Task	Example	Typical Use	Strengths	Limitations	E.g.
LM /Generation?	The boy by the boats [is/*are] smiling.	Syntactic phenomena	No additional training on top of pretrained LM	Often uses ppl, so best for left-to-right language models. Harder to use for newer variants.	Linzen et al. (2016)
Acceptability	The boy by the boats [is/*are] smiling.	Syntactic and semantic phenomena	More flexible than LM across architectures; well studied in ling.	Usually requires additional training on top of LM.	Warstadt et al. (2020)
NLI	The boy is smiling. -> The boy [is/*is not] happy.	Semantics/pragmatics/ world knowledge	Flexible, easy to "recast" many tasks to NLI; long history	Often awkward sentences/confounds; low human agreement	White et al. (2017)
Generation	Dante was born in [Mask]	Semantics/pragmatics/ world knowledge	Can be more natural than NLI; incorporates more context	Hard to auto evaluate, esp. beyond one word/factoid questions	Petroni et al. (2019)
MT	The repeated calls from his mother **should** have alerted us. / Les appels rep´ et´ es de sa m ´ ere **devraient** ` nous avoir alertes.	Multilingual morpho-/lexico-/syntax (e.g. cross-lingual agreement)	Only way of specifically probing cross-lingual systems	Often relies on manual eval (though recent approaches use probabilities similar to in LM tasks)	Isabelle et al. (2017)

Experimental Designs

- Tightly Controlled
- Loosely Controlled
- Adversarial Examples

Experimental Designs: Tightly Controlled

Experimental Designs: Tightly Controlled

- Minimal Pairs/Counterfactuals

Experimental Designs: Tightly Controlled

- Minimal Pairs/Counterfactuals

Gender Bias: Rudinger et al. (2018)

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

Experimental Designs: Tightly Controlled

- Minimal Pairs/Counterfactuals

Gender Bias: Rudinger et al. (2018)

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

Subj.-Verb Agree.: Marvin and Linzen (2018)

a. The farmer that the parents love swims.
b. *The farmer that the parents love swim.

Experimental Designs: Tightly Controlled

- Minimal Pairs/Counterfactuals

Gender Bias: Rudinger et al. (2018)

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

Subj.-Verb Agree.: Marvin and Linzen (2018)

a. The farmer that the parents love swims.
b. *The farmer that the parents love swim.

Veridicality: White et al. (2018)
Someone {knew, didn't know} that a particular thing happened.
Someone {was, wasn't} told that a particular thing happened.
Did that thing happen?

Experimental Designs: Tightly Controlled

- Minimal Pairs/Counterfactuals
- Pros: Few confounds, easier to attribute difference to the phenomena itself
- Cons: Can be hard to generate; may not exist in a way that is natural
- Good for phenomena that manifest neatly in the grammar (SV agreement, gender bias), but less so for complex phenomena ("common sense")

Gender Bias: Rudinger et al. (2018)

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

Subj.-Verb Agree.: Marvin and Linzen (2018)

a. The farmer that the parents love swims.
b. *The farmer that the parents love swim.

Veridicality: White et al. (2018)
Someone {knew, didn't know} that a particular thing happened.
Someone {was, wasn't} told that a particular thing happened.
Did that thing happen?

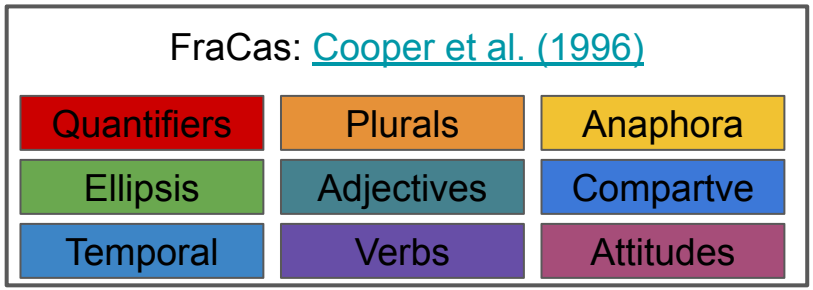
Experimental Designs: Loosely Controlled

