Syntactic dependencies correspond to word pairs with high mutual information

Richard Futrell
University of California, Irvine
rfutrell@uci.edu
@rljfutrell

Peng Qian
MIT
pqian@mit.edu

Evelina Fedorenko
MIT
evelina9@mit.edu

Edward Gibson
MIT
egibson@mit.edu

Idan Blank
University of California, Los Angeles
iblank@psych.ucla.edu

DepLing 2019
2019-08-27

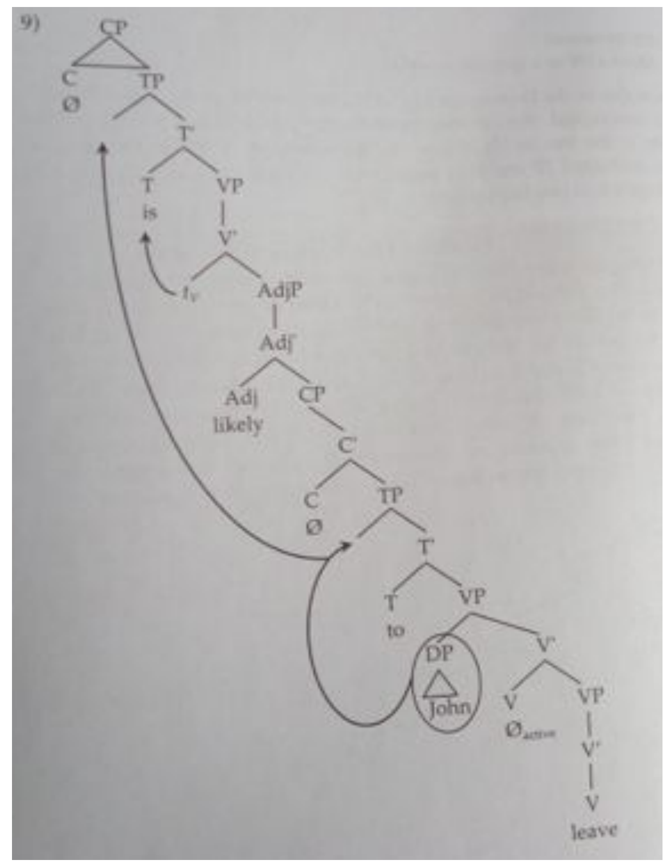
Two Kinds of Structure

Two Kinds of Structure

- 1. Formal syntactic structure:

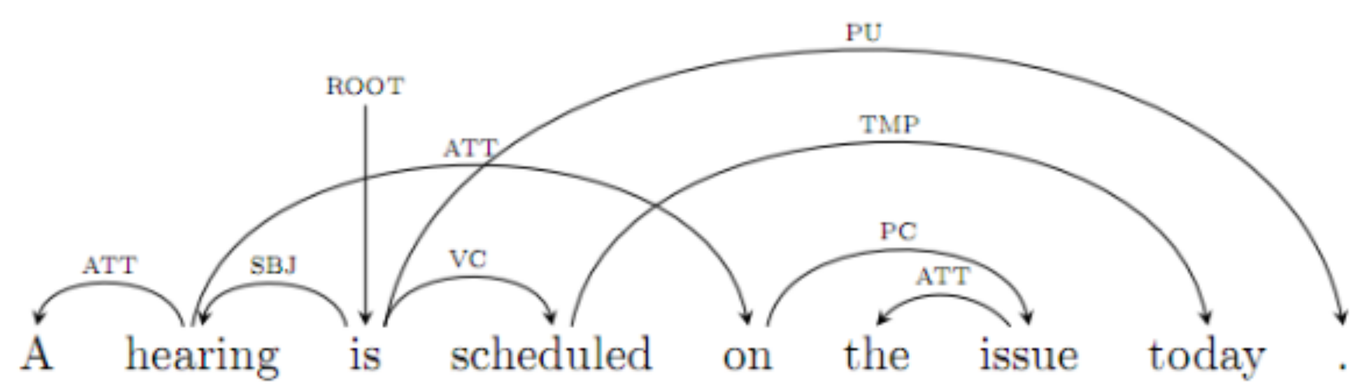
Two Kinds of Structure

- 1. Formal syntactic structure:



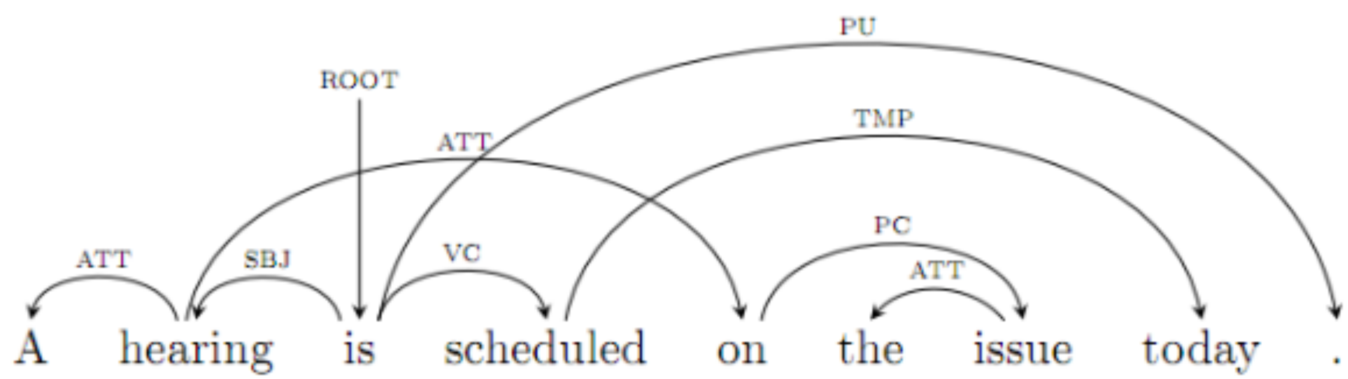
Two Kinds of Structure

- 1. Formal syntactic structure:



Two Kinds of Structure

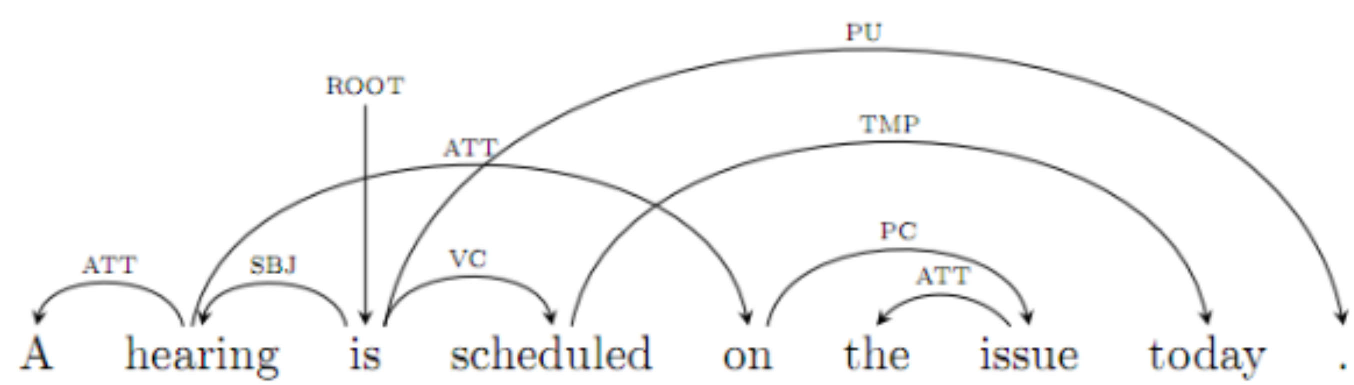
- 1. Formal syntactic structure:



- Goal: Define latent structures required to

Two Kinds of Structure

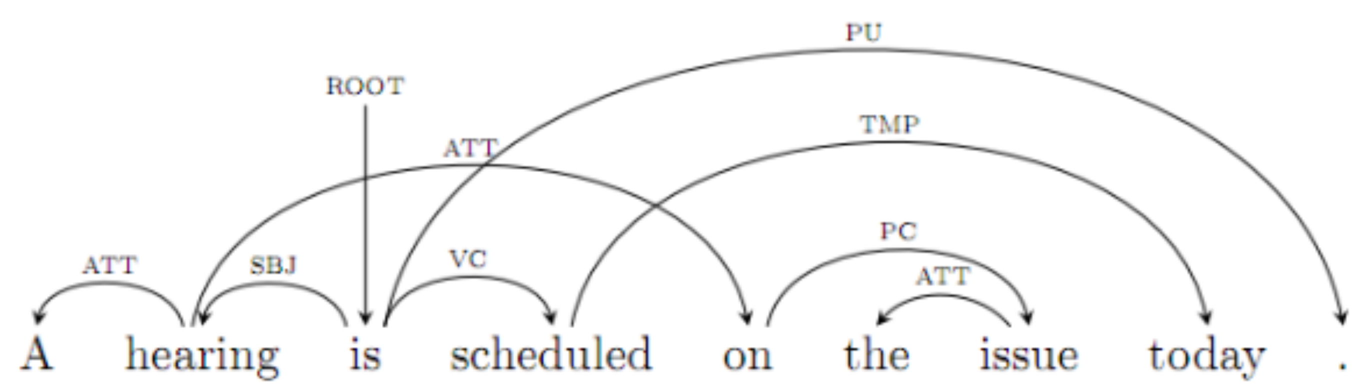
- 1. Formal syntactic structure:



