Paper title: **Design, Development and Field Evaluation of a Spanish into Sign Language Translation System**

Authors: R. San-Segundo, J.M. Montero, R. Córdoba, V. Sama, F. Fernández, L.F. D'Haro.

Affiliation: Departamento de Ingeniería Electrónica. ETSI Telecomunicación. Universidad Politécnica de Madrid

Authors: C. Vicente, A. García.

Affiliation: Fundación CNSE. Confederación Estatal de Personas Sordas.

Contact Information:

Contact Author: Rubén San-Segundo

Mailing Address:

Grupo de Tecnología del Habla
Dpto. Ingeniería Electrónica
ETSI Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria s/n
28040-Madrid Spain

Telephone: +34 915495700 then dial 4228 when prompted

Fax: +34 913367323

Email: lapiz@die.upm.es

Number of pages: 17

Number of figures: 15

Number of tables: 4

Keywords: Deaf people, Spanish Sign Language (LSE), Spoken Language Translation, Sign Animation, Driver's License renew.

DESIGN, DEVELOPMENT AND FIELD EVALUATION OF A SPANISH INTO SIGN LANGUAGE TRANSLATION SYSTEM

ABSTRACT

This paper describes the design, development and field evaluation of a Spanish into Spanish Sign Language (LSE: Lengua de Signos Española) translation system. The developed system is focused on helping Deaf people when they want to renew their Driver's License. The system is composed 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 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 in a hierarchical structure. This paper includes a detailed description of the field evaluation carried out. This evaluation has been performed in the Local Traffic Office in Toledo including real government employees, and Deaf people from Madrid and Toledo. The evaluation includes objective measures from the system and subjective information from questionnaires. The paper reports a detailed analysis of the main problems found and a discussion about how to solve them (some of them specific for Spanish Sign Language).

1. Introduction

In the last 10 years, the European Commission and the USA Government have invested a lot of resources for researching on language translation. In Europe, TC-STAR is the last 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/), financed by European Commission within the Sixth Program, is envisaged as a long-term effort to advance research in 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).

In USA, 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, analyze 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 forms 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 dialogue (not necessarily in natural language). This project has been active for two 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 been also boosted by the machine translation evaluation organised by USA Government, NIST (http://www.itl.nist.gov/iad/mig/tests/mt/).

The best performing translation systems are based on various types of statistical approaches (Och and Ney, 2002; Mariño et al, 2006), including example-based methods (Sumita et al, 2003), finite-state transducers (Casacuberta and Vidal, 2004) and other data driven approaches. The progress achieved over the last 10 years is due to several factors like efficient algorithms for training (Och and Ney, 2003), context dependent models (Zens et al, 2002), efficient algorithms for generation (Koehn, 2003), more powerful computers and bigger parallel corpora, and automatic error measures (Papineli et al, 2002; Banerjee and Lavie, 2005; Agarwal and Lavie, 2008).

Another important effort on machine translation has been the organization of several Workshops on Statistical Machine translation (SMT). In the webpage http://www.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 named Moses available in this web page (http://www.statmt.org/moses/). Moses is a phrase-based statistical machine translation system that allows you to build machine translation system models for any language pair, using a collection of translated texts (parallel corpus).

In recent years, several groups have shown interest in Spoken language translation into Sign Languages, developing several prototypes: example-based (Morrisey and Way, 2005), rule-based (San-Segundo et al 2008), full sentence (Cox et al, 2002) or statistical (Bungeroth and Ney, 2004; Morrissey et al, 2007; SiSi system http://www-03.ibm.com/press/us/en/pressrelease/22316.wss) approaches. Present paper describes the first system that combines and integrates several translation strategies for translating Spanish into LSE and also presents the first field evaluation in real conditions: with real interactions between hearing and deaf people.

In 2007, the Spanish Government accepted Spanish Sign Language (LSE: Lengua de Signos Española) as one of the official languages in Spain, defining a long-term plan to invest resources on this language. One important problem is that LSE is not disseminated enough between hearing people. This problem is the reason of important communication barriers between a deaf person and, for example, a government employee who is providing a personal service. These barriers can make deaf people to have less opportunities or rights. This happens, for example, when people want to renew the Driver's License (DL). Generally, a lot of government employees do not know LSE so a deaf person needs a human interpreter for translating the government employee explanations.

About 3D avatars for representing signs, the VISICAST and eSIGN European Project (Essential Sign Language Information on Government Networks) (http://www.sign-lang.uni-hamburg.de/esign/) (Zwiterslood et al, 2004) have been one of the most important research efforts in developing tools for 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 animations in a rapid and easy way. The system proposed in this paper uses this 3D avatar as it will be shown in section 6.

One of the partners of VISICAST and eSIGN projects is the research group on Virtual Humans at University of East Anglia (http://www.uea.ac.uk/cmp/research/graphicsvisionspeech/vh). This group has been involved in several projects concerning sign language generation using virtual humans: TESSA, SignTel, Visicast, eSIGN, SiSi, LinguaSign, etc.

This paper describes the first translation system from Spanish into LSE evaluated in real interactions between a deaf person and a hearing person without interpreter: government employees that provide a service (Driver's License renewing) and deaf users that want to access to this service. The proposed system translates the government employee explanations into LSE for deaf users.

The papers is organised as follows. Section 2 describes the linguistic study performed to develop the system. Section 3 presents the system architecture. Section 4, 5 and 6 describe the speech recognizer, language translation and sign animation modules respectively. Section 7 presents the system interface. The field evaluation and the main conclusions are described in sections 8 and 9.

