Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2008) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech to sign language translation system for Spanish

R. San-Segundo [a,*], R. Barra [a], R. Córdoba [a], L.F. D'Haro [a], F. Fernández [a], J. Ferreiros [a], J.M. Lucas [a], J. Macías-Guarasa [b], J.M. Montero [a], J.M. Pardo [a]

[a] *Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain*
[b] *Department of Electronics, University of Alcalá, Spain*

Received 6 June 2007; received in revised form 3 December 2007; accepted 3 February 2008

Abstract

This paper describes the development of and the first experiments in a Spanish to sign language translation system in a real domain. The developed system focuses on the sentences spoken by an official when assisting people applying for, or renewing their Identity Card. The system translates official explanations into Spanish Sign Language (LSE: Lengua de Signos Española) for Deaf people. The translation system is made up of a speech recognizer (for decoding the spoken utterance into a word sequence), a natural language translator (for converting a word sequence into a sequence of signs belonging to the sign language), and a 3D avatar animation module (for playing back the hand movements). Two proposals for natural language translation have been evaluated: a rule-based translation module (that computes sign confidence measures from the word confidence measures obtained in the speech recognition module) and a statistical translation module (in this case, parallel corpora were used for training the statistical model). The best configuration reported 31.6% SER (Sign Error Rate) and 0.5780 BLEU (BiLingual Evaluation Understudy). The paper also describes the eSIGN 3D avatar animation module (considering the sign confidence), and the limitations found when implementing a strategy for reducing the delay between the spoken utterance and the sign sequence animation.

Keywords: Spanish Sign Language (LSE); Spoken language translation; Sign animation

1. Introduction

During the last two decades, there have been important advances in the three technological areas that support the implementation of an automatic speech to sign language translation system: sign language studies, spoken language translation and 3D avatar animation.

Sign language presents a great variability depending on the country, even between different areas in the same country. Because of this, from 1960 sign language studies have appeared not only in USA (Stokoe, 1960; Christopoulos and Bonvillian, 1985; Pyers, in press), but also in Europe (Engberg-Pedersen, 2003; Atherton, 1999; Meurant, 2004), Africa (Nyst, 2004) and Asia (Abdel-Fattah, 2005; Masataka et al., 2006). In Spain, there have been several proposals for normalizing Spanish Sign Language (LSE: Lengua de Signos Española), but none of them has been accepted by the Deaf community. From their point of view, these proposals tend to constrain the sign language, limiting its flexibility. In 1991, (Rodríguez, 1991) carried out a detailed analysis of Spanish Sign Language showing its main characteristics. She showed the differences between the sign language used by Deaf people and the standard proposals. This work has been expanded with new studies (Gallardo and Montserrat, 2002; Herrero-Blanco and Salazar-Garcia, 2005; Reyes, 2005).

Spoken language translation has been and is being studied in a number of joint projects such as C-Star, ATR, Vermobil, Eutrans, LC-Star, PF-Star and TC-Star. Apart from the TC-Star project, these projects addressed translation

* Corresponding author. Tel.: +34 915495700x4228; fax: +34 913367323.
E-mail address: lapiz@die.upm.es (R. San-Segundo).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.02.001

tasks within rather limited domains (like traveling and tourism) and medium sized vocabularies. The best performing translation systems are based on various types of statistical approaches (Och and Ney, 2002), including example-based methods (Sumita et al., 2003), finite-state transducers (Casacuberta and Vidal, 2004) and other data driven approaches. The progress achieved over the last 10 years result from several factors such as automatic error measures (Papineni et al., 2002), efficient algorithms for training (Och and Ney, 2003), context dependent models (Zens et al., 2002), efficient algorithms for generation (Koehn et al., 2003), more powerful computers and bigger parallel corpora.

The third technology is the 3D avatar animation. An important community of scientists worldwide is developing and evaluating virtual agents embedded in spoken language systems. These systems provide a great variety of services in very different domains. Some researchers have embedded animated agents in information kiosks in public places (Cassell et al., 2002). At KTH in Stockholm, Gustafson (2002), Granström et al. (2002) and their colleagues have developed several multimodal dialogue systems where animated agents were incorporated to improve the interface. These include Waxholm (Bertenstam et al., 1995) (a travel planning system for ferryboats in the Stockholm archipelago), August (Lundeberg and Beskow, 1999), an information system at the Culture Center in Stockholm, and AdApt (Gustafson and Bell, 2003), a mixed-initiative spoken dialogue system, in which users interact with a virtual agent to locate apartments in Stockholm. Another application for combining language and animated agent technologies in the past has been interactive books for learning. The CSLU Toolkit integrates an animated agent named Baldi. This toolkit has been developed at CSLU (Oregon Graduate Institute OGI) (Sutton and Cole, 1998; Cole et al., 1999) and now it is being expanded at CSLR (University of Colorado) (Cole et al., 2003). This toolkit permits interactive books to be developed quickly with multimedia resources and natural interaction.

The eSIGN European Project (Essential Sign Language Information on Government Networks) (eSIGN project) constitutes one of the most important research efforts in developing tools for the automatic generation of sign language contents. In this project, the main result has been a 3D avatar (VGuido) with enough flexibility to represent signs from the sign language, and a visual environment for creating sign animations in a rapid and easy way. The tools developed in this project were mainly oriented to translating web content into sign language: sign language is the first language of many Deaf people, and their ability to understand written language may be poor in some cases. The project is currently working on local government websites in Germany, the Netherlands and the United Kingdom.

When developing systems for translating speech transcriptions into sign language, it is necessary to have a parallel corpus to be able to train the language and translation models, and to evaluate the systems. Unfortunately, most of the currently available corpora are too small or too general for the aforementioned task. From among the available sign language corpora mentioned in the literature, it is possible to highlight the following. The European Cultural Heritage Online organization (ECHO) presents a multilingual corpus in Swedish, British and The Netherlands sign languages (ECHO corpus). It is made up of five fables and several poems, a small lexicon and interviews with the sign language performers. Another interesting corpus (ASL corpus) is made up of a set of videos in American Sign Language created by The American Sign Language Linguistic Research group at Boston University. In (Bungeroth et al., 2006), a corpus called Phoenix for German and German Sign Language (DGS) in a restricted domain related to weather reports was presented. It comes with a rich annotation of video data, a bilingual text-based sentence corpus and a monolingual German corpus. In Spanish, it is difficult to find this kind of corpus. The most important one is currently available at the Instituto Cervantes; it consists of compound of several videos with poetry, literature for kids and small texts from classic Spanish books. However, this corpus does not provide any text or speech transcriptions and it cannot not be used for our application, a citizen care application where Deaf people can obtain general information about administrative procedures. For this reason, it has been necessary to create a new corpus.

