On Elicited Data in Sign Language Syntax

Jessica Lettieri (Institut Jean-Nicod, Département d'Etudes Cognitives
ENS - EHESS - CNRS, PSL University.)
jecla1995@gmail.com
Mirko Santoro (CNRS, Structures Formelles du Langage.)
mirko.santoro@cnrs.fr
Carlo Geraci (Institut Jean-Nicod, Département d'Etudes Cognitives
ENS - EHESS - CNRS, PSL University.)
carlo.geraci76@gmail.com

Abstract

We present a proof of concept that sign language elicited data are as reliable as experimental data. We use data reported in the recently published reference grammar of Italian Sign Language (Branchini and Mantovan 2020) to create a formal experiment contrasting grammatical sentences with sentences that minimally violate the rule. On a forced-choice task, participants prefer sentences that follow the rule significantly more often than sentences that violate it. The experimental results obtained in this study proves that elicited data are reliable.

Keywords: Research Methods, Playback Method, Experimental Data, Reliability, Elicited Data.

1 Introduction

Data are probably the key aspect of every scientific research. They are essential to validate theories, but most importantly to falsify them. This is why their reliability is so crucial to the scientific method. Linguistic data are typically of three types: elicited data, experimental data, and corpus data. This is true for both signed and spoken languages.

Although elicited data have been extensively used in both fieldwork and theoretical linguistics, their reliability has been heavily questioned (i.a., Hill 1961, Spencer 1973, Gibson and Fedorenko 2013, and Wasow and Arnold 2005). The main point made in these works is that judgments coming from very few consultants do not constitute a representative sample of the actual population. Put it in a more scientific perspective, this boils down to the risk of incurring in Type I error (Sprouse and Almeida 2012). Namely, that a contrast found with few speakers is taken as the base to propose a grammatical rule that is valid for the entire community of users and more generally as part of the grammar of the language, while it can be the case that the contrast is only an idiosyncratic property of that small sample. This is in a nutshell the main problem of elicited data. The sampling problem is magnified by the fact that in addition to having few consultants, the total number of data points is very small and it is generally made without varying the lexical material. This increases the chances that
a particular contrast is dependent on the particular lexical material used to construe the research paradigms. This fact undermines not just reliability but also the replicability of the contrast (e.g., reproducing the contrast changing irrelevant factors), since normally grammatical rules, especially syntactic ones, are taken to be neutral with respect to the lexical nature of the items that compose a sentence. These problems are typical of both traditional fieldwork and armchair-type of data collection as both are based on the intuitions from few consultants and using limited linguistic materials (Sprouse and Almeida 2012).

However, although it seems rudimentary as a methodology, its greatest merit is that with a few data points it provides clear negative evidence of what is part of the competence of the speakers and signers of a language (see Section 2 for an illustration).

Improvement to elicited methods have been proposed in the literature. Mahowald et al. (2016) propose to use at least five consultants and at least five tokens per type per paradigm. In the sign language literature, Davidson (2020) and Kimmelman (2021) analyzed the pros and the cons of the playback method proposed in Schlenker, Lamberton, and Santoro (2013) and the degree of collaboration of Deaf informants, consultants and researchers. Davidson (2020) suggests to adopt practices and procedures normally used in experimental works; Kimmelman (2021) offers an overview of how elicitation works with the additional critical points that sign language elicited data may bring, in particular those related to modalities like the non-manual components. Both authors conclude that the reliability of sign language elicited data must be further explored.

These issues are at the center of a vivid debate also in spoken languages. In fact, several studies have been conducted to independently validate the contrasts obtained with elicited data (i.a., Sprouse and Almeida 2012, Sprouse and Almeida 2017, Marty, Chemla, and Sprouse 2020). The rationale of these bulk of studies is to replicate the contrasts that are reported in the literature using a different method of data collection. Specifically, they identify a number of contrasts and constructions reported in reference grammars, or in linguistic journals and build a formal experiment trying to replicate those contrasts. The main advantage of this replication is that formal experiments control for three major factors that are normally missing in elicited data: the sample of participants is larger than just a couple of informants, the number of tokens is adequate to the experimental standards, and the statistical analyses provide an independent measure of the contrast. So the logic goes, if elicited data are reliable, the same contrasts should be found when the data are collected via formal experiments, or to be more precise, if contrasts can be experimentally replicated, then elicited data are as reliable as experimental data.