2. Database collection for the Driver's License renewing process

The linguistic study was performed in collaboration with the Local Traffic Office in Toledo. The most frequent explanations (from the government employees) and the most frequent questions (from the user) were annotated during three weeks.

Figure 1.Different windows at the Local Traffic Office in Toledo and order number machine

This office is organised in several windows (assisting positions) (Figure 1): information window (for general questions and form collection), cash desk (for paying taxes), driver window (driver specific formalities), vehicle window (vehicle related steps) and driving school window.

During three weeks more than 4000 sentences were annotated and analysed for all the windows. 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) renewing process at the Toledo Traffic Office consists of three steps:

1. First of all, the user has to go to the information window where the user gets the application form to fill and a sheet with a list of documents needed for the process: Identification Card, the old DL, a medical certificate and a photo.
2. Secondly, it is necessary to pay 22 euros at the cash desk.
3. Finally, the user must go to the driver window with all the documentation. The new DL will be sent by mail into the next three months. For driving during this time, the user receives a provisional DL.

In all three steps, the user has to get an order number from a machine (Figure 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 collected: 547 pronounced by government employees and 160 by users. These sentences have been translated into LSE, both in text (sequence of glosses) and in video, and compiled in an excel file. The excel file contains six different information fields: VENTANILLA (window: where the sentence were collected), SERVICIO (service provided when the sentence was collected), if the sentence were pronounced by government employee or user (funcionario or usuario receptively), 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 1.

Government employee sentences	Spanish	LSE
Sentence pairs	547	
Different sentences	513	200
Running words	5714	4,247
Vocabulary	411	237

Table 1. Main statistics of the corpus

3. Spanish into Spanish Sign Language translating architecture

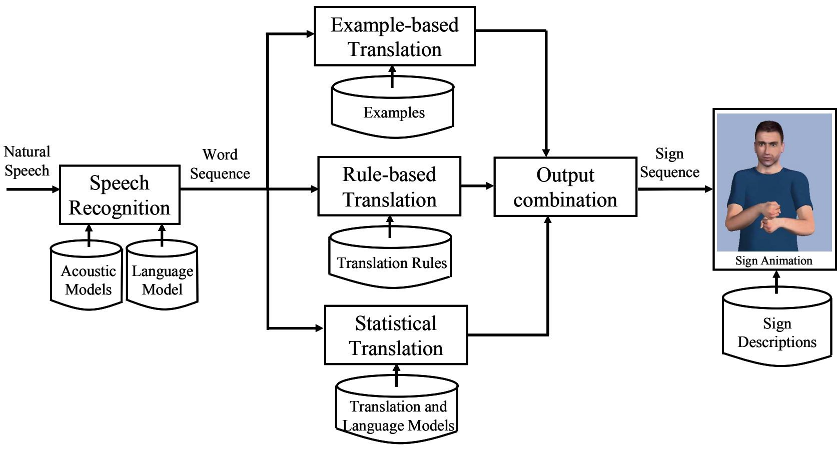

Figure 2. Diagram of the Spanish into LSE translation module

Figure 2 shows the module diagram developed for translating spoken language into Spanish Sign language (LSE). As, it is shown, the main modules are the following:

- The first module, the speech recognizer, converts natural speech into a sequence of words (text). It uses language models and some acoustic models for every allophone.
- The natural language translation module converts a word sequence into a sign sequence. For this module, the paper presents three different strategies that are combined at the output step. The first one consists of an example-based strategy: the translation process is done based on the similarity between the sentence to be translated and the items of a parallel corpus with translated examples. Secondly, a rule-based translation strategy, where a set of translation rules (defined by an expert) guides the translation process. The last one 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 performed by VGuido: the eSIGN 3D avatar developed in the eSIGN project (http://www.sign-lang.uni-hamburg.de/esign/). It has been incorporated as an ActiveX

control. The sign descriptions are generated previously through an advanced version of the eSIGN Editor.

4. Automatic Speech Recognition (ASR)

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

- It is a continuous speech recognition system: it recognizes utterances formed by several words continuously spoken. In this application, the vocabulary size is 533 Spanish words: the corpus vocabulary (with 411 words) was extended with a complete list of numbers (from 0 to 100), week days, months, etc.
- Speaker independency: the recognizer has been trained with a lot of speakers (4000 people), making it robust against a great range of potential speakers without further training by actual users.
- The system uses a front-end with PLP coefficients derived from a Mel-scale filter bank (MF-PLP), with 13 coefficients including c0 and their first and second-order differentials, giving a total of 39 parameters for each 10 msec. frame. This front-end includes CMN and CVN techniques.
- For Spanish, the speech recognizer uses a set of 45 units. The system also has 16 silence and noise models for detecting acoustic sounds (non speech events like background noise, speaker artefacts, filled pauses, etc.) that appear in spontaneous speech. The system uses context-dependent continuous Hidden Markov Models (HMMs) built using decision-tree state clustering: 1,807 states and 7 mixture components per state. These models have been trained with more that 40 hours of speech from the SpeechDat database (Moreno, 1997).
- About 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 recognition system can generate one optimal word sequence (given the acoustic and language models), a solution expressed as a directed acyclic graph of words that may compile different alternatives, or even the N-best word sequences sorted by similarity to the spoken utterance.
- The recognizer provides one confidence measure for each word recognized in the word sequence. The confidence measure is a value between 0.0 (lowest confidence) and 1.0 (highest confidence) (Ferreiros et al, 2005). This measure is important because the speech recognizer performance varies depending on several aspects: level of noise in the environment, non-native speakers, more or less spontaneous speech, or the acoustic similarity between different words contained in the vocabulary.
- The acoustic models can be adapted to one speaker or to a specific acoustic environment using MAP (Maximum a Posteriori)

About the performance of the ASR module in laboratory tests, with vocabularies smaller than 1000 words, the Word Error Rate (WER) is lower than 5%. If this ASR module is adapted to a specific speaker, the WER drops under 2%.

