Translating Italian connectives into Italian Sign Language

Camillo Lugaresi
University of Illinois at Chicago
Politecnico di Milano
clugar2@uic.edu

Barbara Di Eugenio
Department of Computer Science
University of Illinois at Chicago
bdieugen@uic.edu

Abstract

We present a corpus analysis of how Italian connectives are translated into LIS, the Italian Sign Language. Since corpus resources are scarce, we propose an alignment method between the syntactic trees of the Italian sentence and of its LIS translation. This method, and clustering applied to its outputs, highlight the different ways a connective can be rendered in LIS: with a corresponding sign, by affecting the location or shape of other signs, or being omitted altogether. We translate these findings into a computational model that will be integrated into the pipeline of an existing Italian-LIS rendering system. Initial experiments to learn the four possible translations with Decision Trees give promising results.

1 Introduction

Automatic translation between a spoken language and a signed language gives rise to some of the same difficulties as translation between spoken languages, but adds unique challenges of its own. Contrary to what one might expect, sign languages are not artificial languages, but natural languages that spontaneously arose within deaf communities; although they are typically named after the region where they are used, they are not derived from the local spoken language and tend to bear no similarity to it. Therefore, translation from any spoken language into the signed language of that specific region is at least as complicated as between any pairs of unrelated languages.

The problem of automatic translation is compounded by the fact that the amount of computational resources to draw on is much smaller than is typical for major spoken languages. Moreover, the fact that sign languages employ a different transmission modality (gestures and expressions instead of sounds) means that existing writing systems are not easily adaptable to them. The resulting lack of a shared written form does nothing to improve the availability of sign language corpora; bilingual corpora, which are of particular importance to a translation system, are especially rare. In fact, various projects around the world are trying to ameliorate this sad state of affairs for specific Sign Languages (Lu and Huenerfauth, 2010; Braffort et al., 2010; Morrissey et al., 2010).

In this paper, we describe the work we performed as concerns the translation of connectives from the Italian language into LIS, the Italian Sign Language (Lingua Italiana dei Segni). Because the communities of signers in Italy are relatively small and fragmented, and the language has a relatively short history, there is far less existing research and material to draw on than for, say, ASL (American Sign Language) or BSL (British Sign Language).

Our work was undertaken within the purview of the ATLAS project (Bertoldi et al., 2010; Lombardo et al., 2010; Lombardo et al., 2011; Prinetto et al., 2011; Mazzei, 2012; Ahmad et al., 2012), which developed a full pipeline for translating Italian into LIS. ATLAS is part of a recent crop of projects devoted to developing automatic translation from language L spoken in geographic area G into the sign language spoken in G (Dreuw et al., 2010; López-Ludeña et al., 2011; Almohimeed et al., 2011; Lu and Huenerfauth, 2012). Input is taken in the form of written Italian text, parsed, and converted into a semantic representation of its contents; from this semantic representation, LIS output is produced, using a custom serialization format called AEWLIS (which we will describe later). This representation is then augmented with space positioning information, and fed into a final renderer component that performs the signs using a virtual actor. ATLAS focused on a limited domain for which a bilingual Italian/LIS cor-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 270–280,
Sofia, Bulgaria, August 4-9 2013.

pus was available: weather forecasts, for which the Italian public broadcasting corporation (RAI) had long been producing special broadcasts with a signed translation. This yielded a corpus of 376 LIS sentences with corresponding Italian text: this corpus, converted into AEWLIS format, was the main data source for the project. Still, it is a very small corpus, hence the main project shied away from statistical NLP techniques, relying instead on rule-based approaches developed with the help of a native Italian/LIS bilingual speaker; a similar approach is taken e.g. in (Almohimeed et al., 2011) for Arabic.

1.1 Why connectives?

The main semantic-bearing elements of an Italian sentence, such as nouns or verbs, typically have a LIS sign as their direct translation. We focus on a different class of elements, comprising conjunctions and prepositions, but also some adverbs and prepositional phrases; collectively, we refer to them as connectives. Since they are mainly structural elements, they are more heavily affected by differences in the syntax and grammar of Italian and LIS (and, presumably, in those of any spoken language and the "corresponding" SL). Specifically, as we will see later, some connectives are translated with a sign, some connectives are dropped, whereas others affect the positioning of other signs, or just their syntactic proximity.

It should be noted that our usage of the term "connectives" is somewhat unorthodox. For example, while prepositions can be seen as connectives (Ferrari, 2008), only a few adverbs can work as connectives. From the Italian Treebank, we extracted all words or phrases that belonged to a syntactic category that can be a connective (conjunction, preposition, adverb or prepositional phrase). We then found that we could better serve the needs of ATLAS by running our analysis on the entire resulting list, without filtering it by eliminating the entries that are not actual connectives. In fact, semantic differences re-emerge through our analysis: e.g., the temporal adverbs "domani" and "dopodomani" are nearly always preserved, as they do carry key information (especially for weather forecasting) and are not structural elements.

