Understanding Word Order with an Information-Theoretic Approach

Yukti Makhija (2019BB10067)

Prof. Sumeet Agarwal and Prof. Mausam, IIT Delhi

Minor Research Project, July 2021 - May 2022

Motivation

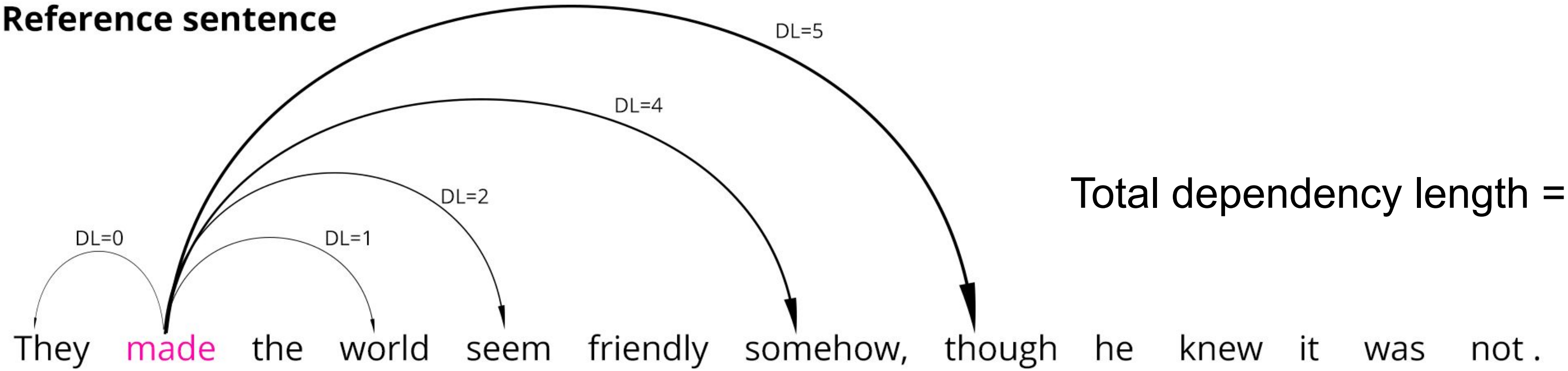
Reference: Clayton lifted him gently into the saddle , like a child .

Variant: Clayton lifted him into the saddle , like a child gently.

Why is it easier to comprehend the reference sentence?

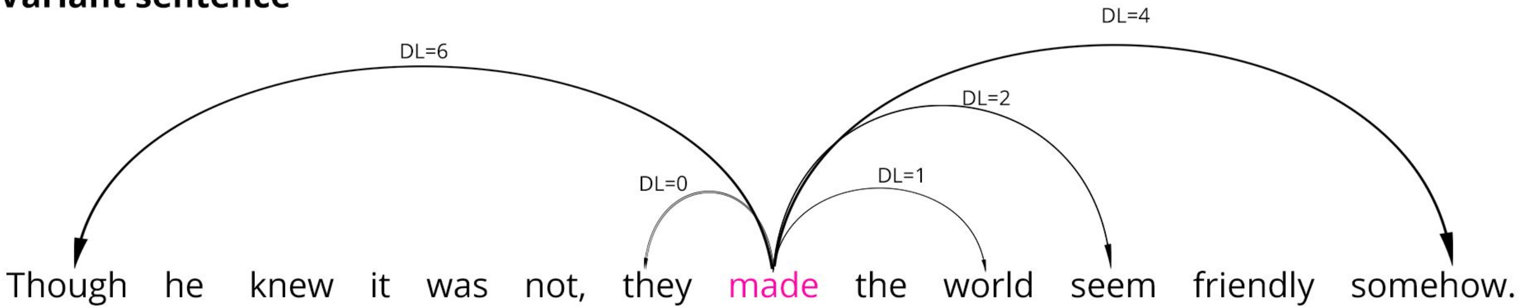
Dependency Length Theory

Reference sentence



Total dependency length = 12

Variant sentence



Total dependency length = 13

Head Dependent Mutual Information (HDMI) Hypothesis (Futrell et al 2019)

$$HDMI = \mathbf{E}[log\frac{p(h,d)}{p(h)p(d)}]$$

- Syntactic dependencies correspond to the word pairs with high mutual information within a sentence.
- Principle of information locality states that an efficient language will minimize the linear distance between elements with high mutual information.

Strength of dependencies based on PMI values (Futrell et al 2019)

- The PMI Values were calculated between Part-of-Speech (POS) Tags of head and dependent of the Universal Dependencies tag-set.
- Computed on Universal Dependencies Treebank 2.5

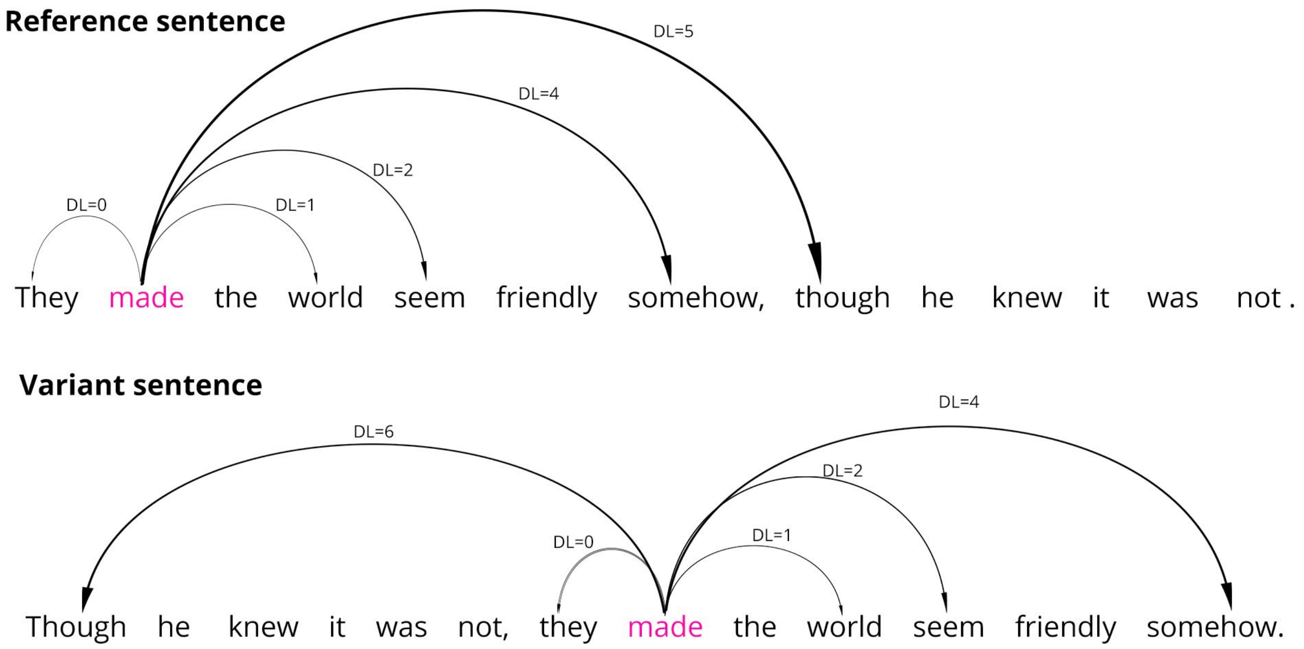
Head	Dependent	PMI Value
VERB	NOUN	0.30144
ADJ	VERB	0.28353

Objectives

- This project aims to create new parameters using Pairwise Mutual Information (PMI) and the Principle of Information Locality, and study their effect on word ordering.
- Use different combinations of PMI, memory and surprisal-based features to train models that distinguish between reference and variant sentences. Also, identify the most performant set of features.
- Look at more refined models of PMI, which may potentially be of more informative of word order than the coarse version with 16 tags being used previously.
- Perform a fine-grained analysis for example using dependency trees to find a correlation between branching pattern and value of PMI based features.

Data

- Brown Corpus - 15 Genres
- Wall Street Journal (WSJ Corpus) - Newswire text only



PMI-based Features

1) PMI between head and it’s nearest dependent.
2) PMI between head and adjacent word.
3) Spearman’s correlation coefficient between the dependency length and PMI value between the root verb and it’s post-verbal dependents
4) Spearman’s correlation coefficient between the linear distance for all root and post-verbal word pairs.

