Analyzing Syntactic Structural Dependency in NLP

Hardjono Thomas *, Rodolfo Patel, Paul Dave

Posted Date: 6 December 2023

doi: 10.20944/preprints202312.0335.v1

Keywords: Syntactic Analysis; Dependency Grammar; Linguistic Typology

Article

Analyzing Syntactic Structural Dependency in NLP

Hardjono Thomas *, Rodolfo Patel and Paul Dave

Briar Cliff University

* Correspondence: thomas@briarcliff.edu

Abstract: This paper introduces a novel graph polynomial approach for differentiating tree structures in dependency grammar. Utilizing this polynomial representation, we develop a metric to assess the similarity in syntax. This approach offers a detailed and inclusive analysis of the dependency structures and relationships in sentence construction. We employ this polynomial method to examine sentence structures across various languages in the Parallel Universal Dependencies treebanks. Our analysis includes comparing the syntax of original sentences and their translated counterparts in diverse languages, alongside a comprehensive study of syntactic typologies within these treebanks. Additionally, we explore the application of our methodology in evaluating the syntactic diversity within language corpora.

Keywords: syntactic analysis; dependency grammar; linguistic typology

1. Introduction

In the realm of linguistic studies, the study of dependency grammar is pivotal, delving into the intricate relationships and the hierarchical arrangement of words within sentences. This is depicted using a model known as dependency trees [1]. The Universal Dependency (UD) project, a significant international collaboration, has pioneered in creating a uniform framework for annotating these trees, thereby assembling a comprehensive collection of linguistic treebanks in numerous languages [2]. These treebanks are crucial for the automation of linguistic analysis and for in-depth syntactic typology research. A specialized section of these treebanks, known as the Parallel Universal Dependency (PUD) treebanks, encompasses the translations of 1,000 sentences from sources like news and Wikipedia, initially in languages such as English, French, German, Italian, or Spanish.

Presently, these treebanks span across 20 different languages, providing an extensive resource for sophisticated syntactic study [3]. Notwithstanding these developments, prevailing techniques for depicting dependency trees tend to neglect all-encompassing syntactic information, focusing instead on aspects like the sequence of words and their inter-dependencies [4–10]. Our research aims to bridge this gap by introducing a holistic polynomial-based model for representing dependency trees, ensuring that two sentences are identified as having the same dependency structure exclusively when their corresponding dependency tree polynomials are identical.

In various mathematical disciplines, particularly in knot and graph theories, structural polynomials have gained considerable interest as tools for characterizing intricate structures. For instance, in knot theory, polynomials such as the Jones polynomial [11] and the HOMFLY polynomial [12] are utilized for analyzing knots and links, shedding light on their characteristics and properties [13–20]. In graph theory, the Tutte polynomial [21] is instrumental in unveiling graph characteristics, including aspects like the number of spanning trees and the variety of possible graph colorings. Recent advancements have seen the development of a unique structural polynomial that effectively characterizes unlabeled trees, establishing a one-to-one correlation between these trees and a set of bivariate polynomials [22]. This novel polynomial has found applications in fields such as the analysis of phylogenetic trees and the study of evolutionary patterns [27], and its scope has been extended to include certain types of phylogenetic networks [28–30]. When compared to conventional tree comparison methodologies, like tree kernels [36,37] and tree edit distances [38], this polynomial-based method has demonstrated enhanced precision and computational efficiency [27].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2023 doi:10.20944/preprints202312.0335.v1

Yet, when it comes to comparing dependency trees, most current methods primarily focus on the local structural components rather than the holistic structure of the trees.

In this paper, we adapt the tree distinguishing polynomial for the purpose of representing dependency trees and formulate a metric for measuring syntactic similarity based on the distances between these polynomials. We then apply our approach to the PUD treebanks, undertaking a comprehensive analysis of sentences that exhibit varying degrees of syntactic divergence and conducting a detailed syntactic typology study across the languages represented in these treebanks. Furthermore, we investigate the utility of polynomial distances as a metric for assessing the syntactic diversity within a corpus, and discuss its potential applications and broader implications.