In performing our analysis, we pursued a different path from the main project, relying entirely on the bilingual corpus. Although the use of statistical techniques was hampered by the small size of the corpus, at the same time it presented an interesting opportunity to attack the problem from a different angle. In this paper we describe how we uncovered the translation distributions of the different connectives from Italian to LIS via tree alignment.

2 Corpus Analysis

The corpus consists of 40 weather forecasts in Italian and LIS. The Italian spoken utterance and LIS signing were transcribed from the original videos – one example of an Italian sentence and its LIS equivalent are shown in Figure 1. An English word-by-word translation is provided for the Italian sentence, followed by a more fluent translation; the LIS glosses are literally translated. Note that as concerns LIS, this simply includes the gloss for the corresponding sign. The 40 weather forecast comprise 374 Italian sentences and 376 LIS sentences, stored in 372 AEWLIS files. In most cases, a file corresponds to one Italian sentence and one corresponding LIS sentences; however, there are 4 files where an Italian sentence is split into two LIS sentences, and 2 files where two Italian sentences are merged into one LIS sentence.

AEWLIS is an XML-based format (see Figure 2) which represents each sign in the LIS sentence as an element, in the order in which they occur in the sentence. A sign's lemma is represented by the Italian word with the same meaning, always written in uppercase, and with its part of speech (*tipoAG* in Figure 2); there are also IDs referencing the lemma's position in a few dictionaries, but these are not always present. The AEWLIS file also stores several additional attributes, such as: a parent reference that represents the syntax of the LIS sentence; the syntactic role "played" by the sign in the LIS sentence; the facial expression accompanying the gesture; the location in the signing space (which may be an absolute location or a reference to a previous sign's: compare `HR` (High Right) and `atLemma` in Figure 2). These attributes are stored as elements grouped by type, and reference the corresponding sign element by its ordinal position in the sentence. The additional attributes are not always available: morphological variations are annotated only when they differ from an assumed standard form of the sign, while the syntactic structure was annotated for only 89 sentences.

(1) (Ita.) Anche sulla Sardegna qualche annuvolamento pomeridiano, possibilità di qualche breve scroscio di pioggia,
Also on Sardinia a few cloud covers afternoon[adj], chance of a few brief downpour of rain,
ma tendenza poi a schiarite.
but trend then towards sunny spells.

"Also on Sardinia skies will become overcast in the afternoon, chance of a few brief downpours of rain, but then a trend towards a mix of sun and clouds".

(2) (LIS) POMERIGGIO SARDEGNA AREA NUVOLA PURE ACQUAZZONE POTERE MA POI NUVOLA
Afternoon Sardinia area cloud also downpour can[modal] but then cloud
DIMINUIRE
decrease

Figure 1: Italian sentence and its LIS translation

```
<Lemmi>
  <NuovoLemma lemma="POMERIGGIO" tipoAG="NOME" ... endTime="2.247" idSign=""/>
  <NuovoLemma lemma="sardegna" tipoAG="NOME_PROPRIO" ... endTime="2.795" idSign="2687"/>
  <NuovoLemma lemma="area" tipoAG="NOME" ... endTime="4.08" idSign="2642"/>
  <NuovoLemma lemma="nuvola" tipoAG="NOME" ... endTime="5.486" idSign="2667"/>
  <NuovoLemma lemma="pure" tipoAG="AVVERBIO" ... endTime="6.504" idSign="2681"/>
  ...
</Lemmi><SentenceAttribute>
<Parent>
  <Timestamp time="1" value="ID:3"/>  <Timestamp time="2" value="ID:2"/>
  <Timestamp time="3" value="ID:3"/>  <Timestamp time="4" value="root_1"/>
  <Timestamp time="5" value="ID:3"/> ...
</Parent>
...
<Sign_Spatial_Location>
  <Timestamp time="1" value=""/>    <Timestamp time="2" value="HR"/>
  <Timestamp time="3" value="HL"/>  <Timestamp time="4" value="atLemma(ID:2, Distant)"/>
  <Timestamp time="5" value=""/> ...
</Sign_Spatial_Location>
...
<Facial>
  <Timestamp time="1" value=""/>                <Timestamp time="2" value="eye brows:raise"/>
  <Timestamp time="3" value="eye brows:raise"/> <Timestamp time="4" value="eye brows:-lwrd"/>
  <Timestamp time="5" value=""/> ...
</Facial>
</SentenceAttribute>
```

Figure 2: Example excerpt from an AEWLIS file

2.1 Distributional statistics for connectives

The list of Italian connectives we considered was extracted from the Italian Treebank developed at the Institute for Computational Linguistics in Pisa, Italy (Montemagni et al., 2003) by searching for conjunctions, prepositions and adverbs. This yielded a total of 777 potential connectives. Of those, only 104 occur in our corpus. A simple count of the occurrences of connectives in the Italian and LIS versions of the corpus yields the following results:

(a) 78 connectives (2068 occurrences total) only occur in the Italian version, for example *ALMENO (at least), CON (with), INFATTI (indeed), PER (for)* .

(b) 8 connectives (67 occurrences total) only occur in the LIS version, for example *CIRCA (about), as in* "Here I am"), *PURE (also, additionally)*.

(c) 25 connectives (925 occurrences total) occur in both versions.