In this paper, spoken language translation and sign language generation technologies are combined to develop a fully automatic speech to sign language translator. This paper includes the first experiments in translating spoken language into sign language in a real domain. This system is the first one developed specifically for Spanish Sign Language (LSE: Lengua de Signos Española). This system completes the research efforts in sign recognition for translating sign language into speech (Sylvie and Surendra, 2005).

This paper is organized as follows: in Section 2, an analysis of the translation problem is described. Section 3 presents an overview of the system including a description of the task domain and the database. Section 4 presents the speech recognizer. In Section 5, the natural language translation module is explained. Section 6 shows the sign playing module using a 3D avatar, and finally, Section 7 summarizes the main conclusions of the work.

2. Main issues in translating Spanish into Sign Language (LSE)

In order to approach the problem of translating Spanish into LSE, an analysis of the relationships between both languages is needed. In order to obtain more significant conclusions, this analysis has been carried out between semantic concepts (extracted from the Spanish text) and signs, instead of considering the relations between words and signs directly (Rodríguez, 1991; Gallardo and

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001

Montserrat, 2002). Bearing this aspect in mind, it is possible to identify four main situations when translating Spanish into LSE. These situations are explained below.

2.1. One semantic concept corresponds to a specific sign

In this case, a semantic concept is directly mapped onto a specific sign. The translation is simple and it consists of assigning one sign to each semantic concept extracted from the text. This sign can be a default translation, independent of the word string, or can differ depending on the word string from which the semantic concept is generated (Fig. 1).

2.2. Several semantic concepts are mapped onto a unique sign

The second situation appears when several concepts generate a unique sign. This situation should be solved by unifying the semantic concepts (resulting in just one concept) to proceed as in the previous situation. This union requires a concept hierarchy and the definition of a more general concept including the original concepts (Fig. 2).

2.3. One semantic concept generates several signs

The third situation occurs when it is necessary to generate several signs from a unique concept. Similar to the previous sections, the sign sequence and its order may depend on the concept and its value, or just the concept. This situation appears in many translation situations:

- VERBS. A verb concept generates a sign related to the action proposed by the verb and auxiliary signs provide information about the action tense (past, present or future), the action subject and the gerund action (Fig. 3).
- GENERAL and SPECIFIC NOUNS. In sign language, there is a tendency to refer to objects with high precision or concretion. As a result of this, there are a lot of domains where several specific nouns exist, but there is no general noun to refer to them collectively. For example, this happens with metals: there are different signs to refer to gold, silver, copper, etc., but there is no a general sign to refer to the concept of metal. The same thing happens when considering furniture: there are several signs for table, chair, bed, etc., but there is no general sign to refer to the concept of furniture. This problem is solved in sign language by introducing several specific signs (Fig. 4).
- LEXICAL–VISUAL PARAPHRASES. Frequently, new concepts (in Spanish, without a corresponding sign representation) appear which do not correspond to any sign in the sign language. In order to solve this problem, Deaf people use paraphrases to represent a new concept with a sequence of known signs. This solution is the first step in representing a new concept. If this concept appears frequently, the sign sequence is replaced by a new sign for reducing the representation time. Some examples of Lexical–Visual Paraphrases are shown in Fig. 5.

The signs are language representations which are more difficult to memorize and distinguish than words.

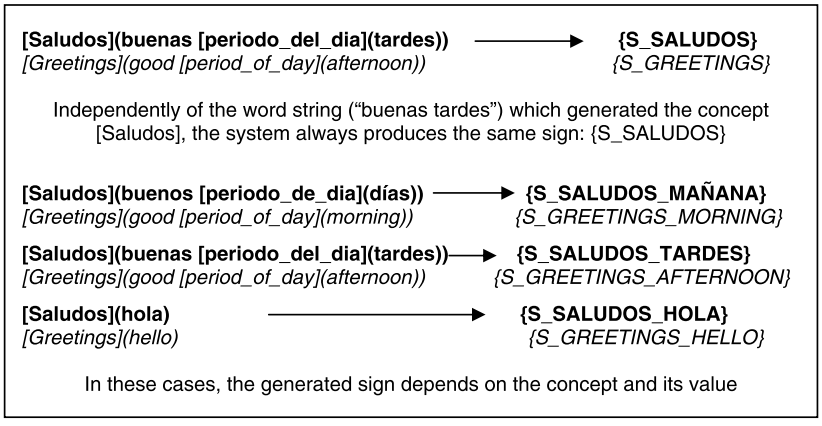

Fig. 1. Examples of assigning a unique sign to a single semantic concept.

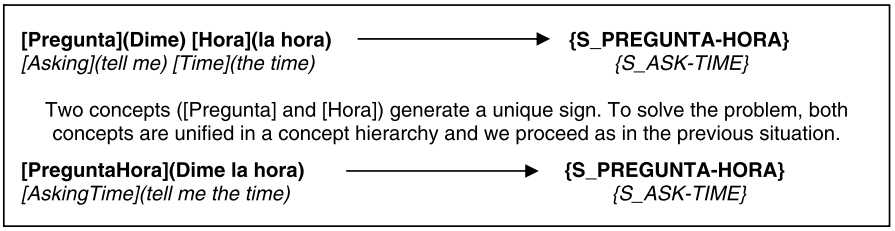

Fig. 2. Examples of assigning a unique sign to several semantic concepts.