5. Natural Language Translation

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

5.1. 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 (from which one is translating) and their corresponding translations in the target language, and translates other similar source-language sentences. In order to determine if one example is equivalent (or at least similar enough) to the text to be translated, the system computes a heuristic distance between them. Defining a threshold on this heuristic distance, it is possible to define how similar must be the example to the text 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 than the example translation. But if the distance is higher, the system can not generate any output. In these circumstances, it is necessary to consider other translation strategies.

In this case, the heuristic distance considered is the well known Levenshtein distance (LD) divided by the number of words in the sentence to be translated (this distance is represented in percentage). Levenshtein Distance is a measure 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 named edit distance. The greater the Levenshtein distance, the more different the strings are. 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 it has been used in this paper). The LD is computed by a dynamic programming algorithm that considers the following costs: 0 for identical words, 1 for insertions, 1 for deletions and 1 for substitutions.

One problem of this distance is that two synonymous are considered as different words (a substitution in the LD) while the translation output can be the same. Currently, the system is 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 behaviours of the two words. These behaviours are obtained from the lexical model computed in the statistical translation strategy (described in section 5.3). For each word (in the source language), a 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.

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

Equation 1. Substitution cost based on the translation behaviour

When both words present the same behaviour (same vectors), substitution cost tends to 0. Otherwise, when there is not any overlapping between translations vectors, substitution cost tends to 1. This improved distance has been incorporated recently and it has not been used in the field evaluation.

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

Spanish: "Veinte euros con diez céntimos" (Twenty euros with ten cents)

LSE: "VEINTE COMA DIEZ EURO"

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 words and phrases included in the class lists, and replacing each occurrence by 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, 4 classes were used: $NUMBER, $PROPER_NAME, $MONTH and $WEEK_DAY.

Figure 3 represents the translation process for the recognised sentence: "catorce euros veinte céntimos". The first step is to categorize the sentence 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. For this final step, it is necessary to specify the solution implemented in these situations.

- If there are several categories of the same type (2 $NUMBER, as in the example presented before). It is supposed that they have the same order in both languages. This assumption is valid considering the two languages involved in the translation process but it can not be valid for other pair of languages.
- If by error (a wrong example selection) there is a category in the selected example that it 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.

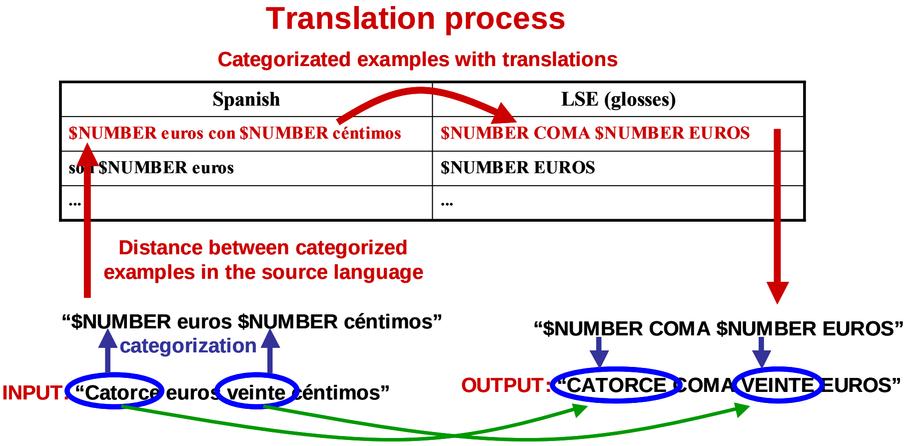

Figure 3. Translation process in an example-based translation system

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 the heuristic distance: 1 minus heuristic distance. The confidence measure will be used to decide if the sign sequence is represented by the avatar or not.

5.2. Rule-based strategy

In this strategy, the translation process is carried out in two steps. In the first one, every word is mapped to one or several syntactic-pragmatic categories (categorization). After that, the translation module applies different rules that convert the tagged words into signs by means of grouping concepts or signs (generally named blocks) and defining new signs. These rules are defined by an expert hand and they 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.

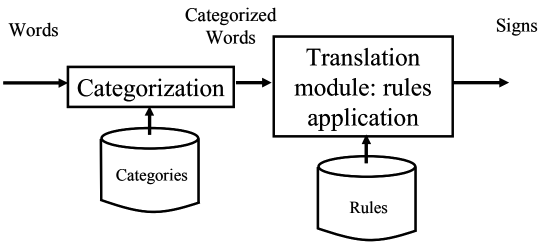

Figure 4. Translation process in a rule-based strategy

In this approach, the translation module and the rules have been implemented considering a bottom-up strategy: the translation analysis is performed starting from each word individually and extending the analysis to neighbourhood context words or already-formed signs (blocks). This extension is done to find specific combinations of words and/or signs (blocks) that generate another sign. The rules implemented by the expert define these relations. Depending on the scope of the block relations defined by the rules, it is possible to achieve different compromises between reliability of the translated sign (higher with higher lengths) and the robustness against recognition errors: when the block relations involve a high number of concepts, one recognition error can cause that the rules are not executed.

The rules are specified in a proprietary programming language consisting of a set of primitives. The rule-based translation module implemented contains 293 translation rules and it uses 10 different primitives. For evaluating the module performance, the following evaluation measures have been considered: SER (Sign Error Rate), PER (Position Independent SER), BLEU (BiLingual Evaluation Understudy; (Papineni, 2002)), and NIST (http://www.nist.gov/speech/tests/mt/), obtaining 21.45%, 17.24%, 0.6823, and 8.213 respectively.

