On the creation and the annotation of a large-scale Italian-LIS parallel corpus

Nicola Bertoldi[1], **Gabriele Tiotto**[2], **Paolo Prinetto**[2], **Elio Piccolo**[2],
Fabrizio Nunnari[3], **Vincenzo Lombardo**[3], **Alessandro Mazzei**[4],
Rossana Damiano[4], **Leonardo Lesmo**[4], **Andrea Del Principe**[5]

(1) Fondazione Bruno Kessler – Trento, Italy – bertoldi@fbk.eu
(2) Dip. di Automatica e Informatica, Politecnico di Torino – Turin, Italy – {gabriele.tiotto,paolo.prinetto,elio.piccolo}@polito.it
(3) Virtual Reality and Multimedia Park – Turin, Italy – {fabrizio.nunnari,vincenzo.lombardo}@vrmmp.it
(4) Dip. di Informatica, Università di Torino – Turin, Italy – {mazzei,damiano,lesmo}@di.unito.it
(5) Centro Ricerche ed Innovazione Tecnologica, RAI – Turin, Italy – andrea.delprincipe@rai.it

Abstract

This paper presents the current development of the first large parallel corpus between Italian and Italian Sign Language (Lingua Italiana dei Segni, LIS). This initiative has been taken within the ATLAS project (Automatic Translation into Sign Languages), that aims at realizing a virtual interpreter, which automatically translates an Italian text into LIS. The Italian-LIS virtual interpreter is implemented by means of two modules interfaced by the ATLAS Extended Written LIS (AEWLIS), which is a translation-oriented representation of LIS: the first module translates the source Italian text into AEWLIS; the second module transforms the AEWLIS content into a coherent LIS sequence, smoothly animated by a virtual character. As no significant amount of electronic data are available for Italian and LIS, we have started building a parallel corpus from scratch in order to train and tune the Italian-AEWLIS translation system, and to compare the resulting virtual animations with human-performed LIS interpretations. The corpus, which will be freely available, actually presents a tri-lingual structure, with the Italian text, the AEWLIS sequence, and the signed LIS video.

1. Introduction

People who were born deaf or acquired deafness in the first years of life -approximately 70,000 in Italy- experience big obstacles to integrate into the society, because they could not properly acquire knowledge of the spoken language, and consequently of the written language, and vice versa hearing people very rarely practice Sign Languages (SLs).

The care of hearing-impaired people progressively grows; the increasing request for SL interpretation in educational, legal, and health contexts is foreseen and soon expected to be extended to culture and entertainments. The depicted scenario makes clear the relevance of the availability of a low cost technology to support the SL interpretation.

ATLAS (Automatic Translation into sign LAnguageS) is a three-year project, funded by the local government of Piedmont, Italy, aiming at providing Italian deaf people with facilities to access broadcast communications, and in particular to follow TV programmes. More specifically, ATLAS aims at developing a virtual interpreter, which automatically translates Italian into LIS.

The virtual interpreter has a modular structure and relies on a translation-oriented symbolic representation of the LIS, called ATLAS Extended Written LIS (AEWLIS). Training and tuning of most components of the virtual interpreter requires a parallel corpus, composed of a large set of Italian sentences, their human-performed LIS interpretations and their corresponding AEWLIS. Furthermore, an excerpt of this parallel corpus is exploited for the component-wise and end-to-end evaluation in terms of both automatic and subjective criteria.

As a significant amount of parallel data is not available yet, we have started building a new corpus from scratch. The corpus actually presents a tri-lingual structure, with the Italian text, the AEWLIS sequence, the signed LIS videos. The first release of the corpus will contain weather forecast bulletins for a total of about 15K Italian running words and about 1.5 hours of LIS videos.

