A Database for Modal Semantic Typology

Qingxia Guo and **Nathaniel Imel** and **Shane Steinert-Threlkeld**
Department of Linguistics, University of Washington
Box 354340, Seattle, WA, 98195-4340
{qg07,nimel,shanest}@uw.edu

Abstract

This paper introduces a database for crosslinguistic modal semantics. The purpose of this database is to (1) enable ongoing consolidation of modal semantic typological knowledge into a repository according to uniform data standards and to (2) provide data for investigations in crosslinguistic modal semantic theory and experiments explaining such theories. We describe the kind of semantic variation that the database aims to record, the format of the data, and a current snapshot of the database, emphasizing access and contribution to the database in light of the goals above. We release the database at https://clmbr.shane.st/modal-typology.

1 Introduction

Modals—expressions used to talk about situations other than the actual one—are ubiquitous in natural language and have been the focus of intense study in the semantics thereof (Kratzer, 1981; Portner, 2009; Matthewson, 2019). An increasingly large body of work has gathered data on the crosslinguistic variation in this domain, i.e. the ways in which languages agree and differ in their mechanisms for expressing modality (Rullmann et al., 2008a; Vander Klok, 2013b; Cable, 2017, i.a.).

This paper introduces and describes a *Modal Typology Database*: a repository that consolidates much of this crosslinguistic knowledge in a format that is uniform and easy both to consume and to produce. Such a resource can play several enabling roles in semantic typology research. For example, it can enable the verification of robust semantic universals (Nauze, 2008; Vander Klok, 2013b; Steinert-Threlkeld et al., 2022) and possibly trigger the formulation of new ones. Similarly, these data and their format can be used in comparison to artificial languages to attempt to *explain* what pressures have shaped semantic typology in the domain of modality, as has been done in several other domains (Kemp and Regier, 2012; Zaslavsky et al., 2018; Steinert-Threlkeld and Szymanik, 2019, 2020; Steinert-Threlkeld, 2021; Denić et al., 2022; Mollica et al., 2021; Uegaki, 2022, i.a.).

After describing some of what is known about the variation in modals (Section 2), we describe a data schema for representing particular axes of variation (namely: force and flavor) in modals crosslinguistically in a relatively theory-netural manner (Section 3.1). We then (Section 4) describe how to access this data, which we make available in two distinct formats: a ‘basic’ format, and one that conforms to the Cross-Linguistic Data Formats (CLDF; Forkel et al. 2018) schema. We illustrate how to use these data to verify a semantic universal in Section 4.3, before explaining how researchers can contribute their own data (Section 5) and providing a snapshot of what data the database currently has (Section 6). We provide a discussion around future directions in Section 7 before summarizing the present work in Section 8.

2 Modal Typology

Modals are expressions that are used to talk about alternative ways the world could be, over and above the way the world actually is. Languages utilize various syntactic forms to express modality. For example, English uses auxiliary verbs like *may* and *must* as modals, in addition to adjectives like *possible*; Javanese makes use of auxiliaries, a main verb, and several adverbs (Vander Klok, 2013a). Since at least Kratzer (1981), the semantics of modals have been explicated in terms of two axes of variation: force and flavor. These axes can be illustrated with the examples listed in table 1.

The *must* examples exhibit strong (i.e. universal) force, but differ in flavor. For example, (1) in the table 1 can be glossed as saying: all of the worlds compatible with my evidence are worlds in which it is raining. The universal quantification represents

Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP (SIGTYP 2022), pages 42 - 51
July 14, 2022 ©2022 Association for Computational Linguistics

	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.

the force, and the domain of worlds (those compatible with my evidence) the flavor, in this case epistemic. (2) exhibits universal force with deontic flavor, roughly saying that all the worlds in which you follow the rules are ones in which you upload a PDF. The examples with *may* in (3) and (4) exhibit weak (i.e. possibility) force: their meaning says that some world satisfies the prejacent. (3) and (4) again differ in flavor, with the former being epistemic and the latter being deontic. In addition to epistemic and deontic flavors, many others have been identified: bouletic (worlds in which desire are fulfilled), teleological (worlds in which goals are satisfied), *et cetera*. Similarly, there are arguably more forces than just weak and strong: for instance, there are weak necessity modals (e.g. *should*, *ought*) which intuitively express universal quantification over a smaller domain of worlds (von Fintel and Iatridou, 2008). See Matthewson 2019 and references therein for further discussion of these two axes. The examples above show that English modals lexically specify modal force (each modal has a fixed quantificational force) but exhibit variability across flavors (the modals can express more than one flavor). We note that such variability does not require that all modals in English can express multiple flavors: for instance, *might* arguably can only be used epistemically. Kratzerian semantics for modals capture this by hard-coding quantificational force into the meaning of a modal but relying on context to determine the flavor.[1]

Not all languages are like English: some exhibit so-called *variable force modals*, which specify flavor but not force. This has been found at least in St'át'icmets (Rullmann et al., 2008a), Nez Perce (Deal, 2011), Old English (Yanovich, 2016), and Pintupi-Luritja (Gray, 2021). We illustrate the phenomenon with elicited examples of St'át'icmets *k'a*:[2]

(5) [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*
FOC INFER
lh(el)-(t)-en-s-wá(7)-(a)
from-DET-1SG.POSS-NOM-IMPF-DET
ptinus-em-sút
think-MID-OOC

'It *must* be from my worrying.'

(6) [Context: His car isn't there.]

plan k'a qwatsáts
already INFER leave

'Maybe he's already gone.'

Example (5) shows *k'a* being used with strong force and epistemic flavor. Example (6) shows *k'a* being used with weak force and epistemic flavor. Further analysis in Rullmann et al. (2008a) shows that *k'a* can only be used with epistemic flavor, so it is an example with lexically specified flavor but variable force. Finally, some languages have modals which exhibit variability along *both* the force and flavor axes. Bochnak (2015b,a) has argued that the modal verb -*eʔ* in Washo can be used in both possibility and necessity contexts with

[1] Typical implementations determine the flavor as the product of two parameters: a modal base and an ordering source. We set aside this distinction for present purposes and focus only on flavor.

[2] These are examples (5c) and and (5e) from Rullmann et al. 2008a, p. 321. See their footnote 5 on p. 320 for the abbreviations.

a range of modal flavors. Similarly, Močnik and Abramovitz (2019) demonstrate that the Koryak attitude verb *ivək* can be used to express both necessity and possibility. For the doxastic flavor, this means that *ivək* can be used to mean roughly 'believe' (necessity) as well as 'allow for the possibility that' (possibility). They also argue that the expression can be used to express both doxastic and assertive flavors, thus demonstrating variability on both axes.[3]

3 Representing Modal Semantics in a Database

A database for cross-linguistic modal semantics should be theory-neutral while still capturing the basic parameters of variation and facts upon which linguists agree. A natural way to proceed is to simply record the flavors and forces a particular modal can be used to express. We elaborate on this analysis in the following subsections.

3.1 General Framework

We assume that force and flavor are fundamentally properties of contexts of use. This reflects current practice in semantic fieldwork as applied to modality (Matthewson, 2004; Bochnak and Matthewson, 2020; Vander Klok, 2021).[4] For example, the modal questionnaire of Vander Klok 2021 consists exactly of discourse contexts designed to isolate a single force-flavor pair. These contexts can be used at least for elicitation, translation, and acceptability tasks. Specifically, we will say that a modal M *can express* a force-flavor pair just in case a bare positive sentence of the form Mp is judged felicitous in a context with that pair.[5] For example, English *must* can express the pair (universal, deontic) because there is a reading for that pair under the context in 1 in table 1. Here we identify a modal as the set of (force, flavor) pairs that it can express. We intend this level of modeling to apply to the expression of modality by diverse syntactic means (as mentioned in the Introduction), and not to be specific to any one syntactic category. A language is (generously) identified as a list of modals.

We adopt this level of generality because it avoids commitment on the exact formal semantics of these expressions, which is often still being debated. For example, we can say that a *variable force modal* is one that can express more than one pair with the same force. This is useful because there are two broad approaches to the semantics of such variable force modals: they actually encode existential quantification but lack a universal scalemate (Deal, 2011) or they encode universal quantification but rely on some mechanism of domain restriction (Rullmann et al., 2008a; Bochnak, 2015a; Močnik and Abramovitz, 2019). On such analyses, the underlying semantics contains one specific quantifier; in the present setting, they will still be considered variable force since bare positive sentences are used in contexts with multiple forces.

This approach to encoding the semantics of modals allows straightforward evalutation of universals, such as proposed by Nauze (2008), Vander Klok (2013b), and Steinert-Threlkeld et al. (2022) which are testable hypotheses and potential targets of explanation. All of these modal semantic universals are formulated constraints on the kinds of sets of (force, flavor) pairs found in any human language. For example, Steinert-Threlkeld et al. (2022) propose the INDEPENDENCE OF FORCE AND FLAVOR (IFF) universal: All modals in natural language satisfy the independence of force and flavor property: if a modal can express the pairs (fo_1, fl_1) and (fo_2, fl_2), then it can also express (fo_1, fl_2) and (fo_2, fl_1). A database that catalogs which force-flavor pairs are expressed by various modals cross-linguistically can thus be used to empirically verify whether this universal holds unrestrictedly or at least very robustly. In Section 4.3 we show how our database can be used in exactly this way.

3.2 Concrete Schema

We can implement the above framework according to the principles of tidy data (Wickham, 2014). Such tabular data has the following properties: every column is a variable, every row an observation, and every cell a value. According to the framework just described, a basic observation in cross-linguistic modal semantics says that a particular modal expression can or cannot express a particular (force, flavor) pair.

[3] There are also apparently bouletic uses of *ivək*, but Močnik and Abramovitz (2019) argues that this flavor does not come from *ivək* alone but from interaction with material in the embedded clause.

[4] In addition to the particular studies already mentioned, see Matthewson 2013; Cable 2017 for more examples of the application of these methods.

[5] We intend 'judged felicitous' to also include the case where such sentences are produced naturally in elicitation tasks, as well as when such sentences are found in naturally-occuring contexts which have a clear force-flavor pair.

Our basic data schema, accordingly, will be a table with four columns (we also record metadata about the language of an expression, in a way detailed in the next section):

1. expression: the name of the particular expression
2. force
3. flavor
4. can_express: a binary variable, with 1 meaning that the expression can express the pair of values in the force and flavor columns, and a 0 meaning that it cannot.[6]

with each row being one observation. For example, we can represent the fact that English *may* can only be used to express weak epistemic and weak deontic combinations as follows:

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

Table 2: Example of our basic data format for English *may*.

A note about possible values of force and flavor: while these are generally thought to be shared cross-linguistically, our data format does not commit to a pre-specified ontology of either. In particular, in order to capture the fact that certain languages make different / finer distinctions than others, we aim to be as liberal as possible in recording featural diversity. The consequences of balancing these goals are that during data collection the list of modal forces or flavors might not be completely exhaustive and disjoint. Later on, features can be collapsed or renamed as necessary, as the database grows, or as particular analysis needs require. For example, the English possibility modal *can* expresses deontic and circumstantial flavors, and so may be considered a "root" modal, but we aim for precision by recording deontic and circumstantial flavors rather than a higher-level grouping. Similarly, it is possible that when recording data from a descriptive grammar, one will find a unique or nonstandard name for a possible flavor. One can record that flavor as given in that grammar, and in a later analysis step, attempt to map that flavor value onto ones that are used in other resources.

On the force side, we are primarily intended in capturing weak, strong, and weak necessity modals, setting aside for the time being the full range of possibilities of graded modality, including probabilistic expressions (Kratzer, 1981; Portner, 2009; Klecha, 2014; Lassiter, 2017). At the present state of theorizing, there is not enough concensus about their typology. That being said, on some approaches to graded modality, the database as currently structured could be easily modified or extended to include some aspects of them: if graded modals are genuinely scalar terms (Klecha, 2014; Lassiter, 2017; Bowler and Gluckman, 2021), then features from the semantics of gradable expressions such as scale-type and the minimum/maximum/relative distinction could be recorded (Kennedy and McNally, 2005; Kennedy, 2007).

4 Accessing the Database

The database may be found at `https://clmbr.shane.st/modal-typology`. This landing page—which will contain more information in the future—will point the reader to a repository containing the data. It is made publicly available in two formats. First, we have a 'raw' format: this is oriented around individual languages and is designed to make it easy for linguists to contribute new data. We describe this format in the next subsection (4.1) and how to contribute in Section 5. Secondly, we have a script to convert the raw format into a Cross-Linguistic Data Formats (CLDF; Forkel et al. 2018) format, which has several benefits of its own that are described in more detail in Section 4.2. We then demonstrate one of these benefits, by showing how to verify the IFF universal using the data in the database (in either format) in Section 4.3.

4.1 Basic Format

The basic format, found in the `basic-format/` sub-directory, contains information both at the language-level and then aggregated across languages. We explain these types of data in turn.

[6] We also will sometimes use a '?' in this column to indicate that it is unclear. As an example, in Tlingit (Cable, 2017), there are some cases where the author writes that it is implausible that an expression can express a particular force-flavor pair, but that there has not been concrete negative evidence to support that judgment. We record cases such as those with a '?'.

To see the data for one language, we will look at Tlingit. The data for this language comes from the fieldwork reported in Cable 2017. To access it, go to the sub-folder named `Tlingit`. There, you will find two files:

1. `metadata.yml`: this contains information about the language and the source(s) from which the modals data was compiled. In particular:
    - Glotto code: this is an ID for the language from Glottlog[7] (Hammarström et al., 2021)
    - Reference: a citation for the source
    - Reference_key: a BibTeX key to a shared bib file (described below)
    - URL: a URL to find the reference
    - Reference_type: the type of source that the reference is
      We note that this will be especially useful in distinguishing languages where the information derives from targeted semantic fieldwork (as in the present case of Tlingit) and from descriptive grammars. The latter tends to lack explicitly *negative* evidence, upon which some analyses may depend, and so those languages may need to be excluded.At present, the values for this field that exist in our database are 'paper_journal' and 'reference_grammar'.
    - Complete_language: whether the reference purports to describe the complete modal system of the language or not.
      Many sources only provide data for some, but not all, modals. Such expression-level data is still very useful, but researchers may wish to exclude incomplete languages from analyses at the language level.
2. `modals.csv`: this is a comma-separated-value (CSV) file, containing the core data in the format described in the previous section

Popping back out to the main `basic-format/` directory, there are several aggregated data files that are generated automatically from the language-specific data:

- `all_observations.csv`: this effectively concatenates `modals.csv` from each language, while also adding columns identifying which language the relevant modal in the observation comes from.
- `all_metadata.csv`: this aggregates the metadata from each language and puts it into one CSV table.
- `all_modals.csv`: this presents a new view of the aggregated data *at the level of indidivudal modals*. In particular, each row corresponds to one expression in one language. In this table, there are columns for each (force, flavor) pair, with the corresponding value from the `can_express` column from the relevant `modals.csv` file in that cell. This allows researchers to see the set of force-flavor pairs that each modal expresses in one place, and may assist analyses that depend on that set.
  Note: If a particular (force, flavor) pair was not annotated for a given modal in a given language, there will be an "NA" as the value in that column in this file. This should be viewed as a distinct value from either 1, 0, or ?.

All of these files are generated by running the R script `combine_data.R`, which also exists in this directory. There are three more files present in this directory: one for forces, one for flavors, and a BibTeX file containing all of the reference material. We will mention these in more detail in Section 5, when explaining how to contribute to the database.

4.2 CLDF Format

While the raw format described above is the easiest for human consumption and for contribution by field linguists (see 5), we have also implemented a script that converts the raw data into a database in the Cross-linguistic Dataset Format (CLDF; Forkel et al. 2018). This dataset format—which underlies resources such as the World Atlas of Language Structures (WALS; Dryer and Haspelmath 2013) and Glottolog (Hammarström et al., 2021)—was designed to make the myriad cross-linguistic data being collected "FAIR": Findable, Accessible, Interoperable, and Reusable. While the raw dataset formats are also based on a set of tables in CSV format, it comes with tools (e.g. the Python library pycldf[8]) for converting those into other formats such as an SQLite database, which can enable

[7] See `https://glottolog.org`

[8] `https://github.com/cldf/pycldf`

researchers to asked detailed questions in a full-powered query language.

Similarly, data in CLDF format can be consumed by the tools from the Cross-Linguistic Linked Data project[9], which can be used for instance to develop interactive web applications to interact with the data. Such an application could for example, provide a graphical interface for research to explore which (force, flavor) pairs are most frequently expressed across the recorded languages, which sets of pairs tend to be expressed by the same morphemes, which languages satisfy certain semantic universals (as they are proposed), and so on. Compared to reading each cited descriptive resource for a given language, these data tools could provide quick initial answers to questions about modal typology that may otherwise take significant time to explore at the same level of detail.

While we refer the reader to the aforementioned reference and their webpage[10] for more information and motivation about this format, we here outline some of its properties in order to highlight novel changes that were necessary for our database. CLDF defines specifications for two types of dataset at the highest level: Wordlist and StructureDataset. A Wordlist is intended to capture lexicon-level information, associating concepts with lexical items in a language (often linking to external resources for the available concepts). The World Loanword Database (WOLD; Haspelmath and Tadmor 2009) is a paradigm example. A StructureDataset primarily captures grammatical features at the language level: a basic entry says that a particular language has a paritcular value for a particular parameter. The World Atlas of Language Structures (WALS; Dryer and Haspelmath 2013) is a paradigm example.

Our data, however, can be seen as a mix of these two types of data: we are recording feature values (e.g. *can_express*), but at the lexical level, not the language level. We have implemented this in the following way: in addition to language-level parameter and value tables (which record which modals exist in which languages), we have also added *unit parameter* and *unit value* tables, which record the exact observations about which modals can express which force-flavor pairs as recorded in the basic-format. We refer the reader to the `README.md` file in the `cldf-format` subdirectory for more information on the exact tables in this dataset. We also note that CLDF was designed with extensbility in mind; it is possible that this dataset format will get added to the standard in the future if more datasets are released with the use of it.[11]

The CLDF Format of the data is automatically generated from the basic format by running the script `./build.sh` in the root directory. This script moves basic format data to the appropriate locations and then executes a CLDFBench (Forkel and List, 2020) script for converting raw data into the relevant CLDF tables. We, the maintainers of the dataset, will run this script whenever a new contribution to the basic format is made, so that the CLDF format stays up-to-date. Future work will explore implementing this via continuous integration, so that the CLDF format is automatically built whenever the basic format is updated, without human intervention.

4.3 Case Study: Verifying the IFF Universal

We here provide a small proof-of-concept of the kind of cross-linguistic semantic research that can be benefited from and enabled by the kind of database that we are releasing here. In particular, we show how to query the data to check whether the IFF universal described in Section 3.1 holds. As more data gets added to the database, we can easily and continuously search for counterexamples to this proposed universal. We provide examples of doing this in both data formats.

4.3.1 Basic Format

Running the file `iff.py` in the `basic-format` directory performs a simple check of `all_observations.csv` for expressions that do not satisfy IFF as stated, and outputs the language, expression, and its corresponding observations for inspection. At the time of writing, there are no counterexamples to the universal in our database.

4.3.2 CLDF Format

One other advantage of the CLDF format and toolkit is that it enables researchers to define custom commands that can be run on the command-line to either manipulate the data or verify certain properties thereof. We have illustrated this functionality by implementing a small command

[9] `https://clld.org/`

[10] `https://cldf.clld.org/`

[11] See discussion here about this dataset format: `https://github.com/cldf/cldf/issues/117`. We are grateful to Robert Forkel for his assistance here.

that checks whether the data supports the IFF Universal described above in Section 3.1. In particular, running `cldfbench modals.iff` from the `cldf-format/` directory will execute a Python script for verifying whether every modal in the database satisfies the IFF universal. (The actual implementation can be found in `cldf-format/modalscommands/iff.py`.)

5 Contributing to the Database

We have designed the database—and the basic format in particular—to be structured in a way that makes it easy for linguists to contribute new data from languages that they are studying. As the primary data resides in a GitHub repository, contributing relies heavily on the mechanism of forking and submitting a pull request; for more information on those specific mechanics, we refer to their documentation.[12] The basic process for contributing data from a new language goes as follows (with further details provided in the file `CONTRIBUTING.md` in the repository):

1. Fork the GitHub repository and edit or create a new folder for your language in the `basic-format` directory of the repository.

2. Add a `metadata.yml` file with the information as described in Section 4.1. You can start by copying an existing such file if desired.

3. Edit `basic-format/sources.bib` with the BibTeX information of the descriptive source of your data. Note that the key used in this entry should exactly match the value for 'Reference_key' in the metadata file.

4. Add a `modals.csv` file to your folder, with the corresponding observations. Columns should be: expression, force, flavor, can_express, notes.

5. Optional: run the combine_data.R script to combine this new data with the existing aggregate data files. (If a contributor does not want to do this step, we are happy to do this upon merging the new data into the main repository.)

6. Submit a pull request to the main repository from your fork.

We will use the pull request interface to note any minor formatting issues and have any necessary discussions of the new data. After that quick process, we will merge your new data into the main database, and run the relevant scripts to join it with the rest of the data, including in the CLDF format version.

6 Snapshot

At the time of writing, we have added data from 17 languages to the database. Some information about these languages, including the reference (and its type) that we used to gather this data, may be found in Table 3. Five of the 17 languages have data coming from detailed semantic fieldwork (the ones with 'paper_journal' as their type), with the rest of the data coming from descriptive grammars. There are at present 435 unique observations in our aggregate data file `all_observations.csv`, each one corresponding to one judgment that a particular modal in a language can or cannot express a particular force-flavor pair.

7 Discussion

Most langauges (12 out of 17) in Table 3 are gathered from descriptive sources, i.e. reference grammars that provide general descriptions of the languages. While these languages add diversity to our typology database, the data often lack negative judgements for the relation between expression forms and force-flavor pairs. In other words, it is very often difficult to tell whether an expression *cannot* express a force-flavor pair (i.e. to categorize any expression form and force-flavor pair with a can_express value being 0) from a reference grammar. Researchers conducting analyses with languages with data from reference grammars should beware of this lack of negative data when proceeding. The data stemming from controlled semantic fieldwork tends to provide more negative and more complete data.

While those data tend to come from understudied languages, the methodologies used could be deployed to generate more conistent data for many 'high-resource' languages by eliciting data through crowdsourcing, which has been shown to produce high-quality semantic typology data (Beekhuizen and Stevenson, 2015). The questionnaire of Vander Klok 2021 provides a template for

[12] In particular, the "Working with forks" and "Creating a pull request from a fork" sub-pages of https://docs.github.com/en/pull-requests/collaborating-with-pull-requests.

Language	Glotto.code	Reference.key	Reference.type	Complete.language
Donmari	doma1258	(Matras, 2012)	reference-grammar	True
Gitksan	gitx1241	(Matthewson, 2013)	paper-journal	True
Goemai	goem1240	(Hellwig, 2011)	reference-grammar	True
Hinuq	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.

the desired crowdsourcing elicitation process. The questionnaire establishes discourse contexts to retreive modal expressions for various force-flavor pairs. It underspecifies the form of targeted tasks to preserve its adaptablity. Future work could investigate applicable crowdsourcing procedures and how to adapt the questionnaire to elicit the expected form of data. This should enable the production of more complete data with negative examples for many languages.

8 Conclusion

This paper introduced the *Modal Typology Database*, a public repository for typological data on the semantics of modals across langauges. It is intended to be a living database for consolidating cross-linguistic knowledge about modal semantic variation and evaluating and explaining modal semantic universals, among other possible uses. As an example, a recent efficient communication analysis of modal typology by Imel and Steinert-Threlkeld (2022) compared artificial languages based on how many modals therein satisfy particular universals; this analysis could be supplemented with the data presented here to directly compare natural and artificial languages. We have presented a simple model for expressing parameters of variation of the semantics of modals in a theory-neutral manner and outlined how the data are structured as well as how anyone (theoretical linguists, fieldworkers, etc.) may contribute new data. We encourage others to both consume and produce these data, and to reach out to discuss any issues that arise therein.

In addition to expanding the core database with more data and encouraging other uses thereof, future work will focus on building visualization and other tools for interacting with the data in a more user-friendly way. (The CLDF format of the data may be especially well-suited to these goals.) The data schema may also be extended to include more information about the syntactic forms of the expression of modality (possibly using elements of the CLDF schema for forms), in addition to the phenomena of gradable modals and other expressions that often partially contribute modality as well (e.g. tense, evidentiality).

References

Barend Beekhuizen and Suzanne Stevenson. 2015. Crowdsourcing elicitation data for semantic typologies. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*. cognitivesciencesociety.org.

M Ryan Bochnak. 2015a. Underspecified modality in Washo. In *Proceedings of the Workshop on Structure and Constituency in Languages of the Americas 18 & 19*, volume 39 of *University of Britisch Columbia Working Papers in Linguistics*, pages 3–17.

M Ryan Bochnak. 2015b. Variable force modality in Washo. In *Proceedings of North-East Linguistic Society (NELS) 45*, pages 105–114.

M. Ryan Bochnak and Lisa Matthewson. 2020. Techniques in Complex Semantic Fieldwork. *Annual Review of Linguistics*, 6(1):261–283.

Margit Bowler and John Gluckman. 2021. Cross-categorial gradability in Logoori. *Semantics and Linguistic Theory*, 30(0):273–293.

Seth Cable. 2017. The expression of modality in tlingit: A paucity of grammatical devices1. *International Journal of American Linguistics*, 83:619 – 678.

George Tucker Childs. 2011. *A grammar of Mani*. Mouton grammar library ; 54. De Gruyter Mouton, Berlin ; Boston.

Amy Rose Deal. 2011. Modals Without Scales. *Language*, 87(3):559–585.

Milica Denić, Shane Steinert-Threlkeld, and Jakub Szymanik. 2022. Indefinite Pronouns Optimize the Simplicity/Informativeness Trade-Off. *Cognitive Science*, 46(5):e13142.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Patience Epps. 2005. *A grammar of Hup*. Ph.D. thesis, University of Virginia.

Sebastian Fedden. 2011. *A grammar of Mian*. Mouton grammar library ; 55. De Gruyter Mouton, Berlin.

Robert Forkel and Johann-Mattis List. 2020. CLDFBench: Give your cross-linguistic data a lift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6995–7002, Marseille, France. European Language Resources Association.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):180205.

Diana Forker. 2013. *A grammar of Hinuq*. Mouton grammar library, 63. De Gruyter Mouton, Berlin ; Boston.

Matthias Gerner, Georg Bossong, and Matthew Dryer. 2013. *A Grammar of Nuosu*, volume 64 of *Mouton Grammar Library [MGL]*. De Gruyter, Inc, Berlin/Boston.

John Gluckman and Margit Bowler. 2020. The expression of modality in logoori. *Journal of African Languages and Linguistics*, 41(2):195–238.

James Gray. 2021. Variable Modality in Pintupi-Luritja Purposive Clauses. *Languages*, 6(1):52.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.5.

Martin Haspelmath and Uri Tadmor, editors. 2009. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Birgit Hellwig. 2011. *A grammar of Goemai*. Mouton grammar library ; 51. De Gruyter Mouton, Berlin ; Boston.

Nathaniel Imel and Shane Steinert-Threlkeld. 2022. Modals in natural language optimize the simplicity/informativeness trade-off. In *Proceedings of Semantics and Linguistic Theory (SALT 32)*.

Charles Kemp and Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

Christopher Kennedy. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30:1–45.

Christopher Kennedy and Louise McNally. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, 81(2):345–381.

Peter Klecha. 2014. *Bridging the Divide: Scalarity and Modality*. Ph.D. thesis, University of Chicago.

Angelika Kratzer. 1981. The Notional Category of Modality. In Hans-Jürgen Eikmeyer and Hannes Rieser, editors, *Words, Worlds, and Context*, pages 38–74. Walter de Gruyter.

