Sistema de Comunicación Oral para Personas Sordas

Verónica López[1], Rubén San-Segundo[1], Raquel Martín[1], David Sánchez[2], Adolfo García[2]
[1]Grupo de Tecnología del Habla-Universidad Politécnica de Madrid
[2]Fundación CNSE
veronicalopez@die.upm.es

***Resumen* — Este artículo describe el desarrollo y la evaluación de un sistema de comunicación para personas sordas en un ámbito de aplicación específico: la renovación del permiso de conducir. El sistema de comunicación desarrollado está compuesto por dos módulos que permiten la comunicación en los dos sentidos. El primer módulo es un traductor de voz en castellano a Lengua de Signos Española (LSE) y está formado por un reconocedor de voz, un traductor de palabras en castellano a una secuencia de signos y un tercer módulo de representación de los signos mediante un agente animado. El segundo módulo es un generador de voz en castellano a partir de una secuencia de signos, y está formado por una interfaz donde se especifican los signos, un traductor (para convertir la secuencia de signos en una secuencia de palabras) y un conversor de texto a voz. En los dos módulos de traducción entre lenguas, se integran tres tecnologías: una basada en ejemplos, una basada en reglas y un traductor estadístico. Este artículo describe la evaluación del sistema llevada a cabo en la Jefatura Provincial de Tráfico de Toledo implicando a funcionarios de dicha jefatura y personas sordas.**

I. INTRODUCCIÓN

Tal como muestran los datos del INE (Instituto Nacional de Estadística) y del Ministerio de Educación, en España, el 92% de las personas sordas tienen serias dificultades para entender y expresarse en castellano escrito. Además, el 47% de las personas sordas mayores de 10 años no tiene un nivel básico de estudios. Estos problemas provienen, principalmente, de la dificultad que tienen dichas personas para conjugar verbos, para aplicar las concordancias de género y número, interpretar conceptos abstractos, y para extraer información semántica de las palabras formando una imagen mental de lo que se les comunica.

Este problema provoca, a su vez, que en ocasiones, las personas sordas tengan menos derechos y oportunidades que el resto de ciudadanos. Por ejemplo, el problema se refleja muy bien cuando quieren acceder a un servicio público como ir a renovar el permiso de conducir, ya que, por lo general, los funcionarios no conocen la Lengua de Signos y, para comunicarse, las personas sordas necesitan intérpretes signantes, cuyo coste es elevado.

En 2007, el Gobierno de España aceptó la Lengua de Signos Española (LSE) como una de las lenguas oficiales españolas, definiendo un plan para invertir en el desarrollo de sistemas que ayuden a la comunicación entre personas sordas y oyentes.

En este artículo se describe el primer sistema de comunicación oral en ambos sentidos castellano-LSE y LSE-castellano, desarrollado por el Grupo de Tecnología del Habla de la Universidad Politécnica de Madrid y la Fundación de la Confederación Estatal de Personas Sordas (Fundación CNSE), que intenta eliminar las barreras de comunicación existentes para las personas sordas en un ámbito de aplicación específico: la renovación del permiso de conducir.

II. ESTADO DE LA CUESTIÓN

La investigación en lengua de signos ha sido posible gracias a los corpus generados por varios grupos de investigación. Algunos ejemplos son los siguientes. En primer lugar destacar el corpus compuesto por más de 300 horas con grabaciones en vídeo de 100 signantes en Lengua de Signos Australiana [1]. La base de datos RWTH-BOSTON-400, que es una base de datos que contiene 843 frases, con alrededor de 400 signos diferentes de 5 signantes en Lengua de Signos Americana, con anotaciones en inglés [2]. Otro ejemplo es el proyecto de generación del British Sign Language Corpus que trata de crear un corpus digital de lectura mecánica y espontánea en Lengua de Signos Británica (BSL), grabando a personas sordas signantes nativas, en todo el Reino Unido [3]. Finalmente, comentar el corpus desarrollado en el Institute for Language and Speech Processing (ILSP) y que contiene partes signadas de narración, así como una considerable cantidad de frases y perífrasis signadas [4].