- Goal: Define latent structures required to
  - Define the well-formedness of sentences (Chomsky, 1957), or

Two Kinds of Structure

- 1. Formal syntactic structure:



- Goal: Define latent structures required to
  - Define the well-formedness of sentences (Chomsky, 1957), or
  - Compute the interpretation of the sentence (Heim & Kratzer, 1998)

Two Kinds of Structure

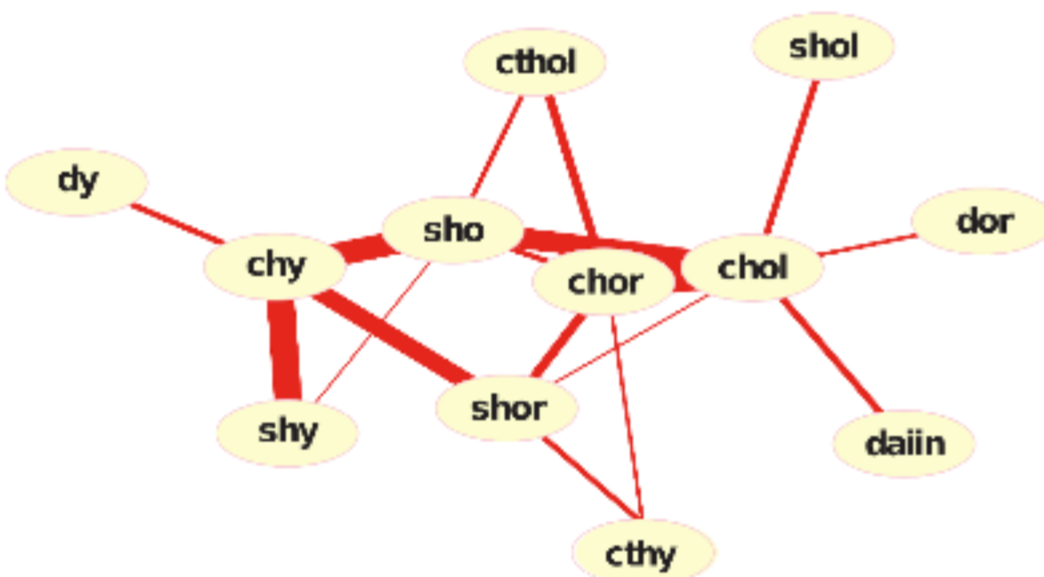
- 1. Formal syntactic structure.

Two Kinds of Structure

- 1. Formal syntactic structure.
- 2. Statistical structure:

Two Kinds of Structure

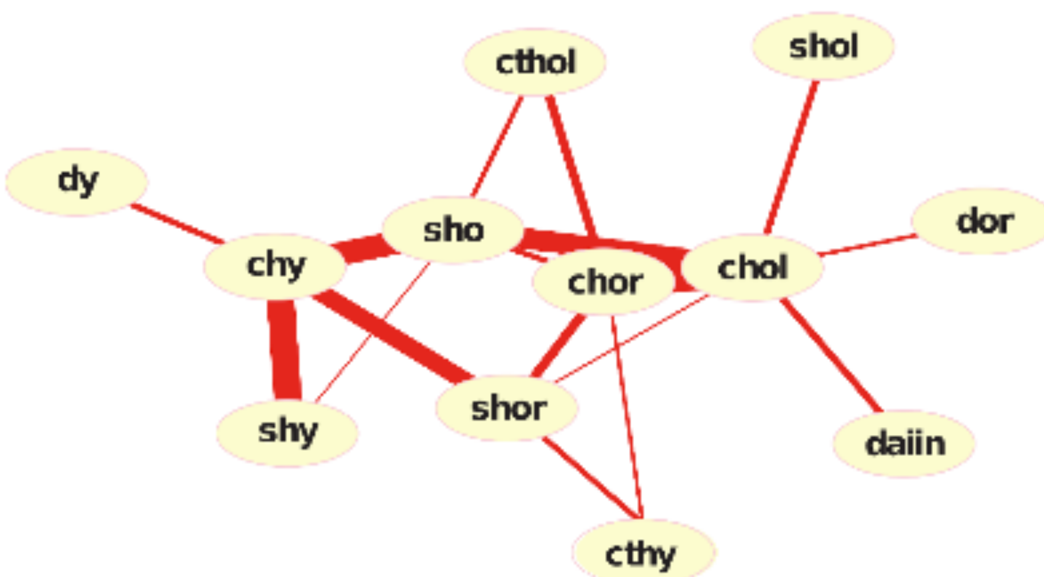
- 1. Formal syntactic structure.
- 2. Statistical structure:



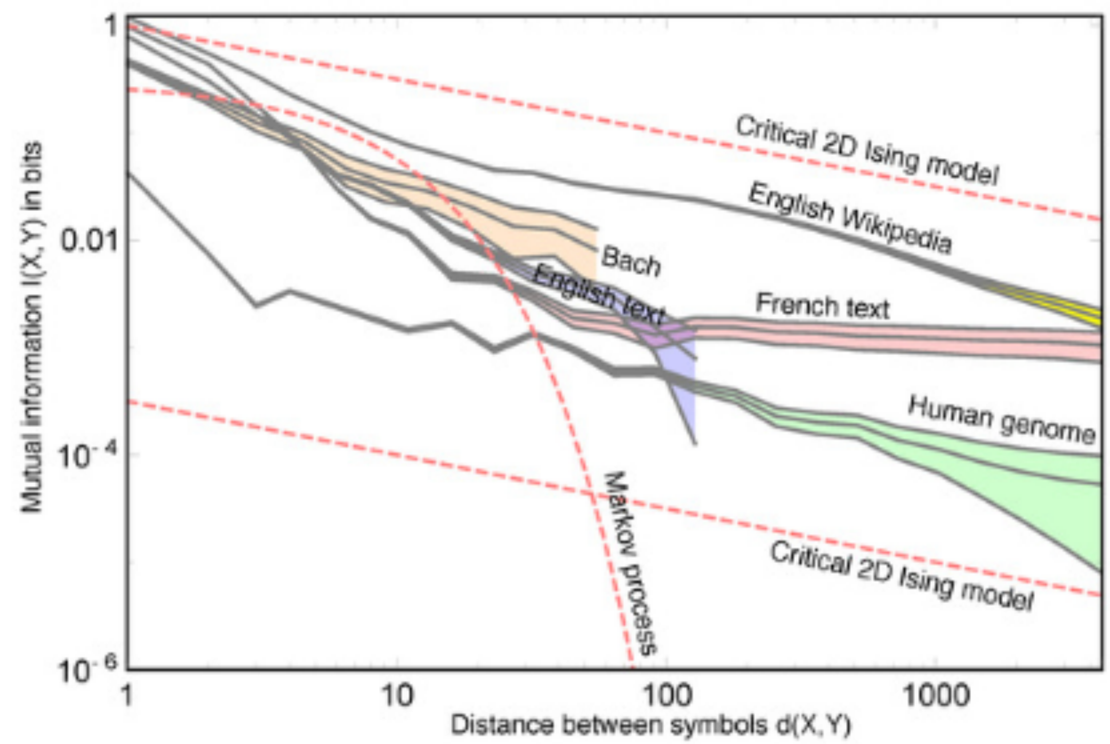
Montemurro & Zanette (2013)

Two Kinds of Structure

- 1. Formal syntactic structure.
- 2. Statistical structure:



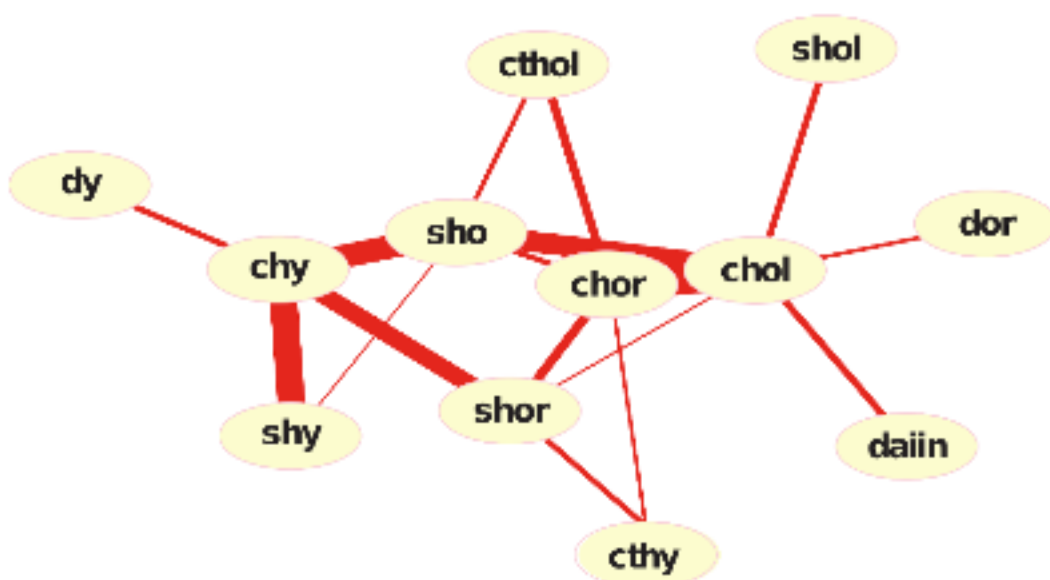
Montemurro & Zanette (2013)