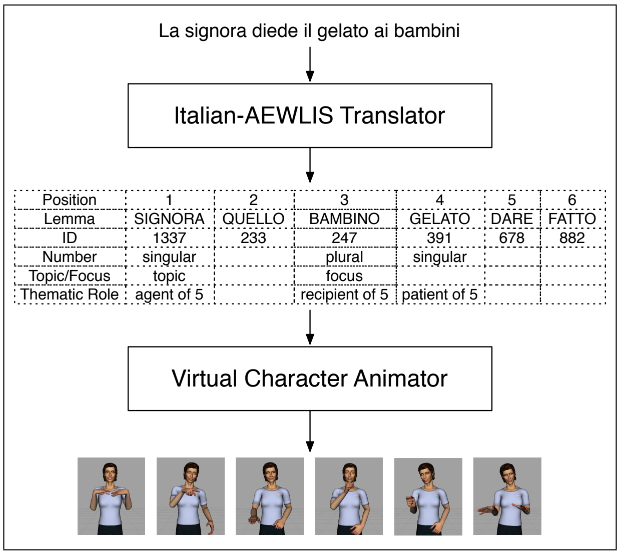
Next Section reports on scientific projects about the automatic translation of Sign Languages around the world. Section 3. briefly overviews the full-fledged virtual interpreter developed within the ATLAS project. Section 4. describes the corpus which the ATLAS partners are building and discusses issues arised during its creation. Section 5. presents AEWLIS, the intermediate artificial language chosen for representing the LIS in a written form. Some conclusions are finally drawn in Section 6..

2. State-of-the-art of the research on SLs

Since early 90's the scientific research on SLs has been constantly growing because of the increasing care of deaf people, their augmenting willingness of integration into the society, and the availability of more and more powerful computing facilities and software which make possible the automatic dealing with SL.

Most research projects on SLs approach American, British, Dutch, and German SLs: (SignWriting, 2009; DictaSign, 2010; eSign, 2009; SignSpeak, 2010; U.DePaul, 2008; U.Boston, 2002; Echo, 2010; NGTCorpus, 2009). They address some (or all) of the following highly-interconnected tasks: SL recognition, 3D character animation, machine translation, production of SL dictionaries and corpora, development of toolkits for SL annotation and transcription like (ELAN, 2010), and integration with existing communication devices, like mobile phones (mobileASL, 2010).

Differently from most oral languages, SLs do not have a natural corresponding written expression. Hence, researcher have proposed many artificial languages to represent them into a written form: gloss-based, (Stokoe et al.,


La signora diede il gelato ai bambini
Italian-AEWLIS Translator
Virtual Character Animator

Position	1	2	3	4	5	6
Lemma	SIGNORA	QUELLO	BAMBINO	GELATO	DARE	FATTO
ID	1337	233	247	391	678	882
Number	singular		plural	singular		
Topic/Focus	topic		focus			
Thematic Role	agent of 5		recipient of 5	patient of 5		

Figure 1: The interpretation process is performed by Donna, the ATLAS virtual interpreter, in two steps. First the Italian text is transformed/translated into AEWLIS, consisting of a sequence of signs and syntactic/semantic relationships among them. Then, a virtual character animates the AEWLIS content into a coherent LIS sequence.

1965), (HamNoSys, 2004), (SignWriting, 2009), (Elliott et al., 2004; Filhol and Braffort, 2006). Each of these transcription methods is tuned to answer the specific need they were developed for: understanding and translating sentence meaning, producing and recognizing isolated sign, producing digital animation, etc.

SLs present many phenomena requiring the adaptation of a sign to the context of the signed sentence: relocation in the signing space, increase or reduction of the “size”, repetition (e.g. for plural), movement of hands through the space from and to context-dependent positions (e.g. for verbs like “to go” or “to give”), the use of hand-shapes as classifiers (Huenerfauth, 2006). Recent research on 3D animation deals with these context-dependent phenomena; through synthetic approach the character is animated with an animation language which is interpreted by a real-time player (Veale et al., 1998; Marshall and Sáfár, 2003).

Both rule-, example-, and, more recently, statistical-based approaches have been adopted for the text-to-SL machine translation. Most MT systems exploit a gloss-based notation of SL. A detailed overview of SL MT can be found in (Morrissey, 2008).

Finally, several research projects focus on collecting dictionaries and corpora related to SLs: (Echo, 2010; eLIS, 2006; SignWriting, 2009; Bungeroth et al., 2006; Bungeroth et al., 2008; BSLCorpus, 2010; NGTCorpus, 2009).

As concerns the LIS, very few and small-size projects have been funded (eLIS, 2006; DizLis, 2010; BlueSign, 2010). The ATLAS project aims at covering this gap, fostering the interest of the research community towards LIS and LIS-related tools. It is worth remarking that all software and linguistic resources developed by ATLAS partners will be made freely available.