2. Probing Test and Study Design

2.1. Elaborating on Dependency Trees

A dependency tree for a sentence, a fundamental concept in syntactic analysis, is a structured model in the form of a rooted, node-labeled tree. It meticulously illustrates the intricate grammatical relationships between words in a sentence. Each node in this tree represents a word, and every edge delineates a grammatical linkage between two words: the word closer to the root (the 'head') and its corresponding 'dependent'. Remarkably, a head can have several dependents, but a dependent is uniquely linked to a single head. The label on a dependent node explicitly specifies its grammatical relationship to the head. The root node, signifying the main head of the sentence, does not function as a dependent and is labeled to indicate its root status. Interestingly, identical grammatical relationships can occur multiple times within a sentence, leading to dependents in the tree having the same labels. We illustrate this concept with the dependency trees of an English sentence and its Chinese translation, where the nodes are labeled with numbers indicating head-dependent grammatical relations as described in Table 1. All dependency trees in this study are meticulously constructed under the universally acknowledged Universal Dependencies (UD) framework [2].

2.2. Exploring Parallel Universal Dependencies

Our analysis extends to the dependency trees found in the Parallel Universal Dependencies (PUD) treebanks, originating from a task set by the Conference on Computational Natural Language Learning (CoNLL 2017) [3]. The PUD treebanks were composed by randomly selecting 1,000 sentences from various online news sources and Wikipedia articles, predominantly in English, German, French, Italian, and Spanish. These sentences were subsequently translated by expert linguists into multiple languages. Each PUD treebank encompasses 1,000 dependency trees, representing either the original sentences or their translations in a specific language. As of now, 20 distinct PUD treebanks are available, showcasing dependency trees in languages ranging from Arabic to Turkish, including English, Chinese, Czech, Finnish, French, German, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish.

2.3. Insights into the Tree Distinguishing Polynomial

The graph polynomial that uniquely identifies unlabeled trees, as introduced in [22], is a cornerstone of our study. Every rooted unlabeled tree T is associated with a distinct bivariate polynomial $P(T, x, y)$. To calculate $P(T, x, y)$ for any given unlabeled tree T, we initiate a bottom-up approach, assigning polynomials to each node, starting from the leaves and progressing to the root. The polynomial at the root then represents $P(T, x, y)$. Let $P(n, x, y)$ denote the polynomial at node n. For a leaf node n, we set $P(n, x, y) = x$. Consider an internal node m with k children $n_1, n_2, ..., n_k$; the polynomial at m is given by $P(m, x, y) = y + \Pi_{i=1}^{k} P(n_i, x, y)$. This process effectively captures the 'topology' of a dependency tree, which refers to the tree's structural layout minus any labels. We delve into the methodology for computing these polynomials, demonstrating how they uniquely represent the topologies of dependency trees. The isomorphism of two unlabeled trees is assured if and only if

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2023 doi:10.20944/preprints202312.0335.v1

their corresponding polynomials match. Each term within an unlabeled tree's polynomial correlates to a specific subtree, offering interpretability and detailed insights into the tree's structure. For a comprehensive understanding of the tree distinguishing polynomial, and methodologies based on this polynomial for analyzing tree structures, refer to [22,27].

2.4. Generalizing Polynomial for Dependency Trees

In this study, we extend the scope of the tree distinguishing polynomial to aptly represent dependency trees. Unlike mere tree topologies, dependency trees come with node labels, and in our case, there are 37 distinct labels depicting head-dependent relationships as listed in Table 1. These labels can appear in both leaf and internal nodes of a dependency tree. Therefore, we adapt the polynomial to include 74 variables, categorized into two sets: $X = \{x_1, x_2, ..., x_{37}\}$ and $Y = \{y_1, y_2, ..., y_{37}\}$. The generalized polynomial for a dependency tree T is

denoted as $P(T, X, Y)$. Similar to the original method, we calculate $P(T, X, Y)$ recursively from the leaf nodes to the root. If node n^ℓ is a leaf with label ℓ, we assign $P(n^\ell, X, Y) = x_\ell$. For an internal node m^ℓ with label ℓ and k children $n_1, n_2, ..., n_k$, the polynomial at m^ℓ is $P(m^\ell, x, y) = y_\ell + \Pi_{i=1}^{k} P(n_i, x, y)$. We demonstrates this process, showcasing the computation of generalized polynomials for the two dependency trees. This generalization ensures that two dependency trees have identical generalized polynomials if and only if they are isomorphic with matching node labels. Thus, two sentences are deemed to have the same dependency structure if their dependency trees yield identical generalized polynomials. For brevity, we refer to this generalized polynomial as the 'dependency tree polynomial' of a sentence.