En los últimos años, ha habido numerosos proyectos de investigación sobre traducción de habla natural. En Europa: C-Star, ATR, Vermobil, Eutrans, LC-Star, PF-Star y, el más ambicioso, TC_STAR. En Estados Unidos está el programa GALE, cuyo objetivo es desarrollar y aplicar tecnologías de procesado de habla y lenguaje natural para analizar e interpretar enormes volúmenes de voz y texto en varios idiomas. También, en cuanto a sistemas de traducción a lengua de signos se refiere, varios grupos de investigación han mostrado su interés desarrollando varios prototipos: basados en ejemplos [5], en reglas escritas por un experto [6], frases completas [7] o métodos estadísticos ([8]; [9]; sistema SiSi de IBM).

Este artículo describe el desarrollo y evaluación del primer sistema de comunicación oral para personas sordas españolas en un dominio de aplicación real: la renovación del permiso de conducir.

III. BASE DE DATOS

Para desarrollar el sistema, primero es necesario generar una base de datos con un corpus paralelo de frases en castellano y LSE. Esta base de datos se obtuvo con la colaboración de la Jefatura Provincial de Tráfico de Toledo y está formada por frases generadas en el dominio de aplicación mencionado: la renovación del permiso de conducir. Se recopilaron frases tanto de las explicaciones de los funcionarios como de las preguntas formuladas por los usuarios.

La oficina de la Jefatura Provincial de Tráfico de Toledo está organizada en varias ventanillas (para hacer trámites diferentes: información, caja, vehículos, conductores y autoescuelas) en las cuales fueron recogidas más de 4000 frases durante un periodo de tres semanas, aunque luego se seleccionaron únicamente las frases empleadas en la renovación del permiso de conducir para definir el sistema en esa aplicación específica.

Las frases en castellano se obtuvieron con la colaboración de los funcionarios y son las frases típicas del día a día en este tipo de procesos burocráticos, dando las explicaciones más frecuentes a los usuarios. Por otro lado, las frases correspondientes a los usuarios son, por lo general, interrogativas, solicitando una cierta información.

Finalmente, se seleccionaron 707 frases: 547 pronunciadas por funcionarios y 160 por usuarios. Posteriormente, miembros del Grupo de Tecnología del Habla de la Universidad Politécnica de Madrid ampliaron el número de frases añadiendo variantes en castellano diferentes a las de la base de datos inicial, pero con el mismo significado y traducción a LSE. En la Tabla I pueden verse las principales características del corpus generado.

Tabla I. Principales características de la base de datos

Funcionario	**Castellano**	**LSE**
Pares de frases	1,641	
Frases diferentes	1,413	199
Palabras	17,113	12,741
Vocabulario	527	237
Usuarios	**Castellano**	**LSE**
Pares de frases	483	
Frases diferentes	389	93
Palabras	3,130	2,283
Vocabulario	294	133

La traducción de cada una de estas frases a LSE fue realizada por personas sordas expertas en LSE y conocedoras de la lengua castellana y, posteriormente, expertos en LSE representaron y grabaron estas frases en videos.

Para la representación de los signos se utiliza un agente animado (Vguido desarrollado en el proyecto europeo eSIGN) que representa los signos a partir de su especificación en HamNoSys [13]. En el desarrollo del sistema fue necesario generar una base de datos con más de 400 signos en varias notaciones: glosas (palabras en mayúscula que representan los signos), SEA (Sistema de Escritura Alfabética) [12], HamNoSys[13], y SIGML[14]. En la Figura 1 puede verse un fragmento.