3. The ATLAS virtual interpreter

Figure 1 provides a high-level representation of the full-fledged Italian-LIS virtual interpreter. It is actually implemented by means of two modules interfaced by AEWLIS: the **Italian-AEWLIS Translator** and the **Virtual Character Animator**.

The **Translator** transfers the meaning of an Italian text into AEWLIS by means of both statistical and rule-based techniques. The phrase-based statistical Machine Translation (MT) system relies on the high-performing state-of-the-art toolkit Moses (Koehn et al., 2007). The rule-based MT system exploits linguistic rules to connect the morphology, syntax and semantics of the source language to those of the target language. The linguistic knowledge about Italian used by the translator has already been exploited (Lesmo et al., 2009) During the project, the integration of the two systems will be investigated.

The **Animator** relies on a *signary*, a repository in which each sign is described by an animation language in terms of motion data (motion-captured or hand-made), procedural animations and applicable parameters for size, relocation, repetition, hand-shapes, etc. First, a motion planner transforms the information present in the AEWLIS sequence into a sequence of signs, taken from the signary, whose parametric values are determined according to the actual context. Then, a blending system, a technique widely used in videogame architectures, creates the LIS by smoothly joining in real-time the existing animation clips through interpolation functions.

Animations will be displayed, for both broadcast and on-demand delivering, on a variety of user terminals (including DVB, Web, Mobile Phones). The heavy computational effort (translator and motion planner) will be carried out on a centralized server, and the visual rendering will be performed on the device. The physical appearance of the virtual character responds to criteria that enhance the perception of hand motion and facial expressions, that are fundamental in understanding signs. We have designed two signing characters, Donna and Manuel.

4. Description of the corpus

The module **Italian-AEWLIS Translator** introduced in Section 3. requires the availability of a parallel corpus for its training. As this kind of electronic data are not available for Italian yet, we have started building a parallel corpus from scratch.

The first application domain of the ATLAS project is the automatic interpretation of weather forecast bulletins daily broadcasted by RAI (Radio Televisione Italiana). Hence, we collected a set of 55 bullettins of 2008, containing about 15K Italian words corresponding to about 1.5 hours of Italian audio/video.

According to LIS experts and interpreters, we defined the following procedure to build the parallel corpus. First, the audio of the TV bullettin is automatically transcribed by a speech recognition system (Brugnara et al., 2000) and manually checked to correct transcription errors. Portions of the bullettin which is not strictly related to the weather domain are eventually removed. Then, a LIS expert interprets the content of the cleaned text and a movie of his/her signing

	Italian	AEWLIS
Number of bullettins	55	
Number of sentences	585	
Running terms	15,012	6,000*
Average terms per sentence	25.7	10*
Dictionary Size	1,442	300*
Singletons	614	-

Table 1: Statistics of the first release of the Italian-AEWLIS parallel corpus; asterisks mark estimated statistics.

is recorded using a standard framing. In order to avoid an unnecessary variability of the LIS, the expert is committed to sign a genuine but plain LIS. Finally, the same expert annotates his/her LIS movie according to the AEWLIS annotation guidelines described in Section 5..

As AEWLIS has several independent annotation levels (see Section 5.), they can be marked in successive steps from the least to the most specific. An editor has been developed by ATLAS partners to support the expert in the annotation process, which will become a Computer Assisted Translation tool after the integration with the **Translator**.

Thus, the corpus actually results in a tri-lingual structure, with the Italian text, the AEWLIS sequence, the signed LIS video. Furthermore, we would have also audio/video of the original TV bulletins and the automatic transcription, but at present these data are not exploited in the project.

The corpus is currently under development. we have completed about a third of the expected final size. The creation of the corpus will be presumably finalized by the end of June 2010, and made publicly available to the community. Portions of the corpus will be extracted to create development and test sets for tuning and evaluation purposes. Some statistics of the parallel corpus are reported in Table 1; asterisks mark estimated statistics based on the actual partial corpus. An example of an AEWLIS-annotated sentence is shown and commented in Figure 2. Of course, Italian audio/video and LIS movie are not reported. The AEWLIS annotation is reported here in a simplified human-readable format, and only relevant information are reported.

The generation of the parallel corpus is time-consuming and expensive, because an intensive effort of skilled human is required both for signing and annotating. In order to get the best trade-off between the size of the training corpus and the cost for collecting it, smart solutions have been adopted: split sentences into small segments conveying (self-)consistent content and syntax; avoid duplicate or highly overlapping segments; incrementally collect segments which are more distant (for instance, with respect to Levenshtein distance) from those already gathered.