For the third category, we have computed the ratio of the number of occurrences in Italian over the number of occurrences in LIS; the ratios are plotted in logarithmic scale in Figure 3. 0 on the scale corresponds to an ITA/LIS ratio equal to 1; positive numbers indicate that there are more occurrences in ITA, negative numbers that there are more occurrences in LIS. We can recognize three clusters by ratio:

(c1) 9 connectives occurring in both languages, but mainly in Italian, for example *POCO (a little), PIÚ (more), SE (if), QUINDI (hence)*.

(c2) 13 connectives occurring in both languages with similar frequency, for example *SOLO (only), POI (then), O (or), MA (but)*.

(c3) 3 connectives occurring in both languages, but mainly in LIS: *MENO (less), ADESSO (now), INVECE (instead)*.

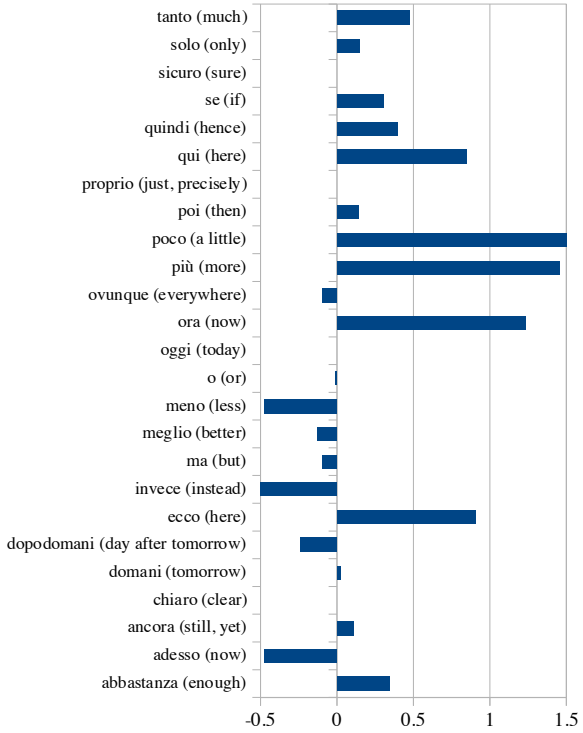

Figure 3: Ratio of ITA/LIS occurrences in logarithmic scale.

3 The effect of the Italian connectives on the LIS translation

From this basic frequency analysis we can already notice that a large number of connectives only appear in Italian, or have far more occurrences in Italian than in LIS. This is unsurprising, considering that LIS sentences tend to be shorter than Italian sentences in terms of number of signs/words (a fact which probably correlates with the increased energy and time requirements intrinsic into articulating a message using one's arms rather than one's tongue). However, our goal is to predict when a connective should be dropped and when it should be preserved. Furthermore, even if the connective does not appear in the LIS sentence as a directly corresponding sign, that does not mean that its presence in the Italian sentence has no effect on the translation. We hypothesize four different possible realizations for a connective in the Italian sentence:

- the connective word or phrase may map to a corresponding connective sign;
- the connective is not present as a sign, but may affect the morphology of the signs which translate words syntactically adjacent to the connective;
- the connective is not present as a sign, but its presence may be reflected by the fact that words connected by it map to signs which are close to each other in the LIS syntax tree;
- the connective is dropped altogether.

The second hypothesis deserves some explanation. The earliest treatments of LIS assumed that each sign (lemma) could be treated as invariant. Attempts to represent LIS in writing simply replaced each sign with a chosen Italian word (or phrase, if necessary) with the same meaning. Although this is still a useful way of representing the basic lemma, more recent studies have noted that LIS signs can undergo significant morphological variations which are lost under such a scheme. The AEWLIS format, in fact, was designed to preserve them.

Of course, morphological variations in LIS are not phonetic, like in a spoken language, but gestural (Volterra, 1987; Romeo, 1991). For example, the location in which a gesture is performed may be varied, or its speed, or the facial expressions that accompany it (Geraci et al., 2008). One particularly interesting axis of morphology is the positioning of the gesture in the signing space in front of the signer. This space is implicitly divided into a grid with a few different positions from left to right and from top to bottom (see HR – High Right, and LH – High Left, in Figure 2). Two or more signs can then be placed in different positions in this virtual space, and by performing other signs in the same positions the signer can express a backreference to the previously established entity at that location. One can even have a movement verb where the starting and ending positions of the gesture are positioned independently to indicate the source and destination of the movement. In other words, these morphological variations can perform a similar function to gender and number agreement in Italian backreferences, but they can also assume roles that in Italian would be performed by prepositions, which are connectives. In fact, as we will see later on, Italian prepositions are never translated as signs, but are often associated with morphological variations on related signs.

