Pattern Anal Applic (2012) 15:203–224
DOI 10.1007/s10044-011-0243-9

INDUSTRIAL AND COMMERCIAL APPLICATION

Design, development and field evaluation of a Spanish into sign language translation system

R. San-Segundo · J. M. Montero · R. Córdoba · V. Sama · F. Fernández · L. F. D'Haro · V. López-Ludeña · D. Sánchez · A. García

Received: 24 February 2010 / Accepted: 15 September 2011 / Published online: 29 September 2011

Abstract This paper describes the design, development and field evaluation of a machine translation system from Spanish to Spanish Sign Language (LSE: Lengua de Signos Española). The developed system focuses on helping Deaf people when they want to renew their Driver's License. The 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 signs). For the natural language translator, three technological approaches have been implemented and evaluated: an example-based strategy, a rule-based translation method and a statistical translator. For the final version, the implemented language translator combines all the alternatives into a hierarchical structure. This paper includes a detailed description of the field evaluation. This evaluation was carried out in the Local Traffic Office in Toledo involving real government employees and Deaf people. The evaluation includes objective measurements from the system and subjective information from questionnaires. The paper details the main problems found and a discussion on how to solve them (some of them specific for LSE).

Keywords Deaf people · Spanish sign language (LSE) · Spoken language translation · Sign animation · Driver's license renewal

1 Introduction

Deafness gives rise to significant communications problems: most deaf people are unable to use written languages, having serious problems when expressing themselves in these languages or understanding written texts. In Spain, 92% of the Spanish deaf population have significant difficulties in understanding and expressing themselves in written Spanish (based on information from the Spanish Statistics Institute [17] and the Ministry of Education [22]). The main problems are related to verb conjugations, gender/number concordances and abstract concept explanations. Because of this, around 47% of the Spanish deaf population,[1] aged more than 10, do not have basic-level studies or are illiterate and only between 1 and 3% have a university level education [22].

Spanish sign language (LSE: Lengua de Signos Española) has been an official Spanish languages since 2007. The Spanish government has defined a long-term plan to invest resources in this language. Because LSE is a young official language, it is not well-known by people who can hear (hereinafter referred to as hearing people), giving rise to

R. San-Segundo · J. M. Montero · R. Córdoba · V. Sama · F. Fernández · L. F. D'Haro · V. López-Ludeña
Departamento de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

R. San-Segundo (✉)
Grupo de Tecnología del Habla, Dpto. Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain
e-mail: lapiz@die.upm.es

D. Sánchez · A. García
Fundación CNSE, Confederación Estatal de Personas Sordas, Madrid, Spain

[1] Along with text, it is necessary to distinguish between "deaf" and "Deaf": the former refers to non-hearing people, and the latter refers to non-hearing people who use a sign language to communicate between themselves, being part of the "Deaf community".

significant communications barriers between a Deaf person and a government employee who is providing a service personally, for example. These barriers lead to Deaf people having fewer opportunities or rights (in practice). This happens, for example, when people want to renew their Driver's License (DL). The cost of this service is defined by the government and it must be less than a money quantity. A lot of government employees do not know LSE so a Deaf person needs a human interpreter to translate the government employee's explanations. In this case, the cost of the service for a Deaf person is much higher. Besides, there is not always an interpreter available, so a Deaf person cannot access this service at the same time as a hearing person does. In Spain, there is a ratio of 221 deaf people to 1 interpreter (statistics from INE) so it is very interesting to develop automatic translation systems for helping hearing and Deaf people to communicate between each other.

This paper describes in detail the design, development and field evaluation of a machine translation system from Spanish to LSE. The paper focuses on three main aspects: integration of speech recognition and language translation algorithms, the generation of the parallel corpus necessary for training the language translation algorithms, and a field evaluation involving Deaf users. As will be presented in next section, there are similar systems for other languages, but the system described here is the first one for translating Spanish into LSE evaluated involving real interactions between Deaf and hearing people without an interpreter: government employees that provide a service (renewing a DL) and Deaf people that want to access this service. The proposed system translates the government employee's explanations into LSE for Deaf people and they can ask questions to the government employee using a spoken Spanish generation system from gloss sequences [34].

This paper is organised as follows. Section 2 presents the state of the art. Section 3 describes the linguistic study carried out to develop the system, including the collection of the parallel corpus. Section 4 presents the system architecture, speech recognition (Sect. 4.1), language translation (Sect. 4.2) and sign animation modules (Sect. 4.3). Section 4 includes a description of the system interface (Sect. 4.4) and the system limitations (Sect. 4.5). Section 5 presents a summary of the spoken Spanish generation system from gloss sequences to allow Deaf people to ask questions to government employees. Finally, the field evaluation and the main conclusions are described in Sects. 6 and 7, respectively.

2 State of the art

In the last 10 years, the European Commission (EC) and the USA Government have invested a lot of resources in language translation research. In Europe, TC-STAR is the latest project of a sequence of them: C-Star, ATR. Vermobil, Eutrans, LC-Star, PF-Star and, finally, TC-STAR. The TC-STAR project (http://www.tc-star.org/) was financed by the EC within the Sixth Program and it was envisaged as a long-term effort to advance research into all core technologies for speech-to-speech translation (SST): automatic speech recognition (ASR), spoken language translation (SLT) and text to speech conversion (TTS) (speech synthesis).

Another important project on language translation funded by the EC is EuroMatrixPlus (http://www.euromatrixplus.net/). This project focuses on creating example systems for every official EU language, and providing other machine translation developers with a baseline infrastructure for building statistical translation models. The EuroMatrixPlus team has organized several Workshops on Statistical Machine Translation (SMT). On the webpages http://www.statmt.org/ and http://matrix.statmt.org/, it is possible to obtain all the information about these events. As a result of these workshops, there is a free machine translation system called Moses and available from this web page (http://www.statmt.org/moses/). Moses is a phrase-based statistical machine translation system that allows machine translation system models to be built for any pair of languages, using a collection of translated texts (parallel corpus).

In the USA, Defence Advanced Research Projects Agency (DARPA) is supporting the GALE program (http://www.darpa.mil/ipto/programs/gale/gale.asp). The goal of the DARPA GALE program has been to develop and apply computer software technologies to absorb, analyse and interpret huge volumes of speech and text in multiple languages. Automatic processing "engines" convert and distil the data, delivering pertinent, consolidated information in easy-to-understand formats to military personnel and monolingual English-speaking analysts in response to direct or implicit requests. GALE consists of three major engines: Transcription, Translation and Distillation. The output of each engine is English text. The input to the transcription engine is speech and to the translation engine, text. The distillation engine integrates information of interest to its user from multiple sources and documents. Military personnel will interact with the distillation engine via interfaces that could include various forms of human–machine dialog (not necessarily in natural language). This project has been active for 2 years, and the GALE contractors have been engaged in developing highly robust speech recognition, machine translation, and information delivery systems in Chinese and Arabic. This program has also been boosted by the machine translation evaluation organised by the USA Government, National Institute of Standards and Technology (NIST) (http://www.itl.nist.gov/iad/mig/tests/mt/).

The best performing translation systems are based on several types of statistical approaches [3, 21, 28, 31], including example-based methods [9, 36], finite-state transducers [7] and other data-driven approaches. The progress achieved over the last years has been thanks to several factors such as efficient algorithms for training [29, 39], context-dependent models [40], efficient algorithms for generation [19, 39], incorporation of more powerful computers and bigger parallel corpora, and automatic error measurements [1, 2, 30].

In recent years, several groups have shown interest in translating spoken language into sign languages, and have developed several prototypes: example-based [24], rule-based [33], full sentence [8] or statistical [6, 25]; SiSi system http://www-03.ibm.com/press/us/en/pressrelease/22316.wss) approaches. Table 1 summarises the main characteristics of the main speech into sign language translation systems, highlighting the contribution of this paper as compared to these previous works. As is shown, this paper describes the first system that combines and integrates several translation strategies for translating Spanish into LSE and also presents the first field evaluation under real conditions: with real interactions between hearing and Deaf people.

As regards 3D avatars for representing signs, the VISICAST and Essential Sign Language Information on Government Networks (eSIGN) European Project (http://www.sign-lang.uni-hamburg.de/esign/) [41] have been two of the most significant research efforts in developing tools for the automatic generation of sign language contents. In this project, the main result has been a 3D avatar with enough flexibility to represent signs from the sign language, and a visual environment for creating sign-language animations quickly and easily. The proposed system uses this 3D avatar: Sect. 4.3 includes more details on it. One of the partners in the VISICAST and eSIGN projects is the research group into Virtual Humans at the University of East Anglia (http://www.uea.ac.uk/cmp/research/graphicsvisionspeech/vh). This group has been involved in several projects concerning the generation of sign language using virtual humans: TESSA, SignTel, Visicast, eSIGN, SiSi, LinguaSign, Dicta-Sign, etc.

The main limitations of using avatars for representing sign languages can be classified into two main characteristics: naturalness and intelligibility [11]. As regards intelligibility, the current 3D avatar technology performs reasonably well and the signs are understandable by many Deaf people. The main problem is naturalness; in this case, avatars are still far from being humans [10]. Representing every sign always in the same way can be boring and artificial. One possibility to avoid this repetition is by developing several avatar animations for the same sign (with very slight modifications not affecting the meaning). Representing every sign in the same way helps Deaf people to adapt themselves to the avatar. In this work, instead of developing several versions of every sign, the authors decided to make a significant effort on sign specification by adding natural lip movements, face expressions and body movements (using all flexibility provided by VGuido).

3 Database collection for the Driver's Licence renewing process

When developing SLT systems, it is important to carry out a detailed linguistic analysis and to collect a sufficient amount of parallel sentences for modeling task knowledge properly. Our linguistic study was carried out in collaboration with the Local Traffic Office in Toledo. Over a period of 3 weeks, the most frequent explanations (from government employees) and the most frequent questions (from the customers) were obtained by recording and transcribing spoken conversations. These conversations were recorded between hearing people and government employees, assuming that Deaf people ask the same questions when accessing the same personal service.

This local traffic office is organised in several windows (assistance positions) (Fig. 1): information window (for general questions and form collection), cash desk (for paying taxes), driver window (driver-specific formalities), vehicle window (vehicle-related procedures) and driving school window.

Over a period of 3 weeks, more than 5,000 sentences from all of the windows were collected and analysed. This analysis showed that including the information from all windows, the semantic and linguistic domain was very wide and the vocabulary very large. In order to define the specific domain for developing the system, the service of renewing the driver's licence was selected. The Driver's Licence (DL) renewal process at the Toledo Traffic Office consisted of three steps: first of all, the customer had to go to the information window where he or she got the application form to fill in and a sheet with a list of documents needed for the process: Identification Card, the old DL, a medical certificate and a photo. Secondly, it is necessary to pay €22 at the cash desk. Finally, the customer had to go to the driver window with all the documentation. The new DL was sent by mail within the next 3 months. To drive during this period, the customer received a provisional DL. In all three steps, the customer had to get an order number from a machine (Fig. 1). For generating the corpus, it was necessary to pick up sentences from the three different windows involved in the process.

