Recovery of Adjective Hierarchy through Unsupervised Learning

by

Run Chen

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Molecular Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author . . . . . .
Department of Electrical Engineering and Computer Science
May 15, 2020

Certified by . . . . . .
Robert C. Berwick
Professor
Thesis Supervisor

Accepted by . . . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Recovery of Adjective Hierarchy through Unsupervised Learning

by

Run Chen

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Molecular Biology

Abstract

To understand the cognitive processes for natural language acquisition, we must differentiate between prior and acquired knowledge of language. We take steps towards identifying some of this prior knowledge by applying a computational approach to the *Cartographic Hypothesis*, a linguistic hypothesis that postulates a universal hierarchical syntactic structure for adverb and adjective sequences such that we prefer “little black (purse)" (169/169) over “black little (purse)" (0/169). Specifically, the adjectives are *clustered* and *ordered*. We consider English adjective bigrams in the Google Books Ngram corpus and attempt to recover the clusters, or syntactic groups of adjectives, based on relative order frequencies through unsupervised learning models. Low accuracy in the clustering results (0.45) strongly implies the information in the corpus is insufficient for speakers to acquire the linguistic intuition, and that the mechanisms needed to learn these syntactic structures may be prenatal as opposed to gleaned from the statistical regularity of the adjectives themselves.

Thesis Supervisor: Robert C. Berwick
Title: Professor

Acknowledgments

I would like to thank Professor Robert Berwick for all his help during my time as a Master of Engineering student at MIT. Besides his guidance on my thesis topic, he has taught me ways of conducting academic research and thinking about the fundamental questions in the field of computational linguistics. His advice and feedback has been extremely important for preparing me for future independent research.

I am grateful to have been part of the Berwick lab and met the wonderful people in this community. I would like to thank Sagar, Héctor Javier, Beracah and Ruowang for the helpful insight and discussions.

I would like to thank my parents who has given me endless love and support, and to whom I can attribute all my success and accomplishments in life.

It has been a great pleasure to meet the amazing and supportive friends who have been an integral part of my life at MIT.

Finally, I would like to thank Bright who has comforted and encouraged me during the tough time amid the pandemic.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cartography

The cartography hypothesis asserts a universal hierarchical syntactic structure for all languages at every level. Specifically, the adjectives are ***clustered*** and ***ordered***. For example, native speakers of English prefer “little black (purse)" (169/169) over “black little (purse)" (0/169), because the adjective category describing SIZE is hypothesized to precede that of COLOR.

Although cartography has been present in the literature for long, to the best of our knowledge, no systematic big data analysis on the language corpora has been carried out. We examine (i) whether proposals of adjective hierarchies are valid; (ii) whether the knowledge of the adjective hierarchy is learnable, given information present in the corpus. The corpus data is expected to conform to cartography if the hypothesis is robust. Specifically, we test whether different clustering algorithms results in the same fine-grained clusters and all adjectives fall into one of the clusters. Furthermore, we verify whether such clustering confirms the correct ordering as we see in corpus data.

In the argument of nature versus nurture, if there is sufficient information in the corpus, it is possible for infants to acquire the linguistic knowledge as a blank slate. On the other hand, if no sufficient information is available to speakers, they need to have at least some

degree of prior knowledge of the adjective clusters before language acquisition begins.

Chapter 2

Theory

2.1 The Linguistic Theory

The cartography hypothesis postulates a universal hierarchical syntactic structure for adverb and adjective sequences preferences. The universal structure is supported by cross-linguistic evidence. In Mandarin, “xiao hong hua" (small red flower) is categorically preferred over “hong xiao hua" (red small flower). Preferential adjective order has reported to be respected in languages including Arabic, Dutch, French, Greek, Irish, Japanese, Kannada and Thai [18].

It may be suggested that the adjective order preference is established by the input frequency, that is an adult prefers “little black" because they hear such order more often. However, as with any frequency-based hypothesis on language acquisition, it fails to address the fact that speakers are able to generate and offer judgment on strings of which they have never heard [8]. One popular explanation is that adjectives are hierarchically ordered by their lexical semantic class, which is determined by abstract syntax [3, 18]. Other proposes that the order is directed by adjective subjectivity, inversely correlated to its proximity to the modified noun [15].