Same 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 primitive generates the confidence for the elements it produces. For example, in the case of primitives that check for a specific sign sequence existence and they generate a new one, the primitive usually assigns to the newly created element the average confidence of the original sign sequence. 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. The confidence measure will be used for controlling the sign sequence representation.

5.3. Statistical translation

For 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 Workshops on Statistical Machine Translation (http://www.statmt.org). The translation process uses a translation model based on phrases and a target language model. The phrase model has been trained following these steps (Figure 7):

- Word alignment computation. At this step, the GIZA++ software (Och and Ney, 2000) has been used to calculate the alignments between words and signs. In order to establish word alignments, GIZA++ combines the alignments in both directions: words-signs and signs-words (Figure 5).

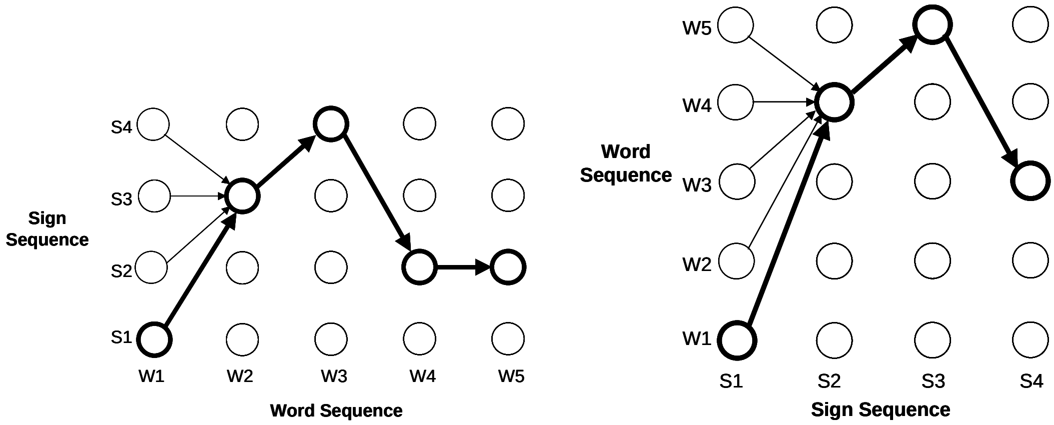

Figure 5. Alignments in both directions: words-signs and signs-words.

  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 (section 5.1).

- Phrase extraction (Koehn et al 2003). 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 (Figure 6). The maximum size of a phrase has been fixed to 7.

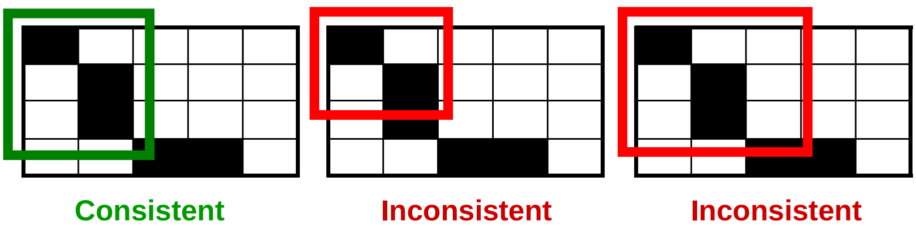

Figure 6. Examples of phrase extraction.

- Phrase scoring. In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

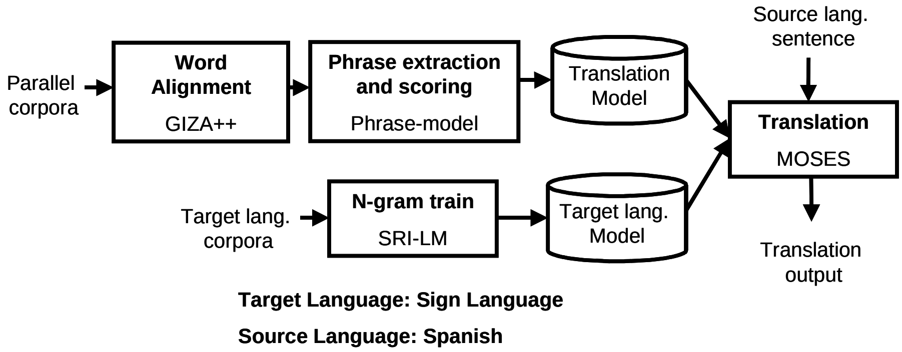

Figure 7. Diagram of the phrase-based translation module

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 (Stolcke, 2002).

The translation based on SFST is carried out following the diagram shown in Figure 8.

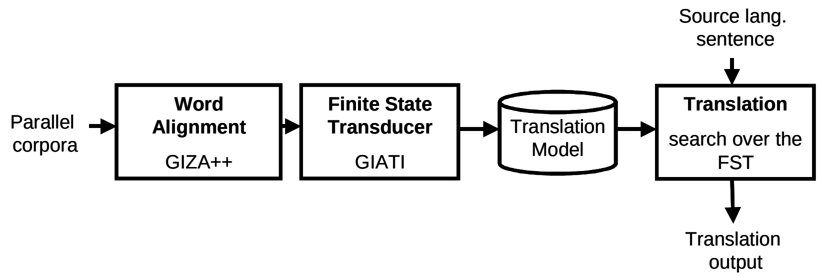

Figure 8. Diagram of the FST-based translation module

The translation model consists of a SFST composed of aggregations: subsequences of source and target words aligned. The SFST is inferred from the word alignment (obtained with GIZA++) using the GIATI (Grammatical Inference and Alignments for Transducer Inference) algorithm (Casacuberta and Vidal, 2004). 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 generate the same confidence measure for the whole sign sequence. 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 composed 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 can not deal with some parts of the sentence. The measure proposed in this case is the portion of signs generated (they are not words passed directly to the output): # of signs generated/ # of tokens in the output. This measure performs very well in restricted domain translation problems for detecting out of vocabulary sentences.