Finally, 707 sentences were selected: 547 pronounced by government employees and 160 by customers. These sentences have been translated into LSE, both in text

Table 1 Spoken language into Sign language translation systems

Reference	Translation technology	Sign language	Translation performance	Limitations	Our approach in comparison
Cox et al. [8]	Full sentence: the system only recognizes a reduced number of pre-translated sentences	British Sign Language (BSL)	Not reported	• It only translates fixed sentences	• Higher flexibility in the sentences to be translated • Combination of different translation technologies
Bungeroth and Ney [6]	Phrase-based model	German Sign Language (DGS)	Translation rate <50%	• Very small database for the experiments • No field evaluation	• A larger database with Cross-validation test • Combination of different translation technologies • Field evaluation
Morrissey and Way [24]	Example-based	Irish Sign Language (ISL)	Translation rate >60%	• No field evaluation	• Combination of different translation technologies • Field evaluation
SiSi system	Phrase-based model	BSL	Not reported	• No field evaluation	• Combination of different translation technologies • Field evaluation
Morrissey et al [25]	Example-based and Phrase-based	ISL and DGS	BLEU >0.5	• No field evaluation	• Field evaluation
San-Segundo et al. [33]	Rule-based translation	Spanish Sign Language (LSE)	BLEU >0.5	• Very small database • A costly translation technology • No field evaluation	• A larger database with cross validation • Combination of different translation technologies • Field evaluation
This paper	Combination of several translation technologies: example-based, rule-based and phrase-based technologies	Spanish Sign Language (LSE)	BLEU >0.7 Translation Rate >90% (see Sects. 4.2.5 and 6.2)	• Focused on a specific domain (see Sect. 4.5 for more details)	

(sequence of glosses) and in video, and compiled in an excel file. This corpus was increased to 2,124 sentences by incorporating different variants for Spanish sentences (maintaining the LSE translation). The translation was made by two LSE experts in parallel. When there was any discrepancy between them, a committee of four people who knew LSE took the decision: select one of the LSE expert proposals, propose a new translation alternative, or consider both proposals as alternative translations. The committee was made up of one Spanish linguist, two Deaf LSE experts and a Spanish linguist expert in LSE. The excel file contains six different information fields (Fig. 2): VENTANILLA (window: where the sentence was collected), SERVICIO (service provided when the sentence was collected), if the sentence was pronounced by the government employee or customer (*funcionario* or *usuario*, respectively), sentence in Spanish (CASTELLANO), sentence in LSE (sequence of glosses), and a link to the video file with LSE representation. For the system development, only the sentences pronounced by government employees were considered.

The main features of the sentences pronounced by government employees are summarised in Table 2.

There are different ways of writing a sign (sign-writing), traditionally the sign has been written using words (in capital letters) in Spanish (or English in the case of BSL,

Fig. 1 Different windows at the local traffic office in Toledo and order number machine

VENTANILLA	SERVICIO	TIPO	CASTELLANO	LSE	VIDEO
CAJA	Decir la cantidad	Funcionario	ahora tiene que ir a la ventanilla de conductores	AHORA TU VENTANILLA ESPECÍFICO PERSONA CONDUCTOR IR-ALLÍ	1.wmv
CAJA	Decir la cantidad	Funcionario	ahora vaya a la ventanilla de conductores	AHORA TU VENTANILLA ESPECÍFICO PERSONA CONDUCTOR IR-ALLÍ	2.wmv
CAJA	Decir la cantidad	Funcionario	catorce con veinte	CATORCE COMA VEINTE EUROS	3.wmv
CAJA	Decir la cantidad	Funcionario	catorce euros con veinte céntimos	CATORCE COMA VEINTE EUROS	4.wmv

Fig. 2 Example of database content

Table 2 Main statistics of the corpus

Government employee sentences	Spanish	LSE
Sentence pairs (including repetitions)	1,641	
Different sentences (without repetitions)	1,413	199
Running words (Spanish) or signs (LSE) including repetitions	17,113	12,741
Vocabulary: words (Spanish) or signs (LSE) without repetitions	527	237

British Sign Language) with a similar meaning to the sign meaning. They are called glosses (i.e. 'BED' for the sign 'bed'). It is important to highlight that it is not always to select a gloss to represent a sign because signs are closer to a semantic concept than to an isolate word. In many cases, one sign must be described with several words in Spanish. In the last 20 years, several alternatives, based on specific characteristics of the signs, have appeared in the international community: HamNoSys [32], SEA (Sistema de Escritura Alfabética) [16] and SignWriting (http://www.signwriting.org/). HamNoSys and SignWriting require defining a specific picture font to be used by computers. SignWriting includes face features in the notation system but HamNoSys and SEA do not include them. All of these alternatives are flexible enough for dealing with different sign languages including LSE. However, in this work, glosses have been considered for writing signs because it is the most familiar and extended alternative according to the Spanish Deaf Association. These glosses include non-speech indicators (i.e. PAY or PAY? if the sign is localized at the end of an interrogative sentence) and finger spelling indicators (i.e. DL-PETER that must be represented letter by letter P-E-T-E-R).

4 Spanish into LSE translation architecture

This section presents the system description and the laboratory translation experiments, comparing the different translation strategies considered in this work.

Figure 3 shows the module diagram developed for translating spoken language into LSE. The first module, the speech recognizer, converts natural speech into a sequence of words (text). It uses both language and acoustic models for every allophone. The natural language translation module converts a word sequence into a sign sequence. For this module, the paper presents and combines three different strategies. The first consists of an example-based strategy: the translation process is carried out based on the similarity between the sentence to be translated and the examples of a parallel corpus (examples and their corresponding translations). The second is a rule-based translation strategy, where a set of translation rules (defined by an expert) guides the translation process. The last is based on a statistical-translation approach where parallel corpora are used for training language and translation models.

At the final step, the sign animation is made using VGuido: the eSIGN 3D avatar developed in the eSIGN project (http://www.sign-lang.uni-hamburg.de/esign/). VGuido has been incorporated as an ActiveX control. The sign descriptions have been generated using a new version of the eSIGN Editor, as it is described in Sect. 4.3.

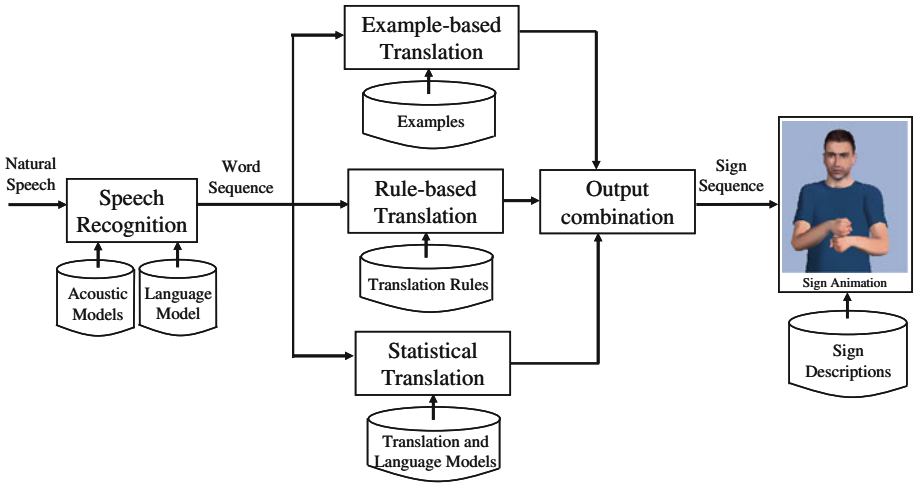

Fig. 3 Diagram of the Spanish into LSE translation module

4.1 Automatic speech recognition

The speech recognizer is a state-of-the-art speech recognition system developed at GTH-UPM (http://lorien.die.upm.es). It is a Hidden Markov Model (HMM)-based system able to recognize continuous speech: it recognizes utterances made up of several continuously spoken words. In this application, the size of the vocabulary was 653 Spanish words: the corpus vocabulary (with 527 words) was extended with a complete list of numbers (from 0 to 100), weekdays, months, etc. The recognizer has been trained by using more than 40 h of speech from the SpeechDat database [23]. This database includes speech from 4,000 people with a studied balance in age, gender, and geographical localization within the Iberian Peninsula. This aspect makes the recognizer robust against a great range of potential speakers without the need for further training (speaker independency). The system uses a front-end with perceptual linear predictive (PLP) [15] coefficients derived from a Mel-scale filter bank (MF-PLP). This front-end includes Cepstral mean normalization (CMN) and Cepstral variance normalization (CVN) techniques [18].

For Spanish, the speech recognizer uses a set of 45 phonemes. The system also has 16 silence and noise models for detecting acoustic sounds (non-speech events such as background noise, speaker artefacts, filled pauses, etc.) that appear in spontaneous speech. The system uses context-dependent continuous HMMs developed using decision-tree state clustering: 1,807 states and 7 mixture components per state. As regards the language model, the recognition module uses statistical language modelling: 2-gram, as the database is not large enough to estimate reliable 3-grams.

The recognizer provides one confidence value for each word recognized in the word sequence. The confidence measurement is a value between 0.0 (lowest confidence) and 1.0 (highest confidence) [12]. This value 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.

The acoustic models can be adapted to one speaker or to a specific acoustic environment using the maximum a posteriori (MAP) technique [13].

As regards the performance of the ASR module in laboratory tests, with vocabularies of less than 1,000 words, the word error rate (WER) is less than 5%. If this ASR module is adapted to a specific speaker, the WER drops to less than 2%.

4.2 Natural language translation

The natural language translation module converts the word sequence obtained from the speech recognizer into a sign sequence that is animated using the 3D avatar (every sign is represented by a gloss). Three different strategies have been implemented and evaluated for this module: example-based, rule-based and statistical translation.

4.2.1 Example-based strategy

Example-based translation is essentially translation by analogy. An example-based translation system uses a set of sentences in the source language and their corresponding translations in the target language, for translating other similar source-language sentences. In order to determine whether one example is equivalent (or at least, similar enough) to the sentence to be translated, the system computes a heuristic distance between them. By defining a threshold on this heuristic distance, it is possible to define how similar the example must be to the sentence to be translated, in order to consider that they generate the same

target sentence. If the distance is lower than a threshold, the translation output will be the same as the example translation. But if the distance is higher, the system cannot generate any output. Under these circumstances, it is necessary to consider other translation strategies. The heuristic distance considered in this work is a modification of the well-known Levenshtein distance (LD). The heuristic distance is the LD divided by the number of words in the sentence to be translated (this distance is represented as a percentage).