Mathematical Representation of PMI-based Features

Approach 1: PMI between head and nearest dependent

$$HDMI = \log \frac{p(h,d)}{p(h)p(d)}$$

where,
h is the verbal root (head)
d is the nearest dependent of the root verb

Use PMI value between the verbal root (head) and the nearest dependent of this root verb

Corpus	Total number of reference and variant pairs	$PMI_{ref} > PMI_{var}$	$PMI_{ref} = PMI_{var}$	$PMI_{ref} < PMI_{var}$
Brown	8264	666 (8.06%)	7216 (87.32%)	382 (4.62%)
WSJ	19990	1356 (6.78%)	18007 (90.08%)	628 (3.14%)

Approach 2: PMI between head and adjacent word

$$HDMI = \ log \frac{p(h, w)}{p(h)p(w)}$$

where,
h is the verbal root (head)
w is the word adjacent to the root verb

Use PMI value between the verbal root (head) and the word just after the root verb

Corpus	Total number of reference and variant pairs	$PMI_{ref} > PMI_{var}$	$PMI_{ref} = PMI_{var}$	$PMI_{ref} < PMI_{var}$
Brown	8264	881 (10.66%)	6803 (82.32%)	580 (7.02%)
WSJ	19990	2981 (14.91%)	13655 (68.31%)	3355 (16.78%)

Approach 3: Spearman's correlation coefficient between the dependency length and the PMI value between the verbal root and all dependents

$$\rho(PMI, \frac{1}{dlg}) = 1 - \frac{6\sum(rank(\frac{1}{dlg_n}) - rank(PMI_n))^2}{N(N^2 - 1)}$$

where,

ρ is the Spearman's Correlation Cofficient

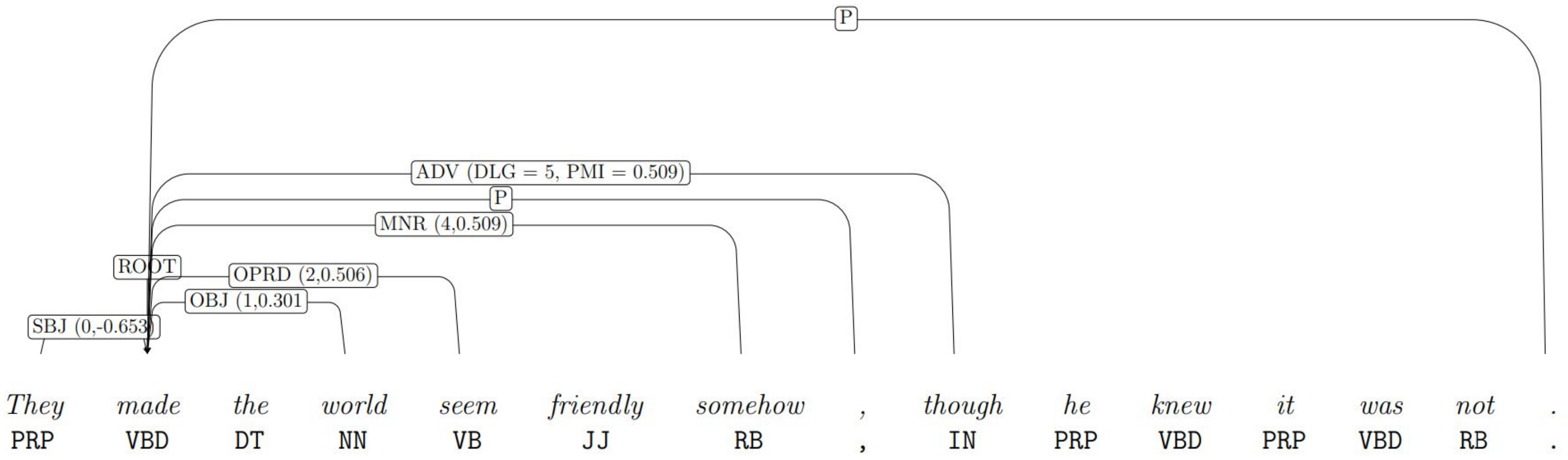
N is the total number of root-dependent pairs in a sentence

PMI_n is the PMI between n^{th} root-dependent pair in a sentence

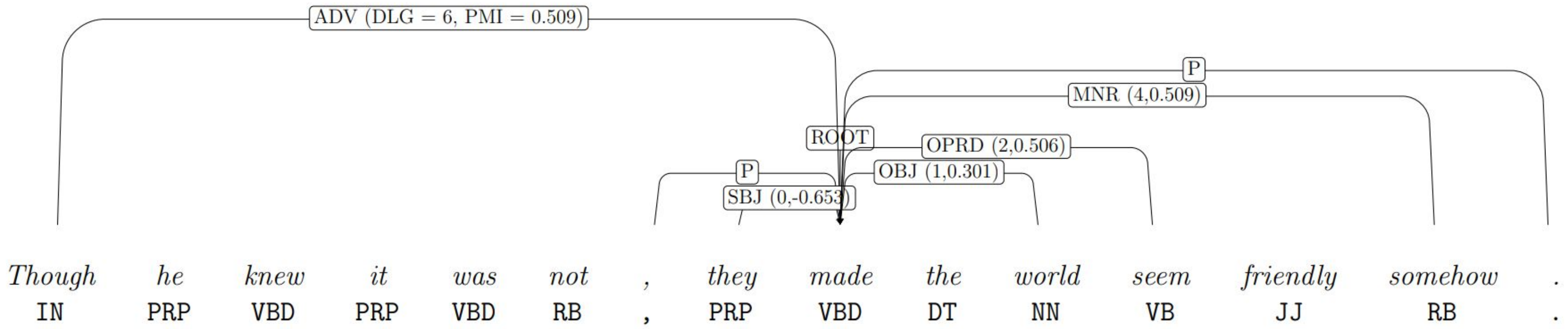
dlg_n is the dependency length for n^{th} root-dependent pair

Note: ρ is calculated separately for pre and post head-dependent pairs

Reference Sentence:



Variant Sentence:



Use Spearman's correlation coefficient between the dependency length and the PMI value between the verbal root and it's dependent for all the dependents after the root verb in the reference and variant sentences.

Corpus	Total number of reference and variant pairs	$PMI_{ref} > PMI_{var}$	$PMI_{ref} = PMI_{var}$	$PMI_{ref} < PMI_{var}$
Brown	8264	1388 (16.8%)	5833 (70.58%)	1043 (12.62%)
WSJ	19990	2978 (14.89%)	14830 (74.19%)	2183 (10.92%)

Approach 4: Spearman's correlation coefficient between the linear distance and the PMI value between the verbal root and all words

$$\rho(PMI, \frac{1}{lg}) = 1 - \frac{6\sum(rank(\frac{1}{lg_n}) - rank(PMI_n))^2}{N(N^2 - 1)}$$

where,

ρ is the Spearman's Correlation Cofficient
N is the total number of root-word pairs in a sentence
PMI_n is the PMI between n^{th} root-word pair in a sentence
lg_n is the distance between root verb and word for n^{th} pair

Note: ρ is calculated separately for pre and post head-dependent pairs

Use Spearman's correlation coefficient between the distance and the PMI value between the verbal root and all the words (taking one at a time) before after the root verb in the reference and variant sentences.

Corpus	Total number of reference and variant pairs	$PMI_{ref} > PMI_{var}$	$PMI_{ref} = PMI_{var}$	$PMI_{ref} < PMI_{var}$
Brown	8264	4174 (50.51%)	1220 (14.76%)	2870 (34.73%)
WSJ	19990	9082 (45.43%)	4249 (21.26%)	6660 (33.32%)

Other Features (Rajkumar et al., 2016)

Information Theoretic (Surprisal) based features:

- Latent Variable PCFG log likelihood (bkpsl)
- n-gram log likelihood (lm)

Memory based features:

- Dependency length (dlg)
- Weighted embedding depth (wt_emb_dep)
- Lexical (1-best) embedding depth (l_dep)

Ranking Model (Rajkumar et al., 2016)

(a) Original data points

Data point label	Feature vector	Feature values		
		Dependency length	PCFG log likelihood	ngram log likelihood
ref	$\Phi(\text{ref})$	30	−137.44	−59.44
var_1	$\Phi(\text{var}_1)$	30	−135.89	−61.16
var_2	$\Phi(\text{var}_2)$	32	−135.79	−58.09

(b) Transformed data points

Data point label	Condition	Feature vector difference	Feature value differences		
			Dependency length	PCFG log likelihood	ngram log likelihood
1	$s_1 = \text{ref}$ $s_2 = \text{var}_1$	$\Phi(s_1) - \Phi(s_2)$	0	−1.55	1.72
0	$s_1 = \text{var}_2$ $s_2 = \text{ref}$	$\Phi(s_1) - \Phi(s_2)$	2	1.65	1.35

Models made using combinations of Memory-based and PMI features

Dependency Length (dlg) + Weighted embedding depth (wtembdep) + Lexical (1-best) embedding depth (ldep) + PMI Features

Testing Accuracy (%)	Dlg(1) + wtembdep (2) + ldep(3)	1+2+3+ Spearman's correlation for all words (Post) (6)	1+2+3+ Spearman's correlation for all words (Post)(6)+ Spearman's correlation for dependents (Post)(7)
Brown	69.1673	69.48208	69.6031
WSJ	69.8914	72.00741	72.04742

Coefficients of features in the best surprisal-based model

Features	Brown Corpus	WSJ Corpus
lm	1.42123	1.59982
bkspl	1.09414	2.24911
PMI (adjacent word)	**-0.01226**	**-0.11325**
PMI (nearest dependent of root verb)	0.16418	**-0.00018**
Spearman's correlation for dependents (Post)	0.03919	0.16650

Models made using combinations of Memory, Surprisal, and PMI features

Brown	Dlg(1) + wtembdep(2) + ldep(3) + lm(4) + bkpsl(5)	All 9 features
Test Accuracy	79.3196	79.48922

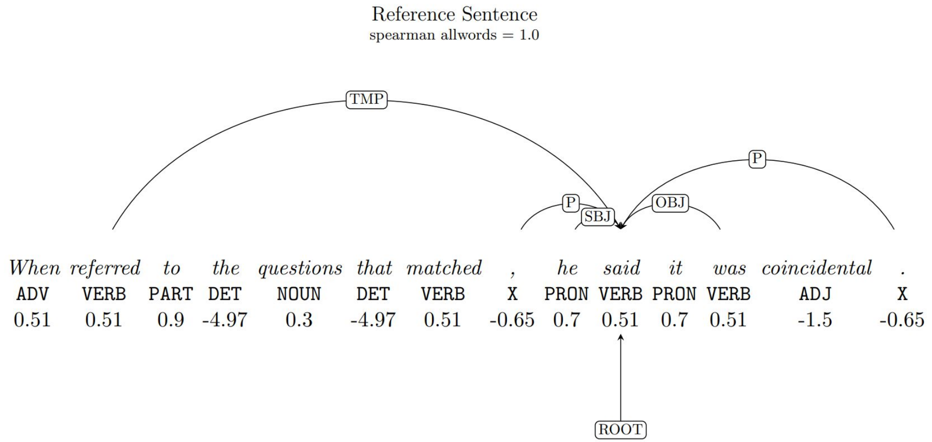
WSJ	Dlg(1) + wtembdep(2) + ldep(3) + lm(4) + bkpsl(5)	1+2+3+4+5+ PMI (adjacent word)(6) + Spearman's correlation for dependents (Post) (9)
Test Accuracy	85.3634	85.38343