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001

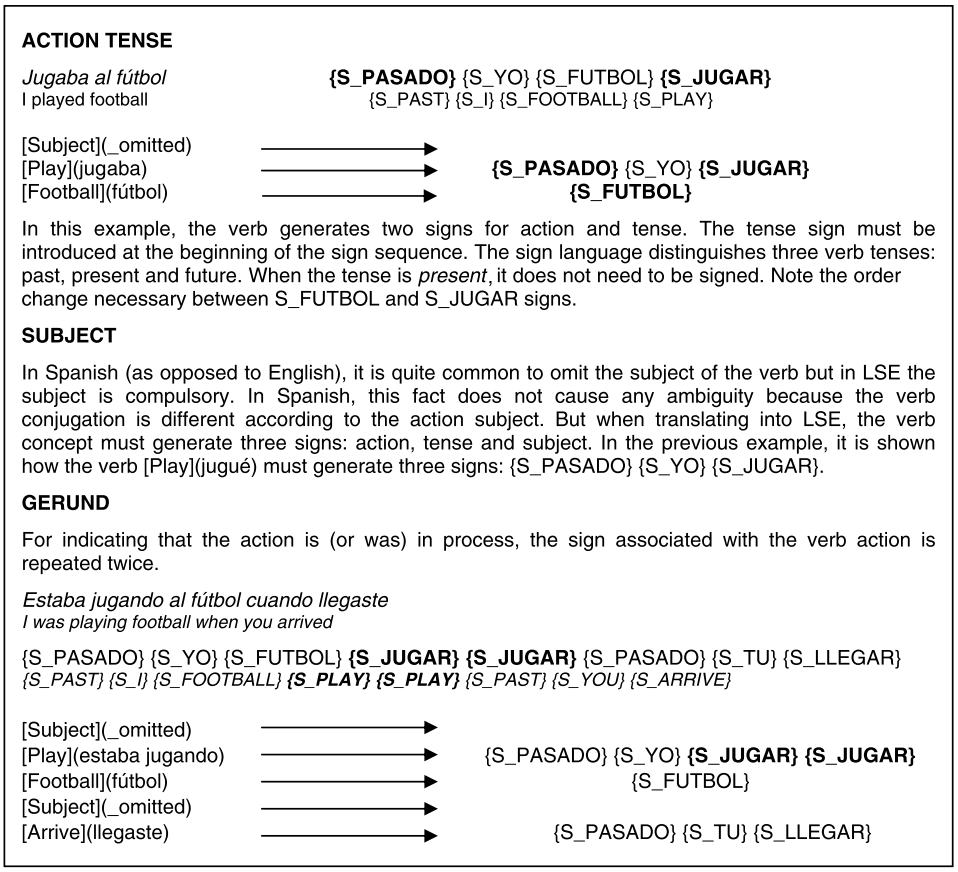

Fig. 3. Type of sign sequences generated by verb concepts.

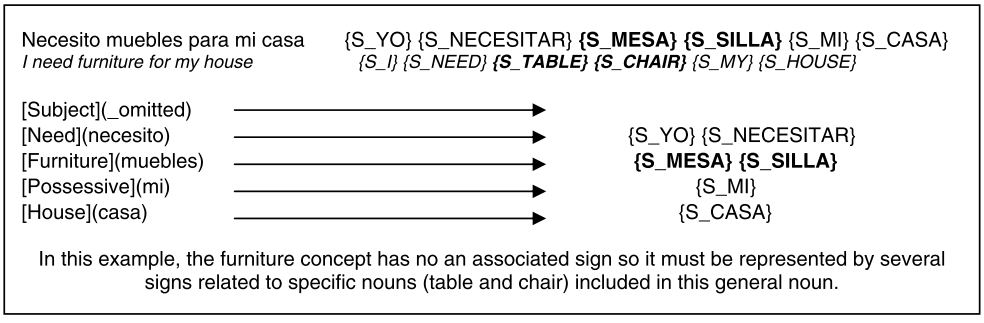

Fig. 4. Signs for general nouns not presented in the sign language.

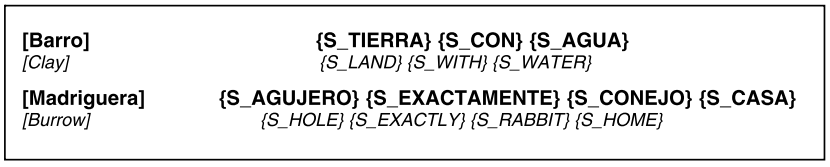

Fig. 5. Examples of Lexical–Visual Paraphrases.

Because of this, the sign dictionary is smaller than the Spanish word dictionary. This fact makes it necessary to combine signs in order to represent other concepts.

- DATE AND TIME. As shown in Fig. 6, a date representation can be made with one or several signs. The time generally requires several signs for a full representation.
- EMPHASIS. When somebody wants to emphasize some aspect of a sentence, this can be done by introducing a new sign (including a face expression) or by repeating

[Fecha](3 de mayo de 2007) *[Date](May 3rd, 2007)*	**{S_TERCERO} {S_MAYO} {S_DOS} {S_MIL} {S_SIETE}** *{S_THIRD} {S_MAY} {S_TWO} {S_THOUSAND} {S_SEVEN}*
[Hora](4:35 am) *[Time](4:35 am)*	**{S_HORA} {S_CUATRO} {S_Y} {S_TREINTA} {S_CINCO} {S_MAÑANA}** *{S_HOUR} {S_FOUR} {S_AND} {S_THIRTY} {S_FIVE} {S_MORNING}*

Fig. 6. Dates and times examples.

[Casa](casas) *[House](houses)*	**{S_CASA} {S_CASA}** *{S_HOUSE} {S_HOUSE}*
[Apple](apples) *[Apple](apples)*	**{S_MANZANA_2MANOS}** *{S_APPLE_2HANDS}*
[Car](cars) *[Car](cars)*	**{S_VARIOS} {S_COCHE}** *{S_SEVERAL} {S_CAR}*

Fig. 7. Plural noun examples.

the associated sign. For example, in order to emphasize the possessive "my" in the sentence "this is my house", the associated sign is repeated: {S_THIS}{S_MY} {S_MY}{S_HOUSE}.

- PLURAL NOUNS. There are several ways of specifying an object in plural (all of them with the same meaning): repeating the sign, introducing an adverbial sign or representing the sign with both hands. Several examples are shown in Fig. 7.
- GENDER. A new sign can be introduced into the sequence to indicate the gender of an object. Usually the gender can be deduced by context and it is not necessary to specify it. This sign appears when the gender is necessary for the meaning or the user wants it to be highlighted.

2.4. Several semantic concepts generate several signs

Finally, the most complicated situation appears when it is necessary to generate several signs from several concepts with dependencies between them. These cases are less frequent than those presented in Sections 2.1, 2.2 and 2.3. Some examples are as follows:

- Verb/Action sign depending on the subject of the action. For example, the verb "fly" is represented with different signs depending on the subject of the action: bird, plane, etc.
- A similar situation arises when the sign associated to an adjective changes depending on the qualified object. For example, the sign for the adjective "good" is different when referring to a person or a material object.