The LD [20] is a measurement of the similarity between two strings (or character sequences): source sequence (s) and target sequence (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t. Because of this, it is also called the edit distance: the greater the LD, the more different the strings. Originally, this distance was used to measure the similarity between two strings (character sequences). But it was already used for defining a distance between word sequences (as has been used in this paper). The LD is computed using a dynamic programing algorithm that aligns both word sequences considering different alignment costs: 0 for identical words, 1 for each insertion, 1 for each deletion and 1 for each substitution. The best alignment between both sequences will provide a distance counting the number of identical words, insertions, deletions and substitutions.

One problem of this distance is that two synonyms are considered as different words (a substitution in the LD) while the translation output can be the same. The system is currently being modified to use an improved distance: the substitution cost between two words (instead of being 1 for all cases) ranges from 0 to 1 depending on the translation behaviors of the two words. These behaviors are obtained from the lexical model computed in the statistical translation strategy (described in Sect. 4.2.3). For each word (in the source language), an N-dimension translation vector ($\hat{w}$) is obtained where the "i" component, $P_w(g_i)$, is the probability of translating the word "w" into the gloss "g_i". N is the total number of glosses (sign language) in the translation domain. The sum of all vector components must be 1: $\sum_{i=1}^{N} P_w(g_i) = 1$. The substitution cost between words "w" and "u" is given by the following equation (substitution cost based on the behaviour of the translation):

$$\text{Subs. Cost}(w, u) = \frac{1}{2}\sum_{i=1}^{N} abs(P_w(g_i) - P_u(g_i)) \quad (1)$$

When both words present the same behaviour (the same vectors), the probability subtraction tends towards 0. Otherwise, when there is no overlap between translations vectors, the sum of the probability subtractions (in absolute values) tends towards 2. Because of this, the 1/2 factor has been included to make the distance range from 0 to 1. This improved distance has been incorporated recently and it was not used in the field evaluation.

The biggest problem with an example-based translation system is that it needs large amounts of pre-translated text to make a reasonable translator. In order to make the examples more effective, it is possible to generalize them [5], so that more than one string can match the same example. Considering the following translation example for Spanish into LSE:

Spanish: "Veinte euros con diez céntimos" (Twenty Euros, ten).
LSE: "VEINTE COMA DIEZ EUROS".

Now, if it is known that "veinte" and "diez" are numbers, it is possible to save this example in the corpus as.

Spanish: "$NUMBER euros con $NUMBER céntimos".
LSE: "$NUMBER COMA $NUMBER EUROS".

where $NUMBER is a word class including all numbers. Notice how it is possible to match many other strings that have this pattern, they are not restricted to these numbers. When indexing the example corpora, and before matching a new input against the database, the system tags the input by searching for words and phrases included in the class lists, and replacing each occurrence with the appropriate token. There is a file which simply lists all the members of a class in a group, along with the corresponding translation for each token. For the system implemented, four classes were used: $NUMBER, $PROPER_NAME, $MONTH and $WEEK_DAY.

Figure 4 represents the translation process for the recognised sentence: "catorce euros veinte céntimos". The first step is to categorize the sentence by obtaining "$NUMBER euros $NUMBER céntimos". The closest example is selected and its translation is proposed. Finally, the categories in the example translation are replaced by the translation of the original words. In this case, numbers are translated directly by putting words in capital letters. If by mistake (a wrong example selection), there is a category in the selected example that does not appear in the input to translate. This category is replaced by a null string and the system will not generate any translated category.

This translation module generates one confidence value for the whole output sentence (sign sequence): a value between 0.0 (lowest confidence) and 1.0 (highest confidence). This confidence is computed as the average confidence of the recognized words (confidence values obtained from the speech recognizer) multiplied by the similarity between this word sequence and the example used for translation. This similarity is complementary of

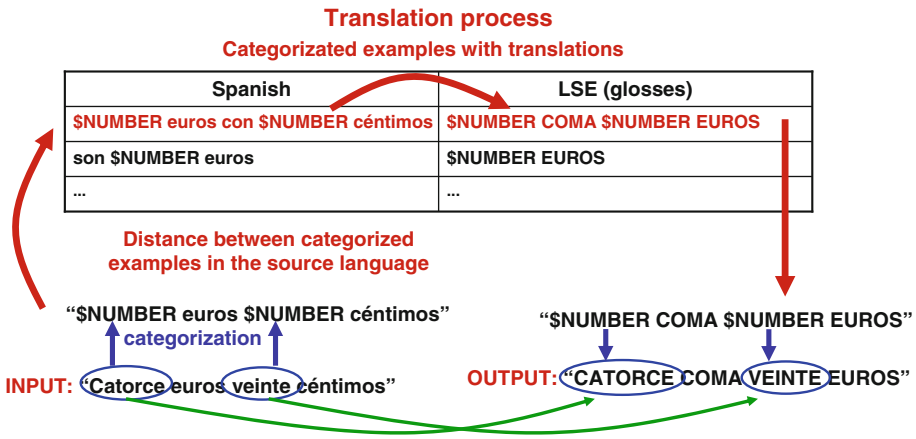

Fig. 4 Translation process in an example-based translation system

the heuristic distance: 1 minus heuristic distance. The confidence value will be used to decide whether the translation output (gloss sequence) is good enough to be presented to a Deaf person. Otherwise, the translation output is rejected and not represented by the avatar. In this case, the government employee must repeat the spoken sentence again.

4.2.2 Rule-based strategy

In this strategy, the translation process is carried out in two steps. In the first one, every word is mapped into one or several syntactic–pragmatic categories (categorisation). After that, the translation module applies different rules that convert the tagged words into signs by means of grouping concepts or signs (generally called blocks) and defining new signs. These rules are defined by experts and can define short and large-scope relationships between 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 (Fig. 5).

In this approach, the translation module and the rules have been implemented by considering a bottom–up strategy: the translation analysis is carried out by starting from each word individually and extending the analysis to neighboring context words or already-formed signs (blocks). This extension aims to find specific combinations

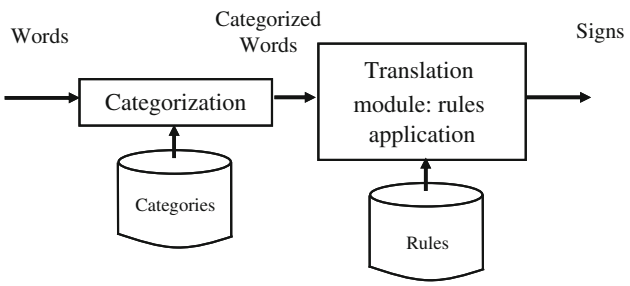

Fig. 5 Translation process in a rule-based strategy

of words and/or signs (blocks) that generate another sign. The rules implemented by the experts define these relationships. In this work, two expert linguists developed all of the rules, forcing an agreement on all of the rules incorporated into the system. If there was no agreement, the rule was not included. The main target was to reduce the linguistic subjectivity.

Depending on the scope of the block relationships defined by the rules, it is possible to reach different compromises between the reliability of the translated sign (greater with greater 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 implemented.

The rules are specified in a proprietary programing language consisting of a set of instructions. The rule-based translation module contains 293 translation rules and uses 10 different instructions. Similar to the example-based translator, this strategy generates one confidence value (between 0.0 and 1.0) but in this case for every sign. This sign confidence is computed by a procedure coded inside the proprietary language. Each instruction generates the confidence for the elements it produces. For example, in the case of instructions that check for the existence of a specific sign sequence and generate a new one, the instruction usually assigns the average confidence of the original sign sequence to the newly created element. In other more complex cases, the confidence for the new elements may be dependent on a combination of confidences from a mixture of words and/or internal or final signs.

4.2.3 Statistical translation

For a statistical translation, two methods have been evaluated: a Phrase-based Translator and a Stochastic Finite State Transducer (SFST). The phrase-based translation system is based on the software released from NAACL

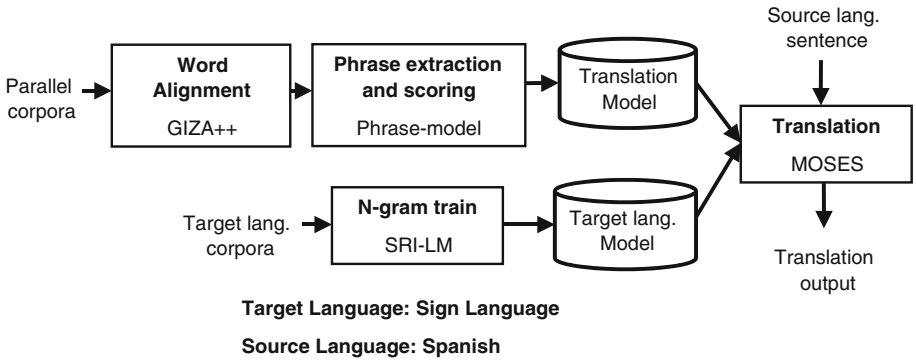

Fig. 6 Diagram of the phrase-based translation module

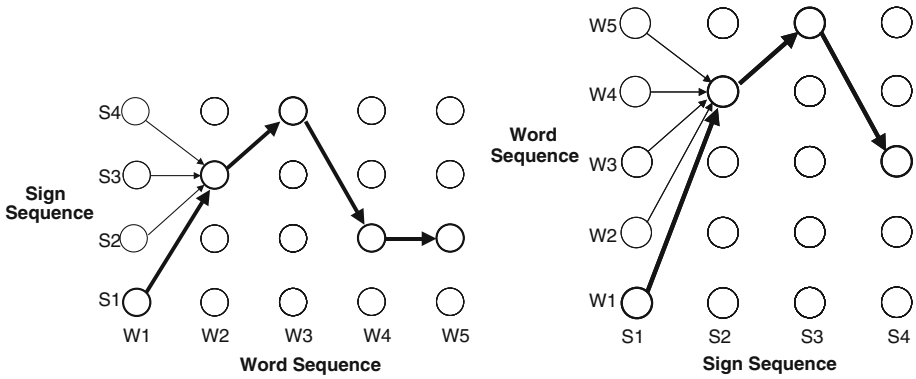

Fig. 7 Alignments in both directions: words signs and signs words

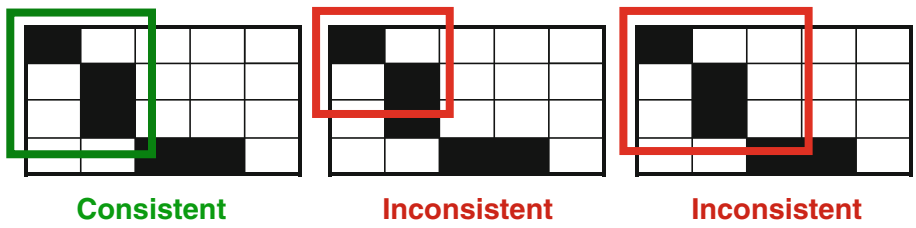

Fig. 8 Examples of phrase extraction

Workshops on Statistical Machine Translation (http://www.statmt.org) in 2008. The translation process uses a phrase-based translation model and a target language model. The phrase model has been trained in accordance with these steps (Fig. 6).

