Developing a methodology for modality type annotations on a large scale

Aynat Rubinstein, Dan Simonson, Joo Chung, Hillary Harner, Graham Katz and Paul Portner

April 20, 2012

GEORGETOWN UNIVERSITY
Georgetown College
Department of Linguistics

Modality types

Modal expressions are traditionally classified according to the kind of possibilities they invoke in a given context.

- Epistemic: based on evidence or knowledge

  Mary must have a good reason for being late; *Hydrangeas might grow here*.

- Circumstantial: based on some circumstances

  Hydrangeas can grow here.

- Dynamic: based on what someone/something can do

  John can swim; *You can see the ocean from here*.

- Teleological: based on the achievement of goals

  You could add some more salt to the soup.

- Bouletic: based on what someone wishes or desires

  You should try this chocolate.

- Deontic: based on what the rules provide

  The rich must give money to the poor.

(Kratzer 1981, Coates 1983, Palmer 2001, . . . , examples from Kratzer 1981 and Portner 2009)

Resolution is difficult

Even examples that seem straightforward aren't so:

> *Many workers have contracts that require them to work in a European Union state different from their country of origin. If they are unfairly dismissed or otherwise have action taken against them they need to know which country it is that they* ***can*** *pursue their claim in. The latest ruling from the European Court of Justice on the issue suggests that it should be the state in which the employee has worked the longest.*

Deontic	40% (2/5)
Circumstantial	40% (2/5)
Dynamic	20% (1/5)

Resolution is difficult

Sometimes resolution is impossible:

> *As well as the remote systems, the escape systems are also dead. Jeff reviews the situation; they **need** to get the escape pod working. He sends Scott and Brains in Thunderbird 1 and Virgil, Alan and Gordon in Thunderbird 2 with Pod 4. Zero-X is descending at 3000 feet a minutes.*

Deontic	20% (1/5)
Teleological	40% (2/5)
Bouletic	40% (2/5)

Motivation

However, it's useful information to have.
It allows us to address questions like:

1. Is there a correlation between the syntactic configuration the modal appears in and its meaning?
2. Do certain form-meaning correlations generalize beyond particular modal words?
3. Correlations between tense and modality type?
4. Correlations between aspect and modality type?

   ⋮

5. Other, new correlations we'll want to investigate one day?

Recent annotation efforts

	[1]	[2]	[3]
Words	150 lemmas	*must* (and others)	*must, have to*
Types	non standard	root/epistemic	root/epistemic
Annotators	expert	expert	expert
Sources	written	written, spoken	written
No. of instances	249	1508 (141)	2426

[1] Baker et al. (2010)
[2] de Haan (2011)
[3] Hacquard & Wellwood (to appear)

Also: Wärnsby (2006), annotations of factuality/veridicality (FactBank and enhancements, Genia biomed).

Common features: expert annotations, two-way, no ambiguity.

Goals of this research

To develop a methodology for modality type annotations that:

(i) distinguishes between multiple modality types;

(ii) supports large scale annotations;

(iii) i.e., by native speakers that are not specialists in modality;

(iv) allows representation of true ambiguity.

Part of larger project of annotating gradability properties of modal expressions.

Agenda for today

Introduction

- Modality type disambiguation
- Relation to previous work

Pilot experiment

- Experimental setup
- Results
  - Quality of gold standard items
  - Quality of corpus items

Methodology for modality type annotations

- Embracing ambiguity
- Applications

Conclusion

Amazon's Mechanical Turk

MTurk is an online labor market where workers are paid small sums of money to complete Human Intelligence Tasks (HITs).

- Various tasks relating to natural language understanding.
- "Fast and cheap".
- "Judgments of many non-experts = quality of experts" (Snow et al. 2008, Callison-Burch 2009).
- Standard Qualifications:
  - Located in the United States.
  - Approval rate $\geq 95\%$.
  - *In the future, require experience: # of approved tasks* > 500 (Akkaya et al. 2010).