3.1 Tree Alignment

Two of our four translation hypotheses involve a notion of distance on the syntax tree, and a no-

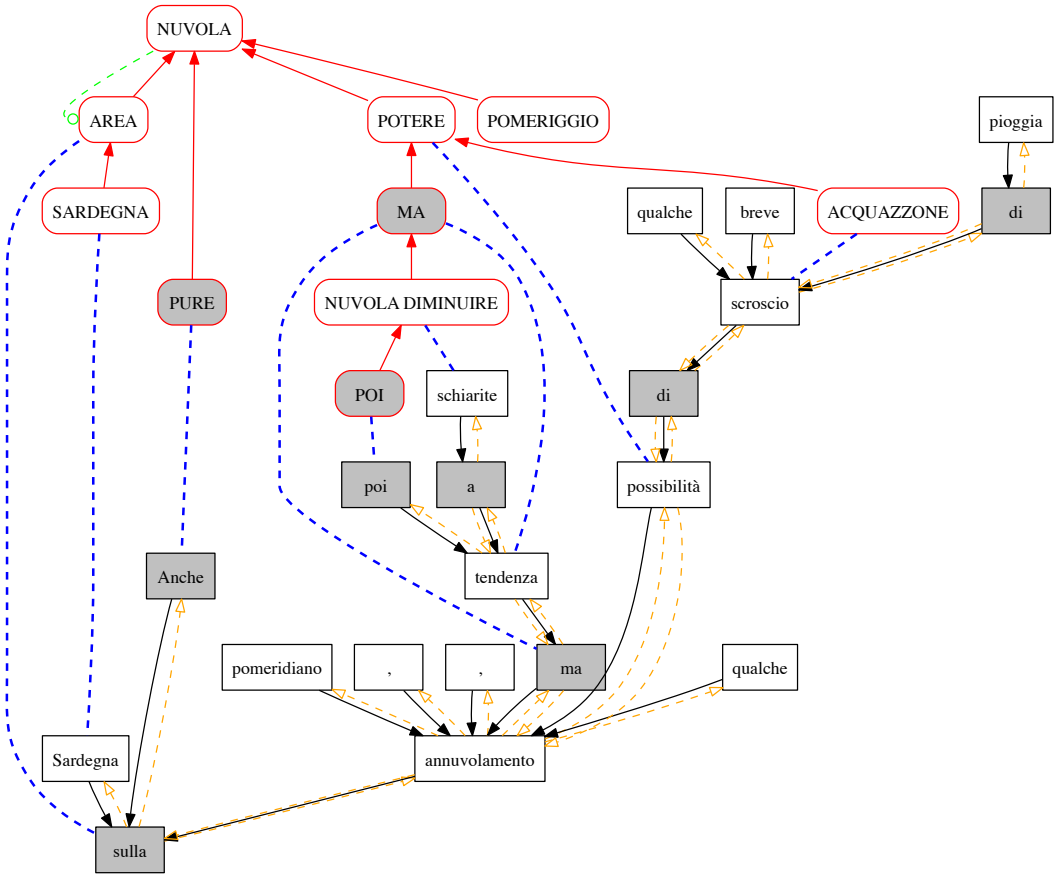

Figure 4: Example of integrated syntax trees.

tion of signs corresponding to words. Therefore, it is not sufficient to consider the LIS sentence and the Italian sentence separately. Instead, their syntax trees must be reconstructed and aligned. Tree alignment in a variety of forms has been extensively used in machine translation systems (Gildea, 2003; Eisner, 2003; May and Knight, 2007). As far as we know, we are the first to attempt the usage of tree alignment to aid in the translation between a spoken and a sign language, partly because corpora that include synctactic trees for sign language sentences hardly exist. (López-Ludeña et al., 2011) does use alignment techniques for translation from Spanish to Spanish Sign Language (SSL), but it is limited to alignment between words or phrases in Spanish, and glosses or sequences of glosses in SSL.

We have developed a pipeline that takes in input the corpus files, parses the Italian sentence with an existing parser, and retrieves / builds a parse tree for the LIS sentence. The two trees are then aligned by exploiting the word/sign alignment. A sample output is shown in Figure 4.

Italian sentence parsing. Since the corpus contains the Italian sentences in plain, unstructured text form, they need to be parsed. We used the DeSR parser, a dependency parser pre-trained on a very large Italian corpus (Attardi et al., 2007; Ciaramita and Attardi, 2011). This parser produced the syntax trees and POS tagging that we used for the Italian part of the corpus.

LIS syntax tree. One of the attributes allowed by AEWLIS is “parent”, which points a sign to its parent in the syntax tree, or marks it as a root (see Figure 2). These hand-built syntax trees are available in roughly 1/4 of the AEWLIS files. Because the size of our corpus is already limited, and be-

cause no tools are available to generate LIS syntax trees, for the remaining 3/4 of the corpus we fell back on a simple linear tree where each sign is connected to its predecessor. This solution at least maintains locality in most cases.

Word Alignment. Having obtained syntax trees for the two sentences, we then needed to align them. For this purpose we used the Berkeley Word Aligner (BWA) [1] (Denero, 2007), a general tool for aligning sentences in bilingual corpora. BWA takes as input a series of matching sentences in two different languages, trains multiple unsupervised alignment models, and selects the optimal result using a testing set of manual alignments. The output is a list of aligned word indices for each input sentence pair. On our data set, BWA performance is as follows: Precision = 0.736364, Recall = 0.704348, AER = 0.280000.