Here, we focus on the reliability of elicited data in sign language, more specifically of Italian Sign Language (LIS). Branchini and Mantovan recently edited a reference grammar of LIS. Most of the data contained in that work come from elicited judgments reported in the literature about LIS. We use these data to build a formal experiment to test their reliability, in a very similar way as Sprouse and Almeida (2012) did for David Adger’s Core Syntax (Adger 2003).

The rest of the paper is organized as follows: In Section 2, we illustrate how the technique

---

1. Specifically, Kimmelman (2021, 562) points out that the simultaneity of manual and non-manual information may require the addressee to split attention between the face and the hands of the signer.

2. These data have been then replicated by the researchers involved in the compilation of the grammar in collaboration with seven Deaf consultants, with whom data have been discussed. Unfortunately, we do not know which particular technique was used. The second author of the paper participated in compiling the Grammar of LIS. His contribution, though, was limited to compound signs and he did not work to any of the phenomena included in this study. Finally, what is described in the next section is probably a strictest variant of elicitation method used in Branchini and Mantovan (2020).
of elicited data works. We present the methodology used to create our own experiment in Section 3. In Section 4, we present the results and in Section 5 we discuss them. Finally, Section 6 concludes the paper.

2 Elicited Methods

The literature on elicited data and fieldwork is vast and discusses all sorts of aspects from the basic how-to-do, to ethical issues (a.o., see Bowern 2008; Schütze and Sprouse 2013). In this section, we briefly describe the playback method (Schlenker, Lamberton, and Santoro 2013), as it is more and more used as a technique to elicit sign language data (Davidson 2020; Kimmelman 2021). The method is designed for hypothesis testing but can be used more generally also for language description purposes (Jaber, Donati, and Geraci 2022). The procedure can be summarized in six steps (see Fig.1):

1. Definition of the paradigm
2. Recording the paradigm from one informant
3. Playing-back the paradigm to the informant(s)
4. Recording the acceptability judgments
5. Discuss possible issues
6. Repeat steps 3-5 at least once

The degree of acceptability of the tokens in a paradigm is then established by taking the average value as computed by all the assessments of the paradigm.

Figure 1: Illustration of the playback method

The definition of the research paradigm consists in the identification of the hypothesis to be tested, which is followed by the creation of a list of minimally different expressions to be used to test the hypothesis. To illustrate, one of the paradigms used to study argument omission in French sign language (LSF) by Jaber, Donati, and Geraci (2022) consisted of the small dialogue in (1-2), where the question asked by signer A is followed by either (2a) or (2b). The paradigm intends to test the possibility of having post-verbal subjects, which are normally
associated to pro-drop in rich agreement languages like spoken Italian and Spanish. If the results are positive, the test would represent an argument to claim that LSF, where null arguments are allowed, is of the pro-drop type like Italian. So, the discourse in (1-2) represents a complete paradigm.

(1) **Signer A:** KID IX-3 CRY WHY

(=33 in Jaber, Donati, and Geraci 2022)

‘Why is the kid crying?’

(2) **Signer B:** a. LEAVE PIERRE

b. PIERRE LEAVE

‘Pierre left.’

The second step consists in video recording the whole paradigm, including possible scenarios. Here, no judgment is elicited from the consultant, who simply has to produce the linguistic expressions as naturally as possible and in the most uniform way. In the case of (1-2), the researchers separately recorded the question and the two possible answers. The third and fourth steps normally happen one or two weeks after the video recording sessions. In the third step, the researchers plays back the two discourse stretches, one with (2a) as the answer to (1) and then (2b) as a possible answer. The consultant is free to watch each example as many times as they wanted. In the fourth step, the consultant is asked whether the sentence in (2a) is acceptable and a possible answer to the question uttered by Signer A, etc. The judgment is normally recorded in written form by the researcher, except for particularly complex cases, where the informant is also asked to provide an explanation in sign language, which is then video-recorded.\(^3\) This discussion basically represents the fifth step, which is done concurrently with the fourth one. The sixth and final step consists of a repetition of steps from 3 to 5 and is used to test for consistency in the judgments.\(^4\) Notice that by itself, this step is already a measure of reliability, because it tests whether the intuition about the contrastive examples is retained over multiple testing occasions. In case of multiple consultants participating to the project, the steps in 1 and 2 are normally conducted with one consultant only, while the steps in 3, 4 and 5 are performed with each consultant.

