Building a Treebank in Universal Dependencies for Italian Sign Language

Gaia Caligiore[1], **Cristina Bosco**[2], **Alessandro Mazzei**[2]
[1]Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Università degli Studi di Torino
[2]Dipartimento di Informatica, Università degli Studi di Torino
gaia.caligiore@edu.unito.it, {alessandro.mazzei|cristina.bosco}@unito.it

CEUR-WS.org/Vol-2769/paper_29.pdf

Abstract

The Italian Sign Language (LIS) is the natural language used by the Italian Deaf community. This paper discusses the application of the Universal Dependencies (UD) format to the syntactic annotation of a LIS corpus. This investigation aims in particular at contributing to sign language research by addressing the challenges that the visual-manual modality of LIS creates generally in linguistic annotation and specifically in segmentation and syntactic analysis. We addressed two case studies from the storytelling domain first segmented on the ELAN platform, and second syntactically annotated using CoNLL-U format.

1 Introduction and research goals

The Italian Sign Language (LIS) is the natural language used by the Italian Deaf community. Signed languages have been extensively studied in the last years (Brentari, 2010). From a theoretical point of view, Signed languages are of interest to the linguistic domain since they are *multi-channels* natural languages, where the coexistence of different articulators (hands, face, lips, posture, feet, etc.) is the test-bed for the formalization of new linguistic theories or objects (e.g. (Huenerfauth, 2006)). From a practical point of view, there is a real necessity to design and realise automatic translators for Deaf communities (Bragg et al., 2019). We want to investigate LIS with the same means used for Vocal Languages (VL) and verify if, in doing so, LIS can be properly represented. In this context, language-specific strategies and resources should be developed. The reference framework in this work is the *Universal Dependencies* formalism[1] (UD), a *de facto* standard for syntactic annotation.

The main goals of this research are three. The first goal is theoretical. We want to investigate the expressiveness of UD tags and its relations with LIS. Signed languages have peculiar forms of lexicons and syntax and we want to experimentally verify the expressive power of the UD formalism in representing this richness. The second goal is theoretical as well. We want to determine the extent of the similarities and differences in syntactic constructions between Italian (as reported in the Italian-UD (Simi et al., 2014)), Swedish Sign Language (SSL, as reported in the SSL-UD treebank (Mesch and Schonstrom, 2018)) and LIS. At the present moment, the SSL is the only sign language that has been annotated on UD. The SSL treebank is comprised of 203 sentences taken from the Swedish Sign Language Corpus (SSLC) (Mesch and Schönström, 2018; Mesch and Wallin, 2015). Being the only reference for the construction of a treebank for a sign language, the SSL treebank was a fundamental resource for the choice of the direction to follow in the annotation of LIS, particularly with regard to the first step of the process, i.e. the segmentation on ELAN (see section 2.1). The third goal is more practical. We want to create a UD compliant resource for the syntactic annotation of LIS: the first LIS-UD treebank. To our knowledge, apart from the domain-specific bilingual corpus developed in the projects ATLAS and LIS4ALL on automatic translation (Mazzei et al., 2013; Geraci et al., 2014; Mazzei, 2015), this is the first attempt to use dependency relations for representing LIS syntax.

For building the corpus, we selected two case studies in the storytelling domain: all of the sentences of two LIS videos, namely the fairy tale *Cappuccetto Rosso* (Little Red Riding Hood) and

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://universaldependencies.org/

the story *I tre fratelli* (The three brothers, written by the Italian writer Grazia Deledda), were collected in the novel treebank that, at the moment, is comprised of 257 dependency trees. While in the *Cappuccetto Rosso* story the signer signs the story without an Italian reference text, in the *I tre fratelli* story the signer is translating from a well defined written Italian text. By using the full original version of these stories, we had to face the challenge represented by unrestricted real data. For instance, very long and complex sentences were annotated and translated into LIS.