3. Translation system overview

Fig. 8 shows the module diagram of the system developed for translating spoken language into Spanish Sign language (LSE). The main modules are as follows:

- The first module, the speech recognizer, converts natural speech into a sequence of words (text). One important characteristic of this module is the confidence measure

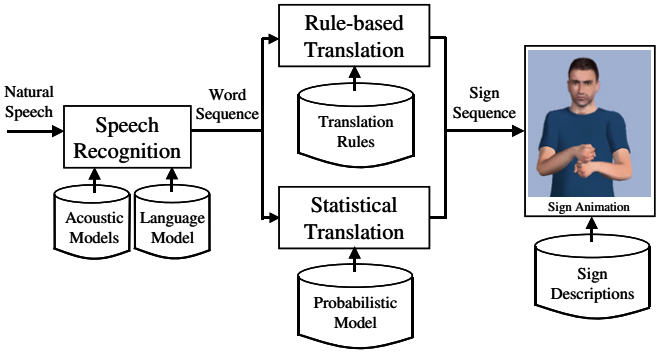

Fig. 8. Spoken Language to Sign Language translation system.

estimation where every recognized word is tagged with a confidence value between 0.0 (lowest confidence) and 1.0 (highest confidence).

- The natural language translation module converts a word sequence into a sign sequence. For this module, the paper presents two proposals. The first one consists of a rule-based translation strategy, where a set of translation rules (defined by an expert) guides the translation process. The second alternative is based on a statistical translation approach where parallel corpora are used for training language and translation models.
- The sign animation is carried out by VGuido: the eSIGN 3D avatar developed in the eSIGN project (eSIGN project). It has been incorporated as an ActiveX control. The sign descriptions are generated previously through the eSIGN Editor environment.

Fig. 9 presents the user interface. In this interface, it is possible to see the virtual agent (Vguido) and other controls and windows for user interaction: a read-only text window for presenting the sequence of recognized words, a text input window for introducing a Spanish sentence, a set of slots for presenting the translated signs and their confidence, etc.

3.1. Domain and database

The experimental framework is restricted to a limited domain that consists of sentences spoken by an official when assisting people who are applying for their Identity Card or related information. In this context, a speech to sign language translation system is very useful since most of the officials do not know sign language and have difficulties when interacting with Deaf people.

The most frequently used sentences have been selected from typical dialogues between officials and users, adding

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001

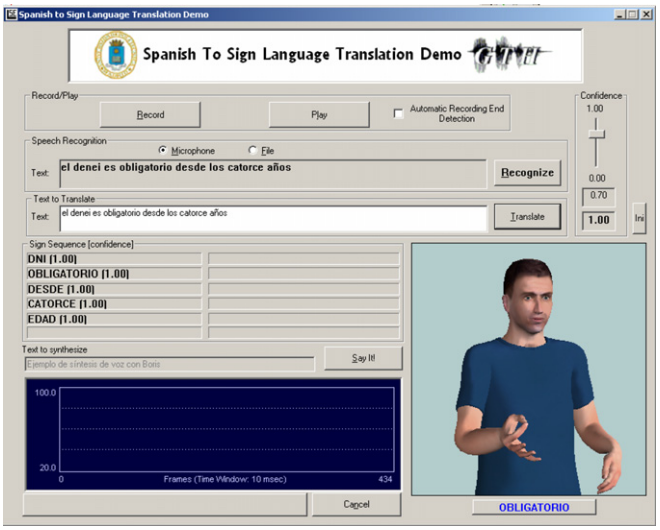

Fig. 9. User Interface for the Spanish to Sign Language Translation System.

up to a total of 416 sentences that contain more than 650 different words. In order to represent the Spanish Sign Language (LSE) in terms of text symbols, each sign has been represented by a word written in capital letters. For example, the sentence "you have to pay 20 euros as document fee" is translated into "FUTURE YOU TWENTY EURO DOC_FEE PAY COMPULSORY".

Once the LSE encoding was established, a professional translator translated the original set into LSE making use of more than 320 different signs. Then, the 416 pairs were randomly divided into two disjoint sets: 266 for training and 150 for testing purposes. The main features of the corpus are summarized in Table 1. For both text-to-sign and speech-to-sign translation purposes the same test set has been used. The test sentences were recorded by two speakers (1 male and 1 female).

As shown in Table 1, the size of the vocabulary compared to the overall amount of running words in the training set is very high (every word appears 6 times on average). In addition, the perplexity of the test set is high considering the small vocabulary. The aforementioned ratio together with the high perplexity show the high data dispersion of this database.

In these circumstances, another important aspect is the large amount of unknown words (OOV words) in the test set. In this task, there are 93 OOV words out of 532 (source language) and 30 OOV signs (target language).

Table 1
Statistics of the bilingual corpus in Spanish and Spanish Sign Language (LSE)

	Spanish	LSE
Training		
Sentence pairs	266	
Different sentences	259	253
Running words	3153	2952
Vocabulary	532	290
Test		
Sentence pairs	150	
Running words	1776	1688
Unknown words, OOV	93	30
Perplexity (3-grams)	15.4	10.7

4. Speech recognition module

The speech recognizer used is a state-of-the-art speech recognition system developed at GTH-UPM (GTH). It is a HMM (Hidden Markov Model)-based system with the following main characteristics:

- It is a continuous speech recognition system: it recognizes utterances made up of several continuously spoken words. In this application, the vocabulary size is 532 Spanish words.
- Speaker independency: the recognizer has been trained with a lot of speakers (4000 people), making it robust against a great range of potential speakers without the need for further training by actual users.
- The system uses a front-end with PLP coefficients derived from a Mel-scale filter bank (MF-PLP), with 13 coefficients including c0 and their first and second-order differentials, giving a total of 39 parameters for each 10 ms. frame. This front-end applies CMN and CVN techniques.
- For Spanish, the speech recognizer uses a set of 45 units: it differentiates between stressed/unstressed/nasalized vowels, it includes different variants for the vibrant 'r' in Spanish, different units for the diphthongs, the fricative version of 'b', 'd', 'g', and the affricates version of 'y' (like 'ayer' and 'cónyuge'). The system also has 16 silence and noise models for detecting acoustic sounds (non-speech events like background noise, speaker artifacts, filled pauses, etc.) that appear in spontaneous speech. The system uses context-dependent continuous Hidden Markov Models (HMMs) built using decision-tree state clustering: 1807 states and seven mixture components per state. These models have been trained with more that 40 h of speech from the SpeechDat database. Although SpeechDat is a telephone speech database, the acoustic models can be used in a microphone application because CMN and CVN techniques have been used to compensate the channel differences. The influence of this aspect in the speech recognition results is small.
- Regarding the language model, the recognition module uses statistical language modeling: 2-gram, as the database is not large enough to estimate reliable 3-grams.
- The recognition system can generate one optimal word sequence (given the acoustic and language models), a solution expressed as a directed acyclic graph of words that may compile different alternatives, or even the *N*-best word sequences sorted by similarity to the spoken utterance. In this work, only the optimal word sequence is considered.