In order to evaluate the different modules, the corpus (including only sentences pronounced by government employees) was divided randomly in three sets: training, development and test, performing a round-robin evaluating process. Table 2 summarizes the results for rule-based and statistical approaches: SER (Sign Error Rate), PER (Position Independent SER), BLEU (BiLingual Evaluation Understudy; (Papineni, 2002)), and NIST (http://www.nist.gov/speech/tests/mt/).

		SER	PER	BLEU	NIST
Statistical approach	**Phrase-based**	39,01	37.05	0.5612	6.559
	SFST-based	34.46	33.29	0.6433	7.700
Rule-based approach		21.45	17.24	0.6823	8.213

Table 2. Result summary for rule-based and statistical approaches

The rule-based strategy has provided better results on this task because it is a restricted domain and it has been possible to develop a complete set of rules with a reasonable effort. Another important aspect is that the amount of data for training is very little and the statistical models can not be trained properly. In these circumstances, the rules defined by and expert introduce knowledge not seen in the data making the system more robust against new sentences. For this corpus the SFST-based method is better than the phrase-based method. For the field evaluation presented in section 8, statistical models have been trained with the whole database.

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. Because of this, the number of deletions is higher. As it is shown in section 8, insertion and substitution errors are the worst type of errors: they produce an important misunderstanding problem.

The example-based module has not been evaluated considering three independent sets because the corpus does not have many similar sentences. Analysing the corpus, the average distance between every example in the corpus and the closest example was computed obtaining a 45%. This number shows that the examples in the corpus are very different. The performed evaluation tried to analyse the influence of the speech recognition errors in the selection of the closest example. All the examples from the corpus were spoken by three different speakers and passed through the speech recogniser obtaining a Word Error Rate lower than 5%. The speech recognition outputs were passed to the example-based module reporting that only in 2% of cases, the recognition errors provoked a wrong example selection for translating.

5.4. Combining translation strategies

The natural language translation module implemented combines the three translation strategies described in previous sections. This combination is described in Figure 9.

The translation module has a hierarchical structure divided into two main steps. At the first step, an example-based strategy is used to translate the word sequence. If the distance with the closest example is lower than a threshold (Distance Threshold), the translation output is the same than the example. But if the distance is higher, a background module translates the word sequence. The Distance Threshold (DT) ranges between 20% and 30%. In the field evaluation, DT was fixed to 30% (one difference is permitted in a 4-word sentence).

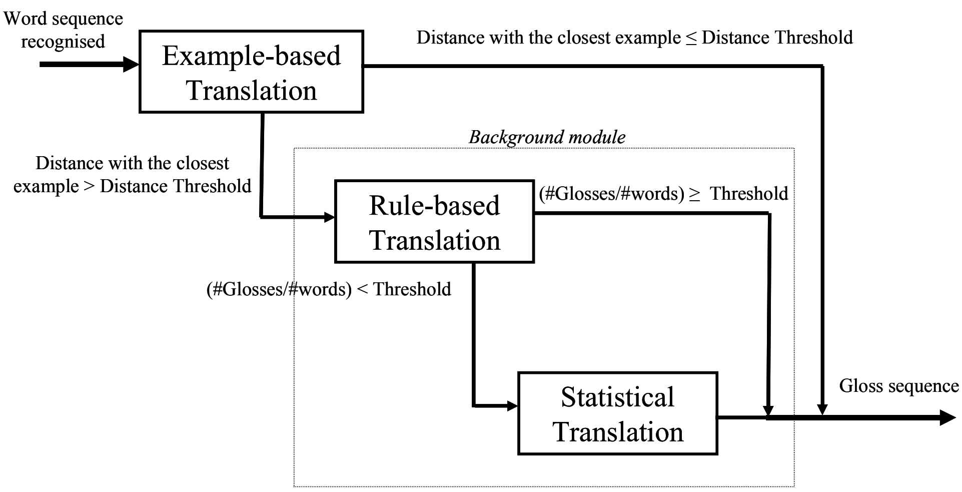

Figure 9. Diagram of natural language translation module combining three different translation strategies

For the background module, a combination of rule-based and statistical translators has been used. Considering the results presented in Table 2, the rule-based strategy would be the best alternative. Anyway, the statistical approach was also incorporated as a good alternative during system development. The main idea is that the effort and time required for developing a statistical translator (in one or two days it was possible to obtain a tuned version) is considerable lower than a rule-based one (it took several weeks to develop all rules). During rule development, a statistical translator was incorporated in order to have a background module with reasonable performance. The relation 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 #glosses/#words ratio is higher than a threshold, the output is the gloss sequence proposed by the rule-based module. Otherwise, if this condition is false, the statistical approach is executed. 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 examples have not been implemented yet (or the sentence is out of domain). During the rule-based system development, the glosses/words ratio mechanism was used to direct (in some cases) the translation process to the statistical approach. The ratio threshold was fixed to 0.5. About the statistical module, both alternatives were incorporated (phrase-based and SFST-based strategies), although only the SFST-based one was used for the field evaluation due to its better performance.

The first idea about the background module was to combine the rule-based module and the two statistical approaches using ROVER (Recognizer Output Voting Error Reduction) (Fiscus, 1997) adapted to translation outputs. The problem of this algorithm is that all translation outputs have the same relevance in the combination process. Because of the best performance of the rule-based strategy, its output was boosted by a hierarchical structure.