Before concluding this section, it is worth noticing two key aspects of this methodology of data collection, when applied to sign languages. One concerns the sociolinguistic background of the informants; the other concerns how the testing examples are recorded (i.e., the shape of the linguistic expressions). While in spoken languages native speakers typically represent the majority of the linguistic community, in the case of sign languages, native signers only represent a very small portion of the signing community. This raises an important methodological issue that concerns the type of language that is described by works that employ native signers only. Of course, there are theoretical and empirical reasons why native users of a language should be the one that contribute to this type of data collection, although from a sociological and anthropological perspective the language described is not the one used by the majority of users.

Coming to video-recordings, these are essential because, no signing community has developed a writing form for their sign language. This is also crucially different from spoken

---

\(^3\) Variants of the paradigm require a more systematic involvement of the consultant by always video-recording the judgments and asking them to type it in written form. The choice of which option is the best depends on several factors including availability of the consultant.

\(^4\) There is no fixed number of iterations. As a rule of thumb, if the judgments are sharp (e.g., full acceptable or clearly unacceptable) and are replicated once, it does not make sense to reiterate the step with the same consultant. However, if the consultant is undecided more iterations could be needed to better understand the nature of difficulties.
languages where written paradigms are frequently used especially in syntactic and semantic works. One potential problem of using video-recorded sentences is that the reason for unacceptability may be due to additional prosodic cues that are extremely difficult to control and manipulate. When the paradigm is presented in the written form, this problem is bypassed. At the moment, there is no technological solution available to linguists to bypass the problem.5

3 Methodology

The methodology used to validate the reliability of elicited data in LIS is very similar to the one used for English in Sprouse and Almeida (2012). Here we use the recently published reference grammar for LIS (Branchini and Mantovan 2020) as the main source of linguistic contrasts. The volume contains several parts describing all components of LIS, we only focused here on sixteen constructions taken from the morphology and syntax parts. These cover broad phenomena like presence vs. absence of directionality (e.g., agreement morphology), canonical word order and basic sentence structure including modals and negation, and typical A'-constructions, like wh-movement, movement of negative quantifiers and across the board movement. The full list of constructions is given in the Appendix. For each construction, we extrapolated the general rule, took the example that illustrates it and created a sentence that minimally violates it. For instance, looking at the rule underlying the distribution of negative signs we created the paradigm in (3).

(3) **Position of negation.** Negation is postverbal in LIS.
   **Type of violation:** Negation is pre-verbal

   a. Grammatical construction:
      
      \[ \text{MARIA CAT SEE NEG} \]

   b. Rule violation:
      
      \[ \text{MARIA CAT NEG SEE} \]

      ‘Maria didn't see the cat.’

For each of the sixteen constructions, we created four pairs of sentences. For each pair, one sentence follows the rule prescribed by the construction, the other violates it. Three additional pairs were also created to be used as training items. A total of 128 testing sentences (16*4*2 sentences) plus 6 for training items were recorded with a Deaf native signer of LIS (second author of the paper).

Each pair was then assembled into a unique video in which each sentence was introduced by a slide indicating that the first or the second video was about to start. The order between sentences that follow and violate the rule has been balanced across the 64 pairs. Four lists of 16 pairs of videos were then created so that each list contained one type of every construction. The four lists plus the 3 pairs of training items were then used to create the experiment on the Labvanced on-line experimental platform.

A forced-choice answer was requested. Participants watched videos with the two sentences and were asked to answer the question “Which sentence is better?” by clicking one of the two buttons. An example of the testing slide is given in Figure 2.

---

5 A potential solution is to use signing avatars to create video sequences of signs. At the time of writing, the technology is not developed enough to be used for this purpose.
The experiment started with a welcome video in LIS, followed by the instruction always in LIS and then by the training items. After this first phase, the actual experiment started. At the end of the experiment, participants were asked to provide some biographical information (age, residence, age of acquisition, etc.) through an on-line questionnaire.

Twenty-four participants took part in the experiment (7 females, age range 74-23). Some of them were recruited through acquaintances and others through contacts with people working in the Calabria regional section of the major Italian Deaf association, Ente Nazionale Sordi. Everyone participated voluntarily; no compensation was given. Thirteen participants were native signers of LIS, 6 have been exposed to LIS before the age of 6 (early learners) and 5 were late learners. They all grew up in the south part of peninsular Italy: 16 in Calabria, 4 in Puglia, 2 in Campania, 2 in Basilicata. Most of them still live in the same areas: 17 in Calabria, 5 in Puglia and 2 in Emilia Romagna, which is the only region not in the South of Italy in our sample.

4 Results

The final dataset consisted of 379 observations. 5 data points have been removed because the software revealed a negative response time. This might have happened because some participants pressed the response button too quickly.