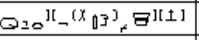
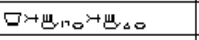
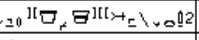
GLOSA	SEA	HAMNOSYS	SIGML
ABAJO	olemuawu		SIGML\ABAJO.txt
ACOMPAÑAR-A_MI	saca íájwe-ye		SIGML\ACOMPAÑAR-A_MI.txt
ACTUAL	s omèawud		SIGML\ACTUAL.txt
ADIÓS	omaudahb		SIGML\ADIÓS.txt
ADJUNTAR	sòaméha òamèug		SIGML\ADJUNTAR.txt
AHÍ	elewe		SIGML\AHÍ.txt

Figura 1. Fragmento de la base de datos de los signos

Además de los signos correspondientes a las traducciones a LSE del corpus paralelo, esta base de datos también incluye descripciones de signos de todas las letras (para los deletreos), los números del 0 al 100, los meses, días de la semana y números para especificación de las horas.

IV. TRADUCCIÓN DE VOZ A LSE

El sistema de traducción de voz en castellano a LSE convierte frases en voz en frases en LSE representadas por un agente animado y está compuesto por tres módulos principales (Figura 2). El primero es un módulo de reconocimiento de voz, que convierte una frase en voz en una secuencia de palabras. El segundo es un módulo de traducción que convierte la secuencia de palabras en una secuencia de signos en LSE. Y el tercer módulo representa los signos mediante un agente animado (el avatar VGuido del proyecto europeo eSIGN).

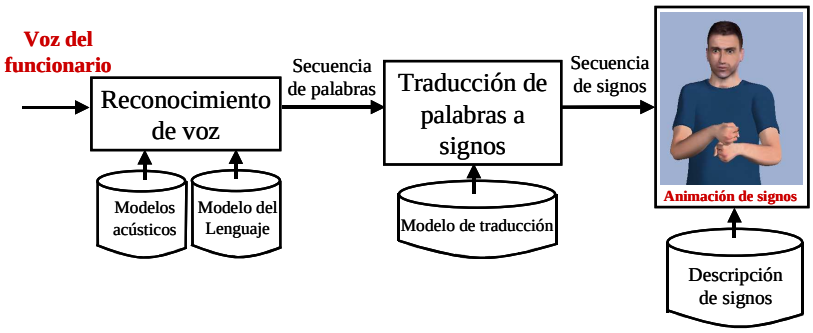


Figura 2. Sistema de traducción de voz a LSE

El reconocedor de voz es un sistema de reconocimiento de voz continuo basado en modelos ocultos de Markov, independiente del locutor, con medidas de confianza y la posibilidad de adaptar modelos acústicos. Este módulo ha sido desarrollado íntegramente en el Grupo de Tecnología del Habla de la UPM.

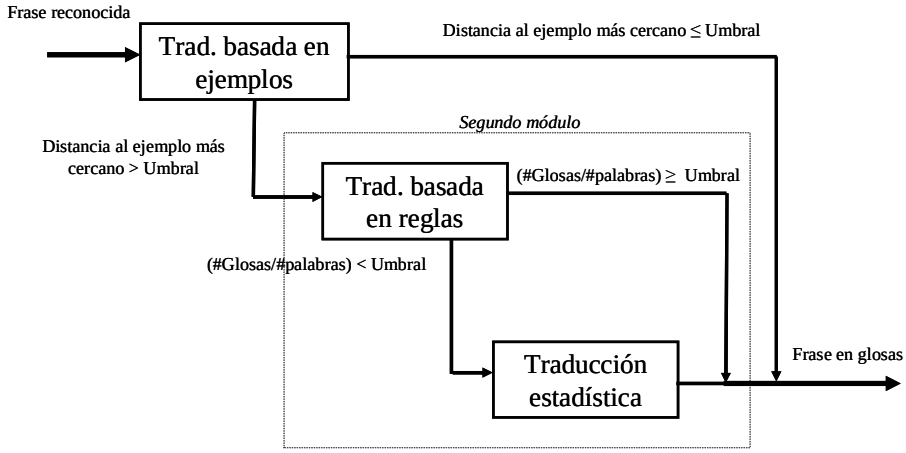


Figura 3. Combinación de las diferentes estrategias de traducción.