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001

- The recognizer provides one confidence measure for each word recognized in the word sequence. The confidence measure is a value between 0.0 (lowest confidence) and 1.0 (highest confidence) (Ferreiros et al., 2005). This measure is important because the speech recognizer performance varies depending on several aspects: level of noise in the environment, non-native speakers, more or less spontaneous speech, or the acoustic similarity between different words contained in the vocabulary.

For the speech recognition experiments, the following three situations have been considered:

Exp 1: In the first situation, the language model and the vocabulary were generated from the training set. This is the real situation. As we can see in Table 2 the WER is quite high. The reasons for this WER are the high number of OOV words (93 OOV words out of 532) in the test set, and the very small number of sentences to train the language model. In order to demonstrate the validity of these reasons, the following two situations were considered.

Exp 2: In this case, the language model was generated from the training set (as in Exp 1) but the vocabulary included all words (training and testing sets). The WER difference between Exp 1 and Exp 2 reveals the influence of the OOV words (as the speech recognizer is not able to handle OOV words).

Exp 3: Finally, the third experiment tried to estimate a top limit in the speech recognizer performance considering all the available phrases for training the language model and generating the vocabulary. In this case, the WER difference between Exp 2 and Exp 3 shows the influence of the small amount of data used for training the Language Model. In this case, the WER is 4.04: a very good value for a 500 word task.

The speech recognition results for this task are presented in Table 2. These results show the outstanding influence of data sparseness (due to the small amount of data) over the decoding process: comparing Exp 1 to Exp 2, it is shown that OOV words are responsible for increasing the WER from 15.04 to 23.50. Comparing Exp 2 to Exp 3, the poor language model makes the WER to increase from 4.04 to 15.04.

Table 2
Final speech recognition results: Word Error Rate (WER)

	WER	Ins (%)	Del (%)	Sub (%)
Exp 1	**23.50**	**2.60**	**6.45**	**14.45**
Exp 2	15.04	1.19	5.43	8.42
Exp 3	4.04	0.66	1.64	1.74

5. Natural language translation

The natural language translation module converts the word sequence, obtained from the speech recognizer, into a sign sequence that will be animated by the 3D avatar. For this module, two approaches have been implemented and evaluated: rule-based translation and statistical translation.

5.1. Rule-based translation

In this approach, the natural language translation module has been implemented using a rule-based technique considering a bottom-up strategy. In this case, the relationship between signs and words are defined by an expert hand. In a bottom-up strategy, the translation analysis is carried out by starting from each word individually and extending the analysis to neighborhood context words or already-formed signs (generally named blocks). This extension is made to find specific combinations of words and/or signs (blocks) that generate another sign. Not all the blocks contribute or need to be present to generate the final translation. The rules implemented by the expert define these relations. Depending on the scope of the block relations defined by the rules, it is possible to achieve different compromises between reliability of the translated sign (higher with higher lengths) and the robustness against recognition errors: when the block relations involve a large number of concepts, one recognition error can cause the rules not to be executed.

The translation process is carried out in two steps. In the first one, every word is mapped to one or several syntactic–pragmatic tags. After that, the translation module applies different rules that convert the tagged words into signs by means of grouping concepts or signs (blocks) and defining new signs. These rules can define short and large scope relationships between the concepts or signs. At the end of the process, the block sequence is expected to correspond to the sign sequence resulting from the translation process.

The rule-based translation module contains 153 translation rules. The translation module has been evaluated with the test set presented in Table 1. For evaluating the performance of the systems, the following evaluation measures have been considered: SER (Sign Error Rate), PER (Position Independent SER), BLEU (BiLingual Evaluation Understudy), and NIST. The first two measures are error measures (the higher the value, the worse the quality) whereas the last two are accuracy measures (the higher, the better). The final results reported by this module are presented in Table 3.

The speech-input translation results obtained from the three experiments mentioned in Section 4 are shown in Table 3. As a baseline (denoted in the Table as REF), the text-to-text translation results (considering the utterance transcription directly) are included. As is shown, the SER is higher when using the speech recognition output instead of the transcribed sentence. The reason is that the

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001

Table 3
Results obtained with the rule-based translation system

	SER	PER	BLEU	NIST
Exp 1	**31.60**	**27.02**	**0.5780**	**7.0945**
Exp 2	24.94	20.21	0.6143	7.8345
Exp 3	18.23	14.87	0.7072	8.4961
REF	**16.75**	**13.17**	**0.7217**	**8.5992**

speech recognizer introduces recognition mistakes that produce more translation errors: the percentage of wrong signs increases and the BLEU decreases.

Analyzing the results in detail, it is possible to report that the most frequent errors committed by the translation module have the following causes:

- In Spanish, it is very common to omit the subject of a sentence, but in Sign Language it is compulsory to use it. In order to deal with this characteristic, several rules have been implemented in order to verify whether every verb has a subject and to include a subject if there is any verb without it. When applying these rules some errors are inserted: typically a wrong subject is associated to a verb.
- Several possible translations. One sentence can be translated into different sign sequences. When one of the possibilities is not considered in the evaluation, some errors are reported by mistake. This situation appears, for example, when the passive form is omitted in several examples.
- In Sign Language, a verb complement is introduced by a specific sign: for example a time complement is introduced with the sign WHEN, or a mode complement is introduced with the sign HOW. There are several rules for detecting the type of complement, but sometimes it is very difficult to detect the difference between a place complement and a time complement. Moreover, when the verb complement is very short (made up of one word: "today", "now", "here", ...), this introductory sign is omitted for simplicity (deaf people do not sign the introductory sign to reduce the signing time). When the system estimates the complement length wrongly an error occurs.
- In the test set, there is a large number of unknown words that generate a significant number of errors.