Coefficients of features in the best model containing all types of features

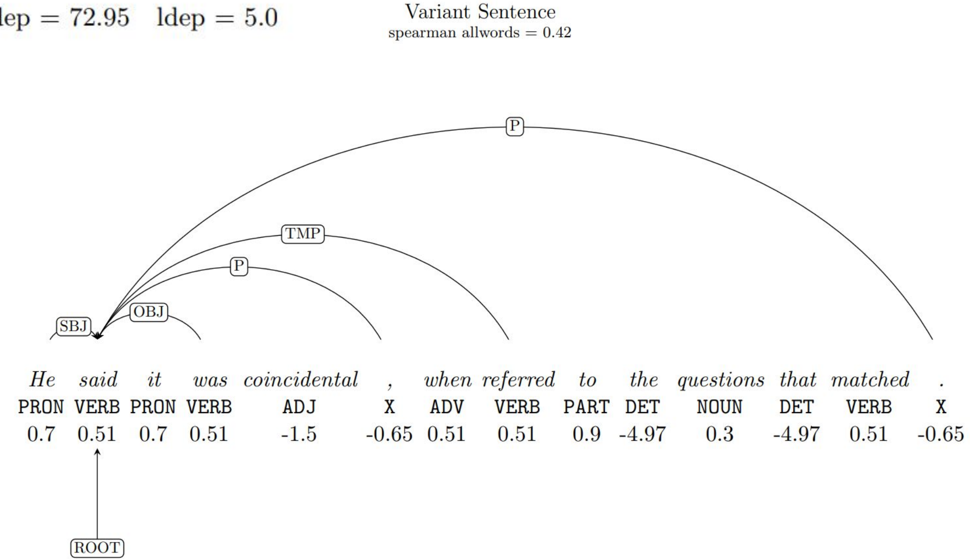
Features	Brown Corpus	WSJ Corpus
dlg	-0.71866	-0.82270
wtembdep	-0.09179	-0.12907
ldep	-0.38311	-0.33854
lm	-0.12663	1.48899
bkspl	0.14699	2.15739
PMI (adjacent word)	0.05673	**-0.17900**
PMI (nearest dependent of root verb)	0.13831	-
Spearman's correlation for all words (Post)	1.36988	-
Spearman's correlation for dependents (Post)	1.04493	0.14991

wsj_0045.24.0

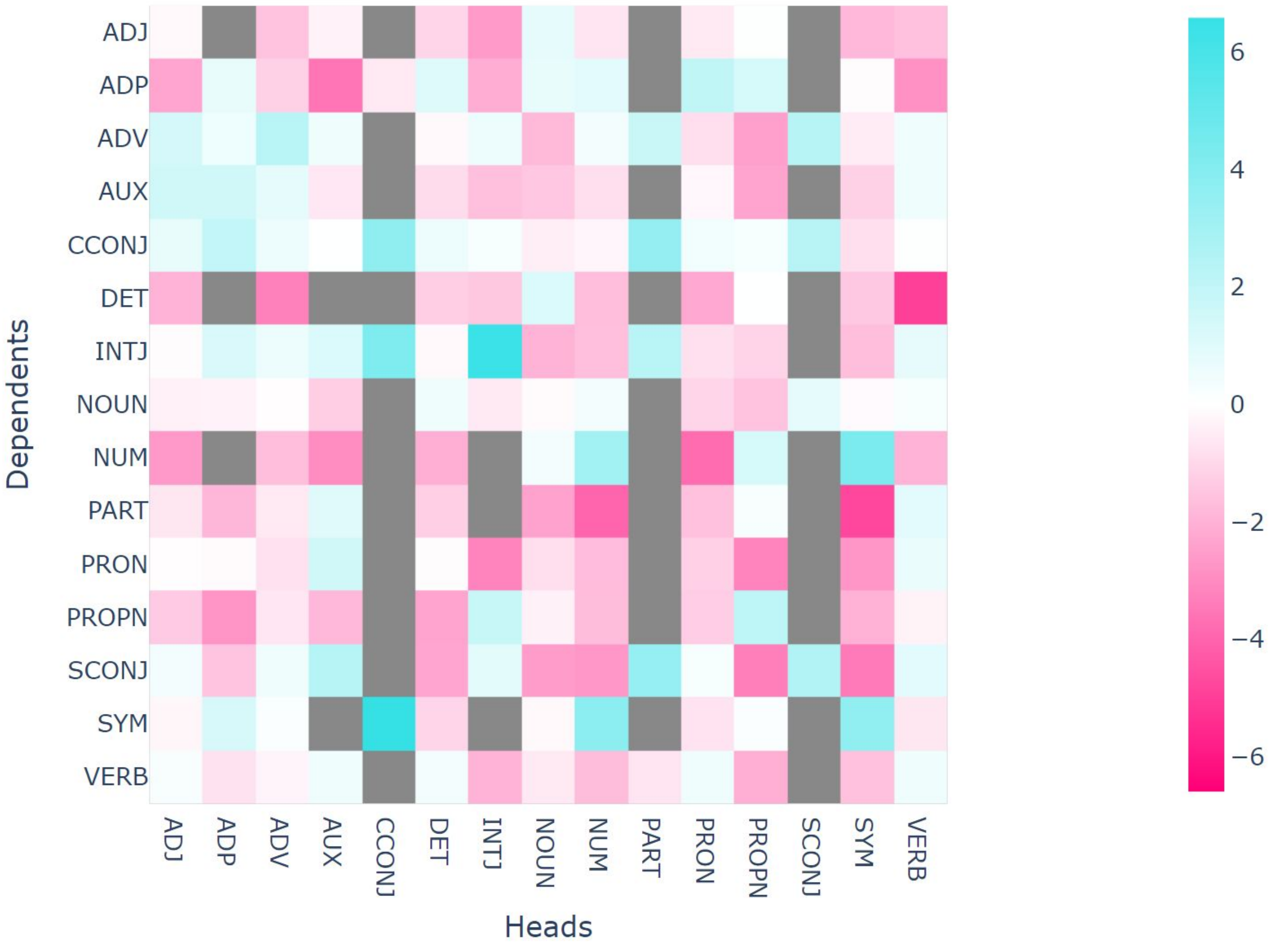
Dependency Tree Comparison



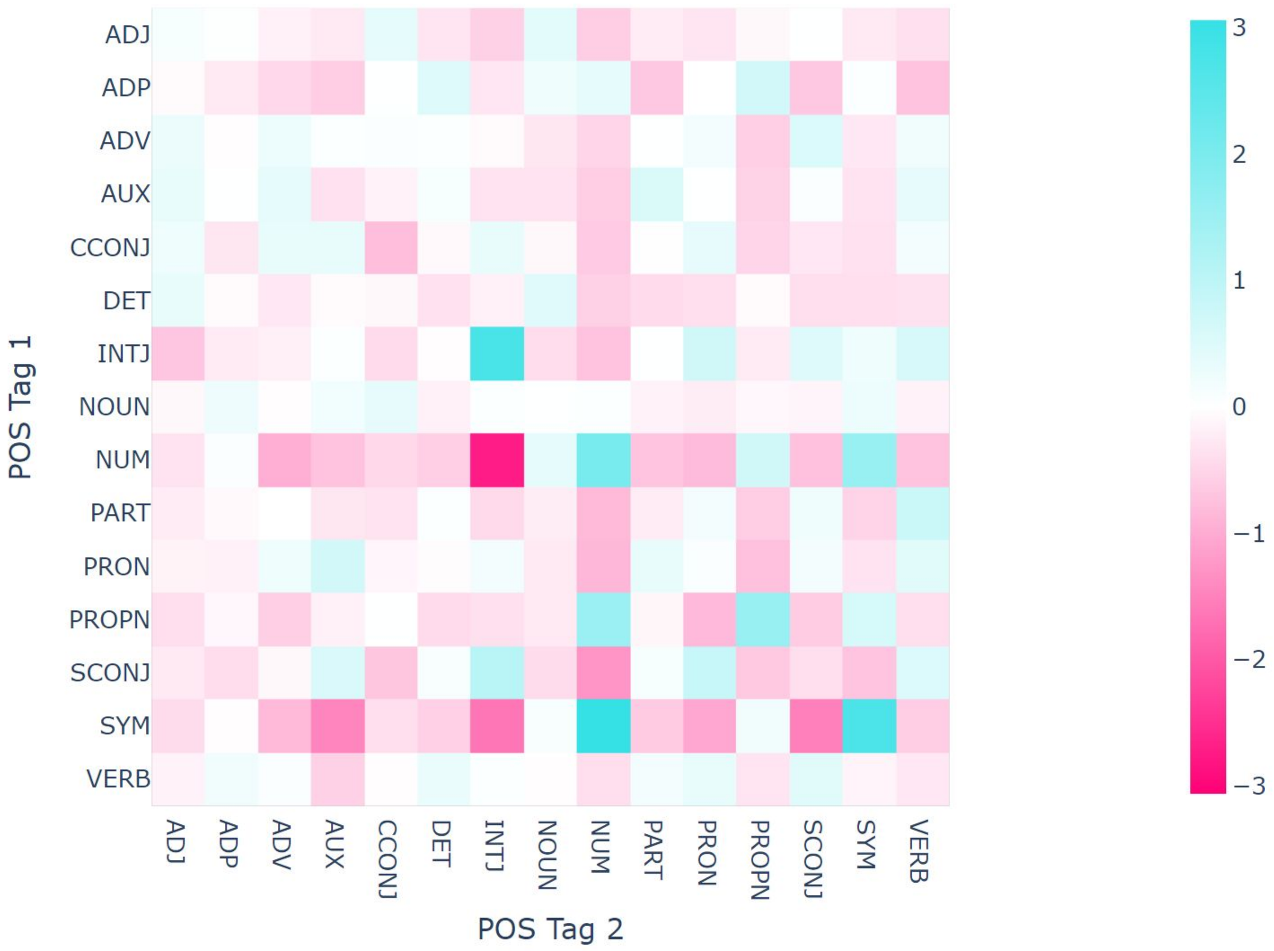
ϕ(ref) - ϕ(var): dlg = 1.0 wtembdep = 72.95 ldep = 5.0