Considering the intrinsic complexity of the task and the novelty of the project, in the preliminary release of data described in this paper we only addressed some of the features of LIS. For instance, the location in space of a sign is only annotated in the portion of analysis carried out on ELAN but was not transferred in the CoNLL-U files. Furthermore, non-manual elements – which are one of the fundamental means used by LIS signers to convey meaning (Volterra, 2004)– are not included in any annotation layer developed in this project. This is the result of a lack of a more appropriate annotation strategy that is specific to sign languages within the UD framework, which is originally developed for the analysis of VLs and, by default, does not include the possibility to annotate the features of a language that go beyond the alphabetical construction of a word (or gloss, in this case).

The paper is organized as follows. In the Section 2, the data collection and the morphological and syntactic annotation processes will be described. In Section 3, language-specific morpho-syntactic phenomena are discussed mainly focusing on pointing signs as Highly Iconic Structures (henceforth HIS). The strategies used to annotate signs and their dependency relations are justified. Section 4 concludes the paper by providing some issue on the future development of the project.

2 Data Annotation

In this Section, we describe the main steps for the realization of the annotation of the UD-LIS, the pre-processing, which consists in the analysis with ELAN, and the application of the tags and relations of the UD format for the generation of the CoNLL-U format of each LIS sentence.

2.1 Analysis on ELAN

Following the same annotation procedure applied for the SSL treebank (Mesch and Schonstrom, 2018), the ELAN platform was used for the identification, segmentation and definition of each sign of our corpus. ELAN (EUDICO Linguistic Annnotator)[2] is a computer software initially released in 2000 by the Max Planck Institute for Psycholinguistics in the Netherlands. ELAN is used to annotate audiovisual files manually and semi-automatically and allows annotators to tag video material frame by frame with information arranged on multiple lines that can be defined by the program itself or personalized by the annotator (Brentari, 2010). It is also a useful tool in multimodality research since it allows the user to manually create multimodal annotations, useful for the analysis of sign languages (Wittenburg et al., 2006). Mesch and Wallin – creators of the SSL treebank– state that "ELAN allows researchers to provide time-alligned annotations of a video file on parallel tiers, making it useful for representing individual articulators on separate tiers as they are used simultaneously to produce a single sign.". For these reasons, the software is considered to be the most used for sign language annotation (Branchini et al., 2013). As a result, four tiers were developed to describe the main qualities of a sign in this context: *segno, luogo, UD POS Tag, traduzione*.

Defining a strategy to gloss the signs included in the ***Segno*** (sign) tier was a challenging task since the final aim was that of providing unambiguous and easily retrievable glosses. The rule that is generally followed when writing down a sign gloss is to write the translation of the sign as it would be normally written in the vocal language, but in capital letters. This writing strategy might cause ambiguity since different variations of a sign can be translated with the same gloss. At this stage of the annotation process, glossing strategies adopted for LIS and for the SSL treebank were compared, in order to minimise ambiguity and create a system that refers to both languages. The developers of the SSL treebank started from the SSL corpus that, in turn, referred to the SSL Dictionary (Ostling et al., 2017). For this reason, it was decided that the most appropriate source for sign gloss annotation would be a well-established dic-

[2] https://archive.mpi.nl/tla/elan/download

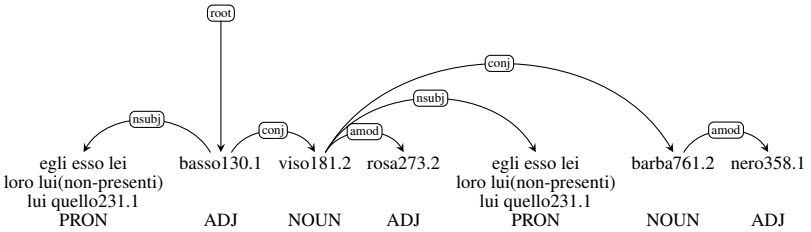

Figure 1: Lui è basso ed ha il viso rosa e la barba nera.