El módulo de traducción tiene una estructura jerárquica dividida en dos pasos (Figura 3). En el primero, se utiliza una estrategia basada en ejemplos para traducir la secuencia de palabras: la traducción se lleva a cabo basándose en la similitud entre la frase que se quiere traducir y los elementos de un corpus paralelo con ejemplos traducidos. Si la distancia de edición con el ejemplo que más se parece es menor que un cierto umbral, la traducción de salida es la misma que la del ejemplo. Sin embargo, si la distancia es mayor que el umbral, un módulo en segundo plano traduce la secuencia de palabras. En este segundo módulo se emplea una combinación de un traductor basado en reglas definidas por un experto y un traductor estadístico (traductor basado en subfrases y transductor de estados finitos). Aunque el traductor basado en reglas ofrece mejores resultados, el traductor estadístico es interesante porque se desarrolla en poco tiempo y con poco esfuerzo (uno o dos días), mientras que para el basado en reglas se necesitó varias semanas para desarrollar todas las reglas. Para elegir un traductor u otro, se tiene en cuenta el número de glosas generadas después de la traducción y el número de palabras de la frase de entrada, de manera que si la relación #glosas/#palabras es mayor que un cierto umbral, se tiene en cuenta la traducción basada en reglas. Y si no se supera dicho umbral, la traducción de salida es la estadística.

Tabla II. Resumen de resultados de las pruebas de laboratorio del sistema voz-LSE

SR-WER	SER	PER	BLEU
6.76	10.11	8.45	0.8019

En la Tabla II pueden verse los resultados obtenidos en pruebas de laboratorio, medidos con las métricas SR-WER (Tasa de error de palabras del reconocimiento de voz), SER (Tasa de error de signos), PER (Tasa de error de signos independiente de la posición) y BLEU (BiLingual Evaluation Understudy). Como se puede observar los resultados son bastante buenos.

V. GENERACIÓN DE VOZ A PARTIR DE LSE

El sistema generador de voz a partir de una secuencia de signos en LSE está compuesto de tres módulos (Figura 4). El primer módulo es una interfaz visual donde el usuario selecciona una secuencia de signos con el ratón. La interfaz incluye distintas herramientas para la especificación de los signos: un agente animado que los representa (para comprobar que la secuencia de signos elegida es la correcta), mecanismos de predicción del siguiente signo que se especificará, calendario y reloj para la especificación de fechas y horas, signos correspondientes a letras para los deletreos, preguntas frecuentes, etc.

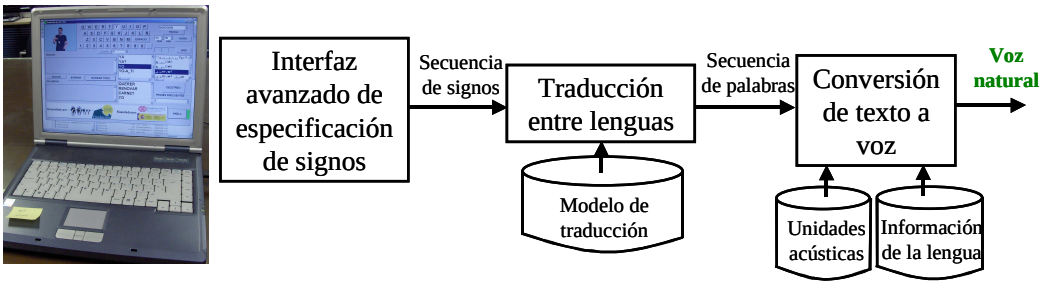


Figura 4. Diagrama de modulo del sistema de generación de voz a partir de LSE.