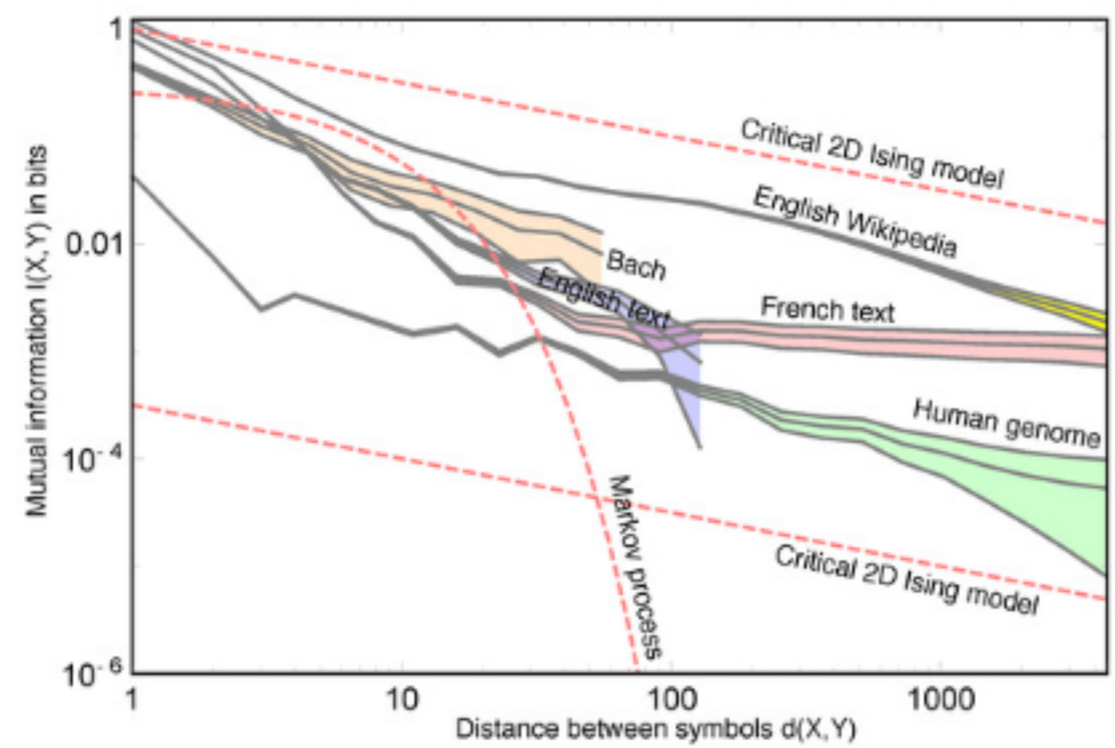
Lin & Tegmark (2017)

Two Kinds of Structure

- 1. Formal syntactic structure.
- 2. Statistical structure:



Montemurro & Zanette (2013)



Lin & Tegmark (2017)

- Goal: Characterize natural language text, as observable in corpora, as a stochastic process.

Linking Syntactic and Statistical Structure

Linking Syntactic and Statistical Structure

- Philosophical & empirical question: **What is the link** between syntactic structure and statistical structure?

Linking Syntactic and Statistical Structure

- Philosophical & empirical question: **What is the link between syntactic structure and statistical structure?**
  - Generative grammarians (e.g. Chomsky, 1957; Adger, 2018): There is **no link at all**.

Linking Syntactic and Statistical Structure

- Philosophical & empirical question: **What is the link between syntactic structure and statistical structure?**
  - Generative grammarians (e.g. Chomsky, 1957; Adger, 2018): There is **no link at all**.
  - Structuralists (e.g. Harris, 1954): Syntactic structure can be defined on top of statistical structure using **discovery procedures**.

Linking Syntactic and Statistical Structure

- Philosophical & empirical question: **What is the link between syntactic structure and statistical structure?**
  - Generative grammarians (e.g. Chomsky, 1957; Adger, 2018): There is **no link at all**.
  - Structuralists (e.g. Harris, 1954): Syntactic structure can be defined on top of statistical structure using **discovery procedures**.
  - Modern grammar induction (e.g. Klein & Manning, 2004, et seq.): Assume syntactic structure is the **trace of a generative process** that generated the data; try to recover the syntactic structure from statistical structure using **Bayesian inference**.

Our Proposal

Our Proposal

- We conjecture a simple information-theoretic link between syntactic and statistical structure: the **Head-Dependent Mutual Information (HDMI) Hypothesis.**

Our Proposal

- We conjecture a simple information-theoretic link between syntactic and statistical structure: the **Head-Dependent Mutual Information (HDMI) Hypothesis.**
    - **Syntactic dependencies correspond to word pairs with high mutual information.**

Our Proposal

- We conjecture a simple information-theoretic link between syntactic and statistical structure: the **<u>Head-Dependent Mutual Information (HDMI) Hypothesis.</u>**
  - **Syntactic dependencies correspond to word pairs with high mutual information.**
  - Explicit or implicit in nearly all previous work on grammar induction (de Paiva Alves, 1996; Yuret, 1998; Klein & Manning, 2004, et seq.), but not yet explicitly tested at scale.

Our Proposal

- We conjecture a simple information-theoretic link between syntactic and statistical structure: the **Head-Dependent Mutual Information (HDMI) Hypothesis.**
    - **Syntactic dependencies correspond to word pairs with high mutual information.**
    - Explicit or implicit in nearly all previous work on grammar induction (de Paiva Alves, 1996; Yuret, 1998; Klein & Manning, 2004, et seq.), but not yet explicitly tested at scale.
- Our contribution: We give **direct empirical evidence** based on a large parsed corpus, and a **new theoretical justification** based on an information-theoretic formalization of basic postulates of dependnecy grammar.

Head-Dependent MI

- Introduction
- Empirical Estimates of HDMI
- Theoretical Arguments for HDMI
- Conclusion

Estimating HDMI

Estimating HDMI

- Claim: Syntactic dependencies are distinguished as word pairs with high mutual information.

Estimating HDMI

- Claim: Syntactic dependencies are distinguished as word pairs with high mutual information.
- Define **head-dependent mutual information** for words *d* and their heads *h* as:

Estimating HDMI

- Claim: Syntactic dependencies are distinguished as word pairs with high mutual information.
- Define **head-dependent mutual information** for words *d* and their heads *h* as:

$$\mathrm{HDMI} = \mathbb{E}\left[\log \frac{p(h,d)}{p(h)p(d)}\right]$$

Estimating HDMI

- Claim: Syntactic dependencies are distinguished as word pairs with high mutual information.
- Define **head-dependent mutual information** for words *d* and their heads *h* as:

$$\mathrm{HDMI} = \mathbb{E}\left[\log \frac{p(h,d)}{p(h)p(d)}\right]$$

- Interpretation: **Amount of information contained in *d* about *h*.**

Estimating HDMI

- Claim: Syntactic dependencies are distinguished as word pairs with high mutual information.
- Define **head-dependent mutual information** for words *d* and their heads *h* as:

$$\mathrm{HDMI} = \mathbb{E}\left[\log \frac{p(h,d)}{p(h)p(d)}\right]$$

- Interpretation: **Amount of information contained in *d* about *h*.**
- Properly, *h* and*d* should be word forms.