The first step is word alignment computation. In this step, the GIZA++ software [27] has been used to calculate the alignments between words and signs. In order to establish word-sign alignments, GIZA++ combines the alignments in both directions: words-signs and signs-words (Fig. 7). Because signs are close to semantic concepts, every sign is frequently aligned to several consecutive Spanish words (a subsequence of the Spanish sentence) instead of just one.

GIZA++ also generates a lexical translation model including the translation probability between every word and every sign. This lexical model is being used to improve the heuristic distance of the example-based translator (see Sect. 4.2.1).

The second step is phrase extraction [19]. All phrase pairs that are consistent with the word alignment are collected. For a phrase alignment to be consistent with the word alignment, all alignment points for rows and columns that are touched by the box have to be in the box, not outside (Fig. 8). The maximum size of a phrase has been fixed at 7 based on development experiments on the validation set (more details are presented in Sect. 4.2.5).

Finally, the last step is phrase scoring. In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward. To estimate the phrase translation probability φ(LSE|Spanish), the system sorts the extracted file. Sorting ensures that all Sign Language phrase translations for a Spanish phrase are next to each other in the file. Thus, it is possible to process the file, one Spanish phrase at a time, collect counts and compute φ(LSE|Spanish) for that Spanish phrase. To estimate φ(Spanish|LSE), the inverted

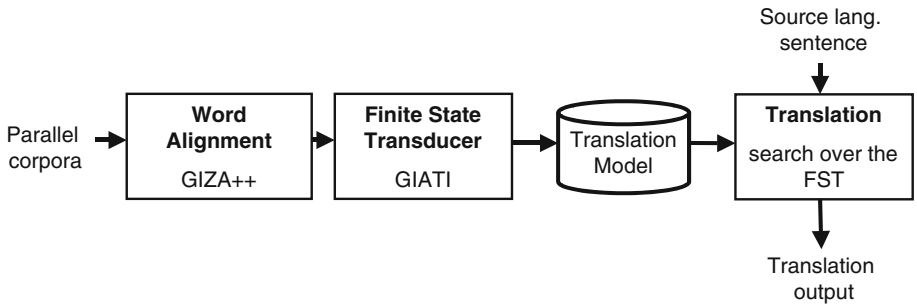

Fig. 9 Diagram of the SFST-based translation module

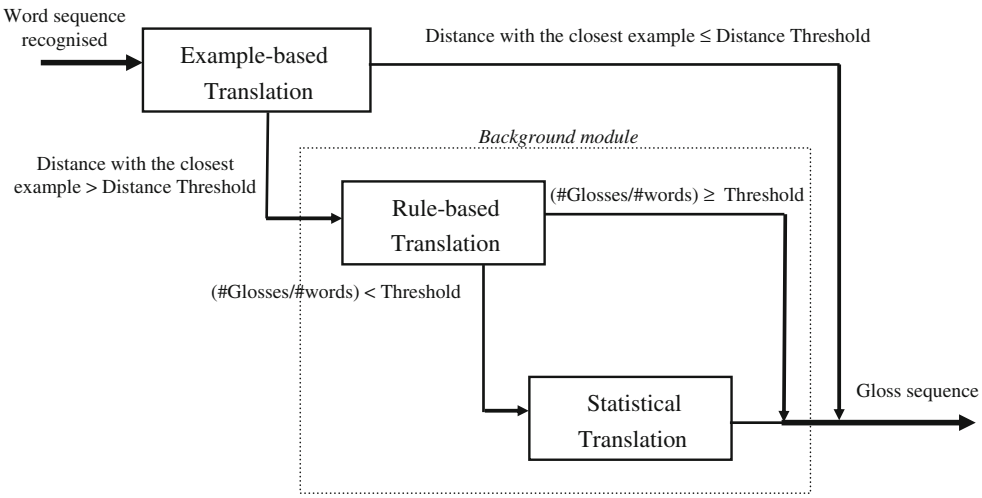

Fig. 10 Diagram of natural language translation module combining three different translation strategies

file is sorted, and then, φ(Spanish|LSE) is estimated for a sign language phrase at a time.

The Moses decoder (http://www.statmt.org/moses/) is used for the translation process. This program is a beam search decoder for phrase-based statistical machine translation models. In order to obtain a 3-gram language model needed by Moses, the SRI language modelling toolkit has been used [35].

The translation based on SFST is made as set out in Fig. 9.

The translation model consists of an SFST made up of aggregations: subsequences of aligned source and target words. The SFST is inferred from the word alignment (obtained with GIZA++, described previously) using the Grammatical Inference and Alignments for Transducer Inference (GIATI) algorithm [7]. The SFST probabilities are also trained from aligned corpora. The software used in this paper has been downloaded from http://prhlt.iti.es/content.php?page=software.php.

Both statistical translation strategies (phrase-based and SFSTs) generate an overall confidence value for the whole gloss sequence. This confidence is computed by following the same process. When a statistical module is not able to translate some words, these words are considered as proper names and they are passed directly to the output. The output sequence is made up of several tokens: signs as a result of translating several words, and other words passed directly to the output. In this domain, there were very few proper names in the corpus so, when the number of words passed directly to the output is high, this fact reveals a poor translating performance: the system cannot deal with some parts of the sentence. The confidence measurement proposed in this case is the portion of generated signs (not words passed directly to the output): # of signs generated/# of tokens in the output. This measurement performs very well as confidence measurements in restricted domain translation problems for detecting out of vocabulary sentences.

4.2.4 Combining translation strategies

The implemented natural language translation module combines the three translation strategies described in previous sections to build a hierarchical translation module that tries to obtain the main advantages from every strategy. This combination is described in Fig. 10.

The translation module has a hierarchical structure divided into two main steps. In the first step, an example-based strategy is used to translate the word sequence in order to look for the best possible match. If the distance with the closest example is lower than a threshold (distance threshold), the translation output is the same as the

Table 3 Result summary for example-based, rule-based and statistical approaches

	SER (%)	±Δ	PER (%)	BLEU	NIST
Statistical approach					
Phrase-based	38.23	0.86	36.67	0.5656	6.599
SFST-based	33.89	0.84	32.29	0.6534	7.789
Example-based approach	35.23	0.84	34.45	0.6012	7.354
Example-based approach (considering a heuristic distance <30%)	5.81	0.42	4.79	0.9112	9.452
Rule-based approach	21.55	0.72	17.14	0.6827	8.243
Combining translation strategies	7.98	0.47	6.75	0.9456	9.745

example translation. But if the distance is higher, a background module translates the word sequence. During the developing tests (see Sect. 4.2.5), the best results were obtained for a distance threshold (DT) ranging from between 20 and 30%. In the field evaluation, the DT was fixed at 30% (one difference is permitted in a 4-word sentence).

For the background module, a combination of rule-based and statistical translators has been used. As will be presented in Table 3, the rule-based strategy is the best alternative, but the statistical approach was also incorporated as a good alternative during rule-based system development. The main idea is that the time and effort required to develop a statistical translator (it was possible to obtain a tuned version in 1 or 2 days, including file format adaptation and tuning experiments) is considerably lower than a rule-based one (it took several weeks to develop all of the rules). In this project, the database collection (described in Sect. 3) required more time than initially foreseen, so there was the risk of not finishing the rule-based module in time for the field evaluation. As a result, during the rule development, a statistical translator was incorporated in order to have a background module with a reasonable performance.

The relationship between these two modules has been implemented based on the ratio between the number of glosses (generated after the translations process) and the number of words in the input sequence. If the #glosses/#words ratio is higher than a threshold, the output is the gloss sequence proposed by the rule-based module. On the other hand, if this condition is false, the statistical approach is carried out. By analysing the parallel corpus, the ratio between number of glosses and number of words is 0.74. When the number of glosses generated by the rule-based approach is very low, it means that specific rules for dealing with this type of example has not yet been implemented (or the sentence is out of the domain). During the rule-based system development, the gloss/word ratio mechanism was used to direct (in some cases) the translation process to the statistical approach. The ratio threshold was fixed to 0.5 in order to add a margin of error (from the average value 0.74). As regards the statistical module, both alternatives were incorporated (phrase-based and SFST-based strategies), although only the SFST-based alternative was used for the field evaluation because of its better performance (see the next section for more details on translation tests).

Finally, it was possible to finish the rule-based translation module before the field evaluation. Although statistical approaches performed worse than the rule-based approach, they have been kept in a hybrid background module in order to facilitate the system scalability and adaptation to other domains. This aspect will be discussed in Sect. 4.5.

4.2.5 Translation results and discussion

In order to evaluate the different translation approaches, the corpus (including only sentences pronounced by government employees: Table 2) was divided randomly into three sets: training (75% of the sentences), development (12.5% of the sentences) and test (12.5% of the sentences), carrying out a cross-validation process. Table 3 summarizes the results for example-based, rule-based and statistical approaches considering several performance metrics: sign error rate (SER) is the percentage of wrong signs in the translation output compared to the reference in the same order. Position independent SER (PER) is the percentage of wrong signs in the translation output as compared to the reference without considering the order. BiLingual evaluation understudy (BLEU) Papineni [30] is an algorithm for evaluating the quality of an automatic translation. The main task is to compare n-grams (sequences of n signs) of the translation output with the n-grams of the reference translation and count the number of matches. These matches are position independent. The more the matches, the better the candidate translation is. BLEU was one of the first metrics to achieve a high correlation with human judgements of quality. BLEU's output is always a number between 0 and 1. This value indicates how similar the candidate and reference sentences are;

values closer to 1 represent more similar sentences. Finally, NIST (http://www.nist.gov/speech/tests/mt/) is very similar to BLEU but using a different method for estimating some weights used in this algorithm. The better the translation, the higher the NIST score. It is important to underline that SER and PER are error metrics (a lower value means a better result) while BLEU and NIST are accuracy metrics (a higher value means a better result).

For every SER result, the confidence interval (at 95%) is also presented. This interval is calculated using the following formula (confidence interval at 95%):

$$\pm\Delta = 1.96\sqrt{\frac{\mathrm{SER}\,(100-\mathrm{SER})}{n}} \tag{2}$$

n is the number of signs used in testing, in this case $n = 12{,}741$. An improvement between two systems is statistically significant when there is no overlap between the confidence intervals of both systems. As shown in Table 3, all improvements between different approaches are higher than the confidence intervals.