PMI Value between Head - Dependent Pairs



PMI Value between Word Pairs



Models made using combinations of Memory-based and PMI features

Dependency Length (dlg) + Weighted embedding depth (wtembdep) + Lexical (1-best) embedding depth (ldep) + PMI Features

Testing Accuracy (%)	**Memory baseline** Dlg(1) + wtembdep (2) + ldep(3)	1+2+3+ Spearman's correlation for all words (Post) (6)	1+2+3+ Spearman's correlation for all words (Post)(6)
Brown	69.167	69.482	69.736
WSJ	69.891	72.007	71.487

Coefficients of features

Feature	Brown Corpus (old)	Brown Corpus (new)
dlg	-1.2344	-1.2640
wtembdep	-0.0842	-0.0914
ldep	-0.0690	-0.0644
Spearman's correlation for all words (Post)	**0.2305**	**0.0487**

Coefficients of features

Feature	WSJ Corpus (old)	WSJ Corpus (new)
dlg	-1.4904	-1.4898
wtembdep	0.1561	-0.1351
ldep	-0.1281	-0.1253
Spearman’s correlation for all words (Post)	**0.4056**	**-0.2756**

SVM Results: WSJ

Model	**Memory baseline** Dlg(1) + wtembdep (2) + ldep(3)	1+2+3+ Spearman's correlation for all words (Post)(4)	1+2+3+ Spearman's correlation for dependents(Post)(5)
Logistic Regression	69.89	**71.49**	70.03
SVM (Linear)	69.72	**72.01**	69.73
SVM (RBF)	71.98	**73.13**	72.14

Maximum Entropy Model for POS Tagging

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)} \qquad \text{defined over } \mathcal{H} \times \mathcal{T}$$

sequence of words $\{w_1, \ldots, w_n\}$

tags $\{t_1, \ldots t_n\}$

$$h_i = \{w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1}, t_{i-2}\}$$

$$f_j(h_i, t_i) = \begin{cases} 1 & \text{if suffix}(w_i) = \text{“}\mathtt{ing}\text{”} \ \& \ t_i = \mathtt{VBG} \\ 0 & \text{otherwise} \end{cases}$$

Word:	the	stories	about	well-heeled	communities	and	developers
Tag:	DT	NNS	IN	JJ	NNS	CC	NNS
Position:	1	2	3	4	5	6	7

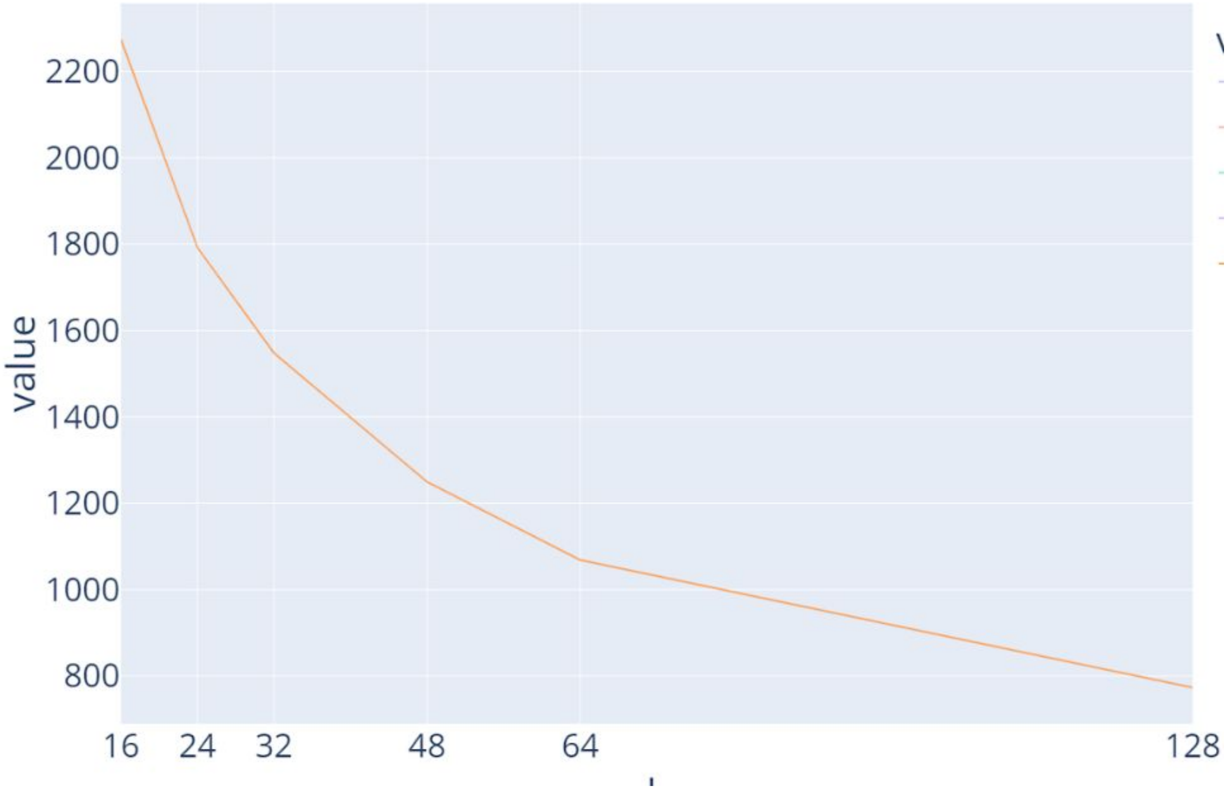
$$
\begin{array}{ll}
w_{i-1} = \texttt{about} & \&\ t_i = \texttt{JJ} \\
w_{i-2} = \texttt{stories} & \&\ t_i = \texttt{JJ} \\
w_{i+1} = \texttt{communities} & \&\ t_i = \texttt{JJ} \\
w_{i+2} = \texttt{and} & \&\ t_i = \texttt{JJ} \\
t_{i-1} = \texttt{IN} & \&\ t_i = \texttt{JJ} \\
t_{i-2}t_{i-1} = \texttt{NNS IN} & \&\ t_i = \texttt{JJ} \\
\text{prefix}(w_i)=\texttt{w} & \&\ t_i = \texttt{JJ} \\
\text{prefix}(w_i)=\texttt{we} & \&\ t_i = \texttt{JJ} \\
\text{prefix}(w_i)=\texttt{wel} & \&\ t_i = \texttt{JJ} \\
\text{prefix}(w_i)=\texttt{well} & \&\ t_i = \texttt{JJ} \\
\text{suffix}(w_i)=\texttt{d} & \&\ t_i = \texttt{JJ} \\
\text{suffix}(w_i)=\texttt{ed} & \&\ t_i = \texttt{JJ} \\
\text{suffix}(w_i)=\texttt{led} & \&\ t_i = \texttt{JJ} \\
\text{suffix}(w_i)=\texttt{eled} & \&\ t_i = \texttt{JJ} \\
w_i \text{ contains hyphen} & \&\ t_i = \texttt{JJ}
\end{array}
$$

Features Generated From h_4 (for tagging **well-heeled**)