6. Sign animation with the eSIGN Avatar

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

A signed animation is generated automatically from an input script in the Signing Sign Markup Language (SiGML) notation. SiGML is an XML application which supports the definition of sign sequences. The signing system constructs human-like motion from scripted descriptions of signing motions. These signing motions belong to "Gestural-SiGML", a subset of the full SiGML notation, which is based on the HamNoSys notation for Sign Language transcription (Prillwitz et al, 1989). The morphological richness of sign languages can be modeled using a sign language editing environment (an advanced version of the eSIGN editor) without the need of manually describing each inflected form.

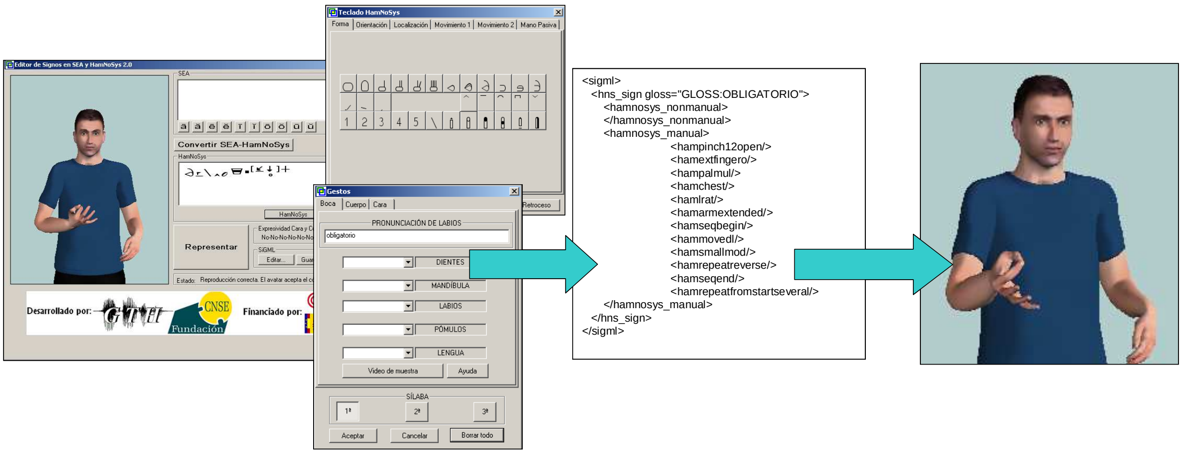

Figure 10. Process to generate signs with the avatar

HamNoSys and other components of SiGML mix primitives for static gestures (such as parts of the initial posture of a sign) with dynamics (such as movement directions). This allows the transcriber to focus on essential characteristics of the signs when describing a sign. This information, together with knowledge regarding common aspects of human motion as used in signing such as speed, size of movement, etc., is also used by the movement generation process to compute the avatar's movements from the scripted instructions. Figure 10 shows the process for specifying a sign from the HamNoSys description.

7. System Interface

The module for translating spoken Spanish into LSE has a visual interface shown in Figure 11.

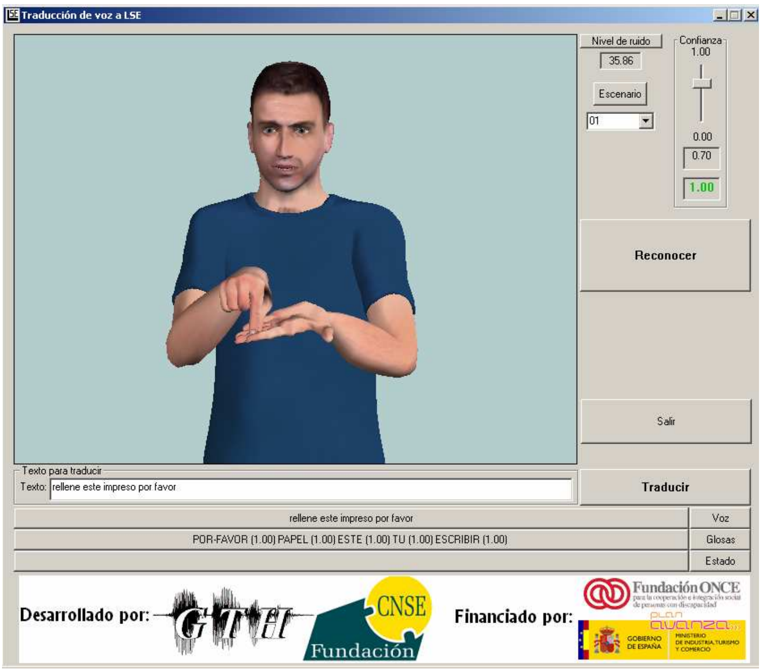

Figure 11. Visual interface of the Spanish into LSE translation module

This interface includes a slider control (in the right-top corner) to define the minimum confidence level of the translation output (sign sequence) in order to represent the signs. If the translation output does not have enough confidence, 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 a confidence value for the whole sign sequence.

When the government employee wants to speak, the "Reconocer" (Recognise) button must be pressed (it is also possible to execute the speech recognizer by pressing the INTRO key in the keyboard). The speech recognition and translation outputs are presented in windows at the bottom side.

The interface also allows translating a word sentence written in one of the controls ("Texto para traducir" text to be translated) by pressing the "traducir" (translate) button. This possibility was implemented for providing an alternative for introducing the word sequence if the speech recognizer would have problems. After all speech recognitions, the recognized output is also copied into the "texto para traducir" (text to be translated) control. This is very useful when the user asks for repetition. In this case, it is not necessary that government employee speaks again. If the previous recognition was OK, by pressing the "traducir" (to translate) button, the system will generate the same sign sequence.