2.1.1 Adjective Hierarchy

Within the lexical semantic class hypothesis, the exact adjective order remains under debate. Several adjective cluster hierarchies have been proposed. Dixon (1982) proposed a order:

VALUE > DIMENSION > PYSICAL PROPERTY > SPEED
> HUMAN PROPENSITY > AGE > COLOR.

The Sproat-Shih Hierarchy (1991) includes six lexical semantic categories (Figure 2-1):

QUALITY>SIZE > SHAPE > COLOR > ETHNIC > SYNCATEGOREMATIC

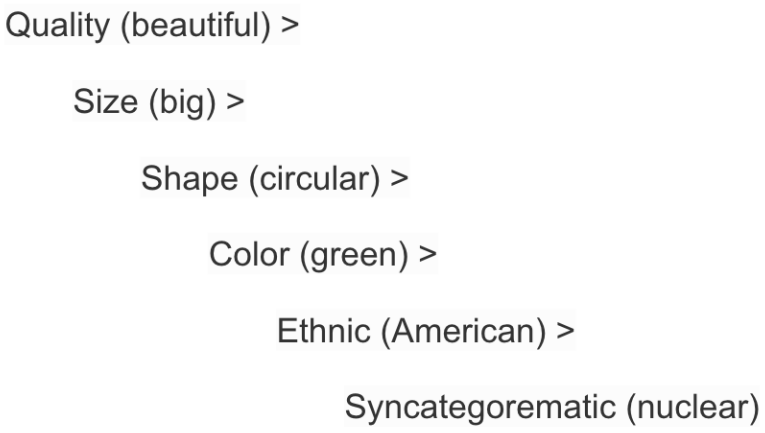

Figure 2-1: Adjective Hierarchy [18]

and Cinque (1994):

POSSESIVE > SPEAKER-ORIENTED > SUBJECT-ORIENTED > MANNER/THEMATIC.

Scott (2002) expands the hierarchy to include more categories such as LENGTH and WEIGHT as shown in Figure 2-2 [16].

SUBJECTIVE COMMENT > EVIDENTIAL > SIZE > LENGTH > HEIGHT
> SPEED > DEPTH > WIDTH > WEIGHT > TEMPERATURE
> WETNESS > AGE > SHAPE > COLOR > ETHNIC
> MATERIAL > COMPOUND ELEMENT > NOUN PHRASE

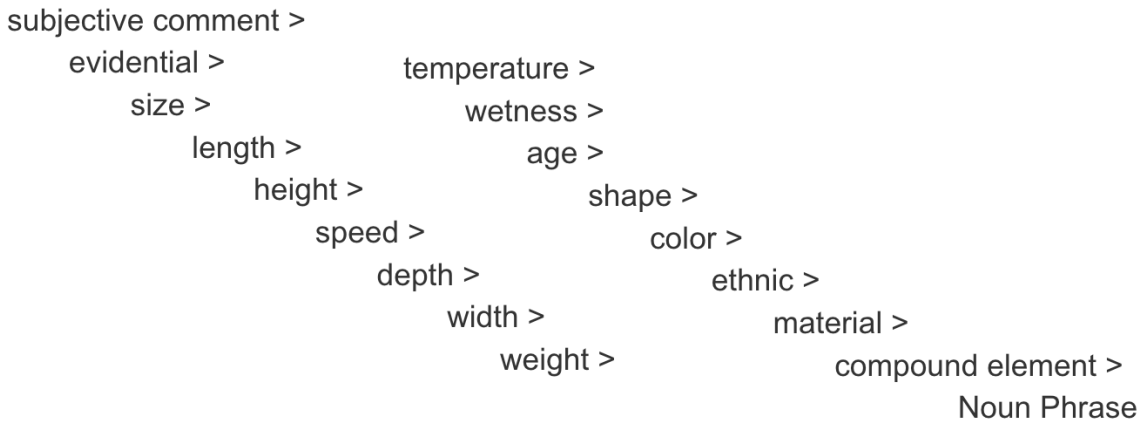

Figure 2-2: Adjective Hierarchy [16]

2.1.2 Exceptions

The adjective hierarchy may no long holds in the presence of

(1) an operator adjective, e.g. a former famous actor;

(2) indefinite superlative, e.g. an Italian shortest student;

(3) phonetic interventions such as comma or focus, e.g. BLACK small purse [18, 19].