As shown in Table 3, the rule-based system obtains better results than example-based and statistical methods probably because the rules defined by experts (based on their wide experience on both languages) introduce translation knowledge (including general translation rules) not seen in the parallel corpus, making the system more robust against new sentences (not considered in the original parallel corpus). For this corpus, the SFST-based and example-based methods are better than the phrase-based method. One important difference between rule-based and statistical approaches is related to the number of insertions and substitutions generated in the gloss sequence. In the case of a rule-based system, these numbers are lower compared to a statistical method. The reason is because most of the rules look for a specific word sequence to generate a gloss sequence: if this sequence does not appear, the gloss sequence is not generated, thus increasing the number of deletions.

The scores obtained with every independent approach are better than those reported in previous similar works (see Table 1). The main reason for this performance is due to working in a restricted domain. In this case, even having a very small amount of data, statistical methods perform reasonable well because this data represents the restricted domain very well. Although the scores are better than previous works, having an SER greater than 20% seems very high for using the system in a field evaluation. In order to improve these results, the authors considered combining the different translation strategies to build a hierarchical translation module. With this target in mind, Table 3 also presents the translation results for the example-based approach for those sentences that have a heuristic distance (with the closest example) lower than 30% (the rest of the sentences were not translated). In this case, the results increase significantly: SER improvement is greater than the confidence intervals (at 95%). Finally, Table 3 presents the results for the combination of several translation strategies: example-based (considering a heuristic distance <30%), rule-based and SFST-based approaches. As is shown, with the hierarchical system it is possible to obtain better results by translating all the test sentences: SER <10%. This module has been used in the field evaluation presented in Sect. 6: statistical models have been trained using the whole database.

4.3 Sign animation with the eSIGN Avatar

The signs are represented by means of VGuido (the eSIGN 3D avatar [41] 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 those characteristics which define additional distortions to be applied to the avatar.

The sign database has been generated using a new version [34] of the eSIGN Editor [14]. The eSIGN Editor was developed in the VISICAST and eSIGN European Projects (Essential Sign Language Information on Government Networks). In a previous work [34], this editor has been adapted to LSE. The new version incorporates the same functionality for defining manual movements (using HamNoSys and SEA) and non-manual aspects such as movements of lips, head, etc. This new editor has three windows (Fig. 11). In the main window, the eSign avatar shows the sign that is currently being designed (using an SEA or a HamNoSys specification). The second window allows HamNoSys characters to be introduced, and the last one permits non-manual gestures to be added (lip movements, facial expressions and body movements). The SEA characters can be introduced using the PC keyboard together with auxiliary buttons.

This new version incorporates a Spanish grapheme to phoneme that, given a Spanish sentence, generates a sequence of phonemes which are represented using Speech Assessment Method Phonetic Alphabet (SAMPA) [37]. This sequence is necessary to make the avatar move the lips according to this pronunciation (see [34] for more details).

Figure 11 shows the process for specifying a sign: defining the manual part (using HamNoSys or SEA), adding non-manual characteristics (lips, head and facial movements) and generating a script in the Sign Gesture Markup Language (SiGML) (XML file). This script is interpreted by VGuido for representing the sign.

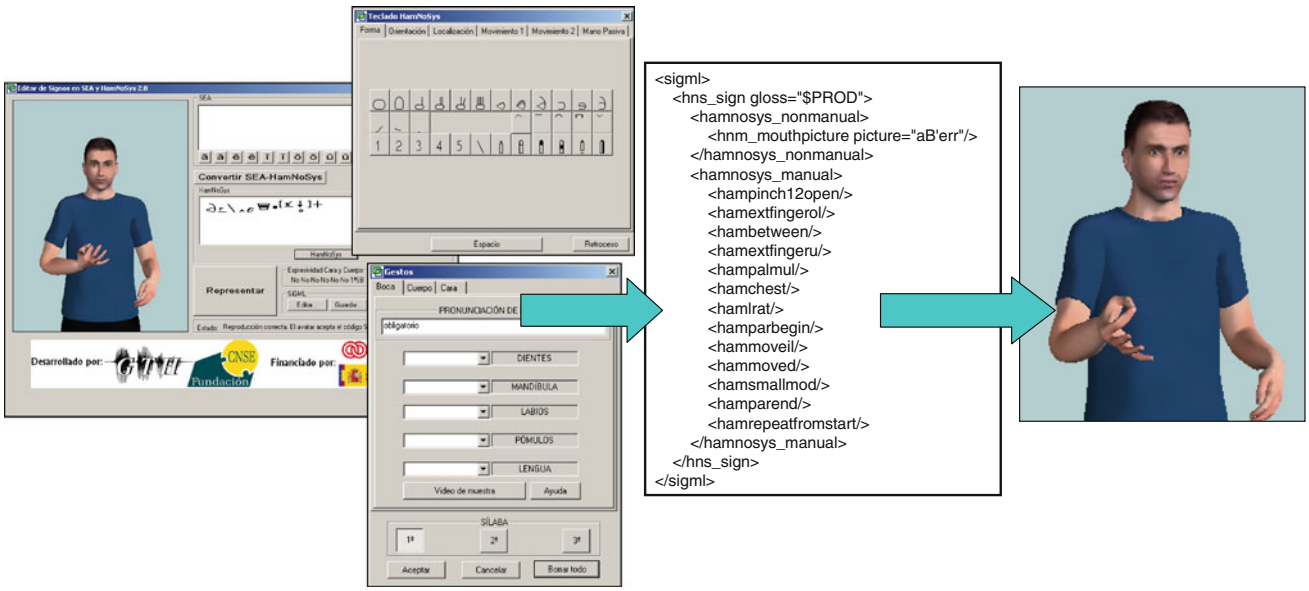

Fig. 11 Process for generating signs with the avatar

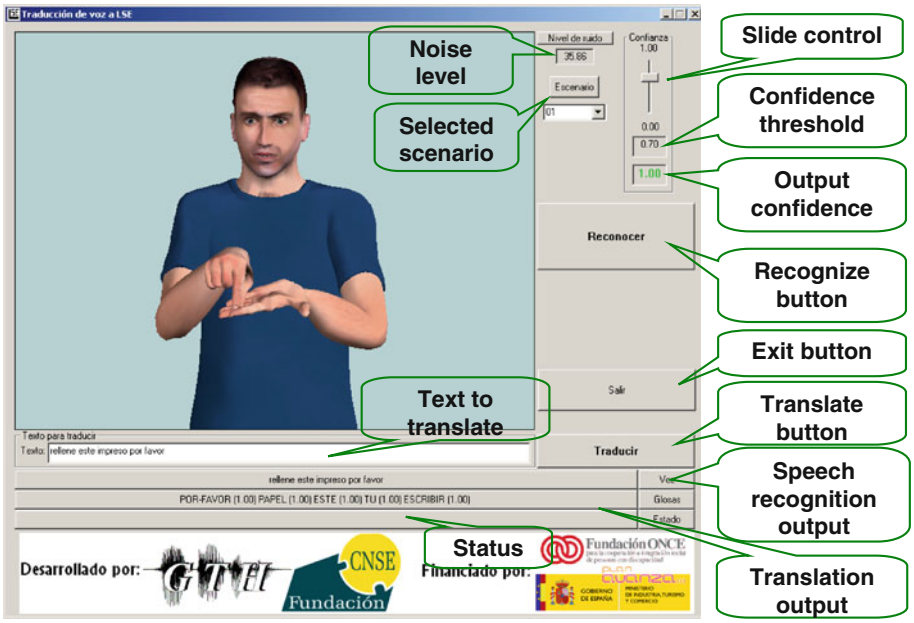

Fig. 12 Visual interface of the Spanish into LSE translation module

4.4 System interface

The module for translating spoken Spanish into LSE has a visual interface shown in Fig. 12.

This interface includes a *slide control* (in the right-top corner) to define the *confidence threshold* of the translation output (sign sequence) to represent the signs. If the translation output does not have enough confidence (the *output confidence* is lower than *confidence threshold*), the sign sequence is not represented. The system uses the whole sign sequence confidence because only the rule-based translation module can generate a confidence value for each sign: example-based and statistical translation modules generate an overall confidence value for the whole sign sequence.

When the government employee wants to speak, the "Reconocer" (*Recognise*) *button* must be pressed (above the *exit button*). It is also possible to give this order by pressing the ENTER key on the keyboard). The system starts listening in order to detect automatically when the government employee starts and finishes speaking. When the end of the speech has been detected the speech recognition sends its output to the translation module. The speech recognition and translation outputs are presented in

windows at the bottom (above the *status* window) only for debugging purposes (speech recognition output and translation output).

The interface also allows a word sentence written in one of the controls ("Texto para traducir" *text to translate*) to be translated by pressing the "traducir" (*translate*) *button*. This possibility was implemented as an alternative to introducing the word sequence if the speech recognizer had problems. After speech recognition, the recognized output is also copied into the "texto para traducir" (text to translate) control. This is very useful when the Deaf user asks for a repetition. In this case, the government employee has to speak again. If the previous recognition was OK, the system will generate the same sign sequence by pressing the "traducir" (*translate*) *button*.

When encountering problems with the environment noise level, the interface allows this *noise level* to be estimated by pressing one button. In the top right-hand corner, there are two controls: a button for estimating the noise level and a text window with the noise level in decibels.

Above the noise level controls, the interface has a list box for selecting the scenario that is being tested at that moment. As will be shown in the evaluation section, six different scenarios have been considered for simulating the six most frequent situations when renewing the driving license. The scenario information is used exclusively for logging: the system does not change its behaviour depending on the scenario. When logging objective measurements for evaluation (see Sect. 6), the scenario information is used to allow a detailed analysis depending on the scenario.

Finally, it is necessary to highlight that the system incorporates two functions to allow the Tablet PC screen to be oriented to the Deaf user: the system feeds back the recognized sentence (with speech synthesis) and generates a beep when the system has finished signing (and is ready for a new turn). The government employee cannot see the Tablet PC screen properly because it is oriented to the Deaf user, so she/he needs some feedback for reporting whether the speech recogniser has committed any error (although errors are very uncommon because the ASR has a good performance with a WER of less than 5%) and a warning when the avatar finishes signing (it is possible to speak again).

4.5 System limitations and scalability

When developing natural language interfaces, a significant amount of resources is required to model task knowledge properly. In this work, the main modelling requirement is speech recognition and language translation. When using example-based and statistical approaches for automatic language translation, it is necessary to have a large parallel corpus including a significant amount of sentences in source and target languages. When considering rule-based approaches, expert interpreters have to spend a lot of time defining the rules of the system. LSE has a tiny fraction of the resources that are available for English or even Spanish.