Estimating HDMI

- Claim: Syntactic dependencies are distinguished as word pairs with high mutual information.
- Define **head-dependent mutual information** for words *d* and their heads *h* as:

$$\mathrm{HDMI} = \mathbb{E}\left[\log \frac{p(h,d)}{p(h)p(d)}\right]$$

- Interpretation: **Amount of information contained in *d* about *h*.**
- Properly, *h* and *d* should be word forms.
- But in that case the MI may be hard to estimate accurately…

Estimating MI is Hard

Estimating MI is Hard

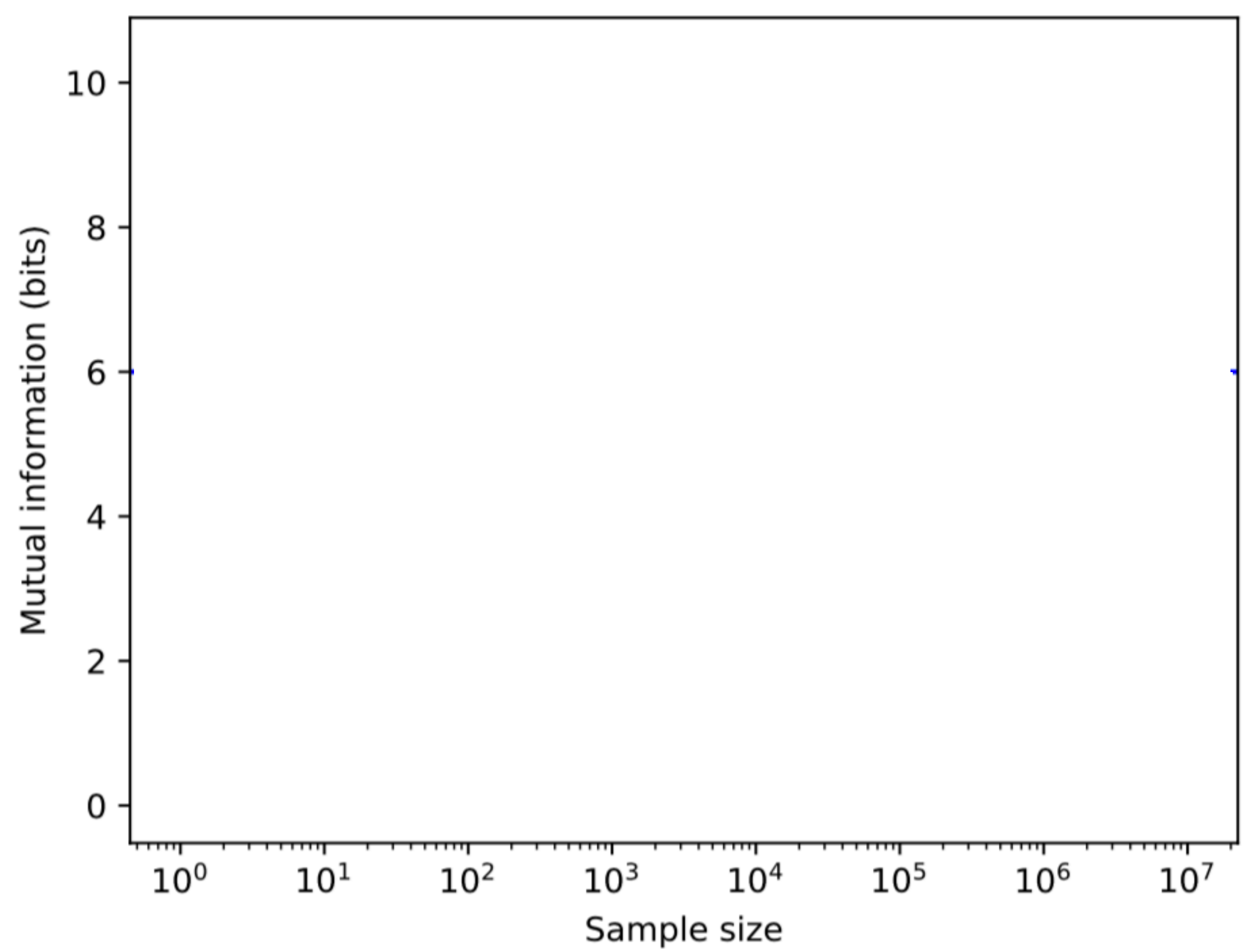
- Demonstration: We simulated a joint distribution of word pairs **known to have exactly 6 bits of MI**.

Estimating MI is Hard

- Demonstration: We simulated a joint distribution of word pairs **known to have exactly 6 bits of MI**.
- We tried to estimate the MI using **maximum likelihood estimation** from a “corpus” drawn from this distribution.

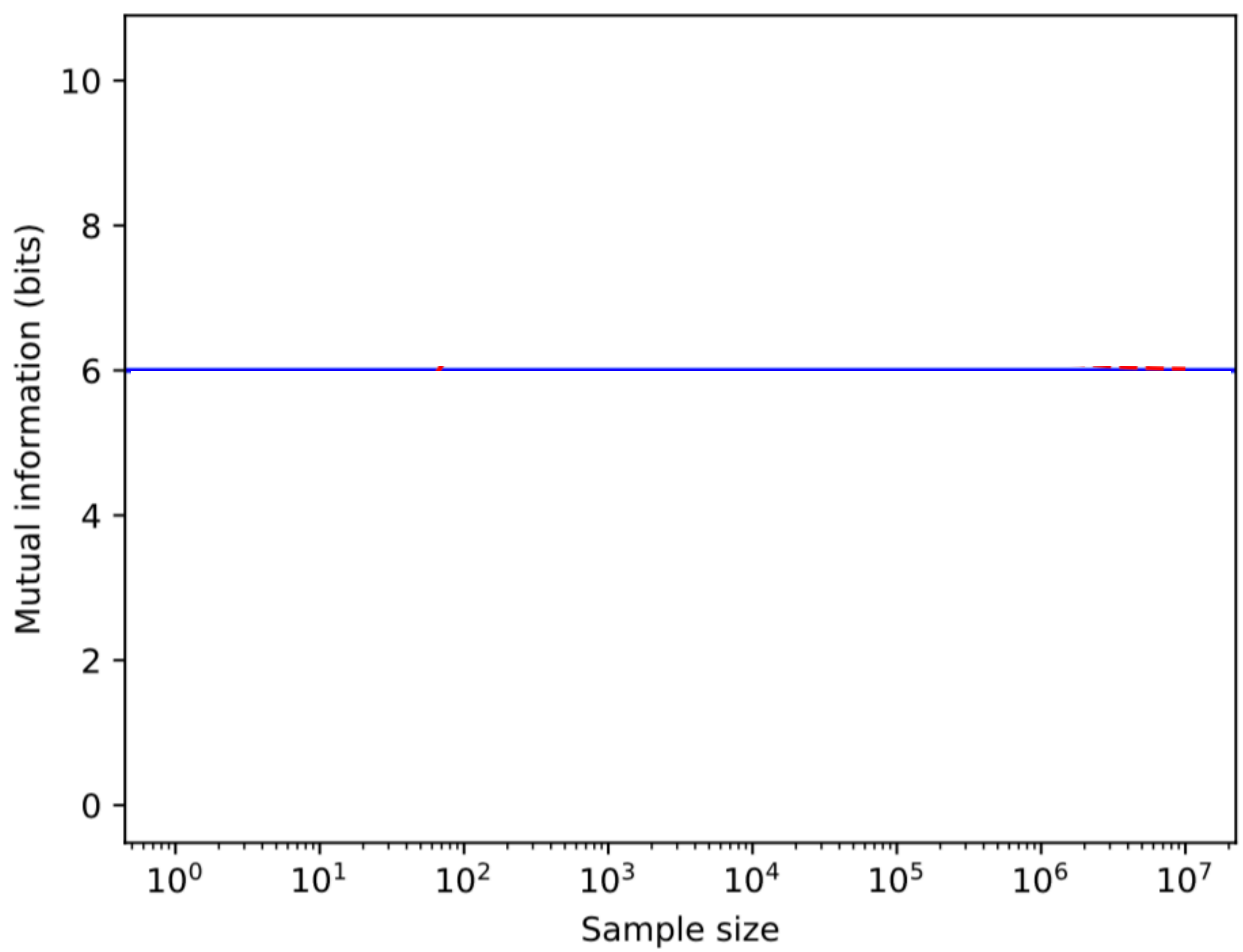
Estimating MI is Hard

- <u>Demonstration</u>: We simulated a joint distribution of word pairs **known to have exactly 6 bits of MI**.
- We tried to estimate the MI using **maximum likelihood estimation** from a “corpus” drawn from this distribution.