Sample item

2. *Many workers have contracts that require them to work in a European Union state different from their country of origin. If they are unfairly dismissed or otherwise have action taken against them they need to know which country it is that they **can** pursue their claim in. The latest ruling from the European Court of Justice on the issue suggests that it should be the state in which the employee has worked the longest.*

QUESTION #1: In your opinion, which of the descriptions below *best describes* the meaning of the highlighted word in the context of the passage? (choose ONE only)

- ○ 1. What someone KNOWS or CONCLUDES (on the basis of information).
- ○ 2. What someone or a set of rules REQUIRES or PERMITS.
- ○ 3. What someone DESIRES.
- ○ 4. What is involved in ACHIEVING a GOAL.
- ○ 5. What someone (or something) has the ABILITY TO DO.
- ○ 6. What the circumstances DETERMINE or ALLOW.

Question #2: In addition to the answer that you chose, were there any other answers that you considered? (choose AT LEAST ONE)

- ☐ Didn't consider any alternative.
- ☐ 1. What someone KNOWS or CONCLUDES (on the basis of information).
- ☐ 2. What someone or a set of rules REQUIRES or PERMITS.
- ☐ 3. What someone DESIRES.
- ☐ 4. What is involved in ACHIEVING a GOAL.
- ☐ 5. What someone (or something) has the ABILITY TO DO.
- ☐ 6. What the circumstances DETERMINE or ALLOW.

Examples: The meaning of the underlined word may have to do with:

1. What someone KNOWS or CONCLUDES (on the basis of information).
   e.g., *It **must** be an address in England (because of its postal code).*
2. What some authority or set of rules REQUIRES or PERMITS.
   e.g., *You **must** put all garbage in the brown box (because otherwise you will get a fine).*
3. What someone DESIRES.
   e.g., *Peter **should** go to the Rolling Stones concert next week (because he likes the Stones).*
4. What is involved in ACHIEVING A GOAL.
   e.g., *To get to the University, you **should** take the bus at the corner.*
5. What someone (or something) has the ABILITY TO DO.
   e.g., *He is **able** to lift 200 lbs (because he has been training).*
6. What the CIRCUMSTANCES DETERMINE or

Modality type characterizations in the instructions

1. What someone KNOWS or CONCLUDES (on the basis of information).
   e.g., *The bridge is **bound** to collapse (because it is poorly designed).*
2. What some authority or set of rules REQUIRES or PERMITS.
   e.g., *You **must** put all garbage in the brown box (because otherwise you will get a fine).*
3. What someone DESIRES.
   e.g., *Peter **should** go to the Rolling Stones concert next week (because he likes the Stones).*
4. What is involved in ACHIEVING A GOAL.
   e.g., *To get to the University, you **should** take the bus at the corner.*
5. What someone (or something) has the ABILITY TO DO.
   e.g., *He is **able** to lift 200 lbs (because he has been training).*
6. What the CIRCUMSTANCES DETERMINE or ALLOW.
   e.g., *It **must** be an address in England (because of its postal code).*