Integration. The result is an integrated syntax tree representation of the Italian and LIS versions of the sentence, with arcs bridging aligned word/sign pairs. Since some connectives consist of multi-word phrases, the word nodes which are part of one are merged into a super-node that inherits all connections to other nodes. Figure 4 shows the end result for the Italian and LIS sentences in Figure 1 (the two sentences are repeated for convenience at the bottom of Figure 4). The rectangular boxes are words in the Italian sentence, while the rounded boxes are signs in the LIS sentence. The Italian tree has its root(s) at the bottom, while the LIS tree has its root(s) at the top. Solid arrows point from children to parent nodes in the syntax tree. Gray-shaded boxes represent connectives (words or signs, as indicated by the border of the box). Bold dashed lines show word alignment. Edges with round heads show relationships where a sign has a location attribute referencing another sign. Arrows with an empty triangular head trace the paths described in the next section.

3.2 Subtree alignment and path processing

At this point individual words are aligned, but that is not sufficient. Our hypotheses on the effect of connectives on translation requires us to align a tree fragment surrounding the Italian connective with the corresponding tree fragment on the LIS side - where the connective may be missing. In effect, since we have hypothesized that the presence of a connective can affect the translation of the two subtrees that it connects, we would like to be able to align each of those subtrees to its translation. However, given the differences between the two languages, it is not easy to give a clear definition of this mapping - let alone to compute it.

Instead, we can take a step back to word-level alignment. We make the observation that, if two words belong to two different subtrees linked by a connective, so that the path between the two words goes through the connective, then the frontier between the LIS counterparts of those two subtrees should also lie along the path between the signs aligned with those two words. If the connective is preserved in translation as a sign, we should expect to find it along that path; if it is not, its effect should still be seen along that path, either in the form of morphological variations to the signs along the path, or in the shortness of the path itself.

The first step, then, is to split the Italian syntax tree by removing the connective. This yields one subtree containing the connective's parent node, if any, and one subtree for each of the connective's children, if any. The parent subtree typically contains most of the rest of the sentence, so only the direct ancestors of the connective are considered. Then, each pair of words belonging to different subtrees is linked by a path that goes through the connective in the original tree. Of these words, we select the ones that have aligned signs, and then we compute the path between each pair of signs aligned to words belonging to different subtrees. This gives us a set of paths to consider in the LIS syntax tree.

For example, let us consider the connective "di" between "possibilità" and "scroscio" in Figure 4.

- This node connects two subtrees: a child subtree containing "qualche, breve, scroscio, di, pioggia", and a parent subtree containing the rest of the sentence.
- From each subtree, a set of paths is generated: all paths extending from the connective to the leaves of the child subtree (for example "scroscio, qualche" or "scroscio, di, pioggia"), and the path of direct ancestors in the parent tree ("sulla, annuvolamento, possibilità").
- Iterate through the cartesian product of each

[1] http://code.google.com/p/berkeleyaligner/

Table 1: Translation candidates for connectives with more than 10 occurrences

Connective	ITA Occurrences	Sign	Location	Close	Missing
domani	71	67 (94.37%)	1 (1.41%)	2 (2.82%)	4 (5.63%)
dopodomani	15	14 (93.33%)	0 (0.00%)	0 (0.00%)	1 (6.67%)
mentre	28	26 (92.86%)	5 (17.86%)	0 (0.00%)	1 (3.57%)
o	37	37 (100.00%)	2 (5.41%)	6 (16.22%)	0 (0.00%)
però	10	9 (90.00%)	1 (10.00%)	1 (10.00%)	1 (10.00%)
ancora	72	44 (61.11%)	1 (1.39%)	3 (4.17%)	25 (34.72%)
invece	17	9 (52.94%)	1 (5.88%)	2 (11.76%)	6 (35.29%)
ma	51	29 (56.86%)	1 (1.96%)	2 (3.92%)	21 (41.18%)
poi	22	10 (45.45%)	2 (9.09%)	0 (0.00%)	10 (45.45%)
abbastanza	11	4 (36.36%)	1 (9.09%)	0 (0.00%)	6 (54.55%)
anche	89	33 (37.08%)	5 (5.62%)	1 (1.12%)	53 (59.55%)
ora	17	6 (35.29%)	1 (5.88%)	1 (5.88%)	10 (58.82%)
proprio	11	5 (45.45%)	0 (0.00%)	0 (0.00%)	6 (54.55%)
quindi	35	9 (25.71%)	1 (2.86%)	0 (0.00%)	25 (71.43%)
come	16	0 (0.00%)	1 (6.25%)	1 (6.25%)	14 (87.50%)
dove	28	0 (0.00%)	1 (3.57%)	0 (0.00%)	27 (96.43%)
generalmente	13	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (100.00%)
per quanto riguarda	14	0 (0.00%)	0 (0.00%)	1 (7.14%)	13 (92.86%)
piuttosto	13	0 (0.00%)	0 (0.00%)	0 (0.00%)	13 (100.00%)
più	57	0 (0.00%)	3 (5.26%)	2 (3.51%)	52 (91.23%)
poco	63	2 (3.17%)	3 (4.76%)	0 (0.00%)	58 (92.06%)
sempre	13	1 (7.69%)	0 (0.00%)	0 (0.00%)	12 (92.31%)
soprattutto	16	1 (6.25%)	1 (6.25%)	0 (0.00%)	14 (87.50%)
a	111	0 (0.00%)	18 (16.22%)	30 (27.03%)	66 (59.46%)
con	91	0 (0.00%)	20 (21.98%)	11 (12.09%)	62 (68.13%)
da	97	0 (0.00%)	26 (26.80%)	18 (18.56%)	62 (63.92%)
di	510	2 (0.39%)	92 (18.04%)	140 (27.45%)	312 (61.18%)
e	206	17 (8.25%)	34 (16.50%)	25 (12.14%)	140 (67.96%)
in	168	6 (3.57%)	37 (22.02%)	16 (9.52%)	113 (67.26%)
per	120	0 (0.00%)	7 (5.83%)	35 (29.17%)	82 (68.33%)
su	327	4 (1.22%)	121 (37.00%)	38 (11.62%)	190 (58.10%)
verso	18	0 (0.00%)	6 (33.33%)	1 (5.56%)	12 (66.67%)