Features Generated using h_i

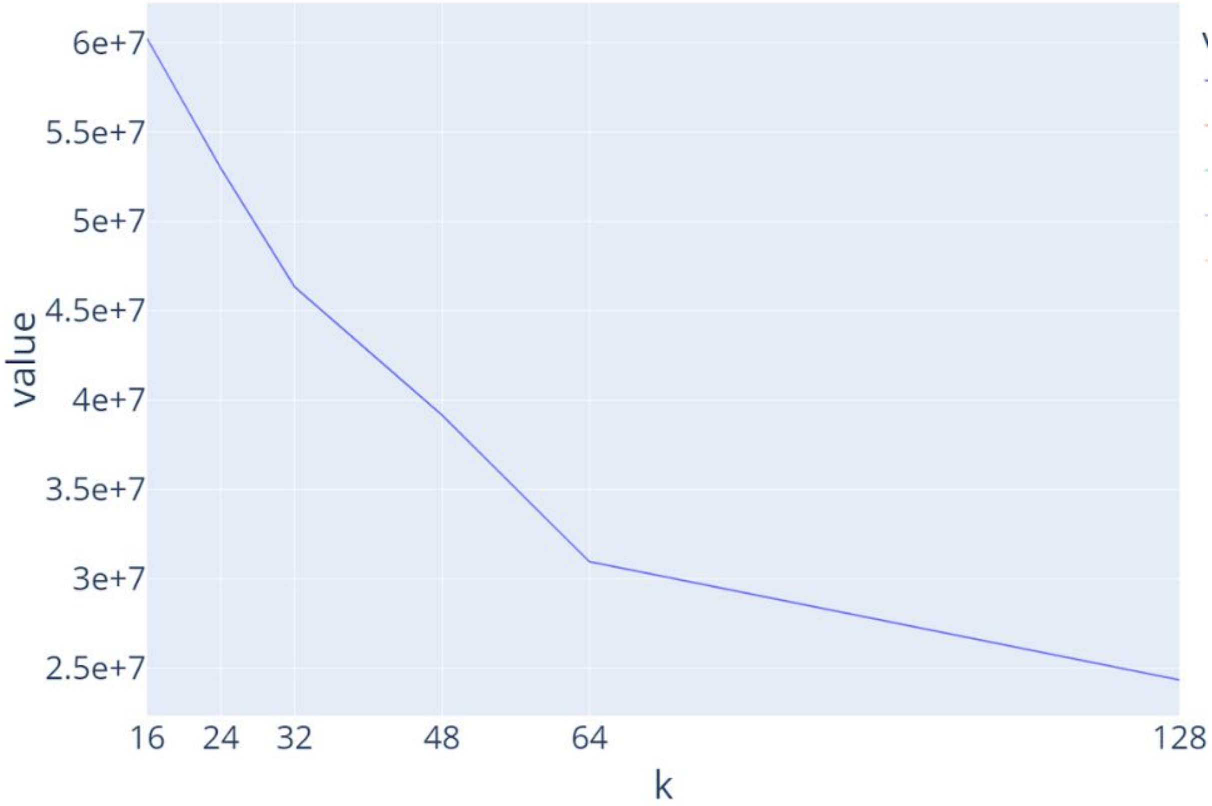
Condition	Features	
w_i is not rare	$w_i = X$	$\& \; t_i = T$
w_i is rare	X is prefix of w_i, $\|X\| \leq 4$	$\& \; t_i = T$
	X is suffix of w_i, $\|X\| \leq 4$	$\& \; t_i = T$
	w_i contains number	$\& \; t_i = T$
	w_i contains uppercase character	$\& \; t_i = T$
	w_i contains hyphen	$\& \; t_i = T$
$\forall \; w_i$	$t_{i-1} = X$	$\& \; t_i = T$
	$t_{i-2}t_{i-1} = XY$	$\& \; t_i = T$
	$w_{i-1} = X$	$\& \; t_i = T$
	$w_{i-2} = X$	$\& \; t_i = T$
	$w_{i+1} = X$	$\& \; t_i = T$
	$w_{i+2} = X$	$\& \; t_i = T$

K-means Clustering Results: Calinski Harabasz score



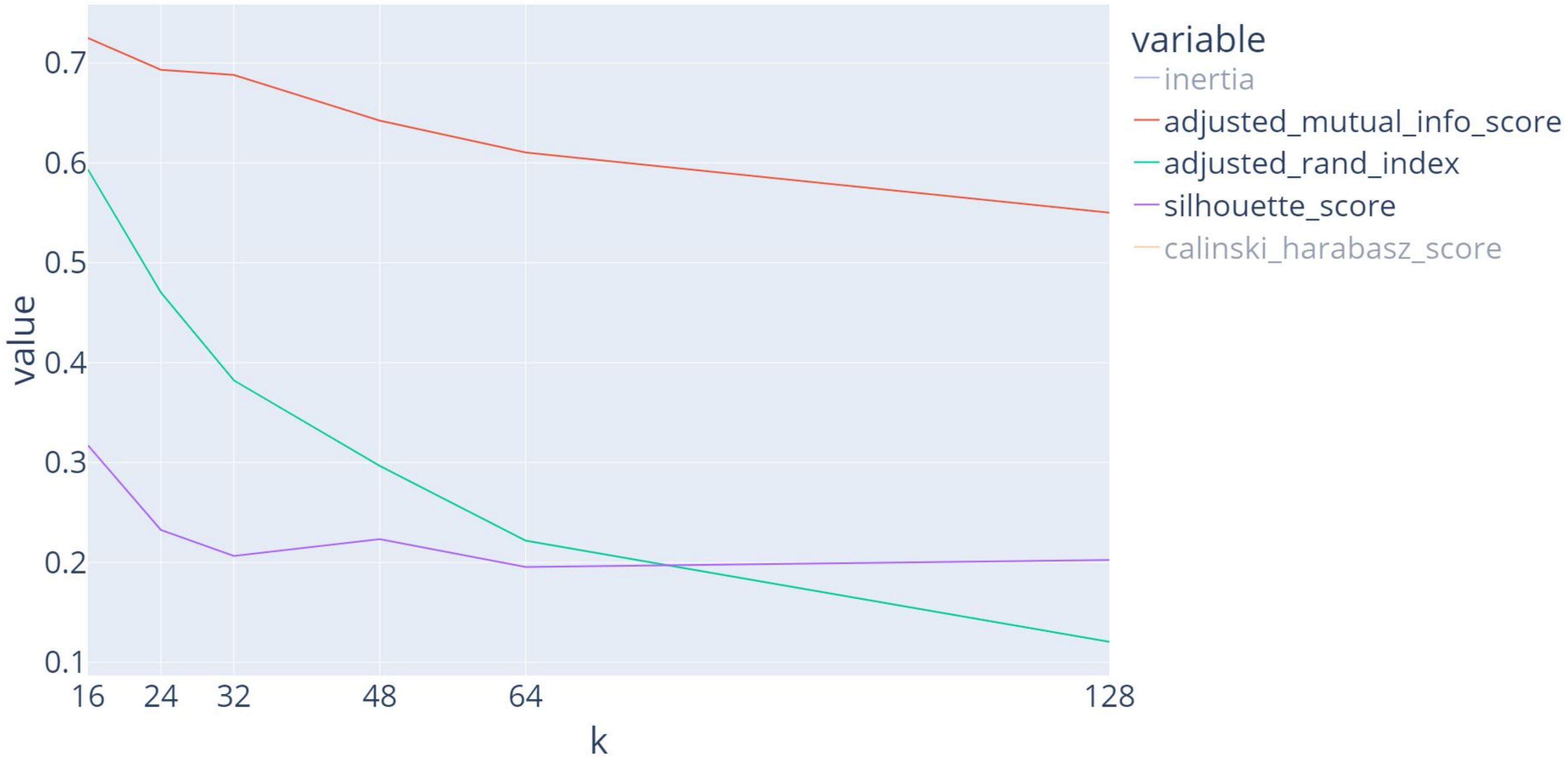
Ratio of sum of inter-cluster dispersion and intra-cluster dispersion for all clusters