Table 1. Descriptions of head-dependent relations, which are defined in the Universal Dependencies (UD) [2].

Index	Dependency Arc	Index	Dependency Arc
1	Adjectival clause modifier	20	Fixed multiword expression
2	Adverbial clause modifier	21	Flat multiword expression
3	Adverbial modifier	22	Goes with
4	Adjectival modifier	23	Indirect object
5	Appositional modifier	24	List
6	Auxiliary	25	Marker
7	Case marking	26	Nominal modifier
8	Coordinating conjunction	27	Nominal subject
9	Clausal complement	28	Numeric modifier
10	Classifier	29	Object
11	Compound	30	Oblique nominal
12	Conjunct	31	Orphan
13	Copula	32	Parataxis
14	Clausal subject	33	Punctuation
15	Unspecified dependency	34	Overridden disfluency
16	Determiner	35	Root
17	Discourse element	36	Vocative
18	Dislocated elements	37	Open clausal complement
19	Expletive		

2.5. Quantifying Polynomial Distance in Dependency Trees

In the polynomial representation of an unlabeled tree, hierarchical information is encapsulated in the coefficients and exponents of each term. In the case of dependency trees, syntactic details are primarily embedded in the term exponents due to the incorporation of additional variables. We introduce a novel metric to compare dependency tree polynomials, and consequently, the dependency trees themselves. The polynomial $P(T, X, Y)$ for a dependency tree T is expressed term-by-term. Each term is presented as a 75-entry vector $t = [e_{x_1}, e_{x_2}, ..., e_{x_{37}}, e_{y_1}, e_{y_2}, ..., e_{y_{37}}, c]$, where e_{x_i} and e_{y_i} are the exponents of variables x_i and y_i, respectively, and c is the term's coefficient. Such a vector is termed a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2023 doi:10.20944/preprints202312.0335.v1

'term vector' of the polynomial $P(T, X, Y)$. Let P and Q be two dependency tree polynomials, with M_P and M_Q representing their respective sets of term vectors. The count of term vectors in M_P (or M_Q) is denoted by $|\_P|$ (or $|M_Q|$). Given two term vectors s and t, the Manhattan distance between them, denoted as $\|s-t\|_1$, forms the basis of our 'polynomial distance' metric for a pair of dependency tree polynomials P and Q, calculated using Formula (1).

$$d(P,Q) = \frac{\sum_{s\in M_P} \min_{t\in M_Q} \|s-t\|_1 + \sum_{t\in M_Q} \min_{s\in M_P} \|s-t\|_1}{|M_P|} + |M_Q| \tag{1}$$

Given the one-to-one correlation between polynomials and dependency trees, the defined distance metric for dependency tree polynomials also applies directly to the trees themselves. Thus, throughout this paper, any reference to the polynomial distance between dependency trees implicitly pertains to the distance between their corresponding dependency tree polynomials. Additionally, since each sentence in the PUD treebanks has a unique dependency tree constructed under the UD framework, references to polynomial distances between sentences implicitly refer to distances between their dependency tree polynomials.

2.6. Experimental Design

In our comprehensive study, we categorize the 1,000 sentences from the Parallel Universal Dependencies (PUD) treebanks into five distinct datasets. These datasets are named using the upper-case ISO 639-2/B language codes, reflecting the original language of the sentences in each dataset. Specifically, we refer to these datasets as ENG, GER, FRE, ITA, and SPA, corresponding to English, German, French, Italian, and Spanish, respectively.