This section describes the main aspects that must be considered when increasing the scope of the system (being able to translate more sentences) or when applying this system to another domain or language. The necessary changes affect all the modules that make up the system: speech recognition, language translation and sign representation.

For the speech recognizer, it is necessary to update the vocabulary with all words (to be recognised) and the language model by considering the sentences spoken by government employee when providing the service.

In the case of the language translation module, it is necessary to update the following components: for the example-based translation module, the examples of the database. These examples consist of spoken utterances and their corresponding translation (a parallel corpus). The rule-based translation module would need to develop new rules for translating new sentences. This is a time-consuming task because several experts must develop the rules by hand. Some of these rules (approximately 40%) are general translation rules and can be used in other domains, but there are a lot of them specific to this domain. In order to give an idea, the rules used in the proposed system were developed by two people over a period of 4 weeks. For the statistical translation, it is necessary to update the translation models: these models are obtained automatically from a parallel corpus (Sect. 3).

For the sign representation module, it is necessary to update the list of glosses. For adding a new sign, the system needs to have a new file in a specific path, named with the sign gloss: i.e. DRIVER.txt. This file contains the sign description in SiGML, which must be represented by the avatar. When a new file is detected in the path, the interface updates the list of glosses with a new gloss (file name) and the avatar can represent it. Sign specification is the most time-consuming because every sign file must be generated by hand (Sect. 4.3). For example, one person working for 1 month was necessary to generate 715 signs. This task requires a great effort but it has significant advantages as compared to the alternative of recording all possible sign sentences using videos. The first advantage is that when using an avatar, it is only necessary to define every sign independently and the system can generate any sentence containing these signs by concatenating all the signs and automatically providing smooth transitions between signs. In the case of video recordings, it would be necessary to

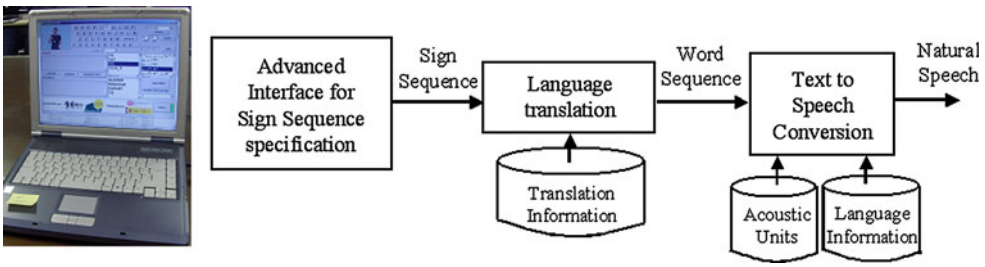

Fig. 13 Diagram module for the Spoken Spanish generation system [34]

record all possible sentences in the specific domain, in order to guaranty smooth transitions between consecutive signs. Another advantage comes about when it is necessary to add a new sign: in the case of using an avatar it takes around 15 min, while considering video recordings, it would be necessary to contact the same person who recorded the previous videos, and to record all possible sentences including this new sign.

It is possible to conclude that many aspects can be updated automatically from a parallel corpus, including sentences (in Spanish and LSE) related to the domain: except SiGML sign specifications and translation rules. For the case of translation rules, the authors are considering working on the statistical translation module in order to increase its performance. The statistical translator and the rule-based translator are combined in a hybrid background translation module. If it is possible to achieve the same performance with both systems, the rule-based translation could be removed from this structure, avoiding the need to generate new rules when adapting the system to a new domain.

5 Spoken Spanish generation from gloss sequences

In order to allow Deaf people to ask questions to the government officer, a spoken Spanish generation system has been used [34]. With this system, the deaf person can specify a sign sequence (gloss sequence) with a set of visual tools. It is not necessary to type any gloss: the gloss sequence can be easily specified by clicking on the screen. This sign sequence is automatically translated into Spanish words that will be spoken by pressing the speak button. This system is based on three module architecture (Fig. 13). The first module consists of an advanced visual interface for sign sequence specification. This interface includes several tools for sign specification: avatar for sign representation (to verify that sign corresponds to the gloss), prediction mechanisms, calendar and clock for date or time definitions, two sign-search methods based on glosses or based on HamNoSys features, a list with the most frequent sign sentences (such as greetings) and the possibility to introduce proper names by spelling a gloss. Secondly, a natural language translation module converts a sign sequence into a word sequence. Finally, the word sequence is converted into spoken Spanish using a commercial text to speech converter (Loquendo) [4].

Because sign recognition technology is not mature enough, this paper describes an advanced interface where a Deaf person can specify a sign sequence (gloss sequence) with a set of visual tools. This sequence is translated into words that will be spoken. This solution allows the Deaf person to specify a sentence in their first language (LSE) and avoid errors from sign recognition. As is shown in the summative evaluation, this interface has been well assessed by the users.

6 Field evaluation and discussion

This section includes a detailed description of the field evaluation carried out in the Local Traffic Office in Toledo. The advanced communication system was used for renewing the Driver's Licence. Both government employees and Deaf people were involved in the evaluation, which includes objective measurements from the system and subjective information from user questionnaires.

6.1 Method

The Driver's Licence (DL) renewing process at the Toledo Traffic Office consists of three steps: obtaining the form, payment, and handing over of the documents. Following an idea suggested from the head of the Toledo Traffic Office, instead of installing three systems at the three windows involved in the process (see Sect. 3 for more details), one new assistance position (Fig. 14) was created where a Deaf person can carry out all three steps to save resources.

The evaluation was carried out over 2 days. On the first day, the assistance position was installed and a 1-h talk about the project and the evaluation process was given to both the government employees and the Deaf users involved in the evaluation. This talk consisted of an overall project presentation describing the main scientific objectives and involved partners (10 min), a detailed explanation of the evaluation process including the different simulated scenarios and the questionnaire presentation (20 min), a demonstration of the system (5 min), some time for

Fig. 14 Assistance position preparation and speech recognizer adaptation

questions (10 min), and finally, there was a short period to practice (15–20 min). The government employee practiced with the speech into sign language translation system while the Deaf users practiced with the spoken language generation from gloss sequences. The evaluation process was carried out with one Deaf user after another; so the first Deaf user only had 15–20 min to practice. The remaining Deaf people had the opportunity to practice later (with an additional PC) while the evaluation process was being carried out. The 1-h talk was given in a meeting room but the evaluation was carried out at a different desk where the government employee and the Deaf user interacted without the help of any interpreter, but under the supervision of one researcher. This researcher was an expert in LSE and was collecting comments from Deaf users.

Half of the users evaluated the system on the first day, leaving the other half for the next day. On the first day, the speech recognizer was adapted to the two government employees involved in the evaluation. For this adaptation, 50 sentences spoken by every government employee (1–2 s) were recorded.

For the evaluation, the Deaf users were asked to interact with government employees using the system developed for renewing the DL. Six different scenarios were defined in order to simulate the most frequent real situations: in one scenario, the Deaf user simulated having all the necessary documents, three other scenarios in which the Deaf user simulated not having one of the documents: Identification Card, a photo or the medical certificate, one scenario where the Deaf user had to fill in some information in the application form, and finally, a scenario where the Deaf user wanted to pay with credit card but it is not allowed, it must be in cash.

The system was evaluated by 10 Deaf users who interacted with 2 government employees at the Toledo Traffic Office using the developed system. These 10 people (six males and four females) tested the system in almost all of the previously described scenarios, generating 48 dialogs between government employees and Deaf users: 12 dialogues were missing because several Deaf people had to leave the evaluation session before completing all of the scenarios. The ages of the Deaf users ranged from between 22 and 55 with the average being 40.9. All of the Deaf users use LSE as the primary communication language. All of the Deaf users said that they used a computer every day (8 Deaf users) or every week (2 Deaf users), and only half of them (5 Deaf users) had a medium–high understanding level of written Spanish and the other half had a low or very low level of understanding Spanish. As regards their experience of renewing their DL, eight Deaf users were drivers and six had renewed their license at least once. All of the Deaf users had experience of interacting with government employees (with the help of an interpreter) in similar services (Fig. 15).

Fig. 15 Different photos of the evaluation process at the Toledo Traffic Office

Table 4 Objective measurements for evaluating the Spanish into LSE translation system

Agent	Measurement	Value
System	Word error rate	4.8%
	Sign error rate (after translation)	8.9%
	Average recognition time per sentence	3.3 s
	Average translation time per sentence	0.0013 s
	Average signing time	4.7 s
	% of cases using example-based translation	94.9%
	% of cases using rule-based translation	4.2%
	% of cases using statistical translation	0.8%
	% of turns translating from speech recognition	92.4%
	% of turns translating from text	0%
	% of turns translating from text for repetition	7.6%
	# of government employee turns per dialogue	8.4
	# of dialogues	48
	Average time for each dialogue	7.34 min

6.2 Results and discussion

The evaluation results include objective measurements from the system and subjective information from both Deaf user and government employee questionnaires. A summary of the objective measurements obtained from the system are shown in Table 4.

The WER for the speech recognizer is 4.8%, higher than the results obtained in laboratory tests for cases in which the speech recognizer was adapted to one speaker: 2%. This WER was small enough to guarantee a low SER in the translation output: 8.9%. The time needed for translating speech into LSE (speech recognition + translation + signing) is around 8 s per sentence. This time allows for a reasonably agile dialogue between government employees and Deaf people. Table 5 presents an analysis of the translation errors (8.9% in total) including an error classification, main causes and impact on the system.

As regards the different translation strategies, the example-based translation has been used in more than 94% of the cases showing the reliability of the linguistic study carried out (corpus collection). In this study, the most frequent sentences were recorded, obtaining a representative corpus in this kind of dialogue. Some of the sentences translated using the rule-based or the statistical-translating modules (they were not similar enough to one of the examples in the corpus) were sentences spoken as a result of the Deaf person having to go to different windows: all the renewing process was carried out at the same window instead of several.

Almost all government employee turns included speech recognition. Only for some repetitions (7.6% of turns), the system translated a text sentence (without using speech recognition) but using the speech recognition output from the previous turn, not editing a new sentence. This result shows that the speech recogniser is working well enough to be the main means of interaction. A 7.6% rate of repetition turns is very good, given that in spoken conversations this rate is around 5% [38]. Considering 8.4 government employee turns per dialogue, a 7.6% rate of repetitions means 2 repetitions every 3 dialogues.

