Italian Sign Language (LIS) Corpus

Mirko Santoro & Carlo Geraci
CNRS, Institut Jean-Nicod

Roadmap

- The IJN Sign Language Group
- Our (ID-)Practice
- Comparing (ID-)Practices
- Conclusions

The IJN SL Group

- Carlo Geraci
  - LIS Corpus (Linguistic Corpus)
  - Atlas (Avatar project: Politecnico of Turin & Univ. of Turin)
  - LIS4ALL (Avatar project: Politecnico of Turin & Univ. of Turin)
  - ELISIR (Avatar project: University of Turin & Venice)
- Valentina Aristodemo & Mirko Santoro (& Lara Mantovan, Ca' Foscari University, Venice)
  - LIS Corpus (Linguistic Corpus)
- Yann Cantin
  - Corpus-LSF-Paris (Linguistic Corpus under constrution)

The LIS Corpus project

- The PRIN 2007 project, ‘Dimensions of Variation in Italian Sign Language’ (PI Caterina Donati)
- 165 participants (from 10 cities) 1h. of recordings each
- **Three age groups**
  - young group between (18-30 years old)
  - medium group between (31-54 years old)
  - old group between (over 55 years old)
- **Task/data type**
  - Free conversation (≈45 minutes)
  - Wh-question elicitation task (≈ 5 minutes)
  - Spontaneous narration (≈10 minutes)
  - Picture naming task
- **No sign Bank (yet)**

Our template

(In collaboration with Kyle Duarte)

- Utterance
    - (ID-)Gloss tier 1 = Dominant hand
        - Phonology
        - Morphology
        - Syntax
        - Semantics
    - Dominant hand phonetics
    - (ID-)Gloss tier 2 = Non Dominant hand
        - …
        - …
    - Non-Dominant hand phonetics

My first 100 signs

Text type

- Narration

Number of glosses at the ID-gloss tier

- The first 100 signs for each signer (16500 tokens)

No sign bank

- Once additional tiers specify phonological and morphological properties

Roadmap

- The IJN Sign Language Group
- Our (ID-)Practice
- Comparing (ID-)Practices
- Conclusions

What do we have in mind?

- A research project (short-term perspective)
    - Collect some data
    - Quick and dirty results
    - Publish or perish

- A tool for research (long-term perspective)
    - Something that can be re-used
    - Something that we can add knowledge to
    - No need to publish soon

Conflicting perspectives

The researcher view

- expert linguist (at least in one field)
- the more information the better
- I want it yesterday

The annotator view (for the ID-GLOSS level)

- not necessarily a linguist (student, informant, signer, Deaf)
- few information
- maybe tomorrow

Data analyser view

- Possibly a linguist
- Columns & cells with non-overlapping values

ID-glossing

- Gloss tier 1 = Dominant hand
- Gloss tier 2 = Non-Dominant hand

Why?

- It is a phonological criterion (happy linguist :-)
- The annotator does not have to switch tier
- Data extraction can be done only ones

What if I am interested in handedness switching?

- (ID-)glosses are not suited for that.
- Other tiers are needed

ID-Glossing = memory task?

Some rules of thumb showing that the ideal world is not so perfect after all:

1. **The task must be simple (few specific knowledge required)**
   - No long training, no long term memory overload
2. **Avoid complex procedures (few things at a time)**
   - No procedural memory overload
3. **Avoid conventions (the less number of symbols the better)**
   - No short term memory overload
4. **Avoid ambiguities (conflict with avoid conventions)**
   - No short term memory overload

The task must be simple

Mechanical tasks

1. Select tier

4. Select the duration

6. Enter basic annotation

8. Add extra symbols

Linguistic tasks

2. Identify the sign
3. Apply criteria for sign boundaries

5. Remember basic symbols

7. Remember extra symbols

Identify the sign

The criteria are theoretically based after Brentari (1998).

We look at the dynamic component of the sign:

- HS change
- Or change
- Loc change

In case of complex movements, we use the more proximal movement as the reference movement.

Symbols (i)

What theory of the lexicon?

- Brentari and Padden (2000)

Core signs

- Italian word: MAMMA (mummy)

ID-Glossing?

- No special coding for lexical or phonological variants at this stage
  - We need further levels of phonology and morphology to be spelled out
- MACCHINA (car) ≠ GUIDARE (drive)

Symbols (ii)

Special signs

- Pronouns = IX-+ number of person (IX-1, IX-2, IX-3)
- Buoys = IX-LOC (+ additional info on a separate tier)
- Classifiers = PASSARE-CL (meaning + symbol: want more? more tiers)
- Fingerspelling = C-I-A-O

Symbols (iii)

Extra Phonological information

- no extra information is added at the (ID-)gloss level
- everything is added in dependent tier(s) under phonology

Extra Morphological information

- Pointing sign: Person & Locative function is added (IX-1, IX-2, IX-3, IX-LOC). Is it really relevant? (maybe not, definitely redundant)
- Negative incorporation: -NEG is added
- Compounds: “-” separates the two (or more) stems

Conventions (i)

Basic conventions imposed by Italian morphology

- no inflection on verbs (infinitival forms ***guid* vs. guidare**)
- adjectives always in masculine singular
- nouns always singular
- MACCHINA (car) ≠ GUIDARE (drive)

Conventions (ii)

Special conventions/symbols

- CL
- segno-nome (= name-sign)
- IX
- ?
- NEG
- -

Avoid ambiguities vs. avoid conventions

is “-” ambiguous in our notation language?

