A Database for Modal Semantic Typology

Qingxia Guo, Nathaniel Imel, Shane Steinert-Threlkeld
Overview

● What are modals? What do modals look like in natural languages?
● What does our database do?
● How to collect data and contribute to the database?
● What are some use cases of this database?
● Some reflections and possible future works
Semantics Universal of Modals

- Modals as ability to talk about possible worlds
- Semantics of modals have been explicated in terms of two axes of variation: force and flavor (Kratzer 1981).
- Force: strong (necessity), weak (possibility), ...
- Flavor: epistemic, deontic (…)
# Examples of English Modals

| Context                                                                 | Expression         | Axes Values       |
|-------------------------------------------------------------------------|--------------------|-------------------|
| (1) A friend walks in and shakes off a wet umbrella. You say:           | It must be raining.| strong epistemic  |
| (2) You are reading the specifications of a homework assignment. It partially reads: | You *must* upload your homework as a PDF. | strong deontic  |
| (3) A friend is leaving and grabs an umbrella on the way out, saying:   | *It may* be raining | weak epistemic    |
| (4) A mother offers a treat to a child for finishing an assignment, saying: | You *may* have a cookie | weak deontic    |

Table 1: Examples of force and flavors in English.
Examples of St’t’at’icmets Modals

(3)  a.  [Context: You have a headache that won’t go away, so you go to the doctor. All the tests show negative. There is nothing wrong, so it must just be tension.]

   nilh k’a lh(el)-(t)-en-s-wá(7)-(a) ptinus-em-sút
   FOC INFER from-DET-1SG.POSS-NOM-IMPF-DET think-MID-OOC

   ‘It must be from my worrying.’

   b.  [Context: His car isn’t there.]

   plan k’a qwatsáts
   already INFER leave

   ‘Maybe he’s already gone.’

source: Rullmann, H., Matthewson, L., & Davis, H. (2008)
Database for Modals in Natural Languages

- Ideally, we would like to derive a full paradigm of modal expression and (force, flavor) pair with can_express value in natural languages so that we can acquire the meaning representation of modals.

| expression | force  | flavor     | can_express |
|------------|--------|------------|-------------|
| may        | weak   | epistemic  | 1           |
| may        | weak   | deontic    | 1           |
| may        | strong | epistemic  | 0           |
| may        | strong | deontic    | 0           |

Table 1: Example of our basic data format for English $m$

| Modal | Meaning representation |
|-------|------------------------|
| may   | $e \quad d \quad c \quad t$ |
|       | $\exists \quad \checkmark \quad \checkmark$ |
|       | $\forall$             |
Database for Modals in Natural Languages

- 17 languages, 5 from semantics fieldwork, 12 from reference grammar
- https://github.com/CLM BRs/modal-typology

| Language          | Glotto.code | Reference.key                  | Reference.type          | Complete.language |
|-------------------|-------------|--------------------------------|-------------------------|-------------------|
| Donmari           | doma1258    | (Matras, 2012)                 | reference-grammar      | True              |
| Gitksan           | gtx1241     | (Matthewson, 2013)             | paper-journal          | True              |
| Goemai            | goem1240    | (Hellwig, 2011)                | reference-grammar      | True              |
| Hiniq             | hinu1240    | (Forker, 2013)                 | reference-grammar      | True              |
| Hup               | hupd1244    | (Epps, 2005)                   | reference-grammar      | True              |
| Jamul-Tipay       | kumi1248    | (Miller, 2001)                 | reference-grammar      | True              |
| Javanese-Paciran  | java1254    | (Vander Klok, 2013a)           | paper-journal          | True              |
| Kwaza             | kwaz1243    | (Voort, 2004)                  | reference-grammar      | True              |
| Lillooet-Salish   | lill1248    | (Rullmann et al., 2008b)       | paper-journal          | True              |
| Logoori           | logo1258    | (Gluckman and Bowler, 2020)    | paper-journal          | True              |
| Mani              | bull1247    | (Childs, 2011)                 | reference-grammar      | True              |
| Mian              | mian1256    | (Fedden, 2011)                 | reference-grammar      | True              |
| Nuosu             | sich1238    | (Gerner et al., 2013)          | reference-grammar      | True              |
| Qiang             | nort2722    | (LaPolla and Huang, 2003)      | reference-grammar      | True              |
| Tlingit           | tlin1245    | (Cable, 2017)                  | paper-journal          | True              |
| Tundra-Nenets     | nene1249    | (Nikolaeva, 2014)              | reference-grammar      | True              |
| Vaeakau-Taumako   | pile1238    | (Næss, 2011)                   | reference-grammar      | True              |

Table 3: Snapshot of current metadata in the Modal Typology Database. Note: we have replaced the ‘Reference.key’ column with actual references using those keys.
Data Format

- **Raw format (/basic-format):**
  - `modal.csv`: expressions, force-flavor pair and can_express values for each language
  - `metadata.yml`: contains information about language, source and citation

- **CLDF format (/cldf-format):**
  - Convert raw format into a database in the Cross-Linguistics Dataset Format
  - CLDF format can be later consumed by tools to develop interactive web applications

Contributing to database (more details in `contributing.md`):

- Fork the github repository
- Create a language folder under `/basic-format`
- Add file `metadata.yml`, `modals.csv` to the created folder
- Edit `source.bib` to reflect source of the data
- Submit a pull request to the main repository
Data Collection

There are several ways of collecting modals in natural languages

- **From** descriptive source
  - targeted semantics fieldwork
  - general reference grammar

- **From** elicitation
  - consult native speakers directly with pre-established questionnaire (Vander Klok, 2021)
  - crowdsourcing semantics elicitation (Beekhuizen and Stevenson, 2015)
    - more applicable to dominant languages
Use Case: Efficient Communication Analysis

- Simplicity - informativeness tradeoff in **artificial** languages
- **Naturalness** significantly correlated to **optimality**
- What about actual **natural** languages?

source: Imel, N., & Steinert-Threlkeld, S. (2022).
Use Case: IFF

- Independence of Force and Flavor (Steinert-Threlkeld et al., 2022)
  - all modals in natural languages satisfy the independence of force and flavor property
  - if a modal can express \((fo1,fl1)\) and \((fo2,fl2)\), then it can also express \((fo1,fl2)\) and \((fo2,fl1)\)

- We provide a simple tool `iff.py` to check whether this property holds for languages in the database and output counter examples if there’s any.
Conclusion

- Introduced a database on modal typology
- Collected data in various format and from various sources
- Employed to several use cases (efficient communication & IFF)

Future Work

- Data visualization using CLDF format
- Extend the database using semantics elicitation methods
- Detailed syntactic information about modal expressions
Thank you! Questions?
Selected References

- Imel, N., & Steinert-Threlkeld, S. (2022). Modals in Natural Language Optimize the Simplicity/Informativeness Trade-Off. Proceedings of Semantics and Linguistic Theory (SALT 32)
- Rullmann, H., Matthewson, L., & Davis, H. (2008). Modals as distributive indefinites. Natural Language Semantics, 16(4), 317–357. [http://www.jstor.org/stable/23748121](http://www.jstor.org/stable/23748121)
- Steinert-Threlkeld, S., Imel, N., & Guo, Q. (2022). A Semantic Universal for Modality. Submitted to *Semantics and Pragmatics*.
- Beekhuizen, B., & Stevenson, S. (2015). Crowdsourcing elicitation data for semantic typologies. Cognitive Science.
- Vander Klok, J. (2021). Revised Modal Questionnaire for Cross-Linguistic Use. Unpublished.