Piccolo e roseo, con una gran barba nera incolta.

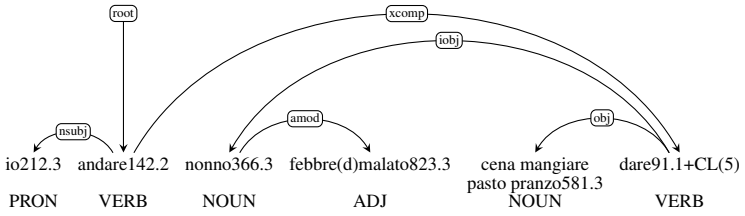

Figure 2: Io vado a dare da mangiare alla nonna malata.

tionary: the *Dizionario bilingue elementare della Lingua dei Segni Italiana LIS* (Radutzky, 1992) in its digital version, which is considered as the most easily retrievable and unambiguous collection of glosses within the context of LIS, including more than 2500 terms signed by native signers. In this dictionary, a gloss – made up of a translation of the sign into Italian and a sequence of numbers– is associated to each sign. For instance, the LIS sign for RED is glossed as "*rosso.202.1*"; for signs not found in the dictionary, glosses are taken from other resources, such as SpreadTheSign[3]. In this case, a translation in caps lock of the sign into Italian is associated to the code "-STS", as in the gloss "*NEMICO-STS*" (ENEMY). Lastly, if a sign is not found in any of the mentioned resources, it is autonomously developed with SignWriting (Renzo et al., 2010) and glossed with the code "SW-", as was done to gloss the sign TO PUT ON a sleeping cap: "*SW-indossare-cuffia*".

Luogo (location) defines where the sign is articulated in the signing space and is based on the 16 sign locations identified by Radutzky (Radutzky, 1992) which are *parte superiore del capo* (upper part of the head), *faccia* (face), *occhi* (eyes), *naso* (nose), *orecchie* (ears), *guancia* (cheeks), *bocca* (lips), *mento* (chin), *spalla* (shoulder), *petto* (chest), *gomito* (elbow), *polso* (wrist), *mano non dominante* (non-dominant hand), *tronco inferiore* (waist). The 16th and most used location of a sign is the Neutral Signing Space that is described by the acronym SN (*Spazio Neutro*).

UD POS Tag includes the UD Part of Speech Tag[4] associated to the sign. A peculiarity of this tier is that the POS Tag might not correspond to the form of the sign gloss. For the annotation of this tier, priority is given to the function of the sign within the sentence, rather than its gloss. In fact, in some cases the gloss associated to a sign will not correspond to the role of the sign in that specific syntactic context. For instance, the gloss of the sign *POESIA* (poesia-STS) suggests that the sign is a NOUN. Yet, in the *I tre fratelli* video, POESIA plays the role of an adjective. In the sign sequence *POESIA TEMPO*, the previously mentioned sign has the meaning of *"tempo poetico"* (poetic time). In this case, the UD POS Tag for the sign *POESIA* will be ADJ and not NOUN. This is because priority was given to the function of the sign within the sentence, rather than its gloss.

Traduzione (translation) provides a translation of the LIS sentence into spoken Italian. If the signs are a direct translation from Italian, as in the *I tre fratelli* video, the information included in this tier will be a word-for-word transcription of the spoken or written text. If the signer is not translating, as in the *Cappuccetto Rosso* video, the information included in the tier will be a translation into Italian that imitates the structure of the sentence in LIS as closely as possible. This tier was included to facilitate the understanding of a signed sentence given that the sequence of sign glosses will look

[3] https://www.spreadthesign.com/it.it/search/

[4] https://universaldependencies.org/u/pos/

fragmented. By providing a linear translation into spoken Italian of each sentence, a non-signer will be able to have a general understanding of the sentence. Furthermore, as mentioned in the previous section, POS Tags might not correspond to sign glosses. Therefore, by providing this translation, any discrepancies between sign glosses and POS Tags will be justified.

