Jessica Lettieri, Mirko Santoro, Carlo Geraci. NELS 54, 2023

Experimental vs. Elicited Data in Sign Language Syntax

Goals. The reliability of elicited data in sign language syntax is assessed via experimental methods.

Background. Scientific hypotheses are tested against the empirical reality which is made of raw data, which are then interpreted following specific theories. One of the most essential aspects of this process is data reliability. If the data are not reliable, the risk of postulating incorrect theories is considerable. A commonly used, but often criticized, method to collect data in theoretical linguistics is via informal elicitation, which is based on acceptability judgments provided by a small number of native users. A growing body of literature has started to look into the methodological weaknesses of elicited data for spoken language in the past decade (i.a., Linzen and Oseki 2019; Schütze and Sprouse 2013; Sprouse and Almeida 2012, 2017). These works analyzed three of the major criticisms, namely small number of informants, reduced number of tested items and lack of quantitative measures. Formal experiments are used to replicate syntactic contrasts that have been previously documented either in reference grammars or in papers coming from a selected number of journals. The working hypothesis is that if formal experiments are able to substantially replicate the same contrasts documented by elicited data, then the methodology of data elicitation is as solid as that of formal experiments. In other words, data replication conducted with an experimental method is used to validate elicited data. Follow-up studies further refined the experimental technique to replicate syntactic contrasts, like evaluating minimal pairs as items, rather than single sentence judgments and separating the contribution of forced choice tasks (e.g., maximizing contrasts) from that of Likert scales (e.g., understanding nuances) or magnitude estimation (i.a., Mahowald et al. 2016; Marty et al. 2020; Smith and Little 2018).

More recently, a similar discussion has also started in the sign language literature, although only from an abstract perspective, among other things suggesting to replicate the method validation for sign language (Kimmelman 2021) and to use elicitation techniques that are more similar to that of formal experiments (Davidson 2020). In this paper, we investigate the reliability of elicited data used to describe the syntax of Italian sign language (LIS) with a formal experiment.

Methodology. *Participants.* 24 participants took part in the experiment (7 females, age range 74-23), recruited at the Deaf association of Catanzaro (South of Italy). 13 participants were native signers of LIS, 6 have been exposed to LIS before the age of 6 (early learners) and 5 are late learners.
Stimuli. Data source is the recently published Grammar of LIS (Branchini and Mantovan 2020). We focused on the Syntax part, where we identified *16* different constructions targeting a variety of structures from basic sign order to A′-movement, subordination, relative clauses, and verb-directionality. For each construction, we recovered the key examples and we extrapolated the underlying rule, when not explicitly reported in the text. We then generated a string of signs that minimally violates the rule. Example (1) illustrates the paradigm of sentential negation, one of the tested constructions.

(1) **Rule:** Negation is normally post-verbal in LIS

a. MARIA CAT SEE NEG. (Branchini and Mantovan 2020: 469)
'Maria does not see the cat.'

b. MARIA CAT NEG SEE. (*Minimal violation*)

A native signer of LIS produced both the string that follows the rule and the string that violates the rule. The two sentences were merged in a single video separated by a 1-sec. black screen indicating the first and the second sentence (Fig. 1a). To ensure lexical variation, four lists of 16 pairs were created so that each list contained one construction type counterbalanced for order (rule vs. violation). The four lists plus three training items were used to create an on-line experiment on Labvanced (Finger et al. 2017).
Procedure. Participants were asked to watch the counterbalanced items containing two sentences, one following the rule and one violating it. After stimuli presentation, they were asked to choose which one they prefer in a forced-choice task (Fig. 1b). The experiment was followed by a questionnaire to collect the relevant metadata. Instructions, consent, training and experiment were administered using LIS.

Experimental Hypothesis: If elicited data are reliable, participants are expected to choose the sentence that follows the rule significantly more often than the sentence that violates it.

Results. The dataset consisted of 379 observations. Participants chose the sentence that followed the rule 70% of the cases (Fig. 1c). A generalized mixed model of the binomial family with *expected result* as fixed effect and *item by participant* as random factor revealed that this difference is significant (estimate for the fixed effect is 2.1906, $p < .001$).

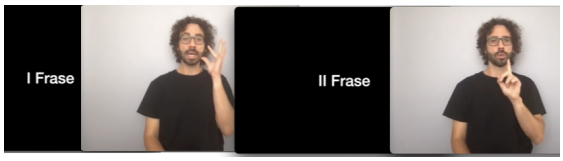

Figure 1: a. Stimulus b. Task c. General Distribution

Qualitative investigation of each construction revealed that the largest contrast was found with *Alternate questions* (96% of expected answers). Two constructions failed to replicate the expected contrast, both involve relative clauses: one construction targeted the ban on externally headed relatives (cf. (2), 42% of expected answers, i.e., reverse pattern), the other targeted the ban on number inflection of the relative pronoun PE (cf. (3), 55% of expected answers).

(2) a. YESTERDAY PAOLO **DOG** FIND PE NOW SLEEP (adapted Branchini and Mantovan 2020)
b. YESTERDAY **DOG** PE PAOLO FIND NOW SLEEP (external head)
'The dog that Paolo found yesterday is asleep now.'