Task specifics

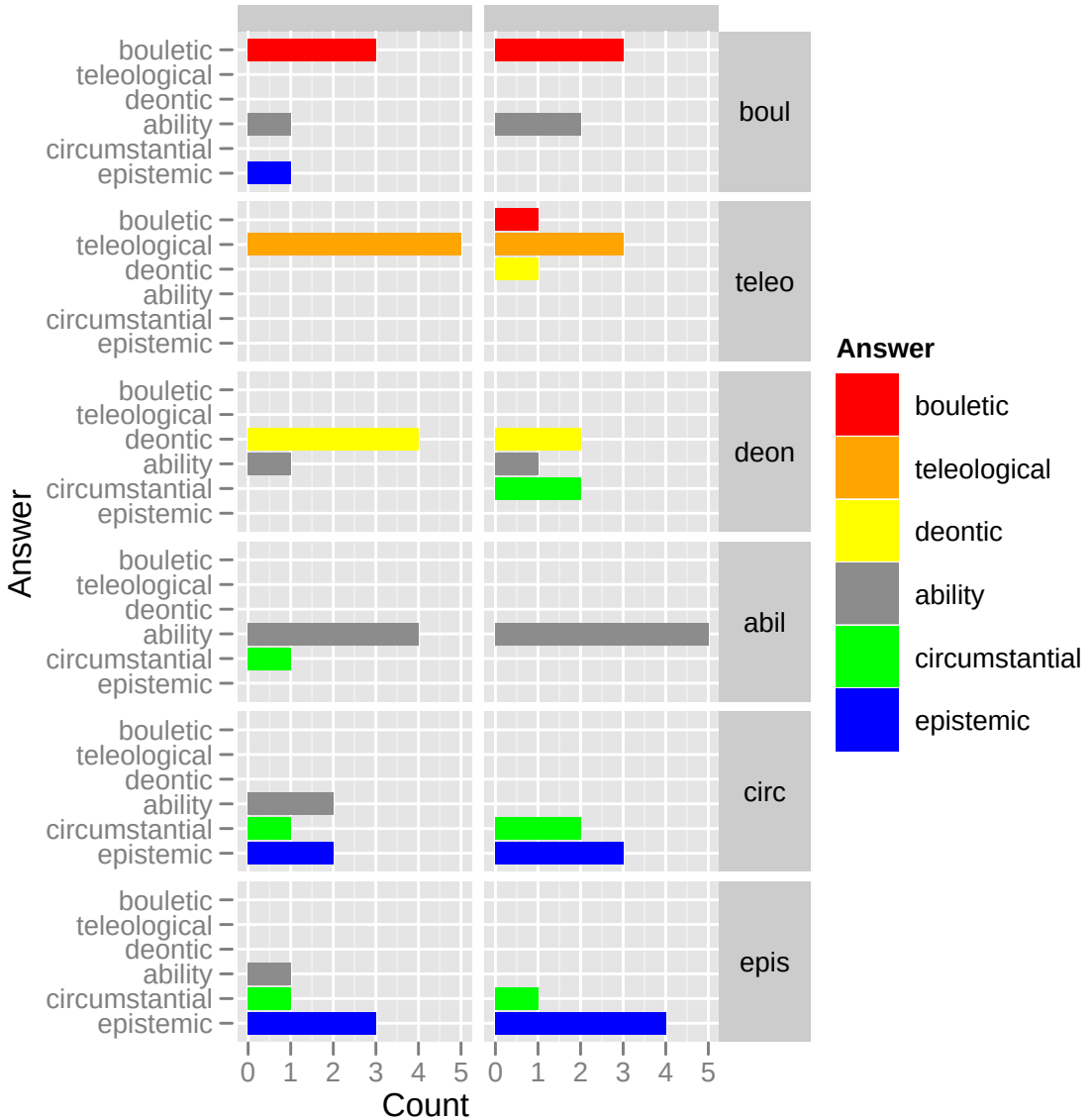
- Minimal context window of 1 sentence before/after the target sentence.
- Five modals: ***can, certain, chance, likely, need***.
  - Necessity and possibility.
  - Different parts of speech.
- 12 gold standard items: 2 for each modality type.
  - Bouletic
  - Teleological
  - Deontic
  - Ability
  - Circumstantial
  - Epistemic
- 48 corpus items: web-derived from the `.uk` domain (Ferraresi 2007).
- 5 annotations for each target sentence.
- We only considered responses from Turkers that completed entire batches (10 paragraphs).

Gold standard items

- Majority response matched expert opinion in 9 out of 12 items (75%).
- Difficulty in the Circumstantial class.