Participants chose the sentence that followed the LIS rule in 70% of the cases, while the violation sentence was chosen in 30% of the cases. A graphical distribution of the percentages is shown in Figure 3.

A generalized mixed model of the binomial family with the expected result as fixed effect and expected result by participant as random factor revealed that this difference is significant. The estimate for the fixed effect is 2.1906, p < .001. The model parameters are: AIC = 454.5, BIC = 478.2, logLik = -221.3.
Unfortunately, given the limited number of participants, no further inferential statistics can be conducted. However, qualitative investigation of the distribution of the answers reveals that we failed to replicate two of the patterns identified in Branchini and Mantovan (2020). Curiously, they both involve relative clauses. We report here the paradigms with the rules and their violations, and the distribution of the participants’ answers for these two cases.

As for number inflection, the relative pronoun PE does not inflect for plurality in the regular case, namely it does not agree in space with the loci where the children have been placed in (4a), while the violation of the rule imposes this type of inflection (4b). The first option was chosen 45% of times while the rule violation was chosen 55% of the time, hence giving a near 50-50 distribution (Figure 4).

(4) **PE number agreement.** No number inflection on the relative pronoun (PE).

*Type of violation*: Number inflection of PE on the head noun location.

a. Grammatical construction: 
   \[
   \text{CHILD}+++	ext{Win PE TEACHER PRIZE GIVE} \quad \text{Branchini and Mantovan (2020, 600)}
   \]

b. Rule violation: 
   \[
   \text{CHILD}+++	ext{Win PE+++ TEACHER PRIZE GIVE}
   \]
   'The teacher gave the prize to the children who won.'

As for the position of the head of the relative clause, the target construction is an object relative clause that is the subject of the matrix clause, as indicated in the expected parsing...
in (5a), while the rule violation generates a string that is compatible with an external head (5b). The first option was chosen 42% of times while the rule violation was chosen 58% of the time, hence giving an opposite pattern than the one we hypothesized (Figure 5).

(5) **Head position in relative clauses**: LIS has internally headed relative clauses.  
**Type of violation**: External position of the head.

a. Grammatical construction:  
\[ ROBERTO \text{ CAT RUN-} \overline{\text{OVER PE}} \text{ WELL } \]  
Branchini and Mantovan (2020, 600)

b. Rule violation:  
\[ \text{CAT } \overline{\text{PE ROBERTO RUN-} \text{OVER }} \text{ WELL } \]  
'The cat that Roberto run over is well.'

Figure 5: Distribution of the answers for internally headed relative clauses

5 Discussion

Data reliability is one of the most essential steps in science. In this work, we asked whether sign language elicited data are reliable. Previous studies on English replicated the contrasts found in scientific journals and in reference grammars with formal experiments (Sprouse and Almeida 2012, 2017). These studies proved that elicited data are reliable, or, as we put it, that elicited data are as reliable as experimental data (but see Linzen and Oseki 2019 for a replication failure on Hebrew and Japanese data).

However, sign languages have their own peculiarities that include the fact that they are minority languages which are often also minoritized (i.e., discriminated), that have a minority of native users compared to a majority of early or late learners, and that they use the visual modality in which simultaneity of information is more prominent than in the spoken modality. These factors may represent further complications on the quality of elicited data, independently from the awareness of the consultants. Therefore, it is important to provide a sign language based proof of concept that elicited data are also as reliable as experimental data.

We used part of the data reported in Branchini and Mantovan (2020), which were collected using a variant of the playback method, to create a formal experiment aiming at replicating the contrasts documented in the literature. The statistical analysis revealed that participants find sentences that follow the rule better than those that do not. This result proves that the playback method provides reliable judgments, at least for almost all the constructions we tested.
However, careful inspection of every individual construction revealed that two constructions do not show the expected distribution. Both cases involve relativization.

A construction that yielded an unexpected distribution is the head-internal relative clause. The position of the head is internal in LIS relative clauses (Branchini 2014). Participants chose the sentence that follows the rule 42% of the time and the sentence that violates the rule 58% of the time. Surprisingly, participants preferred the sentence that violates the rule, namely the one with the external head. We have two possible explanations for this fact.