Finally, it is necessary to comment that the system incorporates two functions thinking on the fact that the Tablet PC screen is oriented to the user (Figure 12): the system feedbacks the recognized sentence (with speech synthesis) and generates a beep when the system has finished signing (and it is ready for a new turn).

8. Field evaluation and discussion

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

8.1. Evaluation setup

The Driver's Licence (DL) renewing process at the Toledo Traffic Office consists of three steps: form obtaining, payment, and handing over of the documents. Following the idea suggested from the head of the Toledo Traffic Office for saving resources, instead of installing three systems at the three windows involved in the process, one new assisting position (Figure 12) was created where a deaf person can do all three steps.

Figure 12. Assisting position preparation and speech recognizer adaptation.

The evaluation was carried out during two days. The first day, the assisting position was installed and a one-hour talk, about the project and the evaluation, was given to government employees and users involved in the evaluation. Half of the users evaluated the system the first day, leaving the other half to the next day. The first day, the speech recognizer was adapted to the two government employees involved in the evaluation. For this adaptation, 50 sentences spoken by the government employee (1-2 seg) were recorded.

For the evaluation, the users were asked to interact with government employees using the system developed for DL renewing. Six different scenarios were defined in order to specify real situations:

- In one scenario, the user simulated to have all the needed documents.
- Three other scenarios where the user simulated not to have one of the documents: Identification Card, a photo or the medical certificate.
- One scenario where the user had to fill in some information in the application form.
- Finally, a scenario where the 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 interact with 2 government employees at the Toledo Traffic Office using the developed system. These 10 users (six males and four females) tested the system in almost all the scenarios described previously, generating 48 dialogues between government employees and deaf users. The user ages ranged between 22 and 55 years old with an overage value of 40.9 years. All the users declared to use a computer every day or every week, and only half of them had a medium-high understanding level of written Spanish.

Figure 13. Different photos of the evaluation process at Toledo Traffic Office

8.2. Results and discussion

The evaluation results include objective measures from the system and subjective information from both user and government employee questionnaires. A summary of the objective measures obtained from the system are shown in Table 3.

AGENT	MEASURE	VALUE
System	Word Error Rate	4.8%
	Sign Error Rate (after translation)	8.9%
	Average Recognition Time	3.3 sec
	Average Translation Time	0.0013 sec
	Average Signing Time	4.7 sec
	% 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

Table 3. Objective measures for evaluating the Spanish into LSE translation system

The WER (Word Error Rate) for the speech recognizer is 4.8%, higher than the results obtained in laboratory tests for cases where the speech recognizer was adapted to one speaker: 2%. Anyway, the WER was small enough to guarantee a low SER (Sign Error Rate) in the translation output: 8.9%. On the other hand, the time needed for translating speech into LSE (speech recognition + translation + signing) is around 8 seconds. This time allows a dialogue between government employee and user.

About the different translation strategies, the example-based translation has been used in more than 94% of the cases showing the goodness of the linguistic study performed (corpus collection). In this study, the most frequent sentences were recorded obtaining a very good representative corpus in this kind of dialogues. 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 change in the assisting position: all the renewing process was performed in the same assisting position instead of in several ones.

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 for being the principal way of interaction.

The subjective measures were collected from questionnaires filled by both: government employees and deaf users. They evaluated different aspects of the system giving a score between 0 and 5. The average results for each aspect are presented in Table 4.

AGENT	MEASURE	VALUE (0-5)
Government employee	System speed	4.0
	Speech Recognition Rate	3.5
	The system is easy to use	3.5
	The system is easy to learn	3.5
	Would you use the system in absence of a human interpreter?	3.5
	GLOBAL assessment	**3.5**
User	The signs are correct	2.1
	I understand the sign sequence	2.2
	The signing is natural	0.8
	Would you use the system in absence of a human interpreter?	2.0
	GLOBAL assessment	**2.2**

Table 4. Subjective measures for evaluating the Spanish into LSE translation system

The evaluation from the government employees is quite positive giving a 3.5 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 to the user (see Figure 14). It is true that the system feedbacks 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 for the future, two screens will be considered.

Figure 14. Government employee speaking to the user with the screen of the Tablet PC turned towards the user.

The user assessment was very low (a global score of 2.2). The worst score was to the sign naturalness (0.8). Although the objective measures were very good (with very good recognition and translation rates) the user did not like the signing. The main causes observed during the evaluation were the following:

- It is true that the avatar naturalness is not comparable to a human signing. It is necessary to keep investing more effort on increasing flexibility, expressiveness and naturalness of the avatar, especially on the face.
- But it is also fair to report that there were discrepancies between users about the correct signing 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 the face expressiveness (i.e. pronouncing a word), aspect that must be improved in the avatar. The sign specification was done based on the normative dictionary generated by Fundación CNSE, DILSE III. These discrepancies showed the need to keep working in the standardization process of the LSE. Although there are not significant data, a high level of agreement between users from Madrid was perceived.
- Another source of discrepancies is the structure of some sign sentences. LSE, as other languages, offers an important level of flexibility. This flexibility some times is not well understood and some of the possibilities are considered as wrong sentences. Some examples are:
    - 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.
    - About 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” (box).
    - 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).
- 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 to the signer. If the avatar is perfectly oriented to the user, the movement to the signer is not perceived properly. In order to solve this problem, the avatar was slightly turned to see movement in all significant directions.
- 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 are. 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 signs, and considering the possibility to use the system in several offices with different distributions, it is necessary to substitute these signs by more specific ones: “VENTANILLA ESPECIFICO CONDUCTOR” (window specific driver).