The reverse order is only valid when the interpretation of the reverse order is different from the original, otherwise, the reversing operation is unlicensed. Considering possible such exceptions that are not directly observable from the analyzed texts, we expect the distribution of adjective orders in the corpus to be less rigid than hypothesized.

Chapter 3

Experiment

3.1 Hypothesis

To understand the cognitive processes for natural language acquisition, we must differentiate between prior and acquired knowledge of language. We take steps towards identifying some of this prior knowledge by focusing on the adjective hierarchy under the cartography hypothesis. Adopting the lexical semantic hypothesis, We test whether the adjectives are *clustered* and *ordered*. We consider English adjective bigrams in the Google Books Ngram corpus and attempt to recover the clusters, or syntactic groups of adjectives, based on relative order frequencies through unsupervised learning models.

Given pairs of prenominal adjectives, the language model computes their semantics and sorts them into respective clusters, and outputs the correct order. The probability of seeing a certain ordered pair of adjectives in the corpus is given by the following:

$$P((i,j)|D) = P(i,j|D) * P((i,j))$$
$$P((j,i)|D) = P(j,i|D) * P((j,i))$$

Unordered list of adjective-meaning pairs are sorted into the pre-defined clusters based on the semantics. We observe different orders with probabilities assigned by the model. The underlying clusters are C_1, C_2, C_3, ... C_n, where each C_i consists of adjectives under such

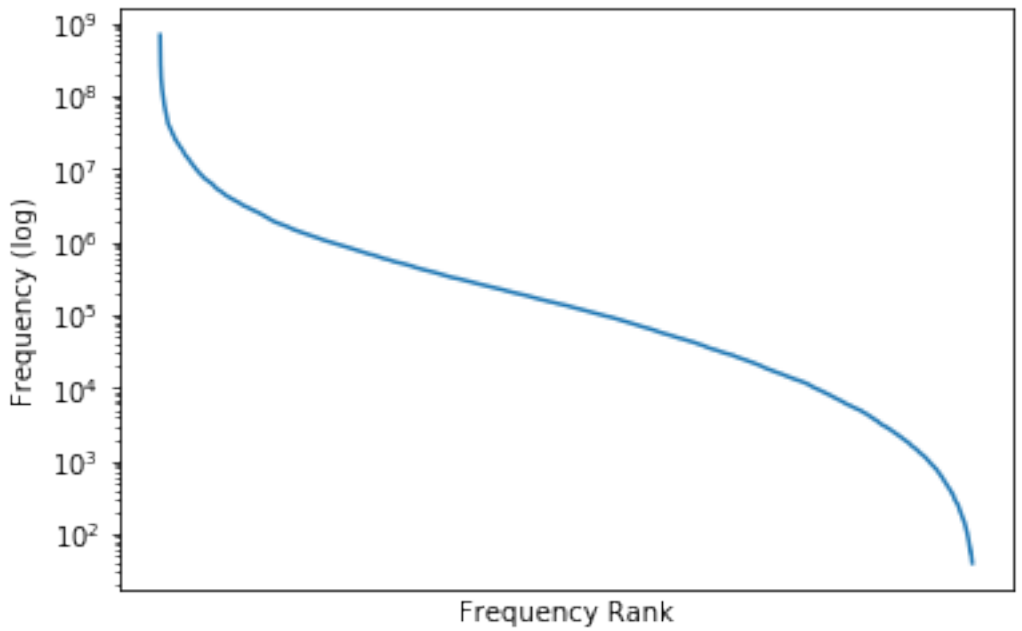

Figure 3-1: Unigram Frequencies on log Scale

semantic category (assuming that the clusters are semantic in nature).