De manera que con esta interfaz una persona sorda puede seleccionar una secuencia de signos para que sea traducida al castellano y posteriormente convertida a voz para una persona oyente. La secuencia de signos puede seleccionarse mediante glosas o con otras notaciones como HamNoSys. (Figura 5)

Figura 5. Interfaz visual del sistema de traducción LSE-voz

El segundo módulo traduce la secuencia de signos a una secuencia de palabras con tres estrategias de traducción combinadas: una basada en ejemplos, una basada en reglas y una última estadística. El procedimiento a seguir para realizar la traducción es el mismo que el realizado en el sistema castellano-LSE. En la Tabla III se resumen los resultados de las traducciones en pruebas de laboratorio empleando las métricas WER (Tasa de error de palabras), PER (Tasa de error de palabras independiente de la posición) y BLEU (BiLingual Evaluation Understudy). Como se puede observar en este caso los resultados son también muy prometedores.

Tabla III. Resumen de resultados de las pruebas de laboratorio del sistema LSE-voz

WER	PER	BLEU
2.36	2.25	0.9113

El último módulo convierte la secuencia de palabras en voz, para lo que se emplea un conversor texto-voz comercial, en este caso el sistema Loquendo (http://www.loquendo.com/en/).

VI. EVALUACIÓN

El sistema completo de comunicación ha sido evaluado en la Jefatura Provincial de Tráfico de Toledo implicando a un funcionario y a cinco personas sordas, de manera que el sistema de traducción de voz a LSE fue utilizado por el funcionario para dar las explicaciones que necesitaban sobre el servicio de renovación del permiso de conducir, y el sistema de traducción de LSE a voz ayudaba a las personas sordas a hacer preguntas al funcionario (Figura 6).

Figura 6. Evaluación en la Jefatura Provincial de Tráfico de Toledo

La evaluación se llevó a cabo en un día, dedicando la primera hora a explicar el proyecto y la evaluación al funcionario y usuarios sordos. Se definieron seis escenarios distintos para evaluar situaciones reales: un escenario donde el usuario simulaba tener toda la documentación necesaria para realizar el trámite; tres escenarios donde se simulaba no disponer de algún documento necesario, como el DNI, una foto o el certificado médico; un escenario donde el usuario debía rellenar algún formulario con cierta información; y un escenario donde el usuario quería pagar con tarjeta de crédito y no estaba permitido.

Cinco usuarios sordos (tres hombres y dos mujeres) interactuaron con un funcionario de la Jefatura Provincial de Tráfico utilizando el sistema desarrollado. Las edades de los usuarios estaban comprendidas entre los 22 y los 55 años, con una media de edad de 39,7 años. Dos de los cinco usuarios utilizaban un ordenador todos los días y los otros tres lo utilizaban muy pocas veces a la semana. Sólo dos de ellos tenían un nivel medio-alto de comprensión del castellano escrito y un buen hábito empleando glosas para especificar los signos de una frase.

Los usuarios y el funcionario utilizaron el sistema en los escenarios descritos anteriormente, tomándose 25 diálogos en total. Para realizar la evaluación se tomaron medidas objetivas del sistema y subjetivas mediante cuestionarios al funcionario y a los usuarios implicados en la evaluación. Las medidas objetivas pueden observarse en la Tabla IV y en la Tabla V.

Tabla IV. Medidas objetivas del sistema de traducción voz-LSE

Medida	Valor
Tasa de error de reconocimiento de habla	**4.6%**
Tasa de error después de la traducción	**8.5%**
Tiempo de reconocimiento	**3.2 sec**
Tiempo de traducción	**0.0012 sec**
Tiempo de signado	**4.6 sec**
% de uso de la traducción basada en ejemplos	**95.0%**
% de uso de la traducción basada en reglas	4.1%
% de uso de la traducción estadística	0.9%
Número de turnos del funcionario	8.4

La tasa de error del reconocedor de voz (4,6%) y la de traducción (8,5%) son bastante buenas, mejorando los resultados de laboratorio. Esto fue posible porque los modelos acústicos del reconocedor de voz fueron adaptados al funcionario, empleando 50 pronunciaciones. Esta adaptación permitió mejorar considerablemente la tasas de reconocimiento, y por tanto la de traducción. El tiempo necesario para realizar la traducción de voz a LSE son unos 4,6 segundos, permitiendo un diálogo fluido. Por otro lado, la traducción basada en ejemplos se utilizó en el 95% de los casos, lo que demuestra la fiabilidad del estudio lingüístico llevado a cabo (la mayoría de las frases pronunciadas tienen un ejemplo bastante parecido en el corpus recopilado).