K-means Clustering Results: Inertia score

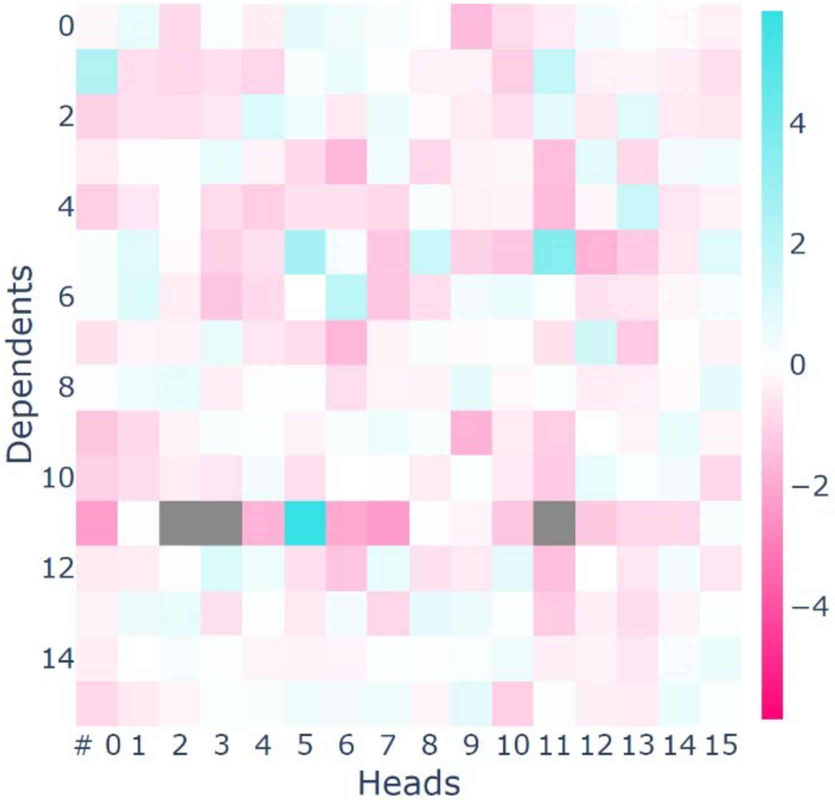


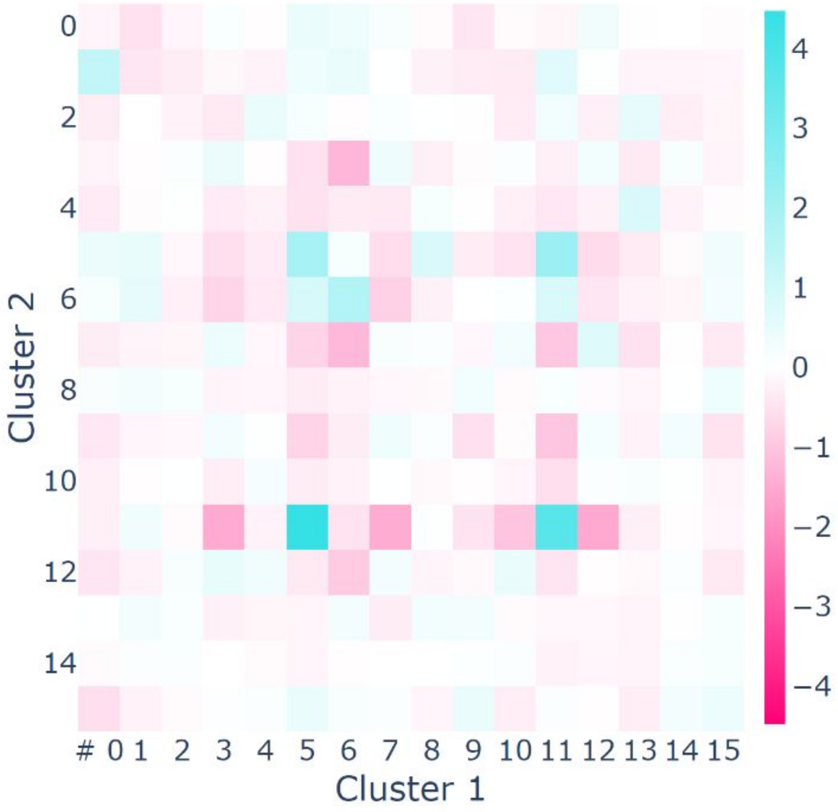
Within-cluster sum of squared distances

K-means Clustering Results: Inertia score

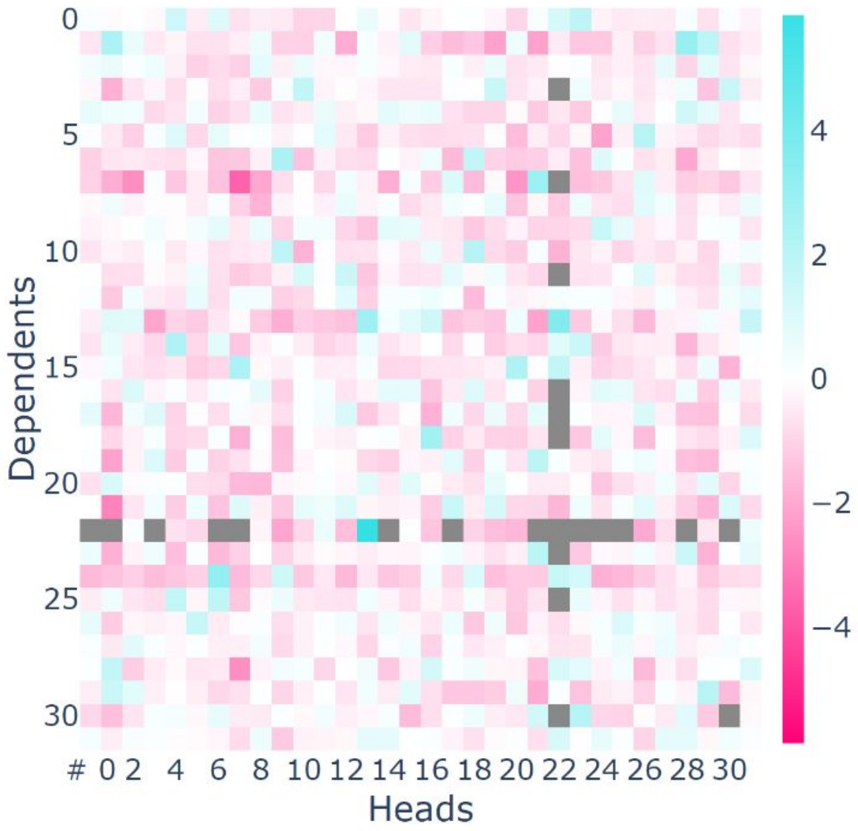


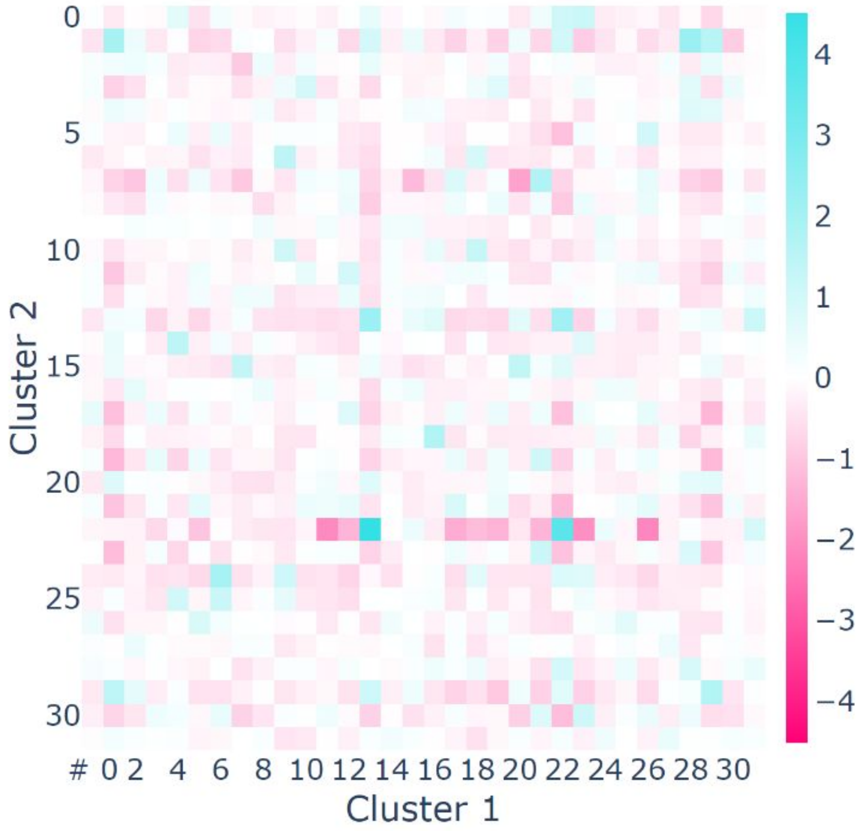
PMI Calculation for K=16





PMI Calculation for K=32





Logistic Regression Results (WSJ)

Memory baseline: Dlg(1) + wtembdep (2) + ldep(3)
Accuracy: 69.891

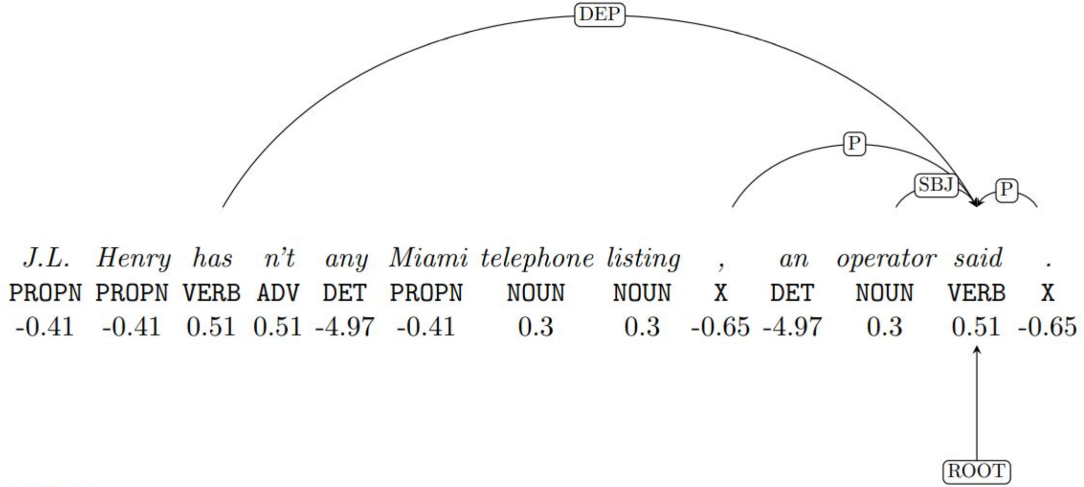
K	1+2+3+ Spearman's correlation for all words (Post)(4)	1+2+3+ Spearman's correlation for dependents(Post)(5)
16	70.275	71.081
32	**71.291**	**73.052**
64	70.978	69.964
128	71.605	70.105

Results for Different Construction Types

Model		Constructions				
		dat	**post**	**pre**	**quote**	**iquote**
SVM (linear)	Memory Baseline	74.7126	80.0513	40.3514	**34.0156**	96.7595
	With spearman_ allwords	73.1211	80.23126	43.29	**55.09582**	89.8237
SVM (rbf)	Memory Baseline	74.7126	81.2505	40.7987	**46.8641**	96.9869
	With spearman_ allwords	73.5632	80.9507	42.3322	**61.5853**	90.847

Logistic Regression	dat	post	pre	**quote**	iquote
Memory Baseline	75.96	80.29	41.77	**45.91**	95.69
With spearman_allwords	71.32	81.95	43.29	**44.45**	91.78