3.2 Method

3.2.1 Data

The adjectives and their frequencies are processed from Google Books Ngram database [10]. We obtained 14045 adjectives in total. Their unigram frequencies range from 40 to 690,317,551. The part-of-speech tagging is verified by checking against the WordNet database [9, 7, 12]. The unigram distribution of adjectives are sorted by frequency of occurrence 3-1.

3.2.2 Encoding

Each adjective is encoded by its relative order frequency to the rest of the adjectives. The bigram matrix A (14045,14045) is constructed from frequencies of pairs of adjectives. We

normalize the entries such that they represent the relative frequencies

$$a_{ij} = [\![\frac{A_{ij}}{A_{ij} + A_{ji}} > 0.5]\!]$$

where A_{ij} is the frequency of the adjective pair (i, j). We say that i categorically precedes j if the likelihood of i preceding j is more than 0.5. Note that if $A_{ij} = A_{ji} = 0, a_{ij} = a_{ji} = 0$.

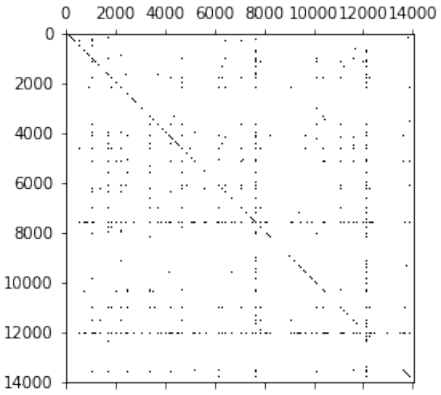

Figure 3-2: The Sparse Bigram Matrix A

The similarity matrix S (28090,28090) is defined as $S = [AA^T]$ such that each row represents an adjective i, where the first half of the columns indicates precedence of i over other adjectives and the second half succession, as shown below.

$$\begin{bmatrix} & A_1 & A_2 & A_3 & \dots & A_1 & A_2 & A_3 & \dots \\ A_1 & - & a_{12} & a_{13} & \dots & - & a_{21} & a_{31} & \dots \\ A_2 & a_{21} & - & a_{23} & \dots & a_{12} & - & a_{32} & \dots \\ A_3 & a_{31} & a_{32} & - & \dots & a_{13} & a_{23} & - & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

We only select the most frequent 1000 adjectives for training. According to the Zipf's law [22], the frequency of a word is inversely proportional to its rank in the frequency table. The adjective unigrams in Figure 3-1 conforms to Zipf's law. The frequency of bigrams drops dramatically beyond the most frequent 1000 adjectives.

Due to the sparsity of the matrix (Figure 3-2), we perform principle component analysis (PCA) to reduce the high feature dimensionality from 28090 to 5.

3.2.3 Clustering

Given the encoding describe in section 3.2.2 and assuming perfect clusters and large dimensions, clustering is a consistent algorithm. Let C_i be a cluster of adjectives. Adjective $A_{ij} \in C_i$ and $A_{ij} \neq A_{kl}$ if $i \neq k$ or $j \neq l$. Then, $\forall i, j$, $C_i \cap C_j = \emptyset$, we call this perfect clusters. Assuming large dimensions, the dimension of the word embedding is much larger than the number of words within the same cluster: $\forall i, \forall j \neq i, |\cap C_j| \gg |C_i|$. We can show that adjectives in the same cluster always have the shortest distance. The distance between any two adjectives within the same cluster is upper-bounded by $|C_i|$, whereas the distance between inter-cluster adjectives is lower-bounded by $|C_j \cap C_i|$. Therefore, given the correct number of clusters, we can always recover the clusters.

We run K-means clustering [1] on the adjective vectors and corresponding pre-trained Word2Vec embeddings for reference [11]. This is based on the implementation of the scikit-learn library [14]. We assign randomly sampled 80% of the data as our training data and the remaining 20% as held-out testing data for the labeled training examples.