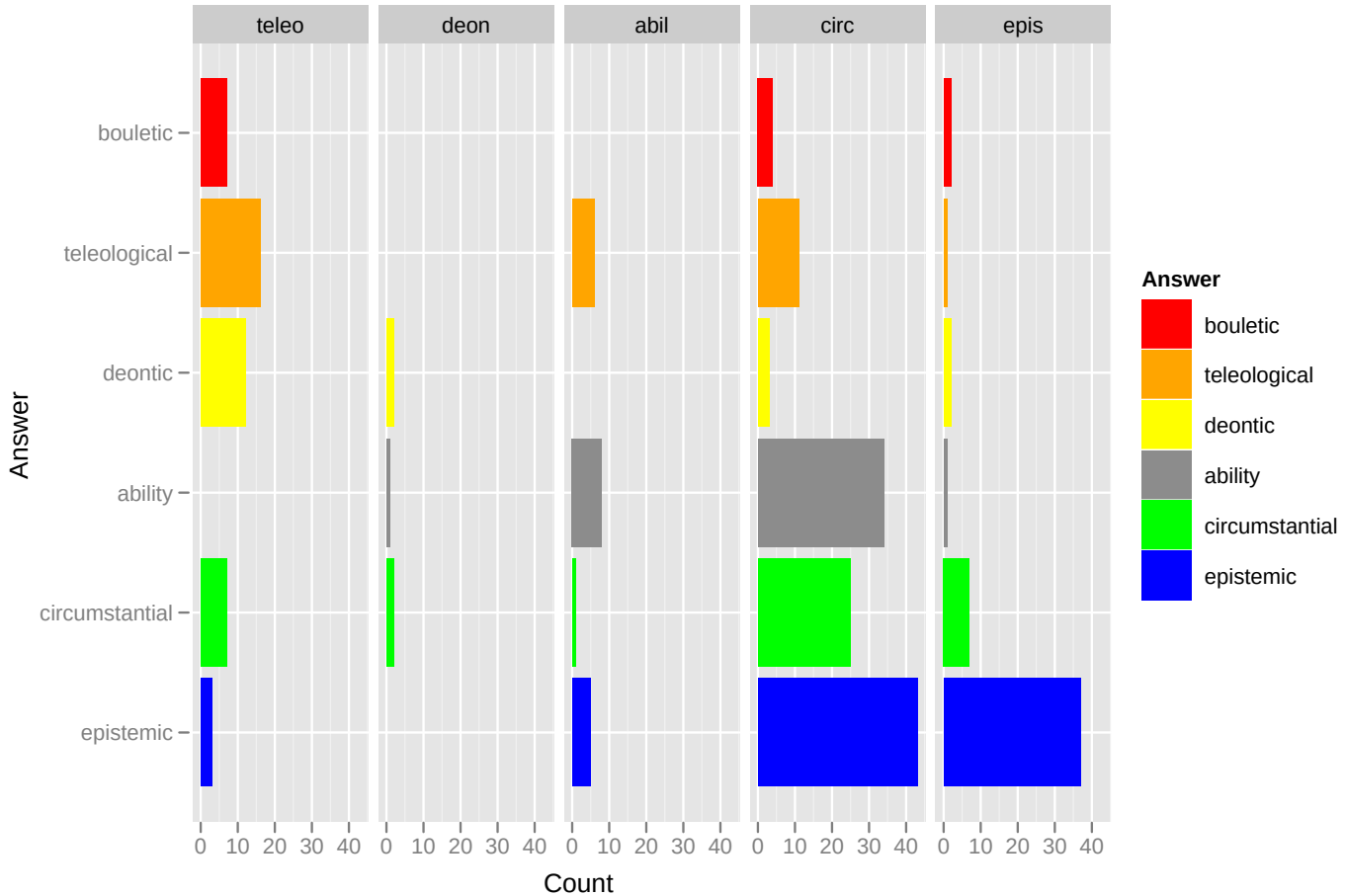
  If you pitch a tent next to the stream and it rains, there's **a possibility** that you will find yourself wet in your sleeping bag in the morning.