Tabla V. Medidas objetivas del sistema de traducción LSE-voz

Medida	Valor
Tasa de error de traducción	**2.56%**
Tiempo de traducción	**0,001 sec**
Tiempo para conversión texto a voz	**1.7 sec**
% de uso de la traducción basada en ejemplos	**92.7%**
% de uso de la traducción basada en reglas	7.3%
% de uso de la traducción estadística	0.0%
Tiempo para definir una secuencia de glosas	16.5 sec
Número de clicks para añadir una glosa	8.5 clicks
Número de glosas por turno del usuario	2.6
Número de turnos del usuario	4.0

En la Tabla V se observa una buena tasa de error de traducción del sistema LSE-voz (2,56%) y poco tiempo para la traducción que hace posible emplear el sistema en condiciones reales. Además, de nuevo, la estrategia basada en ejemplos es utilizada en más del 90% de los casos, mostrando la fiabilidad del corpus paralelo generado.

Las medidas subjetivas se obtuvieron de cuestionarios realizados al funcionario y a los usuarios, donde daban una puntuación de 0 a 5 a cada aspecto valorado. Para el sistema de voz a LSE, la valoración del funcionario fue muy positiva, dando un 3,5 en todos los aspectos considerados. La valoración de los usuarios fue muy baja (sobre 2,6), siendo el principal problema la naturalidad a la hora de representar el signo con el agente animado (1,6). Por otro lado también se detectaron diferencias de criterio entre usuarios sordos a la hora de representar los signos o construir las frases con glosas: por ello, es necesario seguir trabajando en la estandarización de la LSE en España para solucionar este problema.

En el sistema LSE-voz, el funcionario dio un 4,0, valorando muy positivamente la naturalidad e inteligibilidad de la voz generada. Por otro lado, los usuarios dieron una buena puntuación a la interfaz visual (3,2), pero se quejaron de la complejidad

de la misma. Y, de nuevo, un problema importante fue la falta de estandarización en las glosas que representan determinados signos. Este problema se intentará solucionar incorporando imágenes o gifs para seleccionar los signos en versiones futuras.

VII. CONCLUSIONES

En este artículo se ha descrito el desarrollo de un sistema de comunicación oral para personas sordas en un dominio de aplicación específico: la renovación del permiso de conducir. El sistema completo está formado, por un lado, por un sistema de traducción de voz a LSE, compuesto por un reconocedor de voz, un módulo de traducción y un avatar para la representación de los signos; y, por otro lado, un sistema de traducción de LSE a voz, formado por una interfaz visual donde se especifica la secuencia de signos que se quiere traducir, un módulo de traducción y un conversor texto-voz. En cada módulo de traducción se emplean tres tecnologías: una basada en ejemplos, otra basada en reglas y una estadística. Además, en el artículo se ha descrito la evaluación llevada a cabo para probar el sistema y una discusión de los resultados obtenidos.

La evaluación del sistema voz-LSE muestra un buen reconocimiento de voz (4,6% tasa de error) y traducción (8,5% tasa de error en traducción), pero los usuarios no valoraron bien el sistema en los cuestionarios. El problema es la naturalidad del avatar y la poca estandarización de la LSE que provoca que existan discrepancias en la representación de los signos, la gramática, etc., que los usuarios perciben como error del avatar.

Por último, la evaluación del sistema LSE-voz muestra una buena traducción (2,53% tasa de error). Los usuarios valoraron bastante bien el sistema, aunque señalaron algunos problemas con la complejidad de la interfaz (poco tiempo para practicar) y problemas debidos a las discrepancias a la hora de escoger una glosa u otra para representar un signo. Por tanto, es muy importante seguir trabajando en la normalización de la LSE.