5.1.1. Sign confidence measure

The translation module generates one confidence value for every sign: a value between 0.0 (lowest confidence) and 1.0 (highest confidence). This sign confidence is computed from the word confidence obtained from the speech recognizer. This confidence computation is carried out by an internal procedure that is coded inside the proprietary language interpreter that executes the rules of the translation module.

In this internal engine, there are "primitive functions", responsible for the execution of the translation rules written by the experts. Each primitive has its own way of generating the confidence for the elements it produces. One common case is for the primitives that check for the existence of a sequence of words/concepts (source language) to generate some signs (target language), where the primitive usually assigns the average confidence of the blocks which it has relied on to the newly created elements.

In other more complex cases, the confidence for the generated signs may be dependent on a weighted combination of confidences from a mixture of words and/or internal or final signs. This combination can consider different weights for the words or concepts considered in the rule. These weights are defined by the expert as the same time the rule is coded. For example, in Fig. 10, the confidence measures for the signs "DNI" and "SER" (0.7 in both cases) have been computed at the average value of the confidence of "denei" (0.6) and "es" (0.8). The confidence values of the words tagged as GARBAGE are not used to compute sign confidence values. In Ferreiros' work (Ferreiros et al., 2005), it is possible to find a detailed description of confidence measure computation.

This system is one of the few natural language translation modules that generates a confidence measure for signs (target language). Section 6.1 describes the use of sign confidence measures when representing the sign.

5.2. Statistical translation

The Phrase-based translation system is based on the software released to support the shared task at the 2006 NAACL Workshop on Statistical Machine Translation (http://www.statmt.org/wmt06/).

The phrase model has been trained following these steps:

- Word alignment computation. At this step, the GIZA++ software (Och and Ney, 2000) has been used to calculate the alignments between words and signs. The parameter "alignment" was fixed to "grow-diag-final" as the best option.
- Phrase extraction (Koehn et al., 2003). All phrase pairs that are consistent with the word alignment are collected. The maximum size of a phrase has been fixed to 7.
- Phrase scoring. In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

The Pharaoh decoder is used for the translation process. This program is a beam search decoder for phrase-based

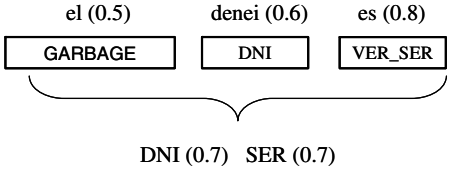

Fig. 10. Example of Sign Confidence computation.

Table 4
Results obtained with the statistical system

	SER	PER	BLEU	NIST
Exp 1	**38.72**	**34.25**	**0.4941**	**6.4123**
Exp 2	36.08	32.00	0.4998	6.4865
Exp 3	34.22	30.04	0.5046	6.5596
REF	**33.74**	**29.14**	**0.5152**	**6.6505**

statistical machine translation models (Koehn, 2004). In order to obtain a 3-gram language model needed by Pharaoh, the SRI language modeling toolkit has been used. The Carmel software was used for the *n*-best list generation.

The speech-input translation results are shown in Table 4 for the three experiments described previously. As a baseline (denoted in the Table as Ref), the text-to-text translation results (considering the utterance transcription directly) are included.

The rule-based strategy has provided better results on this task because it is a restricted domain and it has been possible to develop a complete set of rules with a reasonable effort. Another important aspect is that the amount of data for training is very little and the statistical models cannot be trained properly. In these circumstances, the rules defined by an expert introduce knowledge not seen in the data, making the system more robust with new sentences.

6. Sign animation with the eSIGN avatar: VGuido

The signs are represented by means of VGuido (the eSIGN 3D avatar) animations. An avatar animation consists of a temporal sequence of frames, each of which defines a static posture of the avatar at the appropriate moment. Each of these postures can be defined by specifying the configuration of the avatar's skeleton, together with some characteristics which define additional distortions to be applied to the avatar.

In order to make an avatar sign, it is necessary to send to the avatar pre-specified animation sequences. A signed animation is generated automatically from an input script in the Signing Sign Markup Language (SiGML) notation. SiGML is an XML application which supports the definition of sign sequences. The signing system constructs human-like motion from scripted descriptions of signing motions. These signing motions belong to "Gestural-SiGML", a subset of the full SiGML notation, which is based on the HamNoSys notation for Sign Language transcription (Prillwitz et al., 1989).

The concept of synthetic animation used in eSIGN is to create scripted descriptions for individual signs and store them in a database. Populating this database may take some time but considering a minimum amount of one hundred signs, it is possible to obtain signed phrases for a restricted domain. This process is carried out by selecting the required signs from the database and assembling them in the correct order.

The major advantage of this approach is its flexibility: The lexicon-building task does not require special equipment, just a database. The morphological richness of sign languages can be modeled using a sign language editing environment (the eSIGN editor) without the need of manually describing each inflected form.

HamNoSys and other components of SiGML mix primitives for static gestures (such as parts of the initial posture of a sign) with dynamics (such as movement directions) by intention. This allows the transcriber to focus on essential characteristics of the signs when describing a sign. This information, together with knowledge regarding common aspects of human motion as used in signing such as speed, size of movement, etc., is also used by the movement generation process to compute the avatar's movements from the scripted instructions. Fig. 11 shows the process for specifying a sign from the HamNoSys description.