  . . . many of the regulars weren't there that day, so it was **possible** that Thom would win most of his games.



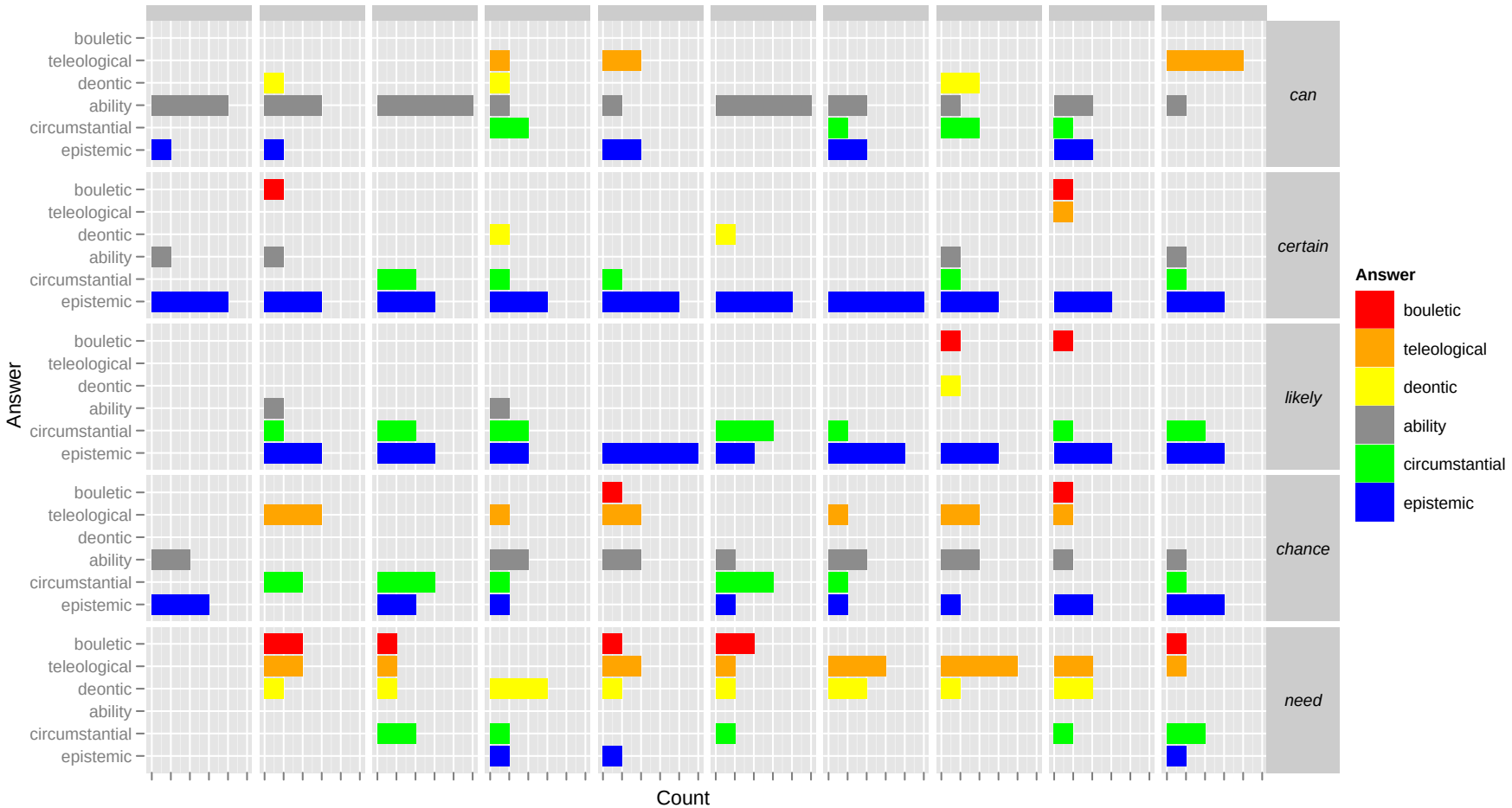
Corpus items

- Majority response matched an expert's opinion in 16 out of 48 items (33%).