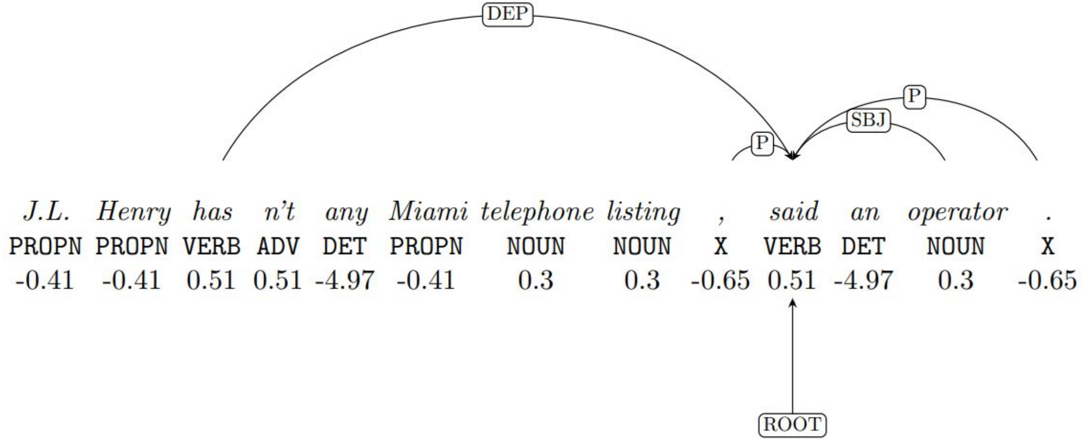
Dependency Trees



wsj_0096.9.1

Variant Sentence

spearman allwords = -1.0



Conclusion

- Addition of PMI features to baseline models increases accuracy, which implies that it successfully captures additional information that is being missed by baseline features.
- Successfully generated PMI Values using maxent followed by k-means clustering on Universal Dependencies Treebank.
- Completed qualitative analysis using dependency trees for different construction types.
- Extended PMI Calculation to cover all post-verbal words.

Future Work

- Try more complex models such as **pre-trained BERT** to obtain word-level embeddings.
- Define **new PMI-based** features and perform similar experiments.
- Currently, only the root verb and it's dependents are used in PMI Features. Try including **other head-dependent relations** present in the sentence.
- Examine the data sparsity problem which occurs when we compute PMI values. PMI values may affected by the frequency of occurrences pair under consideration.

References

Ratnaparkhi A. (EMNLP, 1996) A Maximum Entropy Model for Part-Of-Speech Tagging https://aclanthology.org/W96-0213

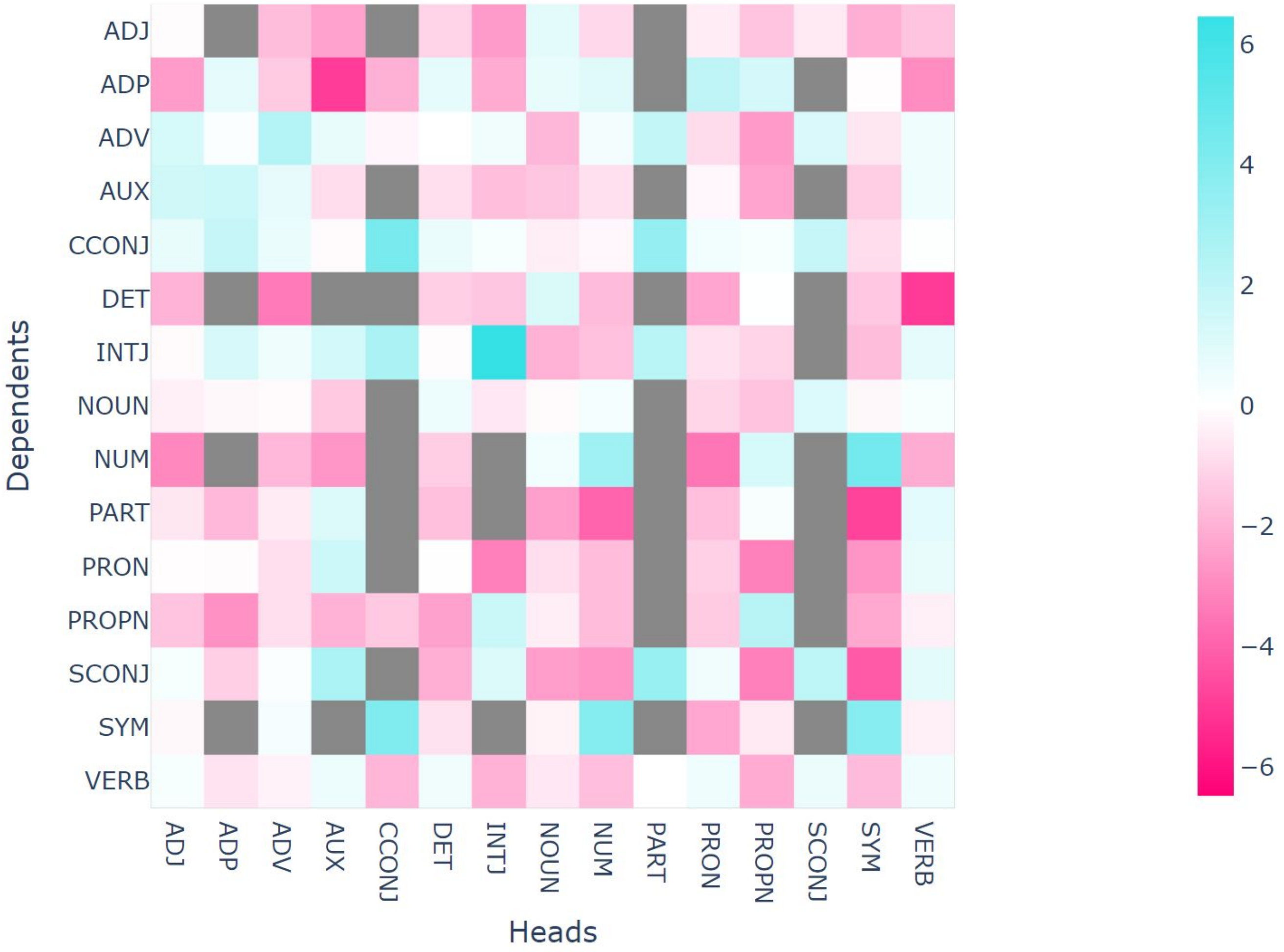
Futrell, R., Gibson, E. and Levy, R.P. (2020), Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. Cogn Sci, 44: e12814. https://doi.org/10.1111/cogs.12814

Rajkumar R, van Schijndel M, White M, Schuler W. Investigating locality effects and surprisal in written English syntactic choice phenomena. Cognition. 2016;155:204-232. https://doi.org/10.1016/j.cognition.2016.06.008

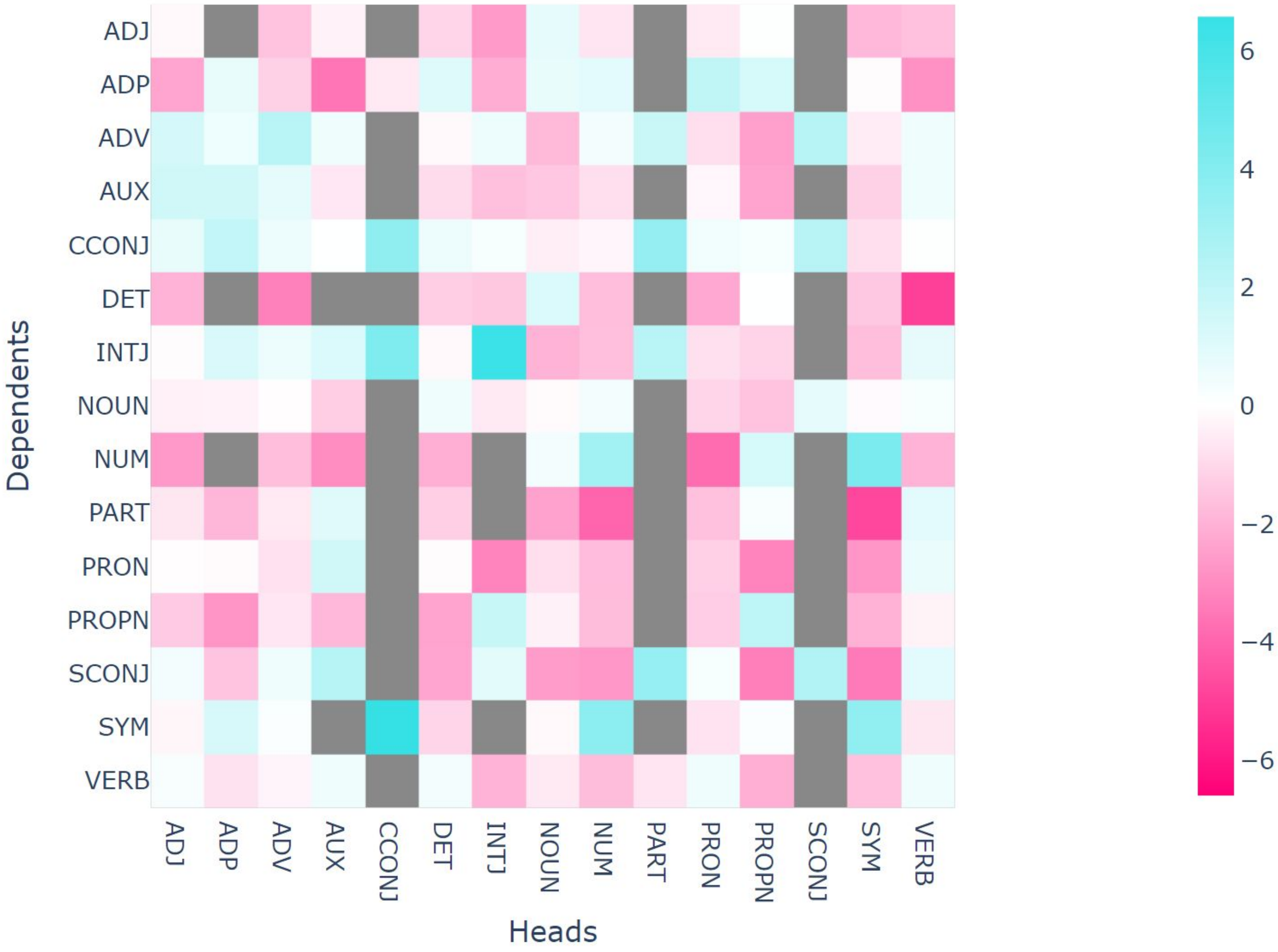
Futrell, R. (2015, August 18). Large-scale evidence of dependency length minimization in 37 languages. PNAS. https://www.pnas.org/content/112/33/10336