- The ENG dataset is the largest, comprising 750 sentences initially penned in English. Each sentence is accompanied by 20 distinct dependency trees, representing translations into the 20 languages included in our study. This results in a total of 15,000 dependency trees within the ENG dataset.
- The GER dataset encompasses 100 German-origin sentences, with each having 20 translated dependency trees, totaling 2,000 trees.
- Similar to the GER dataset, the FRE, ITA, and SPA datasets each contain 50 original sentences in French, Italian, and Spanish, respectively. Each sentence in these datasets also translates into 20 different languages, summing up to 1,000 dependency trees per dataset.

Throughout this paper, we use a color-coded approach to differentiate results obtained from each dataset: blue for ENG, yellow for GER, purple for FRE, green for ITA, and red for SPA.

2.6.1. Dependency Tree Analysis Across Languages

Every sentence in our datasets has been rendered in 20 languages, leading to the construction of 20 corresponding dependency trees for each sentence. This allows for a unique identification of each dependency tree based on the original sentence and its translated language.

We embark on an intricate analysis by computing the polynomials of all dependency trees within each dataset. Our focus is on calculating and examining the pairwise polynomial distances among the 20 dependency trees corresponding to each sentence. This analysis sheds light on the syntactic nuances between different language translations of the same sentence.

For each sentence, the calculated pairwise distances between its 20 dependency trees are organized into a 20×20 matrix, termed the "Translation Distance Matrix." Each element in this matrix represents the "Translation Distance" between two languages for that sentence. To derive broader insights, we aggregate these translation distances across all sentences in a dataset to form the "Language Distance Matrix" of that dataset. This matrix provides a quantifiable measure of syntactic similarity between language pairs based on our corpus.

 doi:10.20944/preprints202312.0335.v1

We further distill this information by presenting the mean and median values of all pairwise language distances. Additionally, we highlight language pairs that exhibit the closest and farthest syntactic relationships. To deepen our analysis, we calculate the "Average Language Distance" for each language, which represents the mean of its syntactic distances when compared with the other 19 languages. This metric allows us to identify languages that are, on average, syntactically closer or more distant from others within each dataset.

2.6.2. Visualizations and Syntactic Typology Study

Utilizing the Language Distance Matrices, we engage in a detailed syntactic typology study of the 20 languages present in the PUD treebanks. We employ Multidimensional Scaling (MDS) [39] and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [40] to construct informative visual representations. These visualizations, including dendrograms, provide diverse perspectives for analyzing syntactic similarities and differences among the languages based on our datasets.

2.6.3. Corpus Analysis and Syntax Diversity Measurement

Each of the 20 language translations in a dataset is considered a separate corpus. We calculate all possible pairwise sentence distances within each corpus, referring to these as "Pairwise Sentence Distances." These distances are then used to analyze the syntactic diversity within each language corpus. The "Diameter" of a corpus, defined as the maximum pairwise sentence distance, serves as a simplistic yet effective measure of syntactic diversity [41]. This approach opens up new avenues for utilizing polynomial methods in assessing and understanding the diversity of syntactic structures across languages.

2.7. Syntax Comparison of Sentences

The introduction of our novel dependency tree polynomial distance metric provides a ground-breaking quantitative approach for analyzing syntax similarities between sentences. When two sentences share an identical dependency structure, their polynomial distance becomes zero, denoting perfect syntactic alignment. Conversely, a larger polynomial distance indicates a greater divergence in their syntactic structures. For instance, the computed distance between two specific dependency trees in our study is 5.06, suggesting a moderate level of syntactic variation.

2.7.1. Detailed Syntax Analysis in Sentence Pairs

We delve deeper into the syntax comparison by examining specific sentence pairs. One illustrative example is the analysis of a sentence from the ENG dataset and its corresponding Chinese translation. The polynomial distance for this pair is strikingly low at 0.43, the lowest in the ENG dataset for English-Chinese comparisons. This minimal distance signifies a strong syntactic resemblance. The notable syntactic similarity lies in their subject-predicate form and the complexity of noun phrases in the subjects. However, a nuanced difference exists in the usage of time adverbials, highlighting the subtleties of syntactic variations across languages.