AGRADECIMIENTOS

Los autores quieren agradecer al consorcio del proyecto eSIGN la posibilidad de utilizar el agente animado VGuido en este trabajo de investigación. Este trabajo ha sido financiado por: Plan Avanza Exp Nº: PAV-070000-2007-567, ROBONAUTA (DPI2007-66846-c02-02) y SD-TEAM (TIN2008-06856-C05-03).

REFERENCIAS

[1] Johnston T. “Corpus linguistics and signed languages: no lemmata, no corpus”. *3rd Workshop on the Representation and Processing of Sign Languages*, June 1. 2008.

[2] Dreuw P., Neidle C., Athitsos V., Sclaroff S., and Ney H. “Benchmark Databases for Video-Based Automatic Sign Language Recognition”. *LREC*, Marrakech, Morocco. 2008a.

[3] Schembri. A. ”British Sign Language Corpus Project: Open Access Archives and the Observer’s Paradox”. Deafness Cognition and Language Research Centre, University College London. *LREC 2008*.

[4] Efthimiou E. and Fotinea, E. “GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI” *LREC 2008*.

[5] Morrissey S., and Way A. “An example-based approach to translating sign language”. In *Workshop Example-Based Machine Translation (MT X–05)*, pages109–116, Phuket, Thailand, September 2005.

[6] San-Segundo R., Barra R., Córdoba R., D’Haro L.F., Fernández F., Ferreiros J., Lucas J.M., Macías-Guarasa J., Montero J.M., Pardo J.M, 2008. “Speech to Sign Language translation system for Spanish”. *Speech Communication*, Vol 50. 1009-1020.

[7] Cox, S.J., Lincoln M., Tryggvason J., Nakisa M., Wells M., Mand Tutt, and Abbott, S., 2002 “TESSA, a system to aid communication with deaf people”. In *ASSETS 2002*, pages 205-212, Edinburgh, Scotland, 2002

[8] Bungeroth J., Ney, H.,: “Statistical Sign Language Translation”. In Workshop on Representation and Processing of Sign Languages, *LREC 2004*, 105-108.

[9] Morrissey S., Way A., Stein D., Bungeroth J., and Ney H., 2007 “Towards a Hybrid Data-Driven MT System for Sign Languages. Machine Translation Summit (MT Summit)”, pages 329-335, Copenhagen, Denmark, September 2007.

[10] Dreuw, P., D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth, and H. Ney. “Spoken Language Processing Techniques for Sign Language Recognition and Translation”. *Journal Technology and Dissability*. Volume 20 Pages 121-133. 2008b.

[11] Dreuw, P., Stein D., and Ney H. “Enhancing a Sign Language Translation System with Vision-Based Features”. *LNAI*, number 5085, pages 108-113, Lisbon, Portugal, 2009.

[12] A. Herrero. “SEA: Sistema de Escritura Alfabética”. Universidad de Alicante. 2004.

[13] Prillwitz, S., R. Leven, H. Zienert, T. Hanke, J. Henning, et-al. “Hamburg Notation System for Sign Languages – An introductory Guide”. *International Studies on Sign Language and the Communication of the Deaf*, Vol. 5. Uni. of Hamburg. 1989.

[14] Zwiterslood, I., Verlinden, M., Ros, J., van der Schoot, S., 2004. "Synthetic Signing for the Deaf: eSIGN". *Workshop on Assistive Technologies for Vision and Hearing Impairment*, CVHI 2004, 29 June-2 July 2004, Granada, Spain. Spoken R.

[15] San-Segundo, R., Pardo, JM., Ferreiros, J. Sama, V. Barra-Chicote R, Lucas, JM, Sánchez, D. García, “A. Spanish Generation from Sign Language” *Interacting with Computers*, Vol. 22, No 2, pp. 123-139, 2009.