pair of sets (in this case we have only one pair), and consider the full path formed by the two paths connected by the connective node (for instance, “sulla, annuvolamento, possibilità, di, scroscio, breve”).

- For each of these paths, take the signs aligned to words on different sides of the target connective, and find the shortest path between those signs in the LIS syntax tree; we call this the aligned path. For example, from “possibilità” and “scroscio” we find “POTERE, ACQUAZZONE”. If this process generates multiple paths, only the maximal ones are kept.

By looking at words within a certain distance of the connective, at their aligned signs, and at the distance between those signs in the aligned path, the program then produces one or more “translation candidates” for each occurrence of a connective:

- Sign: if the connective word is aligned to a connective sign in LIS, that is its direct translation;
- Location: if morphology variations (currently limited to the “location” attribute, see Figure 2) are present on a sign aligned to an It. word belonging to one of the examined paths, and the word is less than 2 steps away from the connective, that morphological variation in LIS may capture the function of the connective;
- Close: if two It. words are connected by a connective, and they map to signs which have a very short path between them (up to 3 nodes, including the two signs), the connective may be reflected simply in this close connection between the translated subtrees in the LIS syntax tree;
- Missing: if none of the above hypotheses are possible, we hypothesize that the connective has been simply dropped.

Table 1 shows the results of this analysis. It includes only connectives with more than 10 occurrences. For each connective and translation hypothesis, the shading of the cell is proportional to

the fraction of occurrences where that hypothesis is possible; this fraction is also given as a percent. Note that Sign, Location and Close candidates are not mutually exclusive: for instance, an occurrence of a connective might be directly aligned with a sign, but at the same time it might fit the criteria for a Location candidate. For this reason, the sum of the percents in the four columns is not necessarily 100.

k-means clustering (MacQueen, 1967; Lloyd, 1982) has been applied to the connectives, with the aforementioned fractions as the features. The resulting five clusters are represented by the row groupings in the table.

The first cluster contains words which clearly have a corresponding sign in LIS, such as "domani" (tomorrow). "Domani" and "dopodomani" are not actually connectives, while "mentre", "o" and "però" are. It is interesting to note that, while a logician might expect "e" (and) and "o" (or) to be treated similarly, they actually work quite differently in LIS: there is a specific sign for "o", but there is no sign for "e". Instead, signs are simply juxtaposed in LIS where "e" would be used in Italian.

The words in the second cluster also have a direct sign translation, but they are missing in the LIS translation around half of the time. Several words represent connections with previous statements or situations, such as "ancora" (again), "invece" (instead), "ma" (but). These appear to be often dropped in LIS when they reference a previous sentence, e.g. a sentence-initial "ma"; or when they are redundant in Italian, e.g. "ma" in "ma anche" ("but also"). Therefore, we think can see two phenomena at play here: a stronger principle of economy in LIS, and a reduced number of explicit connections across sentences.

The third cluster is similar to the second cluster, but with a higher percent of dropped connectives. This is probably related to the semantics of these five words. "Abbastanza" means "quite, enough", and in general indicates a medium quantity, not particularly large nor particularly small. It is no surprise that this word is more likely to succumb to principles of economy in language. "Anche" means "also", and is either translated as "PURE" (also) or dropped. This does not seem to depend on the specific circumstances of its usage; rather, it seems to be largely a stylistic choice by the translator. "Proprio" ("precisely", "just") has a corresponding sign "PROPRIO", but since it does not convey essential information it is a good candidate for dropping. "Quindi", meaning "therefore", has its own sign "QUINDI", but once again the causal relationship it conveys is usually not essential to understanding what the weather will be, and thus it is frequently dropped.