In stark contrast, another sentence pair from the ENG dataset, involving an English sentence and its Chinese translation, exhibits a polynomial distance of 22.93 – the highest in the dataset for this language pair. This significant distance points to a pronounced difference in their syntactic structures. The English sentence displays a right-branching dependency tree, whereas its Chinese counterpart is left-branching. Such branching discrepancies are often observed in longer sentences within the ENG dataset. Additionally, structural differences are evident in their sentence stems, with the English sentence having a double-object structure, unlike the single-object structure in the Chinese translation.

These examples underscore the variability and complexity of syntax across different languages. Shorter sentences typically show less variation in syntax, resulting in smaller polynomial distances. In contrast, longer sentences offer more room for diverse syntactic structures, leading to larger polynomial distances. Further investigations include an analysis of sentences in the ENG dataset that exhibit both

 doi:10.20944/preprints202312.0335.v1

the minimum and maximum polynomial distances when translated to French and Spanish, providing a comprehensive view of syntactic variations across languages.

2.8. Syntactic Similarity of Languages

2.8.1. In-Depth Analysis of Language Distance Matrices

The language distance matrix of the ENG dataset offers an insightful visual representation of syntactic similarities across languages. This matrix, along with its multidimensional scaling (MDS) plot and unweighted pair group method with arithmetic mean (UPGMA) dendrogram, provides a clear picture of how languages cluster together based on syntactic similarities. These clusters generally align with the genealogical classification of languages as per historical-comparative research.

2.8.2. Comparative Study of Language Families

Our analysis across the five datasets reveals fascinating insights into the syntactic relationships between language families:

1. **Italic Languages:** Languages like French, Italian, Portuguese, and Spanish exhibit close syntactic relationships. They form a tight cluster in our analysis, with pairwise language distances consistently below the mean value, indicating a high degree of syntactic similarity.
2. **Balto-Slavic Languages:** The group comprising Czech, Polish, and Russian also demonstrates close syntactic ties. The distances between these languages are notably smaller than the average, signifying a strong syntactic affinity within this group.
3. **Germanic Languages:** The analysis of English, German, and Swedish reveals smaller pairwise distances, suggesting a close syntactic relationship. However, Icelandic, despite being a Germanic language, shows larger distances from English and German, indicating a unique syntactic structure.
4. **Distinct Syntax of Japanese and Thai:** Both Japanese and Thai consistently exhibit large average language distances across datasets, marking them as languages with distinct syntactic structures compared to the others in the study.
5. **Chinese Syntax Analysis:** Chinese shows closer syntactic similarity with Korean and Indonesian but still maintains considerable syntactic distinction from most other languages in the study.

2.8.3. Global Syntactic Landscape

This comprehensive study paints a detailed picture of the global syntactic landscape. It highlights not only the diversity and complexity inherent in language syntax but also the intricate patterns of similarity and divergence across different language families. This analysis serves as a crucial step in understanding the rich tapestry of global linguistic structures.

3. Discussions and Conclusions

In this research, we have innovatively adapted the tree distinguishing polynomial for a detailed representation of dependency trees. This approach has enabled us to establish a novel metric for comparing the syntax of sentences by evaluating the distances between their dependency tree polynomials. This polynomial-based methodology offers a more holistic analysis of dependency grammar compared to traditional methods that focus on the order of words [4,8] or compute dependency distances [6,9]. By encompassing all structural elements and dependency relationships, our approach provides a comprehensive view of syntactic structures.

Applying this method, we meticulously analyzed 1,000 sentences from the Parallel Universal Dependency (PUD) treebanks, which include translations of these sentences across various languages. For a more structured analysis, we categorized these sentences into five distinct datasets based on their languages of origin. Our examination involved a detailed comparison of sentences exhibiting the

 doi:10.20944/preprints202312.0335.v1

smallest and largest polynomial distances in their translations between English, Chinese, French, and Spanish. This comparison highlights the effectiveness of polynomial-based methods in evaluating and contrasting sentence syntax.

We extended our analysis to calculate the average pairwise polynomial distance for every pair of languages within each dataset. This data was instrumental in conducting a comprehensive syntactic typology study of the 20 languages present in the PUD treebanks, covering all five datasets. The findings generally align with the genealogical classifications outlined in Glottolog 4.6 [42]. Our analysis sheds light on intriguing syntactic typologies, including lesser-known connections such as the syntactic similarities between Finnish and Korean and the recently proposed Korean-Turkish link discussed in multidisciplinary studies encompassing genetics, archaeology, and linguistics [43].