The hyper-parameter, the number of clusters k, is chosen as follows. For the labelled *gold clusters* (see section 3.3), k is the number of ground truth clusters chosen. For unlabelled training examples, we perform a grid search for the best value of k.

3.3 Evaluation

Given the categories proposed [16], we construct the *gold clusters* by aggregating synonyms and antonyms. The *gold clusters* consist of categories: *size*, *length*, *height*, speed, *depth*, *weight*, *color*, *ethnic* (Table 3.1). WordNet synsets include synonyms and antonyms for a given word [12, 7]. We perform a closure on all synsets and filter incorrect words, since synsets tend to include loosely relevant words. Web scraping on Thesaurus also re-

trieves synonyms and antonyms for words already in gold clusters [20]. The *color* and *ethnic* clusters are obtained from web scraping Wikipedia, because we cannot expand the list by finding synonyms or antonyms [21].

The clustering algorithm performance is evaluated by the correctness of labeling and separation between the clusters. We use the *gold clusters* as the ground truth class assignments. The accuracy is calculated by measuring the similarity of the predicted assignment of labels and the ground truth assignment. We loop over all possible combinations of numerical ground truth label for the 9 defined categories and report the highest accuracy. When running clustering on all adjectives, the ground truth labels are unknown. We use the Davies-Bouldin index [4] to indicate separation between the clusters. A higher Davies-Bouldin index indicates worse separation and therefore worse performance. Averaged Davies-Bouldin index over 10 epochs is reported.

Order	Cluster	Top Adjectives
1	size	**little**, **small**, **large**, **big**, **huge**, **enormous**, **grand**, **massive**, **minute**, microscopic, gigantic, miniature, monumental, mountainous, colossal, bulky, voluminous, mini, dainty, grandiose, infinitesimal, ponderous, wee, puny, minuscule, smallish, dinky, humongous, bitty, hulking
2	length	**long**, **short**, elongated, stretched, compressed, stringy, longish
3	height	**high**, **short**, **low**, **tall**, elevated, alpine, squat, compressed, stubby, lank, soaring, chunky, rangy, altitudinous
4	speed	**rapid**, **slow**, **quick**, **moderate**, **gradual**, **fast**, **dull**, measured, hasty, inert, fleeting, brisk, accelerated, stagnant, sluggish, ponderous, hurried, leaden, supersonic, flying, expeditious, dilatory, unhurried, fleet, presto, breakneck
5	depth	**low**, **deep**, **shallow**, **superficial**, surface, buried, raised, bottomless, inmost, abysmal
6	weight	**heavy**, **light**, **thin**, **thick**, **fat**,**massive**, **slender**, laden, weighted, meager, bulky, loaded, cumbersome, weighty, airy, lightweight, skinny, ponderous, flimsy, unwieldy, hefty, feathery, portly, corpulent, weightless, chunky, beefy, elephantine
7	temperature	**cold**, **hot**, **warm**, **cool**, **thermal**, **mild**, **tropical**, polar, frozen, temperate, fiery, heated, glacial, burning, icy, bleak, humid, brisk, crisp, boiled, feverish, chilly, arctic, cutting, blazing, chilling, lukewarm, frigid, snug, biting, wintry, frosty, sultry, refrigerated, scorching, torrid, clement, inclement, frosted, searing, sweltering, drafty, baking, summery, sizzling, nippy, algid, brumal, frore
8	shape	**long**, **high**, **short**, **higher**, **wide**, **narrow**, **round**, **square**, **vertical**, **acute**, **plain**, **tall**, **horizontal**, **circular**, **shaped**, **steep**, **angular**, **perpendicular**, **hollow**
9	color	**white**, **black**, **red**, **green**, **blue**, **yellow**, **brown**, **gold**, **grey**, **pink**, **purple**, orange, olive, bronze, coral, crimson, tan, lavender, azure, sapphire, beige, mauve, maroon, jade, amber, indigo, magenta, ultramarine, caramel, amethyst, rose
10	nationality /origin	**american**, **national**, **foreign**, **british**, **english**, **french**, **international**, **german**, **western**, **european**, **indian**, **christian**, **chinese**, **japanese**, **jewish**, **russian**, **greek**, **spanish**, **roman**, **italian**, **irish**, **catholic**, **latin**, **canadian**, **mexican**, **asian**, **dutch**, **australian**, **muslim**, **polish**, **egyptian**, **turkish**, **portuguese**, **korean**, **austrian**, **israeli**, **hindu**

Table 3.1: Gold Clusters
Order by frequency in the Google Books Database. The most frequent 1000 adjectives are in bold.