6.1. Incorporating confidence measures in sign animation

As described above, the result of the natural language translation process is a sign sequence. Every sign in the sequence can be tagged with a confidence measure ranging from 0.0 (lowest confidence) to 1.0 (highest confidence). Depending on its confidence value, each sign is represented in a different way. There are three confidence levels defined:

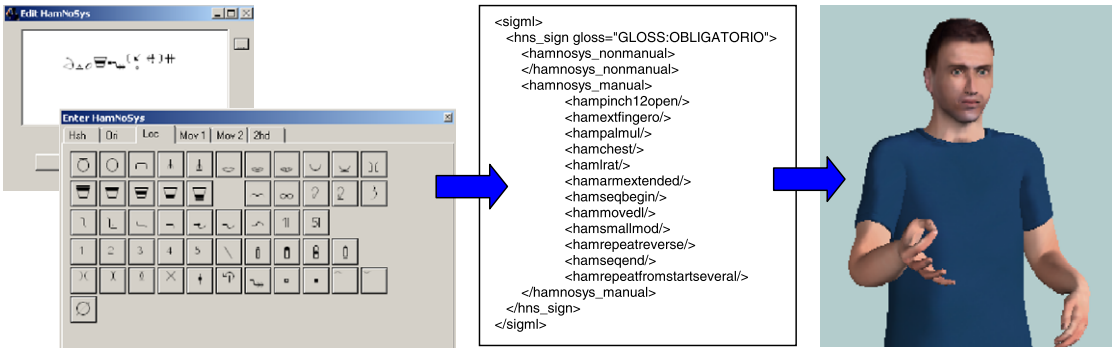

Fig. 11. Process to generate signs with the avatar.

- *High confidence*. Confidence value higher than 0.5 defines a level where all signs are represented in their standard way.
- *Medium confidence*. Confidence value between 0.25 and 0.5. In this case, the sign is represented but an additional low confidence signal is presented: an interrogative mark or a confused avatar face.
- *Low confidence*. Confidence value of less than 0.25. At this level, the sign is not played. During the time associated to this sign an interrogative mark or a confused avatar expression is presented.

6.2. Reducing the delay between the spoken utterance and the sign animation

One important aspect to be considered in a speech to sign language translation system is the delay between the spoken utterance and the animation of the sign sequence. This delay is around 1–2 s and it slows down the interaction. In order to reduce this delay, the speech recognition system has been modified to report partial recognition results every 100 ms. These partial results are translated into partial sign sequences that can be animated without the need to wait until the end of the spoken utterance.

When implementing this solution, an important problem appeared due to the fact that the translation is not a linear alignment process between spoken words and signs. Words that are spoken in the middle of the utterance can report information about the first signs. So until these words are spoken, the first sign is not completely defined and it cannot be represented.

This problem appears in two situations:

- *Verb tense signs*. These signs are FUTURE and PAST (there is not a PRESENT sign because it is the default) and they must appear at the beginning of the sign sequence independently of the verb position. In partial translations, this sign appears when the action verb is spoken and this usually happens approximately in the middle of the utterance.
- *Verb subject signs*. The second aspect is the verb subject. In colloquial Spanish, the subject can frequently be omitted, but in Sign Language, every verb must have a subject. In the translation process, it is necessary to check whether a verb has a subject and the system must include one (at the beginning of the sentence) if there is none.

In order to solve this problem, two restrictions were imposed for representing a partial sign sequence:

- The first sign must be a verb tense or a subject sign: in this case, all sign language sentences start with one of these signs. Considering that these signs are defined when a verb appears in the Spanish sentence, this restriction can be reformulated as follows: the partial sequence should contain a verb sign in order to start the sign representation.
- The first sign should be the same for at least three consecutive partial sign sequences.

With these restrictions, a 40% delay reduction is achieved without affecting the translation process performance.

7. Conclusions

This paper has presented the implementation and the first experiments on a speech to sign language translation system for a real domain. The domain consists of sentences spoken by an official when assisting people who apply for, or renew their Identity Card. The translation system implemented is made up of three modules. A speech recognizer is used for decoding the spoken utterance into a word sequence. After that, a natural language translation module converts the word sequence into a sequence of signs belonging to the Spanish Sign Language (LSE). In the last module, a 3D avatar plays the sign sequence.

In these experiments, two proposals for the natural language translation module have been implemented and evaluated. The first one consists of a rule-based translation module reaching a 31.60% SER (Sign Error Rate) and a 0.5780 BLEU (BiLingual Evaluation Understudy). In this proposal, confidence measures from the speech recognizer have been used to compute a confidence measure for every sign. This confidence measure is used during the sign animation process to inform the user about the reliability of the translated sign.

The second alternative for natural language translation is based on a statistical translation approach where parallel corpora were used for training. The best configuration has reported a 38.72% SER and a 0.4941 BLEU.

The rule-based strategy has provided better results on this task because it is a restricted domain and it has been possible to develop a complete set of rules with a reasonable effort. Another important aspect is that the amount of data for training is very little and the statistical models cannot be estimated reliably. In these circumstances, the rules defined by an expert introduce knowledge not seen in the data making the system more robust against new sentences. The rule-based translation module has presented a very high percentage of deletions compared to the rest of the errors. This is due to the rule-based strategy: when the speech recognition makes an error, some concept patterns do not appear (they do not fit into the defined rules) and some signs are not generated. On the other hand, the statistical translation module has generated greater percentage of insertions and substitutions compared to the rule-based system.

Regarding the 3D avatar module, the eSIGN avatar has been described including the use of sign confidence in sign representation. Finally, the paper has described the problems when implementing a strategy for reducing the delay

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001

between the spoken utterance and the sign animation. With the solution proposed, a 40% delay reduction was achieved without affecting the translation process performance.

Acknowledgements

The authors would like to thank the eSIGN (Essential Sign Language Information on Government Networks) consortium for giving us permission to use of the eSIGN Editor and the 3D avatar in this research work. This work has been supported by the following projects ATI-NA (UPM-DGUI-CAM. Ref: CCG06-UPM/COM-516), ROBINT (MEC Ref: DPI2004-07908-C02) and EDEC-AN (MEC Ref: TIN2005-08660-C04). The work presented here was carried out while Javier Macías-Guarasa was a member of the Speech Technology Group (Department of Electronic Engineering, ETSIT de Telecomunicación. Universidad Politécnica de Madrid). Authors also want to thank Mark Hallett for the English revision.

References

Abdel-Fattah, M.A., 2005. Arabic sign language: a perspective. Journal of Deaf Studies and Deaf Education 10 (2), 212–221.

ASL corpus: <http://www.bu.edu/asllrp/>.

Atherton, M., 1999. Welsh today BSL tomorrow. In: Deaf Worlds 15 (1), pp. 11–15.

Bertenstam, J. et al., 1995. The Waxholm system-A progress report. In: Proc. Spoken Dialogue Systems, Vigso, Denmark.

Bungeroth, J. et al., 2006. A German Sign Language Corpus of the Domain Weather Report. In: 5th Internat. Conf. on Language Resources and Evaluation, Genoa, Italy.