As regards the task performance, it is important to highlight that Deaf users completed the task in all dialogues. Deaf users followed all of the necessary steps for renewing the DL except in those scenarios where the Deaf user simulated not having one of the documents. In these cases, Deaf users got enough information to obtain the necessary document (Identification Card, a photo or the medical

Table 5 Analysis of the errors generated by the translation system

Error description	Percentage	Main causes	Impact
Changes in the sentence structure and substitutions	4.5	Problems in the gloss sentence structure are mainly due to errors in the translation technology, when dealing with sentence structures not seen in the collected corpus	In these cases, the impact is the worst. The Deaf user does not understand anything and the government employee must repeat the information in a different way
Insertions	2.1	These two kinds of errors have their main cause in speech recognition errors: insertions and deletions. Deletions are more frequent when the government employee lowers her/his voice, and they appear at the end of the sentence. Insertions appear when the government employee introduces additional noises into the speech (coughs, breathing, filled pauses "ehmm"). They appear more frequently at the beginning of the sentence	Insertions have a negative impact. Sometimes, the Deaf user understood the Sign Language sentence but in many cases (>70%) the government employee had to repeat it
Deletions	2.3		This is the error with the lowest impact. In many cases (> 80%), the Deaf user understood the overall meaning without the need for repetition

certificate). The main problem was that the required time was 7.34 min per dialogue, while a normal interaction with the help of an interpreter takes around 2 min. Clearly, the proposed system cannot compete against human interpreters but it provides an interesting alternative when a human interpreter is not available. In future work, the authors are considering several strategies for trying to reduce this time. The first idea is to improve the performance of the speech recognition and language translation modules and to increase the naturalness of the avatar so as to reduce the number of times the user asks for a repetition. The second idea consists of trying to reduce the translation time by means of modifying the speech recognition system to report partial recognition results every 100 ms for example. 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. But this idea has to be analysed further because there is a problem related to the fact that the translation is not a linear alignment process between spoken words and signs.

The subjective measurements were collected from questionnaires filled in by both government employees and Deaf users. They evaluated different aspects of the system, scoring them between 1 and 6. The questionnaires were designed by a group of experts made up of one Spanish linguist, two Deaf LSE experts and a Spanish linguist expert in LSE. There are important issues to deal with when designing a questionnaire for Deaf people. The first issue is the language: LSE (using videos) or written Spanish. In this case, the decision was to present the questions in Spanish with translation in LSE (glosses) and having two interpreters for solving any questions. Using interpreters to help to fill in questionnaires has a confidentiality problem. The authors think that this confidentiality problem, added to the fact that Deaf users have problems in writing in Spanish, meant that only one of the Deaf users included subjective comments in the questionnaire: this comment was positive "Good job, congratulations".

Secondly, it was necessary to decide on the aspects to be evaluated and the question design. The first idea was to reuse questionnaires developed for evaluating Speech-based applications [26]. Immediately, experts in LSE (Deaf) reported the problem that in these questionnaires there are concepts and words that have no translation into LSE, so many of these concepts would be difficult for the Deaf to understand. (i.e. questionnaire items that are difficult to translate: *I thought there was too much inconsistency in this system* or *I found the various functions in this system were well integrated*). Because of this aspect, the group of experts decided to reduce the number of questions, designing them based on tangible aspects (easier to explain with examples).

Another important issue when designing a questionnaire is the scale: number of levels and the names for the different levels. For the number of levels, the expert panel decided to define an even number (six in this case) eliminating the neutral level and forcing the user to decide. One reason is that this neutral level is the most common refuge when a user does not understand one of the questions very well. Forcing a user to decide causes this user to ask interpreters more questions to understand all the details. A second reason was that it is very difficult to find Deaf users for evaluating this kind of system and the authors wanted to obtain the best feedback with a small number of users. As regards the label for the different levels, the final decision was to specify six numerical levels providing information for levels 1 and 6 (strongly disagree, strongly agree). Defining labels for all the levels is a difficult problem because the differences between consecutive levels cannot always be described properly using LSE. There is a probability that the nuances were not perceived by a Deaf person, while a numerical scale is easier to understand.

Table 6 Subjective measurements for evaluating the Spanish into LSE translation system

AGENT	MEASUREMENT	Mean (1-6)	Standard Deviation
Government employee	The system is fast	5.0	0.0
	The speech recognition rate is good	4.5	0.7
	The system is easy to use	4.5	0.7
	The system is easy to learn	4.5	0.7
	I would use the system in the absence of a human interpreter	4.5	0.7
	Overall assessment	**4.5**	**0.7**
Deaf user	The signs are correct	3.1	1.2
	I understand the sign sequence	3.2	1.2
	The signing is natural	1.8	0.9
	I would use the system in the absence of a human interpreter	3.0	1.9
	Overall assessment	**3.2**	**1.1**

The measurement column presents the questions presented to the government employee and Deaf users

The average results for each aspect are presented in Table 6. In this table, "measurement" column presents the questions included in the questionnaires.

The evaluation from the government employees is quite positive giving a 4.5 score for all aspects considered. Perhaps the main problem reported by the government employees was that it was very uncomfortable to have the screen of the Tablet PC turned towards the user (see Fig. 16). It is true that the system feeds back the recognized sentence (with speech synthesis) and generates a beep when the system has finished signing (and it is ready for a new turn), but two screens will be considered for the future.

The user assessment was very low (an overall score of 3.2). The worst score was to the naturalness of the sign (1.8). Although the objective measurements were very good (with very good recognition and translation rates) the user did not like the sign language. A significant problem is that the naturalness of the avatar is not comparable to human sign language. It is necessary to keep investing a greater effort in increasing flexibility, expressiveness and naturalness of the avatar, especially as regards the face (in this work, much effort was invested in designing the non-manual sign characteristics but it is necessary to keep working on this aspect). But it is also fair to report that there were discrepancies between Deaf people as to the correctness of some signs (i.e. the "FOTO" (photo) sign, it is represented by moving the index finger from both hands or only from the right hand) or the specific sign used (i.e. using the "FECHA" (date) sign instead of "DÍA" (day) sign). These discrepancies are solved in the real LSE conversations with a facial expression (i.e. pronouncing a word). In spite of the effort invested in this work, this aspect must be improved in the avatar. The sign specification was made based on the dictionary generated by Fundación CNSE, DILSE III. These discrepancies showed the need to keep working on the documentation process of the LSE. LSE is a young language with many variations in the different regions of Spain. Fundación CNSE (Confederación de Personas Sordas) is the national confederation including all local associations; FCNSE is making a significant effort to collect and document all of these variations. With this documentation, a Deaf user can learn these variations improving the communication between Deaf people coming from different regions in Spain. In the future, if LSE is included in TV subtitles, TV could reduce these discrepancies as has happened to other minority languages in Spain.

Another source of discrepancy is the structure of some sign sentences. LSE, as in other languages, offers a high level of flexibility. This flexibility is sometimes not well understood and some of the possibilities are considered as wrong sentences. Some examples are presented in Table 7:

Fig. 16 Government employee speaking to the user with the screen of the Tablet PC turned towards the Deaf user. The Deaf user is interacting to the spoken Spanish generator from gloss sequences

Table 7 Examples of discrepancy in sentence structure

For the question "¿qué desea?" (What do you want?), the translation can be "QUERER QUÉ?" or "TU QUERER?" The system used the first one but some users preferred the second one
Regarding the sign "CAJERO" (cash machine), some of the users think that it must go with the sign "DINERO" (money) or "BANCO" (bank) in order to complement the meaning
Using "FOTO FLASH" for a photo machine box instead of "CABINA" (photo booth)
For the sentence "DNI CARNET CONDUCIR LOS-DOS DAR-A_MI" there was a problem with the meaning of the sign "LOS-DOS": it is not always clear if it is referring to "DNI" (identification card) and "CARNET CONDUCIR" (driver's licence)

In order to deal better with this flexibility in the future, the authors consider changing the behaviour of the system for dealing with the repetitions: when the user asks for repetition, instead of providing the same sentence in LSE, the system should try to generate an alternative sign sentence (with the same meaning) that it could be better understood by users.

Another problem observed is that the avatar represents signs in a very rigid way, making the representation angle important for perceiving some aspects of the signs. For example for the sign "VENIR" (to come), the avatar performs a right hand movement with two displacements: one vertical and one towards the person carrying out the sign language. If the avatar is perfectly oriented to the user, the movement towards the person carrying out the sign language is not perceived properly. In order to solve this problem, the avatar was slightly turned to see the movement in all significant directions. An interesting strategy for reducing this rigidity would be to define several specifications of the same sign with slight differences between them. These differences could represent, for example, variations on linguistic aspects like more or less emphasis depending on if the sign is the main focus of the sentence. This rigidity could be reduced by introducing this useful variability.

Finally, there is a set of signs (déictique signs) that refer to a person, thing or place situated in a specific location. Their representation depends on where the person is, thing or place they are referring to. For example, "esta ventanilla" (this window) is translated into "ESTE VENTANILLA" (this window). The ESTE (this) sign is represented in a different way depending on the window location. In order to avoid this kind of sign language problem, and considering the possibility of using the system in several offices with different distributions, it is necessary to substitute these signs with more specific ones: "VENTANILLA ESPECIFICO CONDUCTOR" (in English: WINDOW SPECIFIC DRIVER). This substitution must be made during the database collection.

Although the reported comments influenced the perception of the sign language the most, the recognition and translation rates can have also a relevant influence on the quality of the system as perceived by users. When the system generates a wrong sentence structure or introduces a wrong sign into the sign language sequence (there is an insertion or a substitution in the translation output), the consequence is very detrimental: the user stops paying attention and asks the meaning of this sign, missing the rest of the signs. If the system deletes one sign by mistake, the user can occasionally understand the sentence meaning. In many cases, the first impulse is to ask the human interpreter. But when they understand that the system evaluation consists of not asking the human interpreter, they try one of these strategies: to ask the government employee for repetition and/or to read the speech recognition output or the gloss sequence. As regard repetitions, there were two repetitions every three dialogues, but it is more complicated to evaluate the number of times the user tries to read the text shown in the system interface. Based on notes taken during the evaluation, it is possible to estimate a rate of one time per dialogue approximately.

As regard the system interface, it is important to highlight several aspects observed during the evaluation. The first is that nobody modified the confidence level: neither the government employee nor the Deaf user. They maintained the default confidence level. The same for the noise level controls. The authors consider that these tools would be interesting if the system were introduced continuously (not just during the field evaluation); the government employee could learn to use the system in a more optimum way. The government employee could realize if the noise level is very high as compared to other days, affecting the speech recognition results drastically. Another example is the adaptation of the confidence threshold based on his or her experience as to how often the system rejects a sentence or not depending on this confidence level. Additionally, the government employee could learn what Spanish sentences are better recognized and translated by the system in order to be used during his/her explanations. In any case, these tools are not useful for Deaf users because they use the system just for a few minutes.

As regards the window controls with the speech recognition output and the gloss sequence, Deaf users tried to read them only when they did not understand signing or when they missed some signs from the avatar signing (signing is volatile while written sentences are permanent). In these cases, Deaf users complained about the small size of these windows. The authors realized that these windows can be useful as a permanent backup if Deaf users do not understand avatar signing or miss some signs. But window size must be increased and the confidence number for each gloss must be removed (these numbers are confusing).