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In Proceedings of the First International Conference on Quantitative Syntax, pages 2–15, Paris Futrell, ACL, 2019

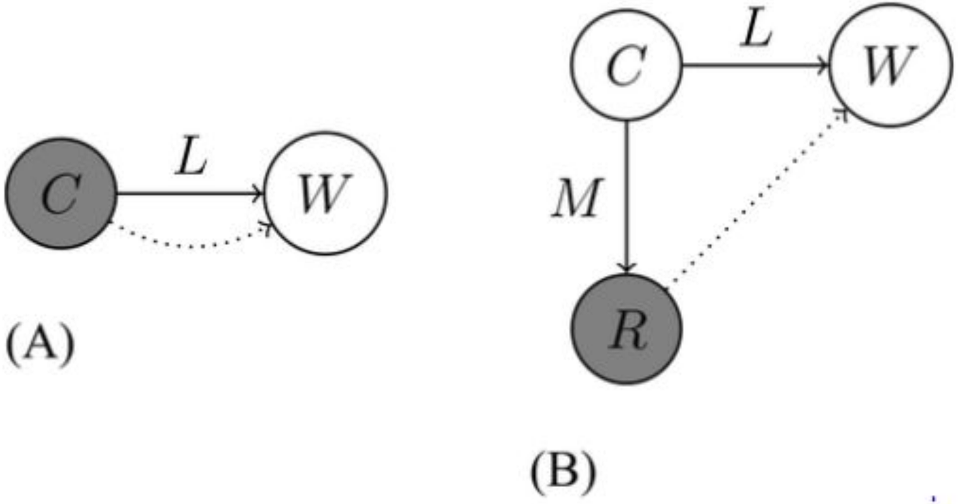
PMI Value between Head - Dependent Pairs **(Futrell et al (2019))**



PMI Value between Head - Dependent Pairs **(My results)**



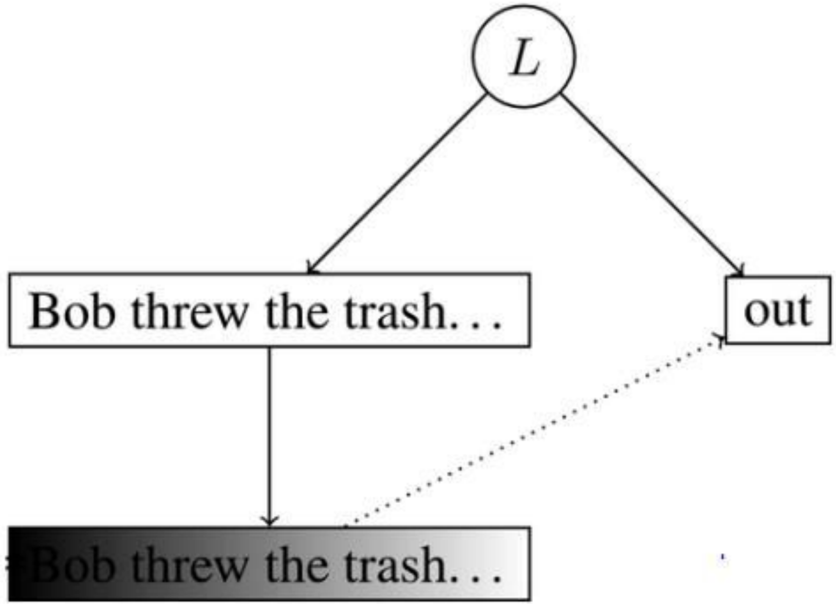
Lossy-Context Surprisal Theory (Futrell, Gibson, Levy (2020))



C: Context
W: Next word
L: Conditional distribution of W given C
R: Random Variable
M: Conditional distribution of R given C

Probabilistic models associated with **Surprisal Theory (A)** and **Lossy-Context Surprisal Theory (B)**

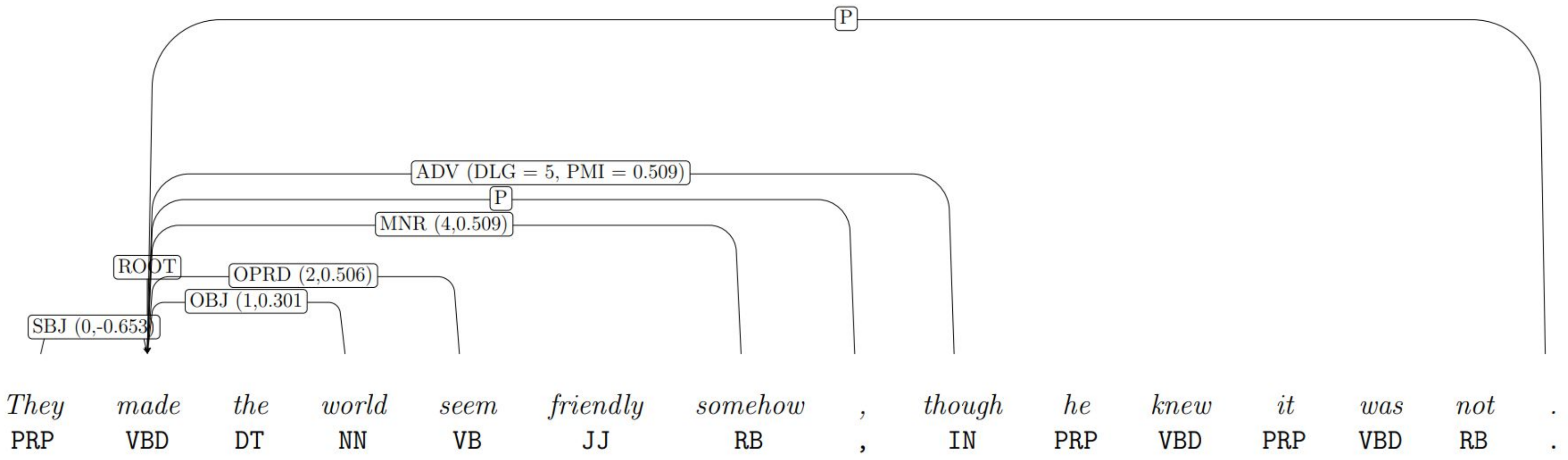
Reference: Futrell, Cognitive Science, 2020



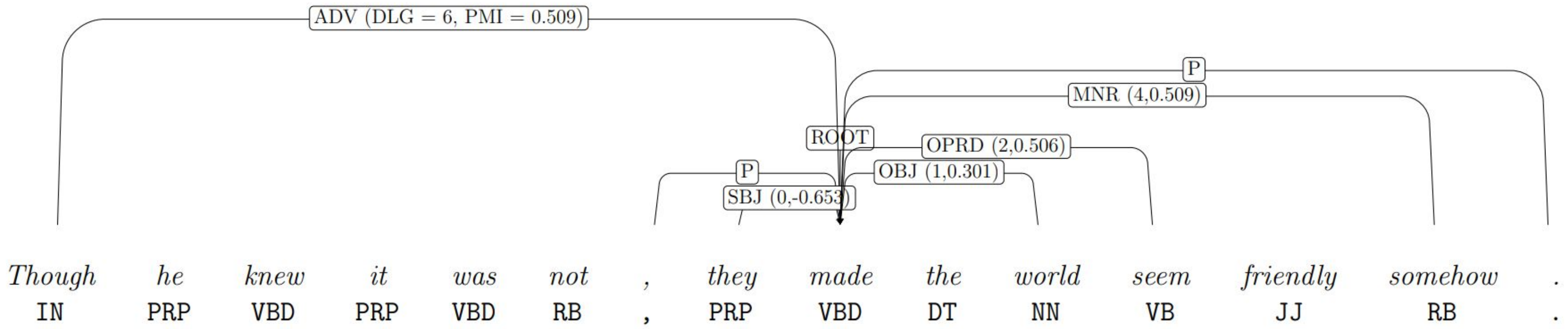
Memory Representation based on Progressive Noise

Reference: Futrell, Cognitive Science, 2020

Reference Sentence:



Variant Sentence:



Models made using combinations of Memory-based and PMI features

Dependency Length (dlg) + Weighted embedding depth (wtembdep) + Lexical (1-best) embedding depth (ldep) + PMI Features

Testing Accuracy (%)	Dlg(1) + wtembdep (2) + ldep(3)	1+2+3+ PMI (adjacent word) (4)	1+2+3 +PMI (nearest dependent of root verb) (5)	1+2+3+ Spearman's correlation for all words (Post) (6)	1+2+3+ Spearman's correlation for dependents (Post) (7)	1+2+3+6+7
Brown	69.1673	68.69548	69.02218	69.48208	69.19156	69.6031
WSJ	69.8914	69.7514	69.75138	72.00741	70.02648	72.04742