Furthermore, we leveraged polynomial distances to measure the syntactic diversity within various language corpora. By examining the distribution of pairwise polynomial distances across all sentence pairs within a corpus, we gained insights into the syntactic richness and variety inherent in these languages. The diameters and average pairwise sentence distances emerged as straightforward yet effective indicators of syntactic diversity. This method holds great potential for diverse applications, including assessing language acquisition progress, evaluating the fidelity of texts generated by artificial intelligence, guiding AI systems in creating syntactically diverse content, analyzing distinctive writing styles, and tracking changes in language syntax over time. As the Universal Dependencies framework continues to annotate more sentences and expand the range of Parallel Universal Dependencies treebanks, we anticipate that our polynomial-based approach will unveil deeper insights into linguistic structures, relationships between different languages and corpora, and open new avenues for linguistic research and exploration.

References

1. Imrényi, A.; Mazziotta, N. *Chapters of Dependency Grammar: A Historical Survey from Antiquity to Tesnière*; Amsterdam/Philadelphia: John Benjamins Publishing Company, 2020.
2. de Marneffe, M.C.; Manning, C.D.; Nivre, J.; Zeman, D. Universal Dependencies. *Computational Linguistics* **2021**, *47*, 255–308.
3. Zeman, D.; Popel, M.; Straka, M.; Hajič, J.; Nivre, J.; Ginter, F.; Luotolahti, J.; Pyysalo, S.; Petrov, S.; Potthast, M.; Tyers, F.; Badmaeva, E.; Gokirmak, M.; Nedoluzhko, A.; Cinková, S.; Hajič jr., J.; Hlaváčová, J.; Kettnerová, V.; Urešová, Z.; Kanerva, J.; Ojala, S.; Missilä, A.; Manning, C.D.; Schuster, S.; Reddy, S.; Taji, D.; Habash, N.; Leung, H.; de Marneffe, M.C.; Sanguinetti, M.; Simi, M.; Kanayama, H.; de Paiva, V.; Droganova, K.; Martínez Alonso, H.; Çöltekin, Ç.; Sulubacak, U.; Uszkoreit, H.; Macketanz, V.; Burchardt, A.; Harris, K.; Marheinecke, K.; Rehm, G.; Kayadelen, T.; Attia, M.; Elkahky, A.; Yu, Z.; Pitler, E.; Lertpradit, S.; Mandl, M.; Kirchner, J.; Alcalde, H.F.; Strnadová, J.; Banerjee, E.; Manurung, R.; Stella, A.; Shimada, A.; Kwak, S.; Mendonça, G.; Lando, T.; Nitisaroj, R.; Li, J. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies; Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 1–19.
4. Chen, X.; Gerdes, K. Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks. Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017); Linköping University Electronic Press: Pisa,Italy, 2017; pp. 54–63.
5. Fei, H.; Ren, Y.; Zhang, Y.; Ji, D.; Liang, X. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics* **2021**, *22*.
6. Chen, X.; Gerdes, K. Dependency Distances and Their Frequencies in Indo-European Language. *Journal of Quantitative Linguistics* **2022**, *29*, 106–125.
7. Fei, H.; Ren, Y.; Ji, D. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management* **2020**, *57*, 102311.
8. Gerdes, K.; Kahane, S.; Chen, X. Typometrics: From implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics* **2021**, *6*.

9. Lei, L.; Wen, J. Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses. *Lingua* **2020**, *239*, 102762.
10. Fei, H.; Ren, Y.; Ji, D. Retrofitting Structure-aware Transformer Language Model for End Tasks. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 2151–2161.
11. Jones, V.F.R. A polynomial invariant for knots via von Neumann algebras. *Bulletin of the American Mathematical Society* **1985**, *12*, 103–111.
12. Freyd, P.; Yetter, D.; Hoste, J.; Lickorish, W.B.R.; Millett, K.; