Cross-linguistic semantics: Methods and results

Lisa Matthewson
University of British Columbia

Boston University
March 7, 2013

Goals of the talk
• Outline the methodologies of cross-linguistic semantics.
• Illustrate the methodologies with respect to a case study on modals.

The goal of a formal semanticist
• Produce a theory which predicts what every sentence in every language can (and can’t) mean.
• A predictive theory must include proposals about what is universal in semantics, and the nature and extent of cross-linguistic variation.
• What aspects of semantics do all languages share? How can they vary?
• This research program requires cross-linguistic data.

The goal of anyone interested in language description
• It’s not just formal semanticists who should be interested in cross-linguistic semantics.
• Any linguist who works on an endangered language, or an under-described language, needs to document that language as fully as possible.
• Whether our goals are formal, functional, descriptive, typological, or some combination of these ... the first step for everyone is always a correct understanding of the facts.
• Semantics is often the most ignored part of language documentation.

Challenges in cross-linguistic semantics
• Conducting semantic fieldwork on a language one doesn’t speak natively brings many logistical and intellectual challenges.
• Meaning is not surface-visible. We can’t directly record meaning; we need to work with clues to extract it.

Rewards of cross-linguistic semantics
• Empirical contribution: 90% of the world’s languages are endangered, and semantics is often the least-studied aspect of these languages.
• Data from languages dissimilar to Indo-European has the potential to impact linguistic theory in dramatic ways.

What kinds of questions do we need to ask?
For any language X, we need to know how X behaves with respect to things like:
• Definiteness
• Specificity
• Plurality
• Distributivity
• Count vs. mass
• Quantification
• Scope
• Polarity
• Comparatives
• Tense
• Lexical aspect
• Viewpoint aspect
• Modality
• Evidentiality
• Mood
• Conditionals
• Counterfactuals
• Attitude verbs
• Negation
• Generics
• Pronouns
• Questions
• Discourse particles
• Adverbials
• Focus
• Topic
• Presuppositions
• Conventional implicatures
• Conversational implicatures
• —
How do we collect cross-linguistic semantic data?

i. State a research question

Brood: Do languages vary in the semantics of their modals?

Bit narrower: For any language X with modals,

1. Do the modals encode distinctions of modality type?
2. Do the modals encode distinctions of modal force?

Use hypothesis-driven testing

- Linguistics is a cognitive science. We should use the scientific method.
- The fundamental principle underlying the scientific method is that research relies on empirically falsifiable hypotheses.

Steps in the method:

i. State a research question
ii. Make a hypothesis
iii. Gather data to test the predictions made by the hypothesis
iv. Analyze the results to see whether they match the predictions
v. If necessary, revise the hypothesis
vi. Repeat steps (iii-v), as often as necessary

Without an initial hypothesis, no coherent research question is formulated and no empirical predictions are made. Data collection is haphazard, yielding at most only surface generalizations.

Modality type in English

- English modal auxiliaries allow varying interpretations: different modality types.
- An epistemic interpretation for must (based on knowledge or evidence):
  (1) Maria must be in her office (given that her door is open).
- A deontic interpretation for must (based on rules or laws):
  (2) Maria must be in her office (from 9 to 5, because she is the receptionist).
- The same word must is used for both epistemic and deontic modality types; the distinction is not lexically encoded (Kratzer 1981, 1991, a.m.o.).

Modal force in English

- Although English modals typically leave modality type up to context, they lexically encode modal force (the strength of the quantification).
- (3) makes a strong claim; must is a necessity modal.
  (3) She must be in her office. **EPISTEMIC OR DEONTIC ONLY Necessity**
- (4) makes a weaker claim; may/might are possibility modals.
  (4) She may/might be in her office. **EPISTEMIC OR DEONTIC ONLY POSSIBILITY**

i. State a research question

Getting double: Do the modals in St̓át’imcets encode distinctions of modality type and modal force?

- St̓át’imcets (a.k.a. Lillooet) is a Northern Interior Salish language, spoken in southwest British Columbia.
- It has fewer than 200 speakers remaining.
ii. Choose an initial hypothesis

- Guiding principle: Choose an initial hypothesis which is easily falsifiable by certain clear necessary and sufficient evidence.
- Don’t worry if your initial hypothesis is one which assumes that the language under investigation will be similar to a familiar language like English. This does not mean that you will miss detecting variation, where it exists (pace Evans and Levinson 2009).
- In the scientific method, there is a crucial distinction between one’s initial hypothesis (step (ii)), and one’s eventual analysis (step (v), after potentially many iterations). An initial hypothesis is not a belief: it is a heuristic.
- In fact, in linguistics just as in other social and physical sciences, researchers are often specifically attempting to disprove their initial hypothesis through empirical testing.

Adopting an initial hypothesis

- The most easily falsifiable hypothesis for each of our modal questions is that the relevant distinction (either modality type, or modal force) is encoded.
- Take modality type. Suppose we observe a certain modal being used in epistemic contexts. We make the initial hypothesis that this modal is strictly an epistemic modal. This initial hypothesis is falsifiable by finding cases where the modal appears in non-epistemic contexts.

Our two initial hypotheses:

- St’át’ímcets modals lexically encode modality type (i.e., are unlike English).
- St’át’ímcets modals lexically encode modal force (i.e., are like English).

iii. Gather data to test the predictions made by the hypothesis

Linguistic data-collection methodologies

- Corpus study
- Fieldwork
- Psycholinguistic experiments
- Neurolinguistic experiments
- A range of methodologies should be applied wherever possible.

Desiderata for methodologies

Desideratum 1: Negative evidence

- Knowing a language entails knowing what one cannot say, as well as what one can, and knowing what things can’t mean, as well as what they can.
- The data must include not only positive examples of what is said, but information about what cannot be said: negative evidence.
- Corpora do not include negative evidence. Therefore, scientific testing of hypotheses about language should if possible not rely exclusively on the study of corpora (although they are a useful tool).

St’át’ímcets modals in corpora

- Modals being investigated: ka, k’a, ko-...-a.
- Tokens in corpus: ka: 4 k’a: 143 ko-...-a: 94

Desideratum 2: Reproducibility

- Core principle of the scientific method: results should be reproducible by other researchers.
- The way to make this possible is to provide full and explicit information about one’s methodology, so that other researchers can attempt to replicate the results.
- Any method of data collection – corpus study, fieldwork, or experimentation – can meet the desideratum of reproducibility.
Fieldwork and experiments

- Targeted one-on-one fieldwork and large-scale experiments both allow hypothesis-testing, provide negative evidence, and allow reproducibility.
- Large-scale experiments are impossible for many minority and endangered languages. Small-scale experiments – i.e., fieldwork – are the only option.

Large-scale experiments in St’át’imcets?

- Fewer than 200 speakers (possibly fewer than 100).
- All speakers elderly, some very elderly.
- From different dialect areas.
- Almost none are literate in the language.
- Large-scale experiments are impossible.

Fieldwork and experiments

- Sprouse et al. (2012) randomly selected 146 two-condition phenomena from articles in Linguistic Inquiry, which were originally gathered using non-experimental methods. They tested each of these data points experimentally using magnitude estimation, and found a replication rate of 95% (with a margin of error of just over 5%).
- Advantage to one-on-one fieldwork: The time spent with each speaker.
- Fieldworkers are able to engage speakers in discussion of why particular examples sound bad, how they could be improved, and so on (see den Dikken et al. 2007:350).

Semantic fieldwork methodologies

- Basic principle: establish the (un)acceptability of utterances in particular discourse contexts, in order to determine truth and felicity conditions.

Elicited Production Task

- Elicit productions as responses either to verbal stimuli, or to visual cues.
- In all cases, productions are elicited in particular discourse contexts, rather than as isolated utterances translated from the contact language.

Semantic fieldwork methodologies

Felicity Judgment Task

- Consultants judge the acceptability of utterances in discourse contexts which are described verbally, or using pictures, props or storyboards.
- Acceptability entails truth in a context (consultants never accept a sentence which is false in the given context), but unacceptability may arise for different reasons, including falsity, presupposition failure, or other pragmatic inappropriateness.
- The researcher often discusses the utterance with the consultant in more detail, eliciting comments on the acceptability level, how the consultant interprets the utterance, why she feels it is unacceptable, how either the sentence or the context can be repaired, etc.

Targeted Construction Storyboards

- A way of obtaining spontaneously produced data, and also facilitating follow-up judgment tasks.
- Pictorial representations of stories, which consultants are asked to tell in their own words. Each story includes at least one targeted context that is used to test hypotheses about the relation between linguistic forms and that context.
- Advantages: Convey complex discourse contexts pictorially, avoiding translation interference and/or speaker confusion; provide more spontaneous and natural speech than direct elicitation; enable easy reproducibility across speakers, languages, and linguists; fun for speakers.
- Disadvantages: Time-consuming to prepare.
- www.totemfieldstoryboards.org; Burton and Matthewson (2013).
Fieldwork results, k’a

- k’a is accepted and offered in epistemic contexts, where the speaker is reasoning based on evidence or knowledge.

(5) waʔ-k’a ɬ-waʔ mól-k’a ɬ-mól k’a-Philotoma k’a-Philomena
   be=MODAL stat-stop in=DET=house POSS=kwaʔ POSS=kwaʔ ‘Philomena must / might be in her house.’ (Rullmann et al. 2008)

(6) 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; it must just be tension.
   niʔ-k’a ɬ-niʔ-k’a ɬ-niʔ-k’a-Philotoma ɬ-niʔ-k’a-Philotoma
   ‘It must be from my worrying.’ (Rullmann et al. 2008)

(7) Context: His car isn’t there.
   plán-k’a ɬ-tuʔ aʔ-k’a-leave
   ‘Maybe he’s already gone.’ (Rullmann et al. 2008)

Fieldwork results, ka-...-a

- ka-...-a is accepted and offered in pure circumstantial and ability contexts (Davis et al. 2009).

(11) naʔ-k’a ɬ-niʔ-k’a ɬ-niʔ-k’a-Philotoma ɬ-niʔ-k’a-Philotoma
   ipfv=3SG=SI ɬ-crc=lift crc=lift crc=reck crc=reg+exs
   ‘I can lift the rock.’ (Davis et al. 2009)

(12) ko-k’a ɬ-ko-k’a ɬ-ko-k’a-Philotoma ɬ-ko-k’a-Philotoma
   going-to=CRC=1SG=SI=crc=wake crc=wake crc=comp=daybreak crc=3SIV
   ‘Are you going to be able to wake up at dawn?’ (Davis et al. 2009)

(13) Context: Gillian had a very bad cough yesterday. Her nose was really plugged up.
   kán-k’a ɬ-kaʔ ɬ-káʔ ɬ-kaʔ ɬ-kiʔ-ko-k’a-nák’-a ɬ-káʔ-ko-k’a-nák’-a
   try-eat crc=sweet bread crc=sneeze crc
   ‘She started to eat some sweet bread, but she had to sneeze.’ (Davis et al. 2009)

Sample results showing St’át’imcets does encode modality type

- Epistemic k’a is rejected in deontic situations. (14) can be interpreted only epistemically; it cannot be used to mean that Henry should knock:

(14) naʔ-k’a ɬ-niʔ-k’a ɬ-niʔ-k’a-Philotoma ɬ-niʔ-k’a-Philotoma
   ipfv=3SG=SI ɬ-crc=lift crc=lift crc=reg+exs
   ‘That’ll be Henry knocking.’ (Rullmann et al. 2008:321)

(15) ko-k’a ɬ-ko-k’a ɬ-ko-k’a-Philotoma ɬ-ko-k’a-Philotoma
   knock.repeatedly:open-ING
   ‘That’ll be Henry knocking.’

- Deontic/irrealis ko cannot be used in epistemic contexts:

(16) niʔ-k’a ɬ-niʔ-k’a ɬ-niʔ-k’a-Philotoma ɬ-niʔ-k’a-Philotoma
   ‘Mary could have cooked this.’ (It tastes like her cooking.) (Rullmann et al. 2008)
Testing using storyboards

Chore Girl and Sick Girl

- Chore Girl and Sick Girl are designed to investigate whether languages distinguish between different types of modality (as well as modal force).
- These storyboards elicit the contrast (or lack of one) between deontic possibility (permission), non-deontic circumstantial possibility (ability), and deontic necessity (obligation).

(17) a. I can come out to play (I am allowed to).  PERMISS
b. I can come out to play (I’m not sick or injured).  ABILITY
c. I have to do my chores.  OBLIGATION

- The basic plot involves a girl whose friends repeatedly ask her to come out and play.
- www.totemfieldstoryboards.org

Sick Girl – sample pictures

- Mary repeatedly injures herself or gets sick, so is unable to go out to play. When she finally recovers from all her ailments, she is physically able to play, but is now _not permitted_ to go out and play, because her mother says she has to do her homework.

Feeding Fluffy

- Feeding Fluffy is designed to elicit epistemic modals with different temporal properties.
- Stacey is pet-sitting for Pat. She visits the store and realizes she doesn’t know what kind of pet Pat has. She thinks it could be a cat, a dog or a rabbit, so she buys food for all those animals. She later finds out that Fluffy is actually a snake. When Pat questions her about why she bought cat-, dog- and rabbit-food, she replies that Fluffy might have been a cat, a dog or a rabbit.

Sample results from storyboards

- As predicted, circumstantial _ko...a_ is used in ability contexts.

(18) s̕əg̡aʔ kʷ=an̕=aʔá kʷ-s̕əʔá laʔ kʷ-s̕əʔá n̕=a s̕əg̡aʔá
  "But I can’t play, I hurt myself and broke my foot."
  (Chore Girl)

- Apparent wrinkle: _ko...a_ is also used in permission contexts.

(19) s̕əg̡aʔ kʷ=an̕=aʔá kʷ-s̕əʔá laʔ kʷ-s̕əʔá say n̕=a s̕əʔá kʷ=an̕=aʔá s̕əg̡aʔá say n̕=a m̕=i=awá
  "My mother said I could go play.”
  (Chore Girl)

- This is just like in many languages: a general circumstantial modal can be used in all circumstantial contexts, which include deontic ones (can vs. be allowed).

- Follow-up elicitation is required to test the prediction that deontic _ko_ is also ok in (19) (and not ok in (18)). Remember the corpus results? _Ko_ is rare in texts.

Feeding Fluffy elicits epistemic _k’a_

- As predicted, _k’a_ appears when Stacey is reasoning that it is epistemically possible that Pat has a dog, etc.

(20) s̕əg̡aʔ kʷ=an̕=aʔá s̕əg̡aʔá s̕əg̡aʔá s̕əg̡aʔá s̕əg̡aʔá
  "It might be a dog,” she thought.”
  (Feeding Fluffy)

- Follow-up elicitation required to obtain negative data; test whether the other modals are felicitous in this context from the story. [Not done yet!]

31
32
33
34
35
36
iv. See whether the St’át’ímcets modal results match our predictions

Our two initial hypotheses:

- St’át’ímcets modals lexically encode modality type.
  - CONFIRMED.
- St’át’ímcets modals lexically encode modal force.
  - FALSE.

v. Revise the hypothesis

- St’át’ímcets is organized in the exact opposite way to that of English: St’át’ímcets modals all lexically encode modality type, but leave modal force up to context.
Cross-linguistic formal results for modal systems

• Some languages encode both type of modality and modal force, e.g. Javanese (Vander Klok 2008), Blackfoot (Reis Silva 2009).

<table>
<thead>
<tr>
<th>Language</th>
<th>DEONTOIC</th>
<th>EPISTEMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>necessity</td>
<td>possibility</td>
</tr>
<tr>
<td>DEONTOIC</td>
<td>must</td>
<td>may</td>
</tr>
<tr>
<td>EPISTEMIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stát'imcets</td>
<td>necessity</td>
<td>possibility</td>
</tr>
<tr>
<td>DEONTOIC</td>
<td>k’a</td>
<td>k’a</td>
</tr>
<tr>
<td>EPISTEMIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Javanese</td>
<td>necessity</td>
<td>possibility</td>
</tr>
<tr>
<td>DEONTOIC</td>
<td>kudu</td>
<td>ento, olèh</td>
</tr>
<tr>
<td>EPISTEMIC</td>
<td>mesthi</td>
<td>mungkin</td>
</tr>
</tbody>
</table>

Some systems mix-and-match, or encode sub-types of modality

• Gitxsan encodes modal force only for non-epistemics (circumstancials). Within the circumstancials, it encodes sub-types of modality (Peterson 2010, Matthewson in press).

<table>
<thead>
<tr>
<th>Modality</th>
<th>necessity</th>
<th>possibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>circumstantial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-deontic</td>
<td>daʔaq(ɨ)xʷ</td>
<td>agi</td>
</tr>
<tr>
<td>deontic</td>
<td>ʔanooq(ɨx*)</td>
<td></td>
</tr>
<tr>
<td>epistemic</td>
<td>plain</td>
<td></td>
</tr>
<tr>
<td>reportive</td>
<td>ʔinaq(ʔa)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gal</td>
</tr>
</tbody>
</table>

Results of hypothesis-driven research

• Cross-linguistically, modals vary in what they choose to encode: modality type, modal force, or both.

Use these techniques for any area of semantics

• Form an initial hypothesis which is falsifiable by necessary and sufficient evidence.

• Test the hypothesis using Felicity Judgment Tasks and Elicited Production Tasks (including storyboards, etc.).

• Establish generalizations, make analyses, test the further predictions …

Examples of semantic diversity in other areas

• Determiners: Encode different semantic distinctions (Matthewson 1998, 1999, Ionin 2006, Gillon 2006, etc.).


• Aspect: Languages vary in the semantics of their lexical aspectual classes, and in the semantics of their viewpoint aspects (Smith 1997, Bar-el 2005, Kiyota 2008, etc.).

• Quantifiers: Some languages lack generalized quantifiers (Jelinek 1995, Davis 2010, etc.).

Examples of semantic diversity in other areas

• Modals: Languages may or may not distinguish necessity from possibility (Rullmann et al. 2008, Peterson 2010, Deal 2011, etc.).

• Evidentials: Encode different distinctions and may have different status in terms of (lack of) propositional content (Izvorski 1997, Faller 2002, McCready & Ogata 2007, Matthewson et al. 2007, Murray 2010, Peterson 2010, etc.).

• Degrees / comparatives: Languages vary in whether they introduce degree arguments, etc. (Kennedy 2007, Beck et al. 2009, Bocznak 2012, etc.).

• Presuppositions: Some languages lack presuppositions of the English type, which require information to be in the common ground at the time of utterance (Davis et al. 2004, Matthewson 2006a, 2009, Davis et al. 2007, etc.).
### The methodological point

- The only way to accurately detect the semantics of modals – the only way to accurately establish the descriptive facts – is to conduct formal, hypothesis-driven research of the type outlined above.
- Large-scale typological research into modals, such as van der Auwera and Ammann (2011) (the World Atlas of Linguistic Structures chapter on modality), cannot give us a complete or accurate picture.
- Examine 207 languages, with the goal of documenting to what extent languages use identical markers for epistemic and non-epistemic modality.
- Generalizations are almost exclusively extracted from secondary sources, of which the vast majority are descriptive grammars.

### Problems with doing it this way

- Data are collected primarily from descriptive grammars, which are not usually written on the basis of targeted semantic fieldwork.
- Information about modals in each of these grammars is sparse (typically 2-3 pages), lacks any negative data, and normally contains simply a list of modal words and some translations of those into English.
- vdA&A often mis- or over-interpret the data in the grammars (see Matthewson 2012b, Matthewson et al. 2012 for examples).
- One example: if we had as much information about English as we do in many of the source grammars for vdA&A, we would mis-categorize English as a language whose modals do lexically encode modality type.
- These problems arise with any attempt to do semantics on the basis of descriptive secondary sources.

### What is needed

- Lots of people doing hypothesis-driven semantic fieldwork.
- Typology is only as good as its primary sources.
- WALS would also be better if it were collaborative and self-improving – like the Syntactic Structures of the World’s Languages website (http://swrl.railsplayground.net/).
- This website does not have a lot of semantic information yet, but it can be expanded, with all of our help.

### Some theoretical points

- Formal analyses and formal typologies shed light on our theoretical understanding of what modals can be like in natural languages.
- The St’át’lmcets system, and others like it which encode modality type but not modal force, challenge our ideas of what it means to be a modal (cf. de Haan 1999, who argues that epistemic modals necessarily encode modal force distinctions).
- The St’át’lmcets system also impacts on the debate about where modality type comes from – the context (Krator 1981, 1991), or as a result party of syntactic position (Hacquard 2006)?
- St’át’lmcets provides empirical support for the standard binary division between epistemic and circumstantial modality (unlike e.g., a three-way primitive distinction between modality types, as argued for by Portner 2009).

### Back to a broad question

- How can languages vary in their semantics?
- In the area of modality, semantic variation seems to be restricted largely to choices about which elements of meaning are lexically distinguished.
- Languages choose whether to lexically encode modality type, and modal force, but the same building blocks are used (quantification over possible worlds, epistemic and circumstantial conversational backgrounds), and the same overall semantic space is covered (exception: Nez Perce, as analyzed by Deal 2011).

### Thanks to

- St’át’lmcets consultants Carl Alexander, Gertrude Ned, Laura Thevarge, the late Beverley Frank and the late Rose Agnes Whitley.
- Henry Davis and Hotze Rullmann; I Neidle and Paul Hagstrom.
- SSHRC grant #410-2011-0431.