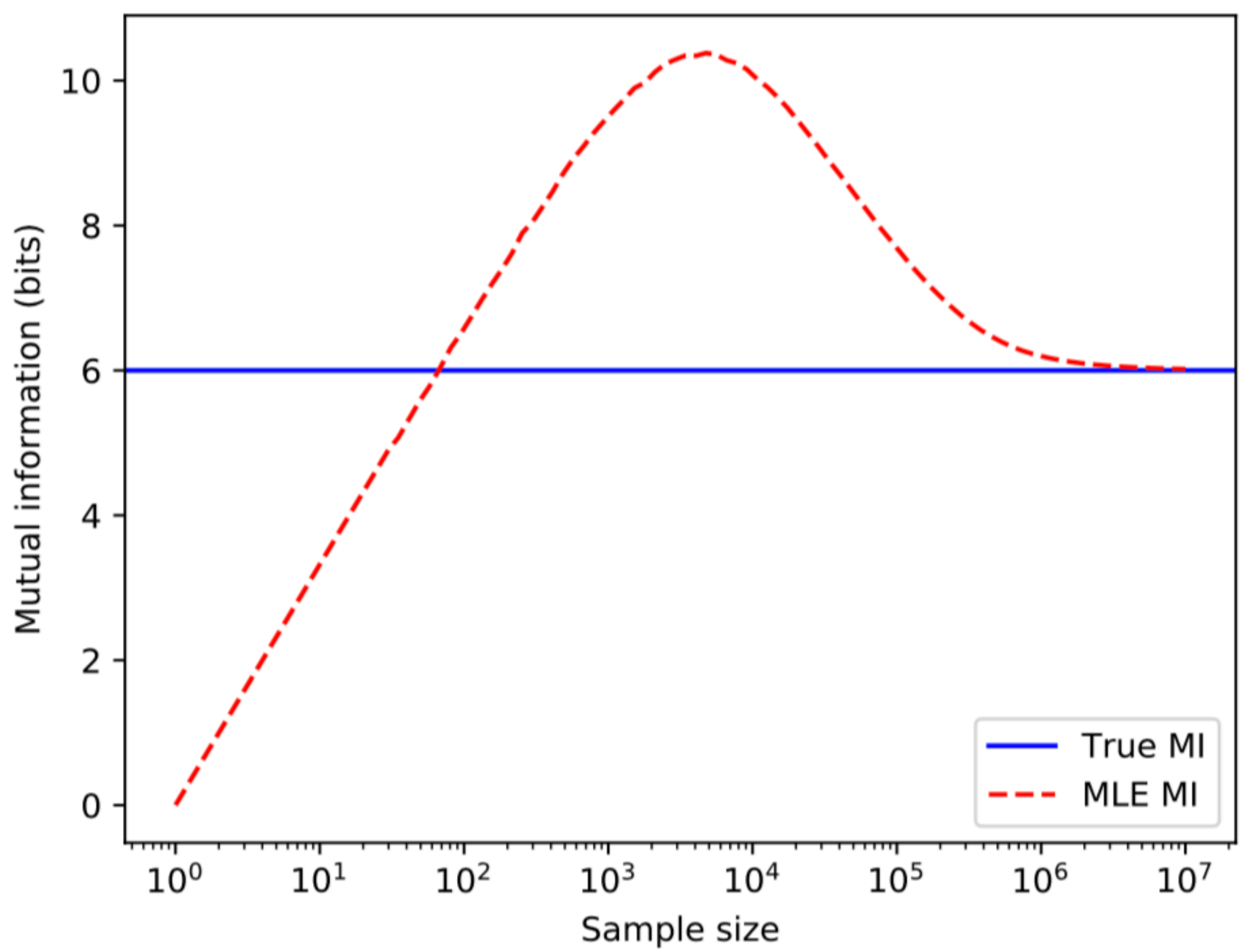
Estimating MI is Hard

- <u>Demonstration</u>: We simulated a joint distribution of word pairs **known to have exactly 6 bits of MI**.
- We tried to estimate the MI using **maximum likelihood estimation** from a “corpus” drawn from this distribution.



Estimating MI is Hard

- <u>Demonstration</u>: We simulated a joint distribution of word pairs **known to have exactly 6 bits of MI**.
- We tried to estimate the MI using **maximum likelihood estimation** from a “corpus” drawn from this distribution.



Data

Data

- Data: **Common Crawl English webtext** as **parsed by SyntaxNet** (Andor et al., 2016).

Data

- Data: **Common Crawl English webtext** as **parsed by SyntaxNet** (Andor et al., 2016).
  - We take 10% of Common Crawl, filtered to contain real utterances (i.e., not “All rights reserved”)

Data

- Data: **Common Crawl English webtext** as **parsed by SyntaxNet** (Andor et al., 2016).
  - We take 10% of Common Crawl, filtered to contain real utterances (i.e., not "All rights reserved")
  - We parse 10% of the filtered data.

Data

- Data: **Common Crawl English webtext** as **parsed by SyntaxNet** (Andor et al., 2016).
  - We take 10% of Common Crawl, filtered to contain real utterances (i.e., not “All rights reserved”)
  - We parse 10% of the filtered data.
  - Result: 320 million parsed tokens.

Data

- Data: **Common Crawl English webtext** as **parsed by SyntaxNet** (Andor et al., 2016).
  - We take 10% of Common Crawl, filtered to contain real utterances (i.e., not “All rights reserved”)
  - We parse 10% of the filtered data.
  - Result: 320 million parsed tokens.
- For evidence for the HDMI Hypothesis from POS tags in hand-parsed UD corpora, see Futrell & Levy (2017).

Comparisons

Comparisons

- Now we **estimate MI between heads and dependents by MLE**.

Comparisons

- Now we **estimate MI between heads and dependents by MLE**.
- We want to show that Head-Dependent MI is **higher than MI of just any word pairs** (HDMI Hypothesis):

Comparisons

- Now we **estimate MI between heads and dependents by MLE**.
- We want to show that Head-Dependent MI is **higher than MI of just any word pairs** (HDMI Hypothesis):
    - So, **compare against a baseline**: Non-dependent word pairs at the **same distance** as the head-dependent pairs.

Comparisons

- Now we **estimate MI between heads and dependents by MLE**.
- We want to show that Head-Dependent MI is **higher than MI of just any word pairs** (HDMI Hypothesis):
    - So, **compare against a baseline**: Non-dependent word pairs at the **same distance** as the head-dependent pairs.

Evaluating Convergence

Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**

Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**
- To do so, we **compare against another baseline** which has analytically 0 MI.

Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**
- To do so, we **compare against another baseline** which has analytically 0 MI.
  - The **permuted baseline**: formed by shuffling the empirical head-dependent pairs.

Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**
- To do so, we **compare against another baseline** which has analytically 0 MI.
  - The **permuted baseline**: formed by shuffling the empirical head-dependent pairs.

Evaluating Convergence

Evaluating Convergence

<u>Head</u>

<u>Dependent</u>

Evaluating Convergence

Head	Dependent
cat	
meowed	
published	
angry	
…	

Evaluating Convergence

Head	Dependent
cat	______________
meowed	______________
published	______________
angry	______________
…	

Evaluating Convergence

Head		Dependent
cat	——	the
meowed	——	cat
published	——	article
angry	——	very
…		…

Evaluating Convergence

Head	Dependent
cat	very
meowed	article
published	cat
angry	the
…	…

Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**
- To do so, we **compare against another baseline** which has analytically 0 MI.
  - The **permuted baseline**: formed by shuffling the empirical head-dependent pairs.

Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**
- To do so, we **compare against another baseline** which has analytically 0 MI.
  - The **permuted baseline**: formed by shuffling the empirical head-dependent pairs.
  - If the permuted baseline shows nonzero MI, it can **only be because of estimation error.**