The fourth cluster consists of connectives which are largely simply dropped. Some of these are elements that just contribute to the discourse flow in Italian, such as "per quanto riguarda" ("concerning"); in fact, this connective mainly occurs in sentence-initial position in the Italian sentences in our corpus and denotes a change of topic from the previous sentence, corroborating our hypothesis of a reduced use of explicit inter-sentence connections in LIS. It may seem strange for comparative and intensity markers such as "più" (more) or "poco" (a little) to be so consistently dropped, but it turns out that intensity variations for weather phenomena are often embedded into a specific sign, for example "NUVOLOSITÀ AUMENTARE" (increasing cloud cover).

The fifth cluster contains all Italian prepositions (with 10 or more occurrences in the corpus), none of which is translated as a sign (the 6 occurrences for "in", the 4 for "su" and the 2 for "di" are due to alignment errors). We can conclude that prepositions do not exist in LIS as parts of speech; however, the prepositions in this cluster are often associated with morphological variations in the spatial positioning of related signs, which suggests that the role associated with these prepositions in Italian is performed by these variations in LIS. The conjunction "e" (and) also ends up in this cluster, although it has 8 legitimate sign alignments with "pure" ("too"); the rest are alignment errors. Unsurprisingly, all connectives in this class also have high ratings for the "close" hypothesis.

4 Rule extraction

We trained a classifier to help a LIS generator determine how an Italian connective should be translated. Because the translation pipeline we plan to integrate with is rule-based, we chose a Decision Tree as our classifier: this allows rules to be easily extracted from the classification model.

In order to identify a single class for each example, we ranked the four possible translation candidates as follows: Sign is the strongest, then Location, then Close, and finally Missing is the

child1_align = None ∩ word = Per_quanto_riguarda ∩ parent_align = None ⇒ Missing
child1_align = None ∩ word = Per_quanto_riguarda ∩ parent_align = PREVEDERE ⇒ Close
child1_align = None ∩ word = o ⇒ Align(O)
child1_align = None ∩ word = su ∩ child2_align_mykind = location ∩ child2_align = SICILIA ⇒ Location

Figure 5: Some rules extracted from the decision tree

weakest. Then, each example is labeled with the strongest translation candidate available for it: thus, for example, if the connective word appears to be translated with a connective sign, and the words it connects are also aligned to signs which are close to each other syntactically, then the class is Sign, not Close.

Our training data suffers from large imbalance between the "missing" class and the others. A classifier that simply labels all examples as "missing" would have an accuracy above 60%, and in fact, that is the classifier that we obtain if we attempt to automatically optimize the parameters of a Decision Tree (DT). We also note that, for connectives where both options are possible, choosing to translate them can make the sentence more verbose, but choosing to drop them risks losing part of the sentence's meaning: the worse risk is the latter. Following accepted practice with unbalanced datasets (Chawla et al., 2004), we rebalanced the classes by duplicating all examples of the Align, Location and Close classes, but not those of the Missing class.

On our data set of connectives with at least 10 occurrences, we trained a DT using AdaBoost (Freund and Schapire, 1997). The features include the word neighboring the connective in the Italian syntax tree, their aligned signs if any, part of speech tags, and semantic categories such as time or location. The resulting tree is very large, but we provide a few examples of the rules that can be extracted from it in Figure 5.

Bootstrap evaluation shows our DT to have an accuracy of 83.58% ± 1.03%. In contrast, a baseline approach of taking the most common class for each connective results in an accuracy of 68.70% ± 0.88%. Furthermore, the baseline classifier has abysmal recall for the Close and Location classes (0.00% and 0.85%, respectively), which our DT greatly improves upon (86.73% and 75.32%).

In order to estimate the impact of the lack of a LIS syntax tree in most of the corpus, we also learned and evaluated a DT using only the 1/4 of the corpus for which LIS syntax trees are available. The accuracy is 81.44% ± 2.03%, versus a baseline of 71.55% ± 1.74%. The recall for Close and Location is 89.22% and 73.58%, vs. 0.00% and 3.51% for the baseline. These results are comparable with the those obtained on the whole corpus, confirming that linear trees are a reasonable fallback.

Both clustering and classification were performed using RapidMiner 5.3. [2]

5 Conclusions and Future Work

The small size of our corpus, with around 375 bilingual sentences, posed a large challenge to the use of statistical methods; on the other hand, having no access to a LIS speaker prevented us from simply relying on a rule-based approach. By combining syntax tree processing with several machine learning techniques, we were able to analyze the corpus and detect patterns that show linguistic substance. We have produced initial results in terms of rule extraction, and we will be integrating these rules into the full Italian-LIS translation system to produce improved translation of connectives.

6 Acknowledgements

This work was supported by the ATLAS project, funded by Regione Piemonte within the "CIPE 2007" framework. Partial support to the authors was also provided by awards IIS 0905593 (from the NSF) and NPRP 5-939-1-155 (from the QNRF). A special thanks to A. Mazzei (ATLAS) for his willingness to answer our email bursts. Thanks to other members of ATLAS, in particular P. Prinetto, N. Bertoldi, C. Geraci, L. Lesmo; and to C. Soria, who extracted the list of potential connectives from the Italian Treebank.