Chapter 4

Results

4.1 Gold clusters

K-means clustering on well-defined categories as show in Table 3.1 yields an accuracy of 0.45 on training data and 0.42 on hold-out testing data. Some clusters have higher consistency, for example, *ethnic* and *color*. Some other clusters have consistent low accuracy, for example, *height*, *weight* and *length*. As a metric for the unlabelled data, the Davies-Bouldin Score is measured as 0.88.

The frequency of a word in the corpus may confound the clustering result. The predicted labels for adjectives are often divided for adjectives in the top 200 and after 200, despite being in the same ground truth cluster. We also observe that the accuracy for clustering improves for most frequent 200 adjectives but this could also be the result of overfitting a much smaller dataset.

4.2 Most frequent 1000 adjectives

We predict the performance for the most frequent 1000 adjectives to be sightly worse than the *gold clusters* because the larger dataset is more noisy. Given the adjectives are largely unlabelled for the most frequent 1000 adjectives, we calculate the Davies-Bouldin Index ranges from 1.20 to 1.23 in the grid search. This is indeed higher than the *gold clusters*.

Training on the Word2Vec embeddings yields in different clusters of adjectives. This is expected because Word2Vec embeddings carry more contextual information than the bigram similarity matrix. Such contextual information could be less helpful for the task. Adjective ngrams are much rarer in the corpus compared to the rest of the adjectives' distribution.

Chapter 5

Discussion

5.1 Learnability

Cartography builds on the Universal Grammar (UG) that assumes a universal structure that applies to all languages at every level. "The study of the feature inventory of UG requires a massive database compiled on the basis of detailed studies of particular grammars. [17]" This project raises the interesting question of how humans acquire a certain language phenomenon. Given any pair of adjectives, a native speaker of English has intuition for the order. Memorization of each lexical entry seems computationally infeasible for the human brain. We check whether such knowledge of English is readily available in the corpus.

The overall performance of the clustering algorithm is relatively low, although it is better than random guessing, which gives an accuracy of about 0.15. Based on the training result, there is no sufficient information in the corpus for machine learning models to learn the underlying clusters and their orders. This poses the question as to how infants acquire these orders. It seems impossible for infants to learn the clusters solely with the language input. Given the poverty of stimulus, they likely have prior knowledge about the correct clustering. Children exhibit development of the acquisition of the adjective hierarchy [8], as older children produce adult-like adjective pairs more frequently in the CHILDES corpus. A probabilistic analysis on child-produced prenominal adjective pairs (Adj Adj N) reveals a larger likelihood of the lexical semantic hypothesis than alternative hypotheses.

5.2 Future Work

We hope to extend the research to more corpora. Google Books Ngram dataset is built on written text. The same experiment on speech may or may not yield similar results. We predict that the speech data tend to be more noisy, and therefore more difficult to generate meaningful clusters.

Another possible direction we could look into is the encoding of the adjectives. Our model fails to address the problem of ambiguity in the presence of homonyms, for example, 'short' can be both in the LENGTH and HEIGHT clusters. A model that accounts for homonyms is expected to distinguish adjectives better. The adjective order is not deterministic. To account for the noisy samples of adjective hierarchy available to children, a categorical order of two adjectives (A_i, A_j) is determined when the number of reverse order pairs follows the Tolerance Principle, $Freq(A_j, A_i) \leq N/\ln N$ where $N = Freq(A_j, A_i) + Freq(A_i, A_j)$ [23].