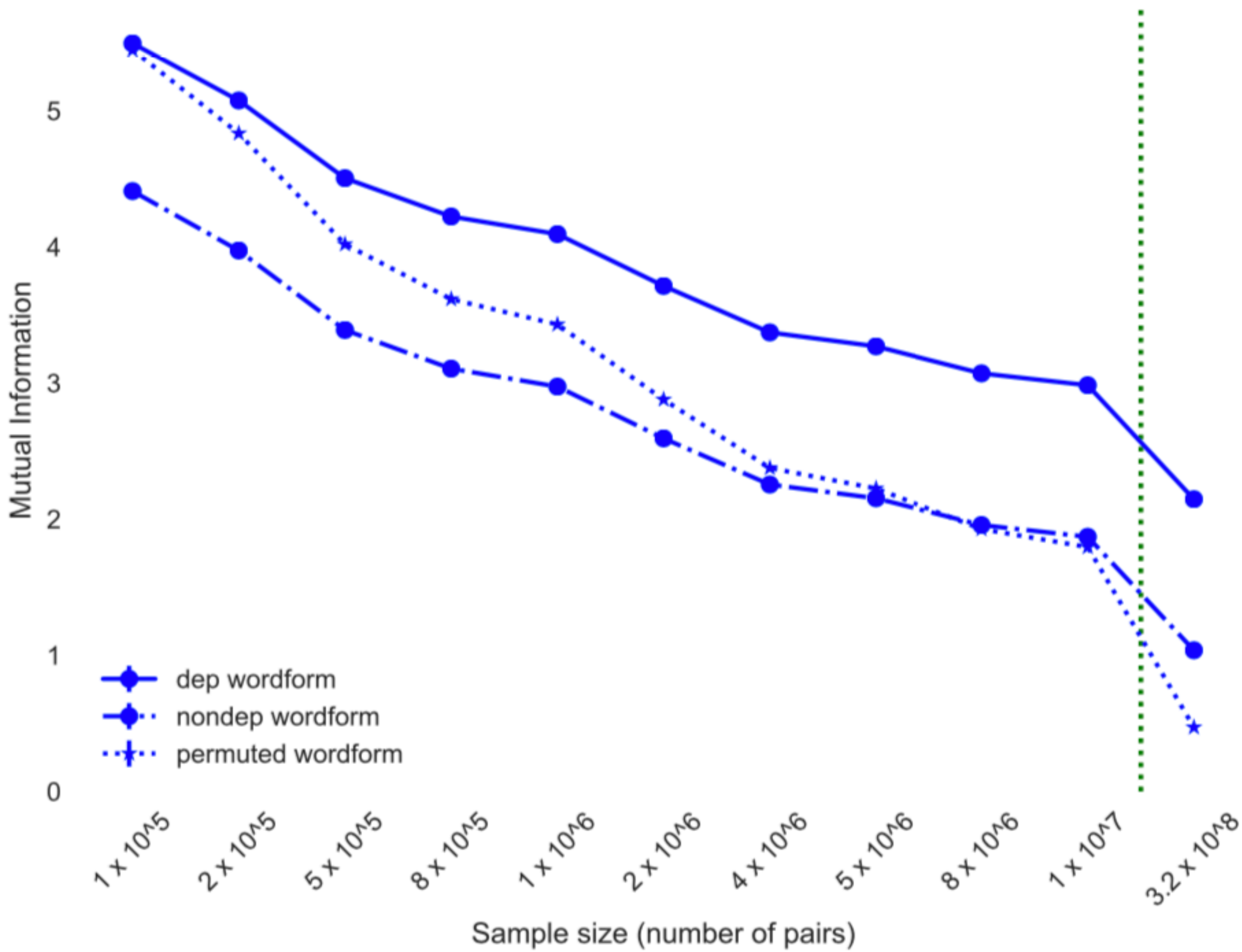
Evaluating Convergence

- We also want to **make sure the MI estimates have converged.**
- To do so, we **compare against another baseline** which has analytically 0 MI.
    - The **permuted baseline**: formed by shuffling the empirical head-dependent pairs.
    - If the permuted baseline shows nonzero MI, it can **only be because of estimation error.**
    - So we want the MI of the permuted baseline to go to zero.

Data & Baselines Summary

dep(endency)	MI of heads and dependents
nondep	MI of words not in a dependency relationship, matched for length with dep
permuted	MI between shuffled heads and dependents (should be zero)

Convergence of MLE Estimates of MI



Convergence?

Convergence?

- Comparison with the permuted baseline suggests **MI estimates have not not converged at 320 million tokens.**

Convergence?

- Comparison with the permuted baseline suggests **MI estimates have not not converged at 320 million tokens.**
- So instead we will measure MI between:

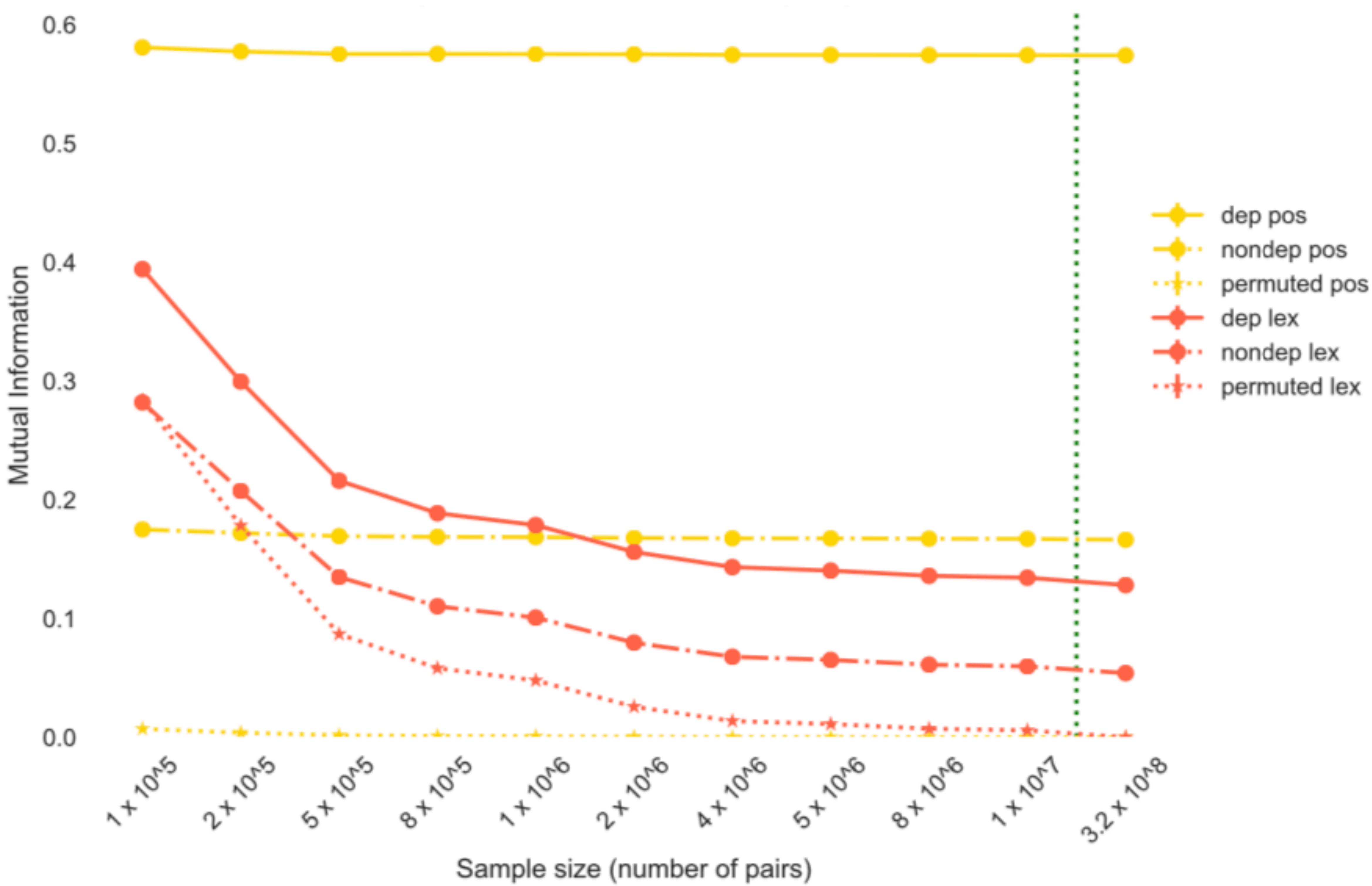
Convergence?

- Comparison with the permuted baseline suggests **MI estimates have not not converged at 320 million tokens.**
- So instead we will measure MI between:
    - **POS tags** (~ a lower bound on the MI between wordforms)

Convergence?

- Comparison with the permuted baseline suggests **MI estimates have not not converged at 320 million tokens.**
- So instead we will measure MI between:
  - **POS tags** (~ a lower bound on the MI between wordforms)
  - **Lexical clusters** derived by a spectral clustering algorithm on GloVe (Pennington et al., 2014) (certainly a lower bound on MI between wordforms).

HDMI between POS tags and Lexical clusters



Empirical Results

Empirical Results

- MI for wordforms **does not yet converge** with 320 million tokens of text.

Empirical Results

- MI for wordforms **does not yet converge** with 320 million tokens of text.
  - Modulo the non-convergence, we have evidence for the HDMI Hypothesis between wordforms.

Empirical Results

- MI for wordforms **does not yet converge** with 320 million tokens of text.
  - Modulo the non-convergence, we have evidence for the HDMI Hypothesis between wordforms.
- MI for POS tags and lexical clusters **does converge**.

Empirical Results

- MI for wordforms **does not yet converge** with 320 million tokens of text.
  - Modulo the non-convergence, we have evidence for the HDMI Hypothesis between wordforms.
- MI for POS tags and lexical clusters **does converge**.
- **Strong evidence** for HDMI Hypothesis for POS tags and lexical clusters.

Empirical Results

- MI for wordforms **does not yet converge** with 320 million tokens of text.
  - Modulo the non-convergence, we have evidence for the HDMI Hypothesis between wordforms.
- MI for POS tags and lexical clusters **does converge**.
- **Strong evidence** for HDMI Hypothesis for POS tags and lexical clusters.

Head-Dependent MI

- Introduction
- Empirical Estimates of MI
- Theoretical Arguments for HDMI
- Conclusion

Theoretical Argument

Theoretical Argument

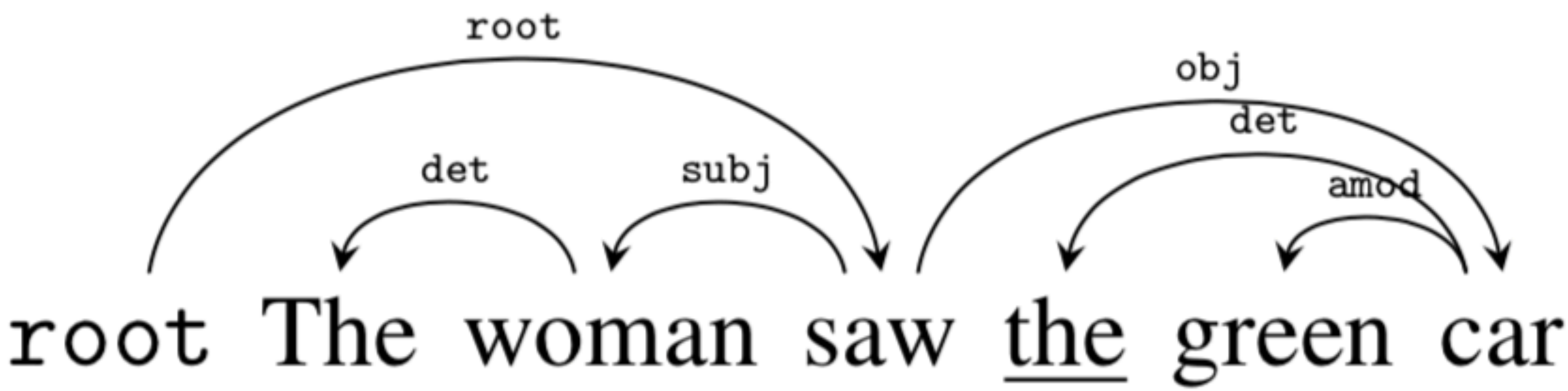
- **Why should the HDMI Hypothesis hold?**