References

Nadeem Ahmad, Davide Barberis, Nicola Garazzino, Paolo Prinetto, Umar Shoaib, and Gabriele Tiotto. 2012. A virtual character based italian sign language dictionary. In *Proceedings of the Conference Universal Learning Design*. Masaryk University.

[2] http://rapid-i.com/

Abdulaziz Almohimeed, Mike Wald, and R.I. Damper. 2011. Arabic Text to Arabic Sign Language Translation System for the Deaf and Hearing-Impaired Community. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 101–109, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Atanas Chanev, and Massimiliano Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 1112–1118.

N Bertoldi, G Tiotto, P Prinetto, E Piccolo, F Nunnari, V Lombardo, A Mazzei, R Damiano, L Lesmo, and A Del Principe. 2010. On the creation and the annotation of a large-scale Italian-LIS parallel corpus. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, CSLT*.

Annelies Braffort, Laurence Bolot, E Chtelat-Pel, Annick Choisier, Maxime Delorme, Michael Filhol, Jérémie Segouat, Cyril Verrecchia, Flora Badin, and Nad'ege Devos. 2010. Sign language corpora for analysis, processing and evaluation. In *Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6.

Massimiliano Ciaramita and Giuseppe Attardi. 2011. Dependency parsing with second-order feature maps and annotated semantic information. In H. Bunt, P. Merlo, and J. Nivre, editors, *Trends in Parsing Technology*, volume 43 of *Text, Speech and Language Technology*, pages 87–104. Springer.

John Denero. 2007. Tailoring word alignments to syntactic machine translation. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007*, pages 17–24.

Philippe Dreuw, Jens Forster, Yannick Gweth, Daniel Stein, Hermann Ney, Gregorio Martinez, Jaume Verges Llahi, Onno Crasborn, Ellen Ormel, Wei Du, et al. 2010. SignSpeak–understanding, recognition, and translation of sign languages. In *The 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010)*, pages 22–23, Valletta, Malta.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 205–208. Association for Computational Linguistics.

Angela Ferrari. 2008. Congiunzioni frasali, congiunzioni testuali e preposizioni: stessa logica, diversa testualità. In Emanuela Cresti, editor, *Prospettive nello studio del lessico italiano, Atti del IX Congresso della Società Internazionale di Linguistica e Filologia Italiana*, pages 411–416, Florence, Italy. Firenze University Press.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Carlo Geraci, Marta Gozzi, Costanza Papagno, and Carlo Cecchetto. 2008. How grammar can cope with limited short-term memory: Simultaneity and seriality in sign languages. *Cognition*, 106(2):780–804.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 80–87. Association for Computational Linguistics.

Stuart Lloyd. 1982. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137.

Vincenzo Lombardo, Fabrizio Nunnari, and Rossana Damiano. 2010. A virtual interpreter for the Italian Sign Language. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'10, pages 201–207, Berlin, Heidelberg. Springer-Verlag.

Vincenzo Lombardo, Cristina Battaglino, Rossana Damiano, and Fabrizio Nunnari. 2011. An avatar-based interface for the Italian Sign Language. In *2011 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pages 589–594. IEEE.

Verónica López-Ludeña, Rubén San-Segundo, Syaheerah Lufti, Juan Manuel Lucas-Cuesta, Julián David Echevarry, and Beatriz Martínez-González. 2011. Source Language Categorization for improving a Speech into Sign Language Translation System. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 84–93, Edinburgh, Scotland, UK, July.

Verónica López-Ludeña, Rubén San-Segundo, Juan Manuel Montero, Ricardo Córdoba, Javier Ferreiros, and José Manuel Pardo. 2011. Automatic categorization for improving Spanish into Spanish Sign Language machine translation. *Computer Speech & Language*.

Pengfei Lu and Matt Huenerfauth. 2010. Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive*

Technologies, pages 89–97, Los Angeles, California, June. Association for Computational Linguistics.

Pengfei Lu and Matt Huenerfauth. 2012. Learning a Vector-Based Model of American Sign Language Inflecting Verbs from Motion-Capture Data. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 66–74, Montréal, Canada, June. Association for Computational Linguistics.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA.

Jonathan May and Kevin Knight. 2007. Syntactic realignment models for machine translation. In *Proceedings of EMNLP*, pages 360–368.

Alessandro Mazzei. 2012. Sign language generation with expert systems and ccg. In *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, pages 105–109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, et al. 2003. Building the Italian syntactic-semantic Treebank. *Treebanks*, pages 189–210.

S. Morrissey, H. Somers, R. Smith, S. Gilchrist, and S. Dandapat. 2010. Building Sign Language Corpora for Use in Machine Translation. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010)*, pages 172–178, Valletta, Malta.

P. Prinetto, U. Shoaib, and G. Tiotto. 2011. The Italian Sign Language Sign Bank: Using WordNet for Sign Language corpus creation. In *2011 International Conference on Communications and Information Technology (ICCIT)*, pages 134–137.

Orazio Romeo. 1991. *Dizionario dei segni: la lingua dei segni in 1400 immagini*. Zanichelli.

Virginia Volterra. 1987. *La lingua italiana dei segni: la comunicazione visivo-gestuale dei sordi*. Il Mulino.