2.2 Annotation in UD :

The information encoded in ELAN was transferred in CONLL-U files and split into its ten columns, except for the LEMMA, DEPS and MISC columns, where no information is included.

2.2.1 Morphological Annotation

Sign glosses and UPOS Tags were included respectively in the FORM and UPOS columns, and coarse-grained tags taken from the Tanl POS Tagetset[5] were included in the XPOSTAG column. Language-specific annotation strategies can be found in the FEATS column where specific symbols and labels taken from different sources were used to provide more information on the peculiarities of a sign or of its production. Based on SSLC tags (Mesch and Wallin, 2015), sign types were marked with *@b* for finger-spelled signs and *@g* for gesture-like signs. Reduplicated signs were signalled with the feature tag *@RDP=true*, adapted from (Mesch and Schonstrom, 2018). Role shift is marked with RS= followed by the symbols ⟨ and ⟩ and the codes 3a, 3b, 1 or 2 which are used to identify the positioning of the signer in the orthogonal signing space (Pfau et al., 1987).

2.2.2 Syntactic Annotation

The annotation of the HEAD and DEPREL columns of the CoNLL-U format is at the very core of the research. As for the definition of a root node, the UD standard indicates that, in Italian, the root is usually a verbal predicate or a noun. If the verbal predicate is not present due to ellipsis, the root is moved to the leftmost dependent of the verbal predicate[6].

The strategy for finding the syntactic structure of each sentence consists in looking at Italian and SSL treebanks, choosing the most appropriate annotation with respect to the specific phenomenon to be annotated or the context where it occurs. For instance, for the annotation of particular signs we were inspired by the solution adopted in SSLC (see 2.2.1). Nevertheless, if no annotation is deemed to be adequate, a novel independent solution that seems most fitting is applied. If a sentence presented a unique combination of signs and no corresponding or similar dependency relations were found in these other treebanks, the construction of the dependency tree was based on an independent solution that is compliant with the general criteria adopted for LIS in the novel treebank.

In the following Section, a small selection of phenomena encountered in the treebank are described.

3 Annotation of Language-specific Morpho-syntactic Phenomena

In this section we discuss two LIS phenomena, i.e. pointing signs and repetitions, which we addressed in the development of the novel resource.

3.1 Pointing signs

In LIS, pointing signs are HIS realized using an extended index finger and carrying out several functions. For example, a pointing sign can function as a pronoun, as a determiner or as an adverb (Cormier et al., 2013) depending on the context in which it is performed. They are in all cases deictics found in PE-clauses[7], used to establish a location in space of a certain referent and create agreement in space. In the annotation of dependency relations, the second repetition of a deictic sign with an anaphoric function was either marked as dependent on the first one, or – as can be seen in figure 1– as dependent on a noun (viso181.2) that, in turn, is attached to the root.

Additionally, pointing signs can also be used by the signer to refer to himself or herself during role shift, that is, while impersonating a character, as in Figure 2. A sub-type of pronominal or determiner pointing signs are demonstrative pointing signs (Cormier et al., 2013, p.232), which could be compared to Italian demonstrative pronouns. When a pointing sign was used as an adverb of

[5] http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

[6] https://universaldependencies.org/it/dep/root.html

[7] PE-clauses are labelled as such by Branchini and Donati (Branchini and Donati, 2009). These clauses can be compared to relative clauses in spoken Italian and include a PE-marker that is a sign "[...] realized manually with the index finger stretched out and shaken downwards [...]" and is "[...] coreferential with an NP within the clause, and this coreferentiality can be realized through agreement in space".