Experimental Designs: Loosely Controlled

- Average over sets with vs. without property of interest

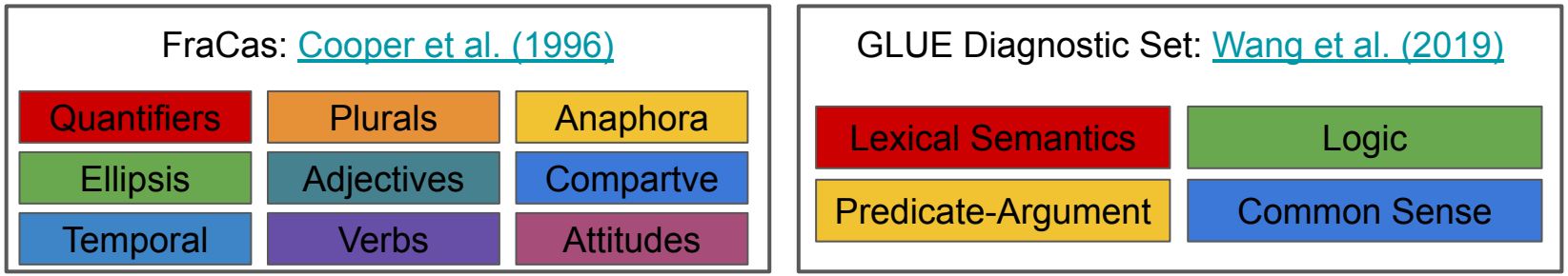
Experimental Designs: Loosely Controlled

- Average over sets with vs. without property of interest



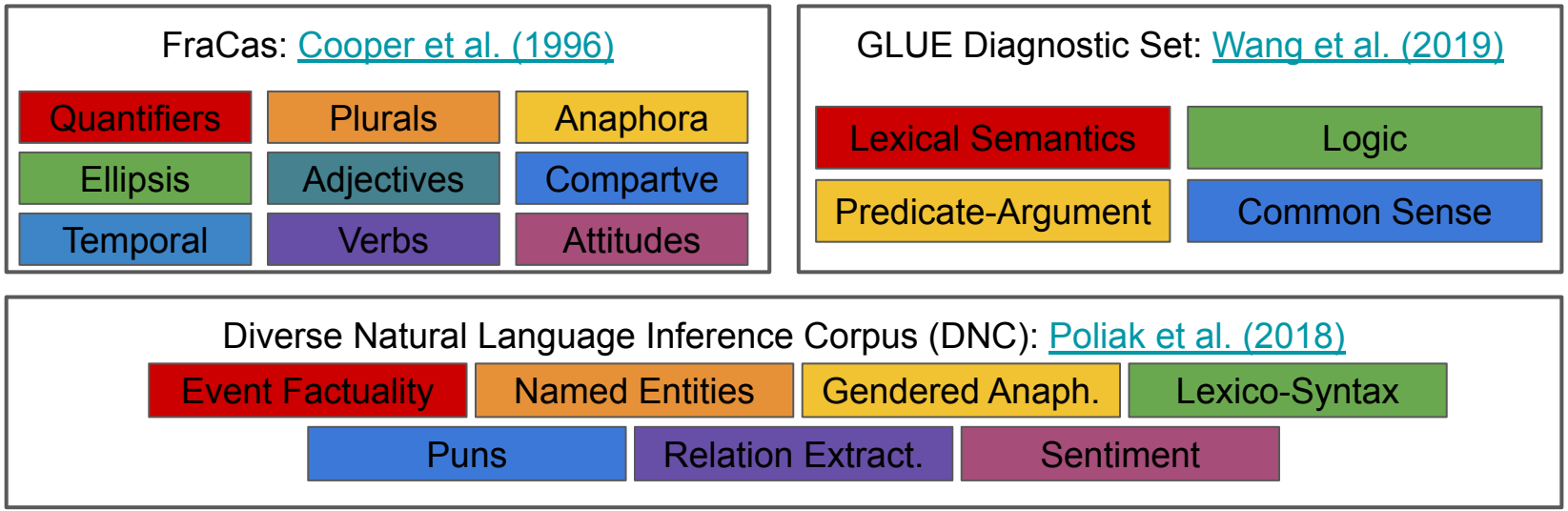
Experimental Designs: Loosely Controlled

- Average over sets with vs. without property of interest



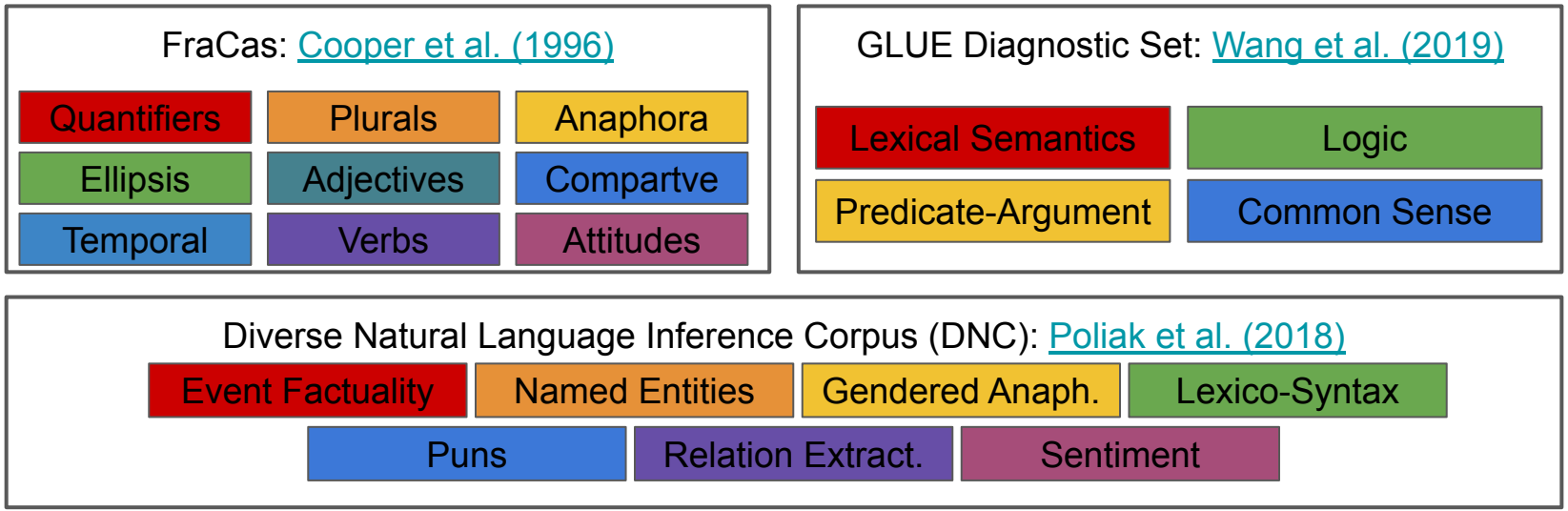
Experimental Designs: Loosely Controlled

- Average over sets with vs. without property of interest



Experimental Designs: Loosely Controlled

- Average over sets with vs. without property of interest
- Pros: Can consist of naturalistic data; can generate larger test sets
- Cons: Contain artifacts, harder to attribute differences to target phenomena



Experimental Designs: Adversarial Examples

Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses

Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses

Jia and Liang (2017)

Article: Super Bowl 50
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”
Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

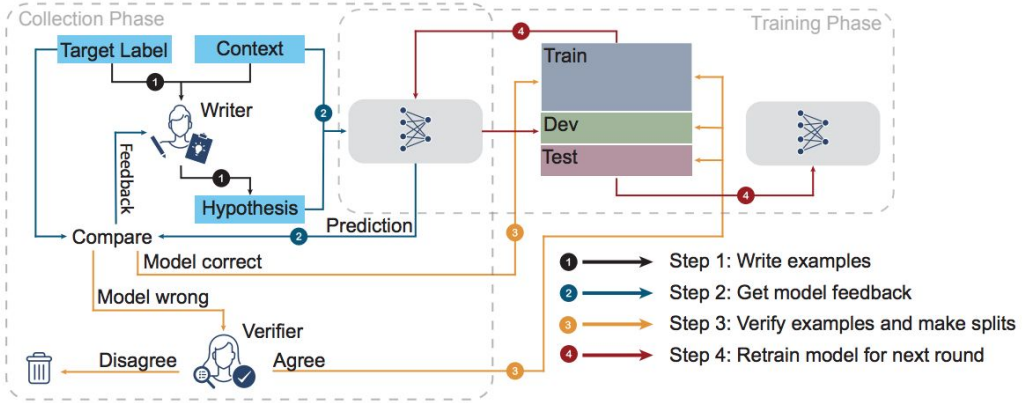
Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses

Jia and Liang (2017)

Article: Super Bowl 50
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”
Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversarial NLI: Nie et al. (2019)



Experimental Designs: Adversarial Examples

- Design data sets (usually using minimal pairs or “perturbations”) that specifically emphasize a model’s weaknesses
- Pros: Practical analysis of failures; can be used as training to improve model
- Cons: Sets age quickly; are model/data specific; “whack-a-mole” approach

Jia and Liang (2017)

Article: Super Bowl 50
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”
Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Adversarial NLI: Nie et al. (2019)