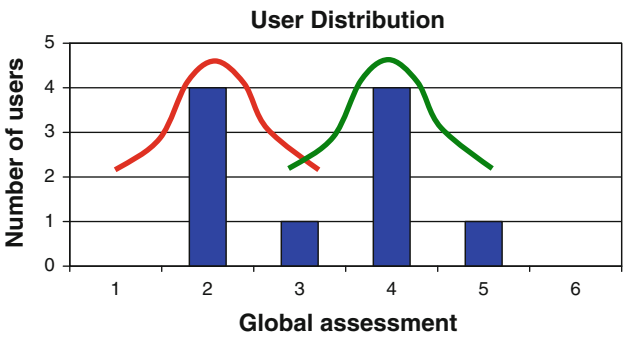

Fig. 17 Distribution of users versus global assessment

Table 8 Analysis of correlations between Deaf user evaluation and their background

Evaluation measurement	Computer experience	Confidence with written Spanish	Confidence with glosses
The signs are correct	0.18 ($p = 0.620$)	0.62 ($p = 0.054$)	0.54 ($p = 0.110$)
I understand the sign sequence	0.38 ($p = 0.281$)	0.61 ($p = 0.060$)	0.62 ($p = 0.058$)
The signing is natural	−0.08 ($p = 0.830$)	−0.03 ($p = 0.930$)	−0.06 ($p = 0.872$)
I would use the system in the absence of a human interpreter	0.43 ($p = 0.219$)	**0.67 ($p = 0.033$)**	**0.68 ($p = 0.023$)**
OVERALL assessment	0.39 ($p = 0.266$)	**0.63 ($p = 0.048$)**	**0.64 ($p = 0.047$)**

Bold values are statistically significant at $p < 0.05$

Finally, in order to report more information on the user assessment, Fig. 17 shows the distribution of the number of users versus the overall assessment provided. As is shown, there are two very different types of user: the first group gave a good overall assessment 4.2, while the second group gave a very negative one: 2.2. This analysis reveals two different user behaviours. This difference can be due to several causes: different perception on the use of new technologies (including an artificial avatar) for generating LSE content, different levels of politeness when evaluating assistance tools involving much development effort, or considering the possible reduction in job opportunities for interpreters. In order to expand this analysis, Table 8 shows Spearman's correlation between Deaf user evaluation and their background: computer experience, confidence with written Spanish, and confidence using glosses. This table also includes p-values for reporting the correlation significance. Because of the very low number of data and the unknown data distribution, Spearman's correlation has been used. This correlation produces a number between −1 (opposite behaviours) and 1 (similar behaviours). A 0 correlation means no relation between these two aspects.

As is shown, only those results in bold are significant ($p < 0.05$): *the use of the system in the absence of a human interpreter* and *the overall evaluation* correlate positively with the *user confidence with written Spanish* and *user confidence with glosses*.

7 Main conclusions

This paper has described the design, development and evaluation of a Spanish into LSE translation system for helping Deaf people when they want to renew their DL. This 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 signs). For the natural language translator, three technological proposals have been evaluated and combined in a hierarchical structure: an example-based strategy, a rule-based translation method and a statistical translator.

In the field evaluation, the system performed very well in speech recognition (4.8% WER) and language translation (8.9% sign error rate), but Deaf users did not positively assess the system. From the user comments and evaluation discussions, the main conclusion obtained is that it is necessary to improve the naturalness of the avatar and to make a greater effort in improving the documentation of the LSE. The discrepancies in sign representation, sign selection or sign sentence grammar are perceived as wrong behaviours of the avatar. When having problems with the system, the users tried to solve the problem by asking the government employee for a repetition or by reading the speech recognition output or the gloss sequence shown in the system interface.

This paper has presented the first field evaluation of a machine translation system from Spanish to LSE by detailing an interesting discussion on the main problems that must be solved in order to improve the system for obtaining a commercial prototype. Although the authors have not made any comparison with other forms of communication without the interpreter, the authors have the impression that the system presented in this paper provides a better communication alternative as compared to writing questions and answers in a paper, traditionally used in this situation. The main reason is because Deaf people have problems understanding written Spanish.

Acknowledgments The authors want to thank the eSIGN (Essential Sign Language Information on Government Networks) consortium for permitting the use of the eSIGN Editor and the 3D avatar in this research work. The authors want to thank discussions and suggestions from the colleagues at GTH-UPM and Fundación CNSE. This work has been supported by Plan Avanza Exp N°: PAV-070000-2007-567, ROBONAUTA (MEC ref: DPI2007-66846-c02-02) and SD-TEAM (MEC ref: TIN2008-06856-C05-03) projects. Authors also want to thank Mark Hallett for the English revision.

References

1. Agarwal A, Lavie A (2008) Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of

Machine Translation Output. In: Proceedings of workshop on statistical machine translation at the 46th annual meeting of the Association of Computational Linguistics (ACL-2008), Columbus, June 2008
2. Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization at the 43th annual meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005
3. Bender O (2010) Robust machine translation for multi-domain tasks. PhD thesis, Aachen, Germany, March 2010
4. Bornardo D, Baggia P (2005) Loquendo White paper Loquendo Report January 2005. Online last access: 1 June 2011 http://www.loquendo.com/en/whitepapers/SSML.1.0.pdf
5. Brown RD (2000) Automated generalization of translation examples. In: Proceedings of the eighteenth international conference on computational linguistics (COLING-2000), pp 125–131. Saarbrücken, Germany, August 2000
6. Bungeroth J, Ney H (2004) Statistical sign language translation. In: Workshop on representation and processing of sign languages, LREC 2004, pp 105–108
7. Casacuberta F, Vidal E (2004) Machine translation with inferred stochastic finite-state transducers. Comput Linguist 30(2):205–225
8. Cox SJ, Lincoln M, Tryggvason J, Nakisa M, Wells M, Mand Tutt, Abbott S (2002) TESSA, a system to aid communication with deaf people. In: ASSETS 2002, pp 205–212, Edinburgh, Scotland
9. Dandapat S, Forcada M, Groves D, Penkale S, Tinsley J, Way A (2010) OpenMaTrEx: a free/open-source marker-driven example-based machine translation system. In: Proceedings of Icetal 2010, the 7th international conference on natural language processing, Reykjavik, Iceland (in press)
10. Efthimiou E, Fotinea SE, Sapountzaki G (2007) Feature-based natural language processing for GSL synthesis. Sign Lang Linguist 10(1):3–23
11. Fels DI, Richards J, Hardman J, Lee DG (2006) Sign language web pages. Am Ann Deaf 151(4):423–433
12. 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, pp 3377–3380. Lisboa. Portugal. September 2005
13. Gauvain JL, Lee CH (1994) Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans SAP 2:291–298
14. Hanke T, Popescu H (2003) Intelligent sign editor. eSIGN project deliverable. D2.3. 2003
15. Hermansky H (1990) Perceptual linear prediction (PLP) analysis for speech. J Acoust Soc Am 87(4):1738–1752
16. Herrero A (2004) Escritura alfabética de la Lengua de Signos Española Universidad de Alicante. Servicio de Publicaciones
17. Instituto Nacional de Estadística de España (INE) Spanish Statistics Institute. Social Analysis Report 2008. Online last access: 1 June 2011. http://www.ine.es
18. Jankowski CR, Hoang-Doan, Jr., Lippmann RP (1995) A comparison of signal processing front ends for automatic word recognition. IEEE Trans Speech Audio Process 3(4):286–293
19. Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127–133, May 2003
20. Levenshtein V (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady
21. Mariño JB, Banchs R, Crego JM, Gispert A, Lambert P, Fonollosa JA, Costa-Jussà M (2006) *N*-gram-based machine translation. Comput Linguist Assoc Comput Linguist 32(4):527–549
22. Ministerio de Educación (MEC): Ministry of Education. National Report (2009) Online last access: 1 June 2011. http://www.mec.es
23. Moreno A (1997) SpeechDat Spanish Database for fixed telephone networks. Corpus Design Technical Report, SpeechDat Project LE2-4001
24. Morrissey S, Way A (2005) An example-based approach to translating sign language. In: Workshop example-based machine translation (MT X–05), pp 109–116, Phuket, Thailand, September
25. Morrissey S, Way A, Stein D, Bungeroth J, 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
26. Möller S, Smeele P, Boland H, Krebber J (2007) Evaluating spoken dialogue systems according to de-facto standards: a case study. Comput Speech Lang 21(1):26–53
27. Och J, Ney H (2000) Improved statistical alignment models. In: Proceedings of the 38th annual meeting of the association for computational linguistics, pp 440–447, Hongkong, China, October 2000
28. 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, pp 295–302
29. Och J, Ney H (2003) A systematic comparison of various alignment models. Comput Linguist 29(1):19–51
30. Papineni K, Roukos S, Ward T, Zhu WJ (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
31. Popović M (2009) Machine translation: statistical approach with additional linguistic knowledge. PhD thesis, Aachen, Germany, April 2009
32. 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
33. San-Segundo R, Barra R, Córdoba R, D'Haro LF, Fernández F, Ferreiros J, Lucas JM, Macías-Guarasa J, Montero JM, Pardo JM (2008) Speech to sign language translation system for Spanish. Speech Commun 50:1009–1020
34. San-Segundo R, Pardo JM, Ferreiros F, Sama V, Barra-Chicote R, Lucas JM, Sánchez D, García A (2010) Spoken Spanish generation from sign language. Interact Comput 22(2):123–139
35. Stolcke A (2002) SRILM—an extensible language modelling toolkit. ICSLP. 2002. Denver Colorado, USA
36. 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. pp171-174. 2003
37. Wells JC (1997) SAMPA computer readable phonetic alphabet. In: Gibbon D, Moore R, Winski R (eds) Handbook of standards and resources for spoken language systems. Mouton de Gruyter, Berlin. Part IV, section B
38. Wong J (2002) Repetition in conversation: a look at 'first and second sayings'. Res Lang Soc Interact 33(4):407–424
39. Zens R (2008) Phrase-based statistical machine translation: models, search, training. PhD thesis, Aachen, Germany, February 2008
40. Zens R, Och FJ, Ney H (2002) Phrase-based statistical machine translation. In: German Conference on Artificial Intelligence (KI 2002). Aachen, Germany, Springer, LNAI, pp 18–32, September 2002
41. Zwiterslood I, Verlinden M, Ros J, van der Schoot S (2004) Synthetic signing for the Deaf: eSIGN. In: Proceedings of the conference and workshop on assistive technologies for vision and hearing impairment, CVHI 2004, 29 June–2 July 2004, Granada, Spain