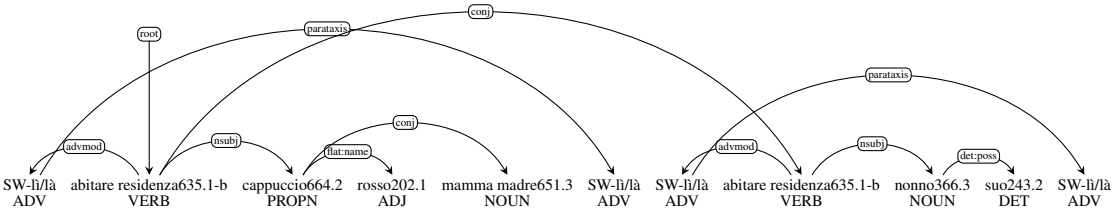

Figure 3: In una casa abitavano Cappuccetto Rosso e sua mamma, nell'altra abitava sua nonna.

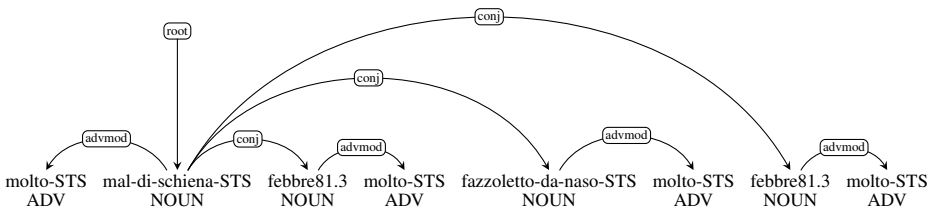

Figure 4: Mi fa molto male la schiena, ho molta febbre, sono molto raffreddata, ho molta febbre.

location, a specific sign was developed with SignWriting and glossed as *SW-lì/là*, as can be seen in Figure 3. A peculiarity found in sentence 3 is that the pointing sign with an adverbial function is reduplicated for each repetition of the verb *abitare residenza635.1-b* (TO LIVE), for a total of four times. To solve the problem that such repetition could pose in the identification of dependency relations, the four occurrences of the pointing sign were divided into two branches. The first occurrence was marked as dependent on the root verb and the second as dependent on the first one by means of the *parataxis* tag. The same strategy was used for the third and fourth occurrences, with the only difference that the third occurrence referred to the second repetition of the verb.

3.2 Repetition of non-pointing signs

The repetition of non-pointing signs by means of *reduplication* can be used to convey that an event is still ongoing or that is happening several times (Borstell, 2011). This strategy is used in the *Cappuccetto Rosso* fairy tale to intensify the action and express that it took place over a long time-span, to the point of excessiveness. For instance, in one of the occurrences of the LIS signs *CAMMINARE* (TO WALK) and *ASPETTARE* (TO WAIT), the signer reduplicated these signs and associated a facial expression characterized by wide-open eyes pointing towards the hands.

Another form of repetition of non-pointing signs occurs in the *Cappuccetto Rosso* fairy tale. In this case, the repetition cannot be classified as reduplication since the sign is not repeated through a circular movement (Borstell, 2011), The sign glossed as *molto-STS* in Figure 4 was repeated four times to emphasize the preceding or following signs referred to the state of health of the impersonated character. Due to the intensifying function of this sign – and the fact that it is found in a signed sentence on the website SpreadTheSign where it is translated into Italian as "*molto*" (very) – the sign was glossed as *molto-STS* and consequently marked with the POS Tag ADV in the UPOSTAG column of CoNLL-U files. In all occurrences of the sign *molto-STS* where repetition is not found, the sign is marked as dependent on the head noun or adjective, see Figure 4.

4 Conclusion

The paper describes the development of the novel resource LIS-UD, namely a treebank for LIS in the UD format[8]. Provided the preliminary stage of the project, the main aim of this work is to discuss the issues raised by the annotation in UD of a sign language. This issue has previously been only partially addressed in one other single project for the development of the SSL treebank, and so the novel resource can be the opportunity for better investigating it. Several future directions can be drawn for the development of our project. In particular, among these directions, we are planning first of all to draft a more detailed document about the annotation guidelines. This can help us in checking the material annotated until now, but also in driving the work of novel annotators.