Randy J LaPolla and Chenglong Huang. 2003. *A Grammar of Qiang: With annotated texts and glossary*, 1. aufl. edition, volume 31 of *Mouton Grammar Library [MGL]*. Mouton de Gruyter, Berlin/Boston.

Daniel Lassiter. 2017. *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press.

Yaron Matras. 2012. *A Grammar of Domari*, 1. aufl. edition, volume 59 of *Mouton Grammar Library [MGL]*. Mouton de Gruyter, Berlin/Boston.

Lisa Matthewson. 2004. On the Methodology of Semantic Fieldwork. *International Journal of American Linguistics*, 70(4):369–415.

Lisa Matthewson. 2013. Gitksan modals. *International Journal of American Linguistics*, 79(3):349–394.

Lisa Matthewson. 2019. Modality. In Maria Aloni and Paul Dekker, editors, *The Cambridge Handbook of Formal Semantics*, pages 525–559. Cambridge University Press.

Amy Miller. 2001. *A grammar of Jamul Tiipay: Amy Miller*, volume 23 of *Mouton grammar library*. De Gruyter Mouton.

Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. 2021. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49).

Maša Močnik and Rafael Abramovitz. 2019. A Variable-Force Variable-Flavor Attitude Verb in Koryak. In *Proceedings of the 22nd Amsterdam Colloquium*, pages 494–503.

Fabrice Dominique Nauze. 2008. *Modality in Typological Perspective*. Ph.D. thesis, Universiteit van Amsterdam.

Irina Nikolaeva. 2014. *A grammar of Tundra Nenets*. Mouton grammar library ; Volume 65. De Gruyter Mouton, Berlin, [Germany] ; Boston, [Massachusetts].

Ashild Næss. 2011. *A grammar of Vaeakau-Taumako*. Mouton grammar library ; 52. De Gruyter Mouton, Berlin ; New York.

Paul Portner. 2009. *Modality*. Oxford University Press.

Hotze Rullmann, Lisa Matthewson, and Henry Davis. 2008a. Modals as distributive indefinites. *Natural Language Semantics*, 16(4):317–357.

Hotze Rullmann, Lisa Matthewson, and Henry Davis. 2008b. Modals as distributive indefinites. *Natural Language Semantics*, 16(4):317–357.

Shane Steinert-Threlkeld. 2021. Quantifiers in Natural Language: Efficient Communication and Degrees of Semantic Universals. *Entropy*, 23(10):1335.

Shane Steinert-Threlkeld, Nathaniel Imel, and Qingxia Guo. 2022. A Semantic Universal for Modality. Submitted to Semantics and Pragmatics.

Shane Steinert-Threlkeld and Jakub Szymanik. 2019. Learnability and Semantic Universals. *Semantics & Pragmatics*, 12(4).

Shane Steinert-Threlkeld and Jakub Szymanik. 2020. Ease of Learning Explains Semantic Universals. *Cognition*, 195.

Wataru Uegaki. 2022. The informativeness / complexity trade-off in the domain of Boolean connectives. *Linguistic Inquiry*.

Jozina Vander Klok. 2013a. Pure possibility and pure necessity modals in pariran javanese. *Oceanic Linguistics*, 52(2):341–374.

Jozina Vander Klok. 2013b. Restrictions on semantic variation: A case study on modal system types. In *Workshop on Semantic Variation*.

Jozina Vander Klok. 2021. Revised Modal Questionnaire for Cross-Linguistic Use. Unpublished.

Kai von Fintel and Sabine Iatridou. 2008. How to Say Ought in Foreign: The Composition of Weak Necessity Modals. In Jacqueline Guéron and Jacqueline Lecarme, editors, *Time and Modality*, volume 75 of *Studies in Natural Language and Linguistic Theory*, pages 115–141. Springer Netherlands.

Hein van der Voort. 2004. *A Grammar of Kwaza*, 1. aufl. edition, volume 29 of *Mouton Grammar Library [MGL]*. Mouton de Gruyter, Berlin/Boston.

Hadley Wickham. 2014. Tidy data. *The Journal of Statistical Software*, 59.

Igor Yanovich. 2016. Old English *motan, variable-force modality, and the presupposition of inevitable actualization. *Language*, 92(3):489–521.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.