One explanation is that the rule violation is preferred because of the influence from spoken Italian, where relative clauses have an external head. In fact, the construction in (5b) has the same word order of an externally headed relative clause in Italian. Interestingly, no such an effect is visible in other cases where the rule violation returns a string that is more similar to the grammar of spoken Italian, like in the case of wh-movement, position of negation and position of modal verbs. However, one could speculate that object relative clauses are notoriously difficult to process (see Hauser et al. 2021 for evidence from LIS), and that in this specific condition the pressure from Italian provides a stronger bias.

Another possible explanation is that both constructions are actual instantiations of internally headed relative clauses. Unfortunately, we realized this possibility only after the experiment was completed. The actual stimuli of the violation we proposed for the head internal relative clause can receive the fully grammatical parse of an internally headed object relative clause in subject position with the object fronted inside the relative clause, as illustrated in (6). Under this legitimate parse, there is no violation of the head-internal rule.

(6) [Relative Clause [CAT PE] ROBERTO ________CAT PE] RUN-OVER WELL

‘The cat that Roberto run over is well.’

Under this specific scenario, what we are witnessing is not a replication failure but just simple error in the experimental design which contains data that are irrelevant with respect to the research hypothesis. However, this accident unveils an interesting aspect of the grammar of LIS, namely that object fronting in object relative clauses have been chosen more often than the in situ alternative. We leave the psycholinguistic and processing ramifications of this finding for future research.

Another construction that yielded an unexpected pattern is the one involving plural agreement of the relative pronoun PE. While we expected no plural morphology, the distribution of the data shows that having plural inflection or not on the relative pronoun is irrelevant on the acceptability of the sentence. One possible explanation of this fact could be that the sample of participants that took the experiment have a more flexible grammar than the one described in Branchini and Mantovan (2020), allowing PE to optionally have plural inflection. The fact that virtually all participants grew up in the same small Region (Calabria), and also still live there, makes this explanation plausible.

6 Conclusions

We addressed a long-standing issue about the reliability of elicited data in Sign Language. By using an independent method of data collection, we replicated acceptability contrasts previously documented with elicited methods. Our proof of concept demonstrates that sign language elicited data are as reliable as experimental data. For the two constructions that failed to replicate the contrast reported in the literature, we found that one is probably due to an error in the stimuli design (head-internal relative clause), and the other can be explained via geographical variation (number inflection on the relative pronoun PE).
Appendix

This is the list of rules used to create the stimuli:

- **Position of negation:** Negation is postverbal in LIS.  
  Type of violation: negation is pre-verbal.

- **Position of sentential complement:** No center embedding with finite clauses.  
  Type of violation: The sentential complement is center embedded.

- **Canonical order:** The SOV is the canonical order in LIS.  
  Type of violation: The order is VSO.

- **Subject Wh-questions:** The position of wh-words is clause final in LIS.  
  Type of violation: The position of wh-words is clause initial.

- **Object Wh-questions:** The position of wh-words is clause final LIS.  
  Type of violation: The position of wh-words is clause initial.

- **If-Then order:** The antecedent clause must precede the consequent clause.  
  Type of violation: Antecedent clause after consequent clause.

- **PE Agreement:** No number inflection of PE on the head noun (PE is the relative pronoun).  
  Type of violation: Number inflection of PE on the head noun location.

- **Head position in relative clauses:** Internal position of the head in relative clauses.  
  Type of violation: External position of the head in relative clauses.

- **Position of subject Negative quantifiers:** Negative quantifiers are postverbal.  
  Type of violation: N-words are in initial and preverbal position.

- **Position of object Negative quantifiers:** Negative quantifiers are postverbal.  
  Type of violation: N-words are in initial and preverbal position.

- **Alternate Questions:** Two or more options are necessary to complete the question.  
  Type of violation: Just one of the two options is necessary to complete the question.

- **Directionality Forward:** The verb agrees with the subject and the object.  
  Type of violation: The verb does not agree with either.

- **Directionality Backward:** The verb agrees with the object and the subject.  
  Type of violation: The verb does not agree with either.

- **Across-The-Board:** A constituent is extracted from both conjuncts.  
  Type of violation: A constituent is extracted only from the second conjunct.

- **Position of modals of possibility:** Modals of possibility are postverbal in LIS.  
  Type of violation: Modal verbs are preverbal.

- **Position of modal of necessity:** Modal of necessity are postverbal in LIS.  
  Type of violation: Modal verbs are preverbal.
Acknowledgments

The research leading to these results received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC H2020 Grant Agreement No. 788077–Orisem (PI: Schlenker). Research was conducted at Institut d’Etudes Cognitives (ENS), which is supported by grants ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC, and ANR-17-EURE-0017 FrontCog.

References