[8]The current version of LIS-UD is available at: `https://github.com/alexmazzei/LIS-UD`

References

Carl Borstell. 2011. Revisiting Reduplication. Toward a description of reduplication in predicative signs in Swedish Sign Language. Master's thesis, Stockholm University, Faculty of Humanities, Department of Linguistics.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2019, Pittsburgh, PA, USA, October 28-30, 2019*, pages 16–31.

Chiara Branchini and Caterina Donati. 2009. Relatively different: Italian sign language relative clauses in a typological perspective. In Anikó Lipták, editor, *Correlatives Cross-Linguistically*, pages 157–191. John Benjamins Publishing Company, Amsterdam.

Chiara Branchini, Carlo Cecchetto, and Isabella Chiari. 2013. La lingua dei segni italiana. In Gabriele Iannaccaro, editor, *La linguistica italiana all'alba del terzo millennio (1997-2010)*, pages 369–404. Bulzoni.

D. Brentari. 2010. *Sign Languages*. Cambridge Language Surveys. Cambridge University Press.

Kearsy Cormier, Adam Schembri, and Bencie Woll. 2013. Pronouns and pointing in sign languages. *Lingua*, Volume 137:230–247.

Carlo Geraci, Alessandro Mazzei, and Marco Angster. 2014. Some issues on Italian to LIS automatic translation. The case of train announcements. In *Proc. of CLiC-it 2014, first Italian conference on computational linguistics*, December.

Matt Huenerfauth. 2006. Representing coordination and non-coordination in American Sign Language animations. *Behav. Inf. Technol.*, 25(4):285–295.

Alessandro Mazzei, Leonardo Lesmo, Cristina Battaglino, Mara Vendrame, and Monica Bucciarelli. 2013. Deep Natural Language Processing for Italian Sign Language Translation. In *AI*IA 2013: Advances in Artificial Intelligence - XIIIth International Conference of the Italian Association for Artificial Intelligence, Turin, Italy, December 4-6, 2013. Proceedings*, pages 193–204.

Alessandro Mazzei. 2015. Translating Italian to LIS in the rail stations. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 76–80, Brighton, UK, September. Association for Computational Linguistics.

Johanna Mesch and Krister Schonstrom. 2018. From Design and Collection to Annotation of a Learner Corpus of Sign Language. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 121–126. European Language Resources Association.

Johanna Mesch and Krister Schönström. 2018. From Design and Collection to Annotation of a Learner Corpus of Sign Language. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*. European Language Resources Association, may.

Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics (IJCL)*, Volume 20:102–120.

Robert Ostling, Carl Borstell, Moa Gardenfors, and Mats Wiren. 2017. Universal Dependencies for Swedish Sign Language. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 303–308, Gothenburg. Linkoping University Electronic Press.

Roland Pfau, Marcus Steinbach, and Annika Herrnann (Eds.). 1987. *A matter of complexity. Subordination in sign languages*. Walter de Gruyter Inc. and Ishara Press, Boston/Berlin and Preston, UK.

Elena Radutzky. 1992. *Dizionario bilingue elementare della lingua italiana dei segni: Oltre 2.500 significati*. Kappa, Roma.

Alessio Di Renzo, Luca Lamano, Tommaso Lucioli, Barbara Pennacchi, Gabriele Gianfreda, Giulia Petitta, Claudia Savina Bianchini, Paolo Rossini, and Elena Antinoro Pizzuto. 2010. *Scrivere la LIS con il Sign Writing. Manuale Introduttivo.* Consiglio Nazionale delle Richerche, Roma.

Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of LREC 2014*, page 83–90.

Virginia Volterra, editor. 2004. *La lingua dei segni italiana*. Il Mulino.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genova, Italy, May. European Language Resources Association (ELRA).