- MOTHER-FATHER (separate compounds)
- C-I-A-O (separate fingerspelling)
- PASSARE-CL (identifies classifiers)
- METTERE-A-POSTO (PUT)

Notice that:

- “-” is not ambiguous. It means: one single gloss is not enough to describe the sign
- Notice that to avoid ambiguity new symbols and new conventions are required

Roadmap

- The IJN Sign Language Group
- Our (ID-)Practice
- Comparing (ID-)Practices
- Conclusions

Comparing Phonological info

	BSL	NGT	LIS	Summary
2 hands vs 1 hand	Y	Y	Y	same
Pointing	ø	Y	ø	LIS and BSL are simpler
Classifiers	Y	Y	ø	LIS simpler

Comparing Morphological info

	BSL	NGT	LIS	Summary
Pointing	Y	Ø	Y	**NGT is simpler**
Compound	^	-	-	**same**
Neg-incorporati	-NOT	-NOT	-NEG	**same**
Directional verb	Ø	only1	Ø	**LIS and BSL are simpler than NGT**
Plurality	Ø	PL	Ø	**LIS and BSL are simpler than NGT**
Classifiers	Y	Y	Ø	**LIS simpler**

Comparing Special signs

	BSL	NGT	LIS	Summary
Buoy	sem.+BUOY	COUNTING-HAND-…	IX-LOC	**LIS is simpler**
Lexical Variants	1, 2, 3, …	A, B, C,…	Ø	**LIS is simpler**
Numbers	ONE	1	ONE	**NGT is simpler**
Fingerspeling	FS: WORD	#:WORD	W-O-R-D	**LIS is more complex**
Pointing	PT:…	PT…	IX-number IX -LOC IX-POSS-number	**LIS is simpler**

Comparing Special signs

	BSL	NGT	LIS	Summary
Classifiers	sym+…	sym+…	-CL	**LIS is simpler**
Gesture+	G:…	%	gesture	**same**
Construed Action	G:CA:…	%	-CL	**LIS is simpler**
Number sequence	NINETEEN^EIGHT^NINE	1989	MILLENOVECENTOOTTANTANOVE	**LIS and BSL are more complex**
Sign-names	…	…	SEGNO-NOME	**LIS is simpler**
Number incorporation	HOUR-FOUR	HOUR-4	QUATTRO-ORA	**LIS and BSL are more complex**

Discussion

- Overall, LIS (ID-)glosses are simpler than BSL and NGT
  - The task of the annotator is simpler (close to 0 interpretation of data or phenomena)
- LIS (ID-)glosses have more in common with BSL than with NGT
- More tiers are required to get the same amount of information
  - The ELAN template is overall more complex (hide tiers is the key feature)

Roadmap

- The IJN Sign Language Group
- Our (ID-)Practice
- Comparing (ID-)Practices
- Conclusions

Conclusions

- Annotation is not an easy task
- Different and conflicting perspectives have to be taken into account even at the very basic level of (ID-)glosses
- Our practice avoid
  - additional phonological info
  - additional morphological info
- Linguistic phenomena are to be glossed at the relevant (dependent) tier
- LIS glosses are more similar to BSL than NGT

Many thanks to …

Fabio Poletti
Kyle Duarte
Lara Mantovan
Valentina Aristodemo
Yann Cantin

Part of this work is realized by the contribution of ERC Advanced Grant FRONTSEM (PI: Philippe Schlenker)
& Labex Funds (PI: Carlo Geraci)

References

Brentari, Diane. 1998. A prosodic model of Sign Language Phonology. Cambridge, MA: MIT Press.

Brentari, Diane & Carol A. Padden. 2000. Native and Foreign Vocabulary in American Sign Language: A Lexicon With Multiple Origins. In Diane Brentari (ed.), Foreign Vocabulary in Sign Language: A Cross-linguistic Investigation of Word Formation, 87–119. Mahwah, NJ: Lawrence Erlbaum Associates.

Discussion: is it simple enough?

- **On the excessive load (10 min.)**
  - Is it possible to find a reasonable compromise?
  - How long is the training of an (ID-)annotator before s/he can provide reliable annotations?
- **Is the “extra” in the ID-Gloss is necessary? (10 min.)**
  - To what extent the use of Regular expressions in “ELAN searches” may help avoiding complex (ID-)practices?
  - Can we shift the burden of complexity on the shoulders of the researcher not the annotator?