Casacuberta, F., Vidal, E., 2004. Machine translation with inferred stochastic finite-state transducers. Computational Linguistics 30 (2), 205–225.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, 2002. MACK: media lab autonomous conversational kiosk. In: Proc. of Imagina: Intelligent Autonomous Agents, Monte Carlo, Monaco.

Christopoulos, C., Bonvillian, J., 1985. Sign language. Journal of Communication Disorders 18, 1–20.

Cole, R. et al., 1999. New tools for interactive speech and language training: using animated conversational agents in the classrooms of profoundly deaf children. In: Proc. ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, London, pp. 45–52.

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., Yan, J., 2003. Perceptive animated interfaces: first steps toward a new paradigm for human computer interaction. IEEE Transactions on Multimedia: Special Issue on Human Computer Interaction 91 (9), 1391–1405.

ECHO corpus: <http://www.let.kun.nl/sign-lang/echo/>.

Engberg-Pedersen, E., 2003. From pointing to reference and predication: pointing signs, eyegaze, and head and body orientation in Danish Sign Language. In: Kita, Sotaro (Ed.), Pointing: Where Language, Culture, and Cognition Meet. Erlbaum, Mahwah, NJ, pp. 269–292.

ESIGN project: <http://www.sign-lang.uni-hamburg.de/eSIGN/>.

Ferreiros, J., San-Segundo, R., Fernández, F., D'Haro, L., Sama, V., Barra, R., Mellén, P., 2005. New Word-Level and Sentence-Level Confidence Scoring Using Graph Theory Calculus and its Evaluation on Speech Understanding. Interspeech 2005, Lisboa, Portugal, Septiembre, pp. 3377–3380.

Gallardo, B., Montserrat, V., 2002. Estudios Lingüísticos sobre la lengua de signos española. Universidad de Valencia. Ed. AGAPEA ISBN: 8437055261. ISBN-13: 9788437055268.

Granström, B., House, D., Beskow, J., 2002. Speech and Signs for Talking Faces in Conversational Dialogue Systems. Multimodality in Language and Speech Systems. Kluwer Academic Publishers, pp. 209–241.

GTH: <http://lorien.die.upm.es>.

Gustafson, J., 2002. Developing multimodal spoken dialogue systems – empirical studies of spoken human–computer interactions. PhD. Dissertation. Dept. Speech, Music and Hearing, Royal Inst. of Technology, Stockholm, Sweden.

Gustafson, J., Bell, L., 2003. Speech technology on trial: experiences from the august system. Journal of Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems, 273–286.

Herrero-Blanco, A., Salazar-Garcia, V., 2005. Non-verbal predictability and copula support rule in Spanish Sign Language. In: de Groot, Casper, Hengeveld, Kees (Eds.), Morphosyntactic expression in functional grammar. (Functional Grammar Series; 27) Berlin [u.a.]: de Gruyter, pp. 281–315.

Instituto Cervantes: <http://www.cervantesvirtual.com/seccion/signos/>.

Koehn, P., Och, F.J., Marcu, D., 2003. Statistical phrase-based translation. In: Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127–133.

Koehn, P., 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. AMTA.

Lundeberg, M., Beskow, J., 1999. Developing a 3D-agent for the August dialogue system. In: Proc. Audio–Visual Speech Processing, Santa Cruz, CA.

Masataka, N. et al., 2006. Neural correlates for numerical processing in the manual mode. Journal of Deaf Studies and Deaf Education 11 (2), 144–152.

Meurant, L., 2004. Anaphora, role shift and polyphony in Belgian sign language. Poster. In: TISLR 8 Barcelona, September 30–October 2. Programme and Abstracts. (Internat. Conf. on Theoretical Issues in Sign Language Research; 8), pp. 113–115.

Nyst, V., 2004. Verbs of motion in Adamorobe Sign Language. Poster. In: TISLR 8 Barcelona, September 30–October 2. Programme and Abstracts. (Internat. Conf. on Theoretical Issues in Sign Language Research; 8), pp. 127–129.

Och J., Ney, H., 2000. Improved statistical alignment models. In: Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, HongKong, China, pp. 440–447.

Och J., Ney. H., 2002. Discriminative training and maximum entropy models for statistical machine translation. In: Annual Meeting of the Ass. For Computational Linguistics (ACL), Philadelphia, PA, pp. 295–302.

Och, J., Ney, H., 2003. A systematic comparison of various alignment models. Computational Linguistics 29 (1), 19–51.

Papineni K., Roukos, S., Ward, T., Zhu, W.J., 2002. BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, pp. 311–318.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J., et al., 1989. Hamburg Notation System for Sign Languages – An introductory Guide. In: International Studies on Sign Language and the Communication of the Deaf, Vol. 5. Institute of German Sign Language and Communication of the Deaf. University of Hamburg.

Pyers, J.E., 2006. Indicating the body: Expression of body part terminology in American Sign Language. Language Sciences 28 (2–3), 280–303.

Reyes, I., 2005. Comunicar a través del silencio:las posibilidades de la lengua de signos española. Universidad de Sevilla, Sevilla, p. 310.

Rodríguez, M.A., 1991. Lenguaje de signos Phd Dissertation. Confederación Nacional de Sordos Españoles (CNSE) and Fundación ONCE, Madrid, Spain.

Stokoe, W., 1960. Sign Language structure: an outline of the visual communication systems of the American deaf. Studies in Linguistics. Buffalo, Univ. Paper 8.

Sumita, E., Akiba, Y., Doi, T., et al., 2003. A Corpus-Centered Approach to Spoken Language Translation. Conf. of the Europ. Chapter of the Ass. For Computational Linguistics (EACL), Budapest, Hungary, pp. 171–174.

Sutton, S., Cole, R., 1998. Universal speech tools: the CSLU toolkit. In: Proc. of Internat. Conf. on Spoken Language Processing, Sydney, Australia, pp. 3221–3224.

Sylvie, O., Surendra, R., 2005. Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6).

Zens, R., Och, F.J., Ney, H., 2002. Phrase-based statistical machine translation. German Conf. on Artificial Intelligence (KI 2002). Springer, LNAI, Aachen, Germany, pp. 18–32.

Please cite this article in press as: San-Segundo, R. et al., Speech to sign language translation system for Spanish, Speech Comm. (2008), doi:10.1016/j.specom.2008.02.001