Theoretical Argument

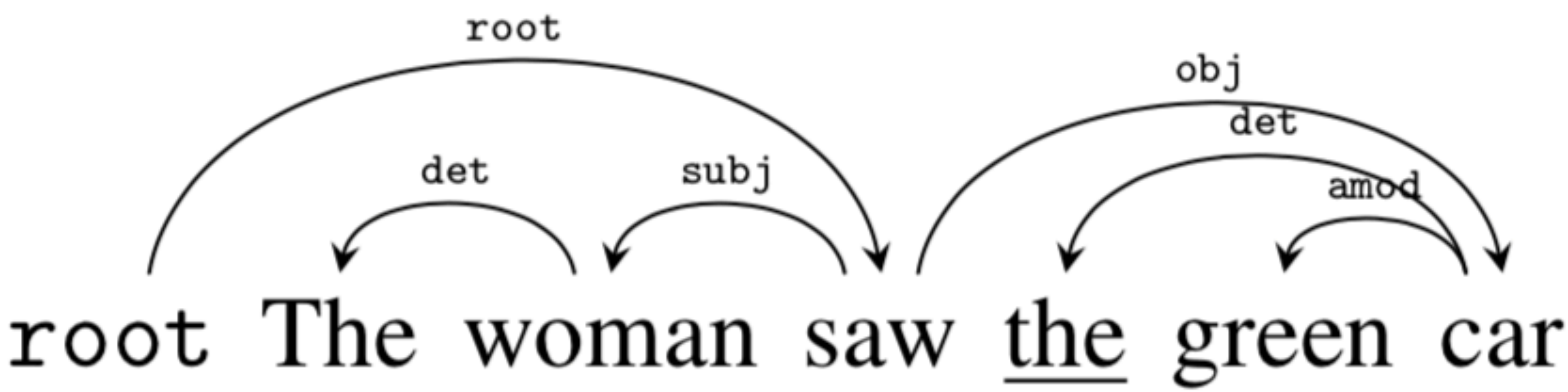
- **Why should the HDMI Hypothesis hold?**
- Question. When linguists are determining the dependency tree for a sentence, what are they doing?

Theoretical Argument

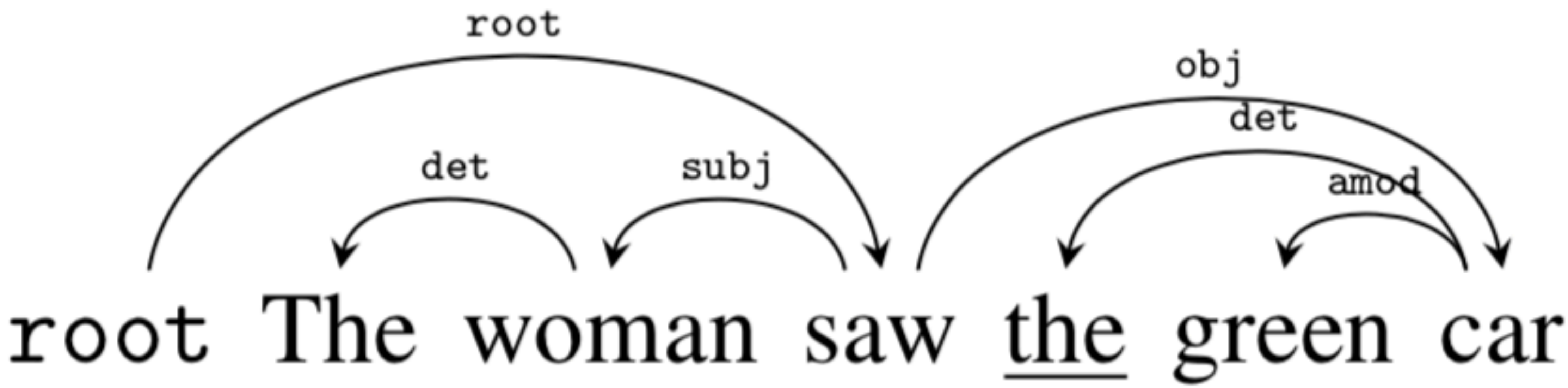
- **Why should the HDMI Hypothesis hold?**
- Question. When linguists are determining the dependency tree for a sentence, what are they doing?



Theoretical Argument

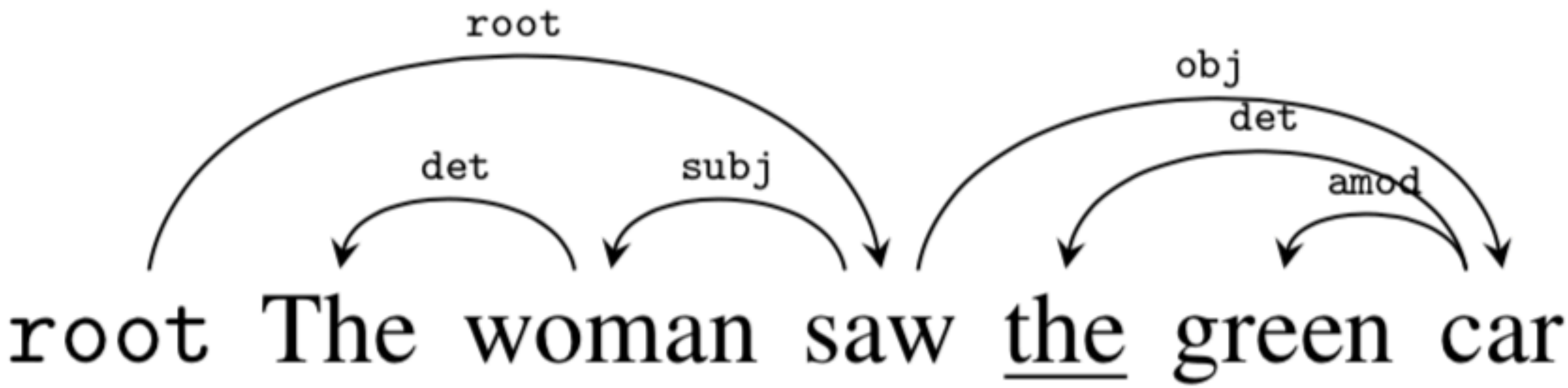


Theoretical Argument



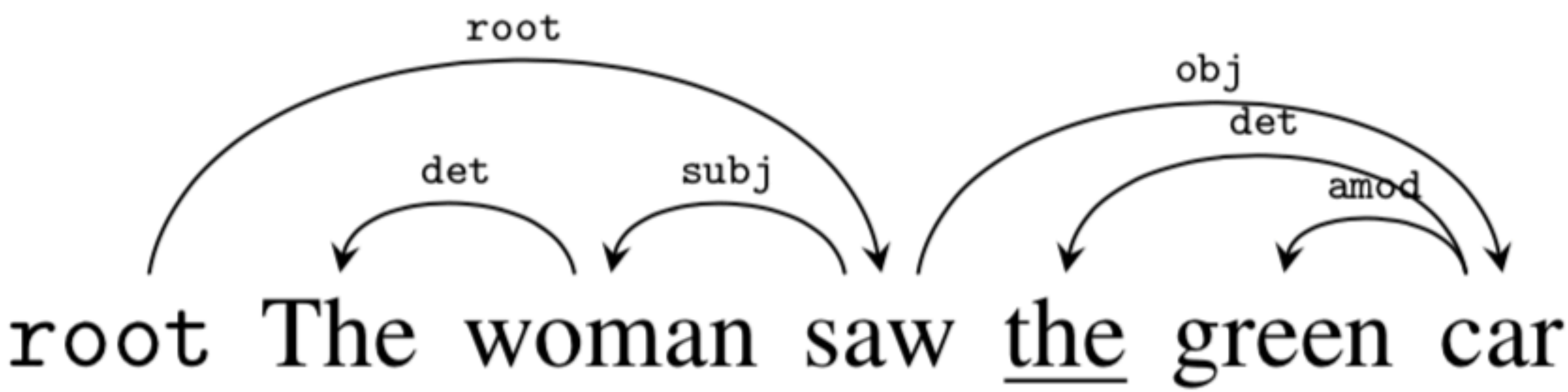
- The head “car” **explains** the distribution (co-occurrence restrictions) of the word “the”.