- Difficulty in the Circumstantial class swamps results:
  - 24 of the 48 items were classified as Circumstantial by the expert.
  - Only 3/24 had a majority vote of Circumstantial.

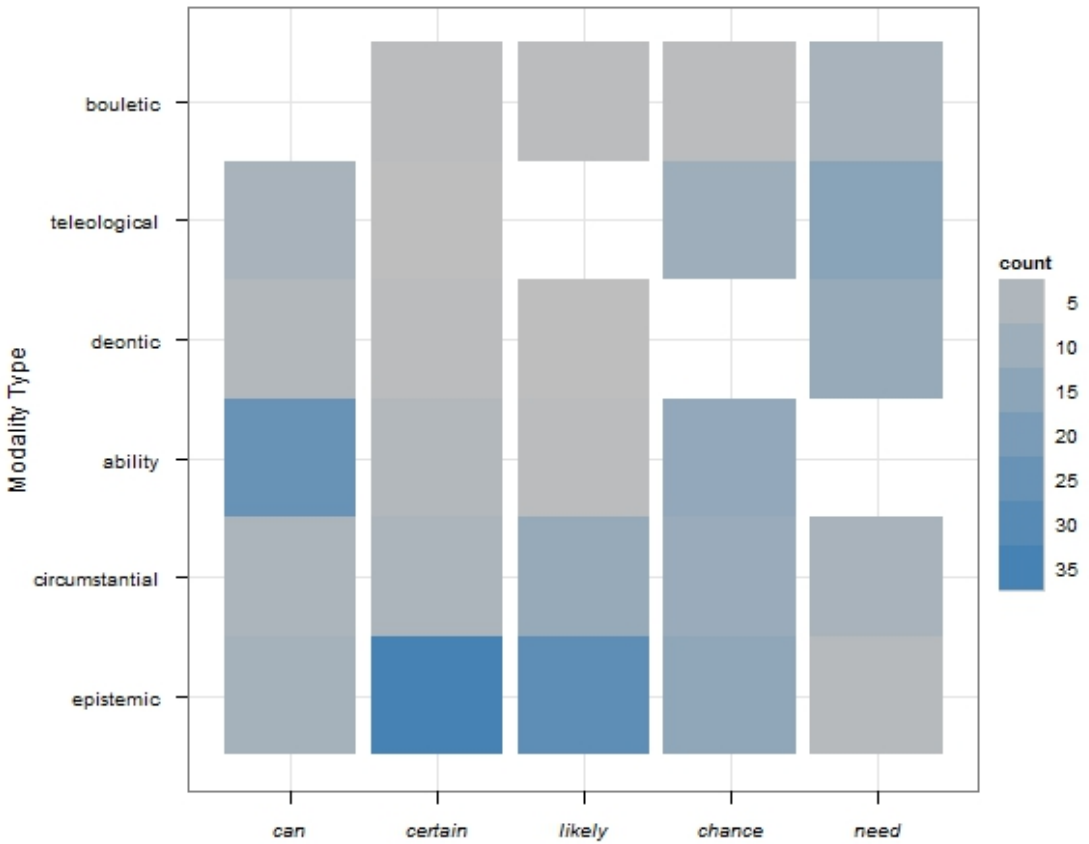
Corpus items (by modal word)



Corpus items (by word, aggregates)

Responses match the modality types associated in the literature.