(3) a. $\text{CHILD}_{a,\,b,\,c}$ WIN PE TEACHER PRIZE GIVE (adaped Branchini and Mantovan 2020: 600)
b. $\text{CHILD}_{a,\,b,\,c}$ WIN $\text{PE}_{a,\,b,\,c}$ TEACHER PRIZE GIVE (PE inflects for number)
'The teacher gives the prize to the children who win.'

Discussion. We provided a proof of concept that elicited data are reliable for sign language by replicating consistent findings reported in the literature for spoken languages (a.o., Sprouse and Almeida 2012). Indirectly, we provided evidence of how robust these data are, since the participants were all from a specific region of Italy whose LIS was never investigated with elicited data before (i.e., Catanzaro in the South of Italy). Furthermore, the method seems to be adequate to address some iconic effects of LIS syntax, like directionality. Differently from other sign languages where agreement seems to be optional, it is more resilient in LIS (for a corpus study see Santoro et al. 2016). Our results confirm this fact because participants consistently preferred the form with overt agreement (71% for forward and 75% for backward predicates), as also described in the Grammar of LIS. In turn, this fact suggests that agreement omission may not be free but could be due to currently unknown factors.

Interestingly, one of the most studied construction of LIS, namely relative clauses, failed to deliver the expected result. This proves the unbiased character of the experiment (with biased stimuli all constructions would have behaved uniformly). A number of possible explanations will be proposed during the talk: a) Influence from spoken Italian (for (2) only), b) genuine isogloss for this particular construction, c) possible Type I error in the elicited data (especially for (3)), d) general difficulties related to relative clauses (Hauser et al. 2021; Zorzi et al. 2022).

Selected References. Branchini & Mantovan (Eds.). 2020. *A Grammar of Italian Sign Language (LIS)*. Davidson. 2020. Is experimental a gradable predicate?, NELS 50: Proceedings, 125–144. Hauser et al. 2021. Asymmetries in relative clause comprehension in three European sign languages. *Glossa* 6(1), 1–36. Kimmelman. 2021. Acceptability Judgments in Sign Linguistics. In *The Cambridge Handbook of Experimental Syntax*, 561–584. Sprouse & Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *J. of Linguistics* 48(3), 609–652. Zorzi et al. 2022. On the Reliability of the Notion of Native Signer and Its Risks. *Frontiers in Psychology* 13, 1–12.

Branchini, Chiara and Lara Mantovan (Eds.). 2020. *A Grammar of Italian Sign Language (LIS)*. Venice: Edizioni Ca' Foscari.

Davidson, Kathryn. 2020. Is experimental a gradable predicate? In Mariam Asatryan, Yixiao Song, and Ayana Whitmal (Eds.), *NELS 50: Proceedings of the Fiftieth Annual Meeting of the North East Lingusitic Society*, pp. 125–144. Amherst, MA: GLSA.

Finger, Holger, Caspar Goeke, Dorena Diekamp, Kai Standvoß, and Peter König. 2017. LabVanced: A Unified JavaScript Framework for Online Studies. In *International Conference on Computational Social Science IC2S2*, pp. 2016–2018.

Hauser, Charlotte, Rita Sala, Giorgia Zorzi, Jordina Sánchez Amat, Valentina Aristodemo, Carlo Cecchetto, Beatrice Giustolisi, Caterina Donati, and Doriane Gras. 2021. Asymmetries in relative clause comprehension in three European sign languages. *Glossa 6*(1), 1–36.

Kimmelman, Vadim. 2021. Acceptability Judgments in Sign Linguistics. In Grant Goodall (Ed.), *The Cambridge Handbook of Experimental Syntax*, Chapter 21, pp. 561–584. Cambridge, MA: Cambridge University Press.

Linzen, Tal and Yohei Oseki. 2019. The reliability of acceptability judgments across languages. *Glossa 4*(1), 1–25.

Mahowald, Kyle, Peter Graff, Jeremy Hartman, and Edward Gibson. 2016. SNAP judgments: a small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language 92*(3), 619–635.

Marty, Paul, Emmanuel Chemla, and Jon Sprouse. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: a journal of general linguistics 5*(1), 1–23.

Santoro, Mirko, Lara Mantovan, Valentina Aristodemo, and Carlo Geraci. 2016. A sociolinguistic view on variable subjects in italian sign language. In *Grammar and Corpora*, Mannheim, Germany. Conference presentation, Nov. 9-11.

Schütze, Carson T. and Jon Sprouse. 2013. Judgment data. In Robert J. Podesva and Dvyani Sharma (Eds.), *Research Methods in Linguistics*, Chapter 3, pp. 27–50. Cambridge University Press.

Smith, Philip L. and Daniel R. Little. 2018. Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin and Review 25*(6), 2083–2101.

Sprouse, Jon and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics 48*(3), 609–652.

Sprouse, Jon and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa 2*(1), 1–32.

Zorzi, Giorgia, Beatrice Giustolisi, Valentina Aristodemo, Carlo Cecchetto, Charlotte Hauser, Josep Quer, Jordina Sánchez Amat, and Caterina Donati. 2022. On the Reliability of the Notion of Native Signer and Its Risks. *Frontiers in Psychology 13*(March), 1–12.