Theoretical Argument



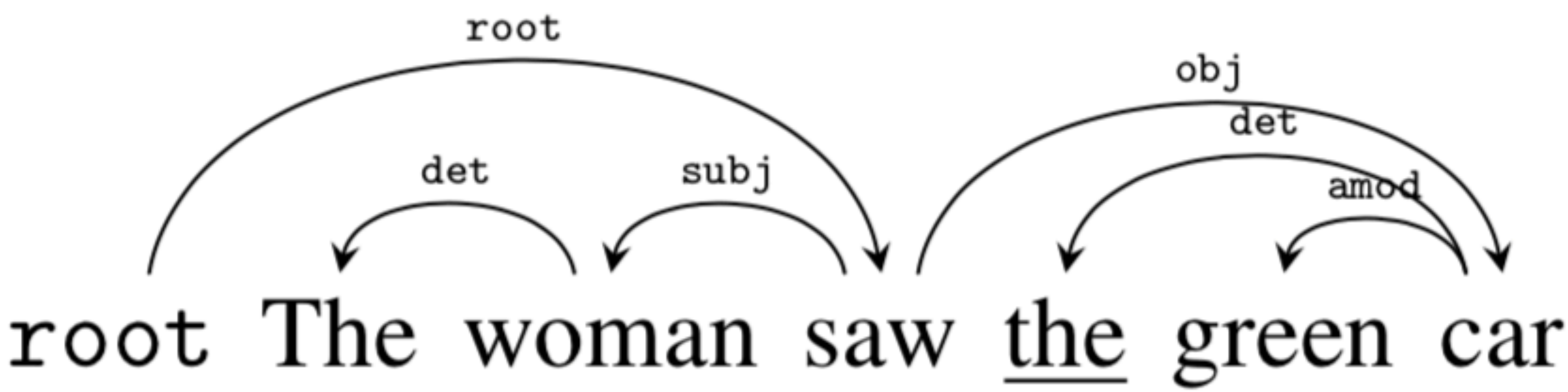
- The head “car” **explains** the distribution (co-occurrence restrictions) of the word “the”.
- A simplification: **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984).

Theoretical Argument



- The head “car” **explains** the distribution (co-occurrence restrictions) of the word “the”.
- A simplification: **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984).
  - No or little need for higher-order groupings as in phrase structure grammar.

Theoretical Argument



- The head “car” **explains** the distribution (co-occurrence restrictions) of the word “the”.
- A simplification: **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984).
  - No or little need for higher-order groupings as in phrase structure grammar.

Theoretical Argument

Theoretical Argument

- **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984, 2010).

Theoretical Argument

- **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984, 2010).
- Translation into the language of statistics: The **head of a word** is a **sufficient statistic** for the **distribution of that word in context.**

Theoretical Argument

- **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984, 2010).
- Translation into the language of statistics: The **head of a word** is a **sufficient statistic** for the **distribution of that word in context.**
- Translation into information theory:

Theoretical Argument

- **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984, 2010).
- Translation into the language of statistics: The **head of a word** is a **sufficient statistic** for the **distribution of that word in context.**
- Translation into information theory:
  - **KL-divergence** D[word | context || word | head] = ε

Theoretical Argument

- **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984, 2010).
- Translation into the language of statistics: The **head of a word** is a **sufficient statistic** for the **distribution of that word in context.**
- Translation into information theory:
  - **KL-divergence** *D*[word | context || word | head] = ε
- ε = 0 means “strong endocentricity”: the head contains 100% of the information you need to determine the distribution of the word. (Obviously too strong!)

Theoretical Argument

- **Dependency grammar** claims that **the distribution of a word in context can be explained *mostly* in terms of exactly one other word**, its head (Hudson, 1984, 2010).
- Translation into the language of statistics: The **head of a word** is a **sufficient statistic** for the **distribution of that word in context.**
- Translation into information theory:
  - **KL-divergence** $D[\text{ word} \mid \text{context} \;\|\; \text{word} \mid \text{head }] = \varepsilon$
- ε = 0 means “strong endocentricity”: the head contains 100% of the information you need to determine the distribution of the word. (Obviously too strong!)
- ε = small is more realistic.

Theoretical Argument

Theoretical Argument

- Translation into information theory:

Theoretical Argument

- Translation into information theory:
  - **KL-divergence** D[word | context || word | head] = ε

Theoretical Argument

- Translation into information theory:
  - **KL-divergence** *D*[word | context || word | head] = ε
- Proposal: When linguists are choosing heads, they are implicitly minimizing the approximation error above.

Theoretical Argument

- Translation into information theory:
  - **KL-divergence** *D*[word | context || word | head] = ε
- Proposal: When linguists are choosing heads, they are implicitly minimizing the approximation error above.
  - **Choosing the head that *best explains* the distribution of each word**, such that the heads and dependents form a tree.

Theoretical Argument

- Translation into information theory:
  - **KL-divergence** *D*[word | context || word | head] = ε
- Proposal: When linguists are choosing heads, they are implicitly minimizing the approximation error above.
  - **Choosing the head that *best explains* the distribution of each word**, such that the heads and dependents form a tree.
- This is also the objective implicitly minimized in grammar induction work based on **head-outward generative models** (Eisner, 1996; Klein & Manning, 2004, et seq.)

Formal Argument

Formal Argument

- Proposition. **The trees that minimize approximation error are found by choosing heads to maximize HDMI.**
  (cf. Chow & Liu, 1968)

Formal Argument

- <u>Proposition</u>. **The trees that minimize approximation error are found by choosing heads to maximize HDMI.**
  (cf. Chow & Liu, 1968)

Formal Argument

- Proposition. **The trees that minimize approximation error are found by choosing heads to maximize HDMI.**
(cf. Chow & Liu, 1968)

$$D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_{\mathbf{t}}(w_i|t_i)) = \mathbb{E}\left[\log \frac{p_L(w_i|\mathbf{w}_{<i})}{p_{\mathbf{t}}(w_i|t_i)}\right]$$

$$= \mathbb{E}\left[\log \frac{p(\mathbf{w}_{<i}|w_i)p(w_i)}{p(\mathbf{w}_{<i})p_{\mathbf{t}}(w_i|t_i)}\right]$$

$$\min_{\mathbf{t}} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_{\mathbf{t}}(w_i|t_i)) = \min_{\mathbf{t}} -\mathbb{E}\left[\log \frac{p_{\mathbf{t}}(w_i|t_i)}{p(w_i)}\right]$$

$$= \min_{\mathbf{t}} -I[W:T]$$

$$= \max_{\mathbf{t}} I[W:T].$$

Conjecture: MI = Syntactic Dependency

Conjecture: MI = Syntactic Dependency

- HDMI provides a way to **translate between syntactic analysis** and **information-theoretic statistics**.

Conjecture: MI = Syntactic Dependency

- HDMI provides a way to **translate between syntactic analysis** and **information-theoretic statistics**.
    - HDMI is a real-valued, statistical analogue to the discrete notion of dependency.

Conjecture: MI = Syntactic Dependency

- HDMI provides a way to **translate between syntactic analysis** and **information-theoretic statistics**.
    - HDMI is a real-valued, statistical analogue to the discrete notion of dependency.
- Could be used to **evaluate syntactic formalisms…**

Conjecture: MI = Syntactic Dependency

- HDMI provides a way to **translate between syntactic analysis** and **information-theoretic statistics**.
  - HDMI is a real-valued, statistical analogue to the discrete notion of dependency.
- Could be used to **evaluate syntactic formalisms…**
  - E.g., content-head vs. function-head dependencies (Osborne & Gerdes, 2019): **Which gives the higher HDMI?**

Conjecture: MI = Syntactic Dependency

- HDMI provides a way to **translate between syntactic analysis** and **information-theoretic statistics**.
  - HDMI is a real-valued, statistical analogue to the discrete notion of dependency.
- Could be used to **evaluate syntactic formalisms…**
  - E.g., content-head vs. function-head dependencies (Osborne & Gerdes, 2019): **Which gives the higher HDMI?**
- Provides a **principled theoretical basis** for corpus linguistics.

Head-Dependent MI

- Introduction
- Empirical Estimates of MI
- Theoretical Arguments for HDMI
- Conclusion

Summary

Summary

- **Syntactic dependencies correspond to word pairs with high information.**

Summary

- **Syntactic dependencies correspond to word pairs with high information.**
    - *Empirically*, in a large automatically-parsed corpora.

Summary

- **Syntactic dependencies correspond to word pairs with high information.**
    - *Empirically*, in a large automatically-parsed corpora.
    - *Theoretically*, according to a formalization of dependency grammar practice.

Summary

- **Syntactic dependencies correspond to word pairs with high information.**
    - *Empirically*, in a large automatically-parsed corpora.
    - *Theoretically*, according to a formalization of dependency grammar practice.
- Provides an empirically strong and theoretically well-grounded link between **syntactic structure** and **statistical structure**.

Thanks all!

- All code is available online at https://github.com/pqian11/mi-hdmi
- Thanks to Roger Levy, Tim O'Donnell, Michael Hahn, and Ryan Cotterell for discussions.
- Thanks to the SyntaxFest reviewers for helpful comments, and thanks to the SyntaxFest and DepLing organizers!