- *Certain* and *likely* are mainly epistemic (Portner 2009)
- *Need* gets primarily root interpretations (Smith 2003)



Interannotator agreement

Interannotator agreement, the K statistic (Fleiss 1971):

	All six	C/E collapse	C/E/A collapse
gold standard	0.37	0.49	0.56
corpus	0.18	0.25	0.26

For comparison:

- Word Sense Disambiguation (baselines): K in the range of 0.36 to 0.67 (French polysemous words, Véronis 1998), K = 0.58 on verb senses (Mihalcea et al. 2004)
- "Ceiling": expert annotations of a root vs. epistemic necessity, $k = 0.84$ (Hacquard & Wellwood, to appear)

Conclusions from the experiment

- Non-experts are sensitive to the difference between priority, ability, and epistemic modality.
- They are also sensitive to distinctions within the class of priority modals (seen in responses to gold standard items).
- There is confusion in distinguishing the Circumstantial and Epistemic classes.
  - Epistemic was characterized as encompassing not just knowledge but also “what can be concluded on the basis of information”.
  - This comes very close to “what the circumstances determine or allow”.

Methodology for modality type annotations

How should we go about annotating modal words for modality type?

- Only one type per word?
- Better: a distribution over modality types (per word).
- Use the individual choices to represent the range of possible interpretations the word can get in the context.

Embracing ambiguity

Markup each modal with a distribution of annotators' responses over the set of modality types (a vector), instead of one type (a tag).

*As well as the remote systems, the escape systems are also dead. Jeff reviews the situation; they **need** to get the escape pod working. He sends Scott and Brains in Thunderbird 1 and Virgil, Alan and Gordon in Thunderbird 2 with Pod 4. Zero-X is descending at 3000 feet a minutes.*

Deontic	20% (1/5)
Teleological	40% (2/5)
Bouletic	40% (2/5)

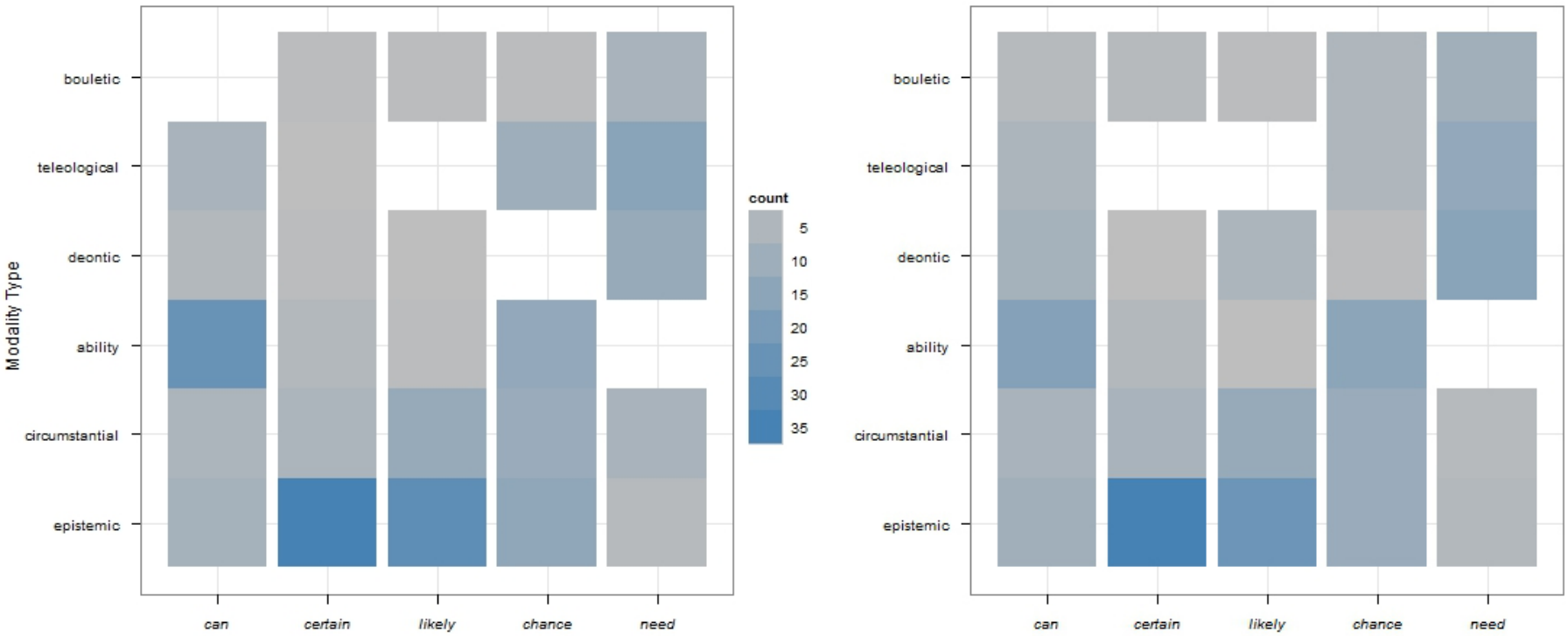
$$need_{modality} = < 0.4, 0.4, 0.2, 0, 0, 0 >$$

