMM: Hidden Markov Models (modelos ocultos de Markov)
LSE: Lengua de Signos Española
LSTM: Long Short-Term Memory (memoria a corto plazo larga)
RALS: Reconocimiento Automático de Lengua(s) de Signos
RGB: Red Green and Blue (rojo verde y azul)
RWTH: Rheinisch-Westfälische Technische Hochschule (universidad técnica Rheinisch-Westfälische)
SDK: Software Development Kit (kit de desarrollo de software)
SP-boosting: Sequential Pattern boosting (realce de patrones secuenciales)
USB: Universal Serial Bus

## Referencias

Agarwal, A. y Thakur, M. K. (2013). "Sign language recognition using Microsoft Kinect". *Sixth Int. Conf. Contemporary Computing* (IC3): 181–185.

Athitsos, V. *et al.* (2008). "The american sign language lexicon video dataset". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*: 1-8.

Brentari, D. (2012). "Phonology". En R. Pfau, M. Steinbach y B. Woll (eds.), *Sign language: An international handbook* (pp. 21-54). Berlin; Boston: De Gruyter Mouton.

Cihan, N. *et al.* (2017). "Subunets: End-to-end hand shape and continuous sign language recognition". *Proc. ICCV*: 3056-3065.

Cooper, H. *et al.* (2012). "Sign language recognition using sub-units". *Journal of Machine Learning Research*, 13: 2205–2231.

DictaSign (2012). *DICTA-SIGN: Sign Language Recognition, Generation and Modelling with application in Deaf Communication*. IEA-LSP. Recuperado de http://www.ilsp.gr/en/infoprojects/meta?view=project&task=show&id=14

Dreuw, P. *et al.* (2007). "Speech recognition techniques for a sign language recognition system," *Proc. Interspeech*: 2513–2516.

España. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales. *Boletín Oficial del Estado*, 6 de diciembre de 2018, núm. 294, pp. 119788-119857.

Economic and Social Research Council (2018). *British Sign Language Corpus Project, Economic and Social Research Council*. ESRC. Recuperado de https://bslcorpusproject.org/

ExTol (2018). *ExTOL: End to End Translation of British Sign Language*. ExTol. Recuperado de http://cvssp.org/projects/extol/

Fundación CNSE (2008). *Diccionario normativo de lengua de signos española* [DVD]. Madrid: Fundación CNSE.

Fundación CNSE (2008). *DILSE III: Tesoro de la lengua de signos española* [DVD]. Madrid: Fundación CNSE.

GRADES-GTM (2019). *Procedimiento para la gestión de los derechos de los interesados: Corpus LSE_Lex40_UVigo: Acceso, Rectificación, Supresión, Oposición, Limitación del Tratamiento y Portabilidad de los datos: Gramática, Discurso e Sociedade y Grupo de Tecnologías Multimedia*. Vigo: Universidad de Vigo. Recuperado de http://gtm.uvigo.es/sites/default/files/pictures/Procedimiento%20Derechos%20de%20los%20Afectados.pdf

GTM (2019). *Descripción del corpus LSE_Lex40_UVIGO: Grupo de Tecnologías Multimedia*. Vigo: Universidad de Vigo. Recuperado de http://gtm.uvigo.es/content/descripcion-lselex40uvigo

Gutierrez-Sigut, E. *et al.* (2016). "LSE-Sign: A lexical database for Spanish Sign Language". *Behavior Research Methods*, 48(1): 123–137.

Herrero, Á. (2009). *Gramática didáctica de lengua de signos española*. Madrid: SM.

Huang, J. *et al.* (2018). "Video-based Sign Language Recognition without Temporal Segmentation". *Proc. The Thirty-Second AAAI Conference on Artificial Intelligence*, (AAAI-18): 2257–2264.

Keskin C. *et al*. (2011). "Real Time Hand Pose Estimation using Depth Sensors". *Proc. ICCV Workshops*: 1228-1234.

Koller, O. *et al.* (2016). "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition". *Proc. of the British Machine Vision Conference* (BMVC).

Koller, O. *et al.* (2015). "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers". *Computer Vision and Image Understanding*, 141:108–125.

López-Ludeña, V. *et al.* (2014a). "Translating Bus Information into Sign Language for Deaf People". *Engineering Applications of Artificial Intelligence*, 32: 258-269.

López-Ludeña, V. *et al.* (2014b). "Methodology for Developing an Advanced Communications System for the Deaf in a New Domain". *Knowledge-Based Systems*, 56: 240-252.

Martínez-Hinarejos, C. D. y Parcheta, Z. (2017). "Spanish Sign Language Recognition with Different Topology Hidden Markov Models". *Proc. of the Interspeech*, 2017: 3349-3353.

San-Segundo, R. *et al.* (2008). "Proposing a speech to gesture translation architecture for Spanish deaf people". *Journal of Visual Languages and Computing*, 19: 523–538.

Schembri, A. y Johnston, T. (2012). "Sociolinguistic aspects of variation and change". En R. Pfau, M. Steinbach y B. Woll (eds.), *Sign language: An international handbook* (pp. 788-816). Berlin; Boston: De Gruyter Mouton.

SignSpeak (2012). *Scientific understanding and vision-based technological development for continuous sign language recognition and translation*. Recuperado de http://www.signspeak.eu/

Starner, T. *et al.* (1998). "Real-time american sign language recognition using desk and wearable computer based video". *IEEE TPAMI 20*, (12): 1371 – 1375.

Stokoe, W. C. (2005). "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf". *Journal of Deaf Studies and Deaf Education*, 10(1): 3-37.

Tilves-Santiago, D. *et al.* (2018). "Experimental framework design for sign language automatic recognition". *Proc. of IberSPEECH*: 72-76.

Van der Hulst, H (1993). "Units in the Analysis of Signs". *Phonology*, 10(2): 209-241.

Vogler, C. y Metaxas, D. (1999). "Parallel hidden markov models for american sign language recognition". *Proc. of ICCV*, 1: 116 – 122.

Wong, S. F. y Cipolla, R. (2005). "Real-time interpretation of hand motions using a sparse bayesian classifier on motion gradient orientation images". *Proc. of BMVC*, 1:379–388.

Yang, H-D. (2015). "Sign Language Recognition with the Kinect Sensor Based on Conditional Random Fields". *Sensors*, 15(1): 135–147.