Although the reported comments influenced to the signing perception the most, the recognition and translation rates can have also a relevant influence over the system quality perceived by users. When the system introduces a wrong sign in the sign sequence (there is an insertion or a substitution in the translation output), the consequence is very bad: user stops paying attention and asks the meaning of this sign, missing the rest of the signs. For these cases, it was necessary to repeat the sentence again. If the system deletes (by error) one sign, sometimes the user can understand the sentence meaning.

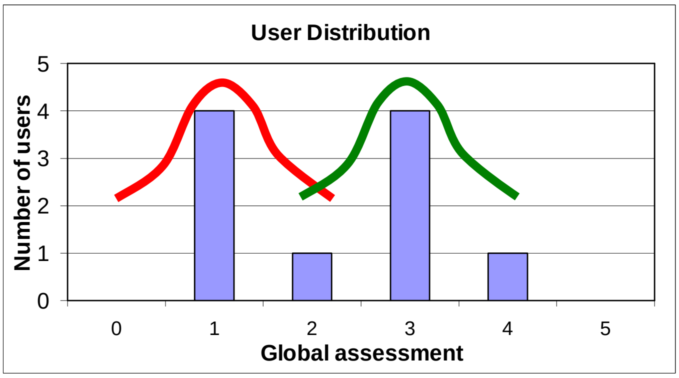

Figure 15. Distribution of users versus global assessment

Finally, in order to report more information about the user assessment, Figure 15 shows the distribution of the number of users versus the global assessment provided. As it is shown, there are two very different types of user: the first group gave a good global assessment 3.2, while the second group gave a very negative one: 1.2. This analysis reveals two different perceptions about the use of new technologies (including artificial avatar) for generating LSE content.

9. Main Conclusions

This paper has described the design, development and evaluation of a Spanish into Spanish Sign Language (LSE: Lengua de Signos Española) translation system for helping Deaf people when they want to renew their Driver's License. This system is composed 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% word error rate) and language translation (8.9% sign error rate), but the users did not assess the system with a very good score in the questionnaires. From the user comments and evaluation discussions, the main conclusion obtained is that it is necessary to improve the avatar naturalness and to invest more effort for increasing the level of standardization for LSE. The discrepancies on sign representation, sign selection or sign sentence grammar are perceived as wrong behaviours of the avatar.

This paper has presented the first field evaluation of a Spanish into LSE translation system reporting an interesting discussion about what are the main problems that must be solved in order to improve the system for obtaining a commercial prototype.

Acknowledgements

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 Nº: PAV-070000-2007-567 and ROBONAUTA (MEC ref: DPI2007-66846-c02-02) and SD-TEAM (MEC ref: TIN2008-06856-C05-03) projects.

References

Agarwal, Abhaya and Lavie, Alon, 2008. "Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output", Proceedings of Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association of Computational Linguistics (ACL-2008), Columbus, June 2008.

Banerjee, S. and A. Lavie, 2005. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", 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.

Bornardo D., and Baggia, P. 2005. “Loquendo White paper” Loquendo Report January 2005. http://www.loquendo.com/en/whitepapers/SSML.1.0.pdf.

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

Casacuberta F., E. Vidal. 2004. “Machine Translation with Inferred Stochastic Finite-State Transducers”. Computational Linguistics, Vol. 30, No. 2, pp. 205-225, June 2004.

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

Ferreiros, J., R. San-Segundo, F. Fernández, L. D'Haro, V. Sama, R. Barra, P. Mellén. 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.

Fiscus, J., 1997. “A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)”. In: Proc. IEEE Automat. Speech Recognition Understand. Workshop, pp. 347-352.

Herrero, Ángel. 2004 “Escritura alfabética de la Lengua de Signos Española” Universidad de Alicante. Servicio de Publicaciones.

Koehn P., F.J. Och D. Marcu. 2003. “Statistical Phrase-based translation”. Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127-133, May 2003.

Mariño J.B., Banchs R., Crego J.M., Gispert A., Lambert P., Fonollosa J.A., Costa-Jussà M., 2006. "N-gram-based Machine Translation", Computational Linguistics, Association for Computacional Linguistics. Vol. 32, nº 4, pp. 527-549.

A. Moreno, 1997. SpeechDat Spanish Database for Fixed Telephone Networks. Corpus Design Technical Report, SpeechDat Project LE2-4001.

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

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

Och J., Ney, H., 2000 "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.

Och J., Ney. H., 2002. “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation”. Annual Meeting of the Ass. For Computational Linguistics (ACL), Philadelphia, PA, pp. 295-302. 2002.

Och J., Ney. H., 2003. “A systematic comparison of various alignment models”. Computational Linguistics, Vol. 29, No. 1 pp. 19-51, 2003.

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

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

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

Stolcke A. “SRILM – An Extensible Language Modelling Toolkit”. ICSLP. 2002. Denver Colorado, USA.

Sumita E., Y. Akiba, T. Doi 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.

Tryggvason J., “VANESSA: A System for Council Information Centre Assistants to communicate using sign language”. School of Computing Science. University of East Anglia. 2004.

von Agris, U., Schneider, D., Zieren, J., Kraiss, K.-F.: 2006. “Rapid Signer Adaptation for Isolated Sign Language Recognition”. In: CVPR Workshop V4HCI, New York, USA, June 2006, p. 159 (2006).

Zens R., F.J. Och, H. Ney. 2002. “Phrase-Based Statistical Machine Translation”. German Conference on Artificial Intelligence (KI 2002). Aachen, Germany, Springer, LNAI, pp. 18-32, Sep. 2002.

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