Review process

- An expert can review items that receive low interannotator agreement scores:
  - make corrections;
  - retain ambiguity where it genuinely occurs.
- Responses could be weighted to reflect the reliability of individual annotators.

Fingerprints of modal inventories

As long as the modality types, and their definitions, stay constant, similarity between modal words from different languages can be compared.



Comparison of modal inventories across languages

The profile of *need* in the judgments of US Turkers:

$$need_{US} = < 0.16, 0.36, 0.27, 0, 0.16, 0.07 >$$

Which modal in GB is $need_{US}$ most similar to?

$need_{GB} = < 0.2, 0.29, 0.36, 0, 0.07, 0.09 >$	corr = 0.88
$can_{GB} = < 0.06, 0.12, 0.16, 0.34, 0.14, 0.18 >$	corr = -0.63
$likely_{GB} = < 0.04, 0, 0.13, 0.02, 0.27, 0.53 >$	corr = -0.36
$certain_{GB} = < 0.06, 0, 0.02, 0.08, 0.14, 0.7 >$	corr = -0.49
$chance_{GB} = < 0.1, 0.1, 0.04, 0.31, 0.22, 0.22 >$	corr =-0.83

Conclusion

Pilot experiments with Mechanical Turk suggest that it is possible to produce annotations of modal words in context that:

- distinguish between multiple modality types;
- are done on a large scale, with native speakers;
- allow for the representation of true ambiguity.

Instead of reducing the distinction just to root/epistemic, we propose to embrace ambiguity in the annotations.

- each individual provides one judgment;
- the judgments are aggregated to provide a distribution over modality types.

The resulting corpus should be a useful resource for crosslinguistic research.

Discussion: what would be most useful for the community?

Acknowledgments

Thanks to Dr. Scott Jackson from the University of Maryland for assistance with creating the graphics.

This research is supported by the National Science Foundation under Grant No. BCS-1053038 to Graham Katz, Paul Portner, and Elena Herburger.

Selected References

Artstein, R. and M. Poesio. 2007. Inter-coder agreement for computational linguistics. Extended version of article submitted to the journal Computational Linguistics, July 5.

Baker, K., M. Bloodgood, B. J. Dorr, N. W. Filardo, L. Levin, and C. Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceeding of LREC 2010*, 1402–1407.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of EMNLP 2009*, 286–295.

de Haan, F. 2011. Disambiguating modals: constructions and *must*. Ms.

Hacquard, V. and A. Wellwood. to appear. Embedding epistemic modals in English: A corpus-based study. *Semantics & Pragmatics*.

Kratzer, A. 1981. The notional category of modality. In *Words, worlds, and contexts*, ed. H.-J. Eikmeyer and H. Rieser, 38–74. Walter de Gruyter.

Portner, P. 2009. *Modality*. Oxford: Oxford University Press.

Smith, N. 2003. Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In *Modality in contemporary English*, ed. R. Facchinetti, M. Krug, and F. Palmer, 241–266. Berlin: Mouton de Gruyter.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, 254–263.