Besides semantics, there could be other latent factors for the adjective orders. For example, it might be the case that speakers prefer frequent adjectives to precede less frequent ones under an information theory approach [13]. Although this hypothesis is not tested in the project, we see effects of frequency on the performance of clustering algorithms.

Bibliography

[1] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. volume 8, pages 1027–1035, 01 2007. doi: 10.1145/1283383.1283494.

[2] T. Bever. *The Cognitive Basis for Linguistic Structures*, pages 279–352. 01 1970. doi: 10.1093/acprof:oso/9780199677139.003.0001.

[3] G. Cinque. *On the evidence for partial N-movement in the Romance DP*, pages 85–110. Georgetown University Press, Washington DC, 1994.

[4] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979. ISSN 1939-3539. doi: 10.1109/TPAMI.1979.4766909.

[5] R. M. W. Dixon. *Where have All the Adjectives Gone? and other essays in semantics and syntax*. De Gruyter Mouton, Berlin, Boston, 1982. ISBN 978-3-11-082293-9.

[6] M. Dye, P. Milin, R. Futrell, and M. Ramscar. Cute little puppies and nice cold beers: An information theoretic analysis of prenominal adjectives. In *CogSci*, 2017.

[7] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[8] G. S. L. P. Galia Bar-Sever, Rachael Lee. Little lexical learners : Quantitatively assessing the development of adjective o rdering preferences. In A. B. Bertolini and M. J. Kaplan, editors, *BUCLD 42: Proceedings of the 42nd annual Boston University Conference on Language Development*, pages 58–71, Somerville, MA, 2018. Cascadilla Press.

[9] Y. Lin, J.-B. Michel, E. Aiden Lieberman, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the Google books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P12-3029`.

[10] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, , J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. ISSN 0036-8075. doi: 10.1126/science.1199644. URL `https://science.sciencemag.org/content/331/6014/176`.

[11] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[12] Princeton University. About wordnet, 2010. URL https://wordnet.princeton.edu/.

[13] M. Ramscar, A. H. Smith, M. Dye, R. Futrell, P. Hendrix, R. H. Baayen, and R. Starr. The 'universal' structure of name grammars and the impact of social engineering on the evolution of natural information systems. In *CogSci*, 2013.

[14] scikit-learn 0.22. 2.3. clustering. URL https://scikit-learn.org/stable/modules/clus

[15] G. Scontras, J. Degen, and N. D. Goodman. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1):53–66, 2017. doi: 10.1162/OPMI_a_00005. URL https://doi.org/10.1162/OPMI$_{a0}$0005.

[16] G.-J. Scott. Stacked adjectival modification and the structure of the nominal phrases. In G. Cinque, editor, *Functional Structure in DP and IP*, pages 91–120. Oxford University Press, 2002.

[17] U. Shlonsky. The cartographic enterprise in syntax. *Language and Linguistics Compass*, 4(6):417–429, 2010. doi: 10.1111/j.1749-818X.2010.00202.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2010.0020

[18] R. Sproat and C. Shih. *The Cross-Linguistic Distribution of Adjective Ordering Restrictions*, pages 565–593. Springer Netherlands, Dordrecht, 1991. ISBN 978-94-011-3818-5. doi: 10.1007/978-94-011-3818-5_30. URL https://doi.org/10.1007/978-94-011-3818-5_30.

[19] A. Teodorescu. Adjective ordering restrictions revisited. pages 399–407, Somerville, MA, 2006.

[20] Thesaurus. Thesaurus. URL https://www.thesaurus.com/. [Online; accessed 13-December-2019].

[21] Wikipedia contributors. Lists of colors — Wikipedia, the free encyclopedia. URL https://en.wikipedia.org/wiki/Lists_of_colors. [Online; accessed 13-December-2019].

[22] A. Wray. Protolanguage as a holistic system for social interaction. *Language Communication*, 18(1):47 – 67, 1998. ISSN 0271-5309. doi: https://doi.org/10.1016/S0271-5309(97)00033-5. URL http://www.sciencedirect.com/science/article/pii/S0271530997000335.

[23] C. Yang. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press, Cambridge, MA, 2016.