The corpus domain has caused the creation of new, intuitively iconic, signs for the human interpreters, for a number of concepts that would have caused long boring paraphrases. This is common practice among LIS speakers; in particular, LIS interpreters need to agree on a limited number of novelties in order to keep the variation among interpreters at a minimum. The LIS signed and annotated in the corpus is not "spontaneous" like in a conversation, but "genuine" like that produced by professionals interpreting, for example, TV programmes.

5. ATLAS Extended Written LIS

AEWLIS is a formal language defined within the ATLAS project and plays a two-fold role: it provides a symbolic representation for the annotation of the LIS corpus, and it is the interchange language between the **Translator** and the **Animator**. AEWLIS format encompasses both functions, but different and possibly overlapping subsets are employed for the two tasks. AEWLIS is translation-oriented in the sense that it contains all information required to (i) convey the meaning of the original Italian sentence and (ii) "instruct" or "pilot" the virtual character to fluently sign it. We know that there is no consensus about the possibility to encode a signed language in a written form. Indeed, we do not assert that AEWLIS is a "linguistic" written form of LIS: AEWLIS contains the necessary (phonologic, syntactic, semantic) information that the virtual character needs to properly realize the LIS sequence.

The annotation is performed at a sentence level, so links to other elements of other sentences of the same (or different) bullettin(s) are not allowed. Each sentence is split into a sequence of Time Slices (TSs), each defined as the time interval needed to perform a sign. A TS is considered atomic in the sign sequence. It is worth noticing that in the annotation phase we do not actually perform a time segmentation of the LIS sequence, but we simply associate a TS with each single sign. Indeed, the goal of the project is not the development of a sign recognizer, which would probably rely on such information.

AEWLIS includes three main kinds of annotation. The first level describes the meaning conveyed by the actual sign assigning a Lemma (or gloss) to each TS. The syntactic number (singular/plural) is possibly reported. A Sign-ID identifies the sign in the signary, if any[1].

The second level independently describes all Communication Channels relevant in LIS: Left and Right Hands, Direction, Body, Shoulder, Head, Facial, Labial, Gaze. Practically, only the modifications with respect to the neutral default for the corresponding sign are annotated. Specific annotation for the Left and Right Hands is given if they realize distinct signs contemporarily[2].

The third level provides a shallow syntactic/semantic structure of the sentence if available, by reporting for each TS its parent and its role. The main thematic roles proposed in (Petukhova and Bunt, 2008) are reported which can have a strong impact on the animation: agent/patient, initial/final location, etc. Topic and Focus (Hajičová et al., 1998) (Lillo-Martin and de Quadros, 2004) of the sentence are possibly annotated; the speech act specifies whether the sentence is declarative, imperative or interrogative.

Furthermore, AEWLIS has been defined as a set of independent annotation levels (*Tags*), which can be filled at different moments, or even left empty. The only mandatory tag is the Lemma. All the annotation Tags are associated to a single TS; thus the AEWLIS sentence can be graphically represented by a matrix, having as many columns as the number of TSs and as many rows as the number of Tags.

[1] For the sake of animation, unknown (out-of-signary) sign can be fingerspelled.

[2] This occurs for instance when the signer keeps the non-dominant hand, until a sign comes that requires both hands.

Italian	Per quanto riguarda i mari, generalmente mossi o molto mossi, poco mosso solo il Tirreno.										
AEWLIS	Position	1	2	3	4	5	6	7	8	9	10
	Lemma	MARE	PROPRIO	ZONA	IONIO	ADRIATICO	MOSSO	MOSSO	ZONA	TIRRENO	MOSSO
	Sign-ID	1349	1875	2100		3002	423	423	2100	3000	423
	Topic/Focus	topic		topic			focus	focus	topic		focus
	Facial							strong			mild

Figure 2: AEWLIS annotation of a sentence "Concerning seas, generally from slight to moderate, smooth Tirrenian sea only". Only the most significant and not empty Tags are reported. The sentence is practically split into three parts, Signs 1-2, 3-7 and 8-10. In each subparts there are specific topics and focuses. The general reference "geenralmente" ("generally") to the seas is interpreted by listing a few exemplars (Signs 4 and 5). The Facial Tag distinguishes between "mosso" and "poco/molto mosso" (empty, "mild", and "strong", Signs 6, 7 and 10, respectively).

6. Conclusion

This paper reported on the work by ATLAS project of defining a translation-oriented symbolic representation of LIS (AEWLIS) and of building a Italian-LIS parallel corpus annotated with AEWLIS. The AEWLIS language has been adopted both as an interchange format among translation and animation modules of the virtual interpreter and as annotation format for the corpus construction.

The first release of the corpus contains weather forecast bullettins, but ATLAS partners intend to significantly increase its size and extend it to other domains in order to make it a benchmark for further research on LIS. The corpus and other related linguistic resources developed as a side-effect of this research will be made freely available.

Acknowledgement

The work presented here has been developed within the ATLAS project, co-funded by Regione Piemonte within the "Converging Technologies - CIPE 2007" framework (Research Sector: Cognitive Science and ICT).

7. References

BlueSign. 2010. http://bluesign.dii.unisi.it/.

F. Brugnara, et al. 2000. Advances in automatic transcription of broadcast news. In *Proc. of ICSLP*, pp II:660–663, Beijing, China.

BSLCorpus. 2010. http://www.bslcorpusproject.org/.

J. Bungeroth, et al. 2006. A German Sign Language Corpus of the Domain Weather Report. In *Proc. of LREC*, pp 2000–2003, Genoa, Italy.

Jan Bungeroth, et al. 2008. The ATIS Sign Language Corpus. In *Proc. of LREC*, Marrakech, Morocco.

J. Chon, et al. 2009. Enabling access through real-time sign language communication over cell phones. In *43rd Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA.

DictaSign. 2010. http://www.dictasign.eu.

DizLis. 2010. http://www.dizlis.it.

Echo. 2010. http://www.let.kun.nl/sign-lang/echo.

ELAN. 2010. http://www.lat-mpi.eu/tools/elan/.

eLIS. 2006. http://elis.eurac.edu.

R. Elliott, et al. 2004. An Overview of the SiGML Notation and SiGMLSigning Software System. In *Proc. of LREC*, pp 98–104, Lisbon, Portugal.

eSign. 2009. http://www.sign-lang.uni-hamburg.de/esign/.

M. Filhol and A. Braffort. 2006. A sequential approach to lexical sign description. In *Proc. of the Workshop on Sign Languages, LREC*, Genoa, Italy.

Eva Hajičová, iet al. 1998. *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer.

HamNoSys. 2004. http://www.sign-lang.uni-hamburg.de/projects/hamnosys.html.

Matt Huenerfauth. 2006. *Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation*. Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

P. Koehn, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL - Demo and Poster Sessions*, pp 177–180, Prague, Czech Republic.

Leonardo Lesmo, et al. 2009. Legal modificatory provisions and thematic relations. In *ICON*, pp 352–357, Hyderabad, India.

Diane Lillo-Martin and Ronice Müller de Quadros. 2004. Structure and acquisition of Focus in ASL and LSB. In *Proc. of TISLR*, Barcelona, Spain, October.

Ian Marshall and Éva Sáfár. 2003. A prototype text to British Sign Language (BSL) translation system. In *Proc. of ACL*, pp 113–116, Morristown, NJ, USA.

mobileASL. 2010. http://mobileasl.cs.washington.edu.

Sara Morrissey. 2008. *Data-Driven Machine Translation for Sign Languages*. Ph.D. thesis, School of Computing, Dublin City University.

NGTCorpus. 2010. http://www.ru.nl/corpusngtuk/.

Volha Petukhova and Harry Bunt. 2008. Lirics semantic role annotation: Design and evaluation of a set of data categories. In *Proc. of LREC*, Marrakech, Morocco.

SignSpeak. 2010. http://www.signspeak.eu/.

SignWriting. 2009. http://signwriting.org.

William C. Stokoe, et al. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Linstok Press., Silver Spring, MD, 2 edition.

U.Boston. 2002. http://www.bu.edu/asllrp.

U.DePaul. 2008. http://asl.cs.depaul.edu/.

Tony Veale, et al. 1998. The Challenges of Cross-Modal Translation: English-to-Sign-Language Translation in the Zardoz System. *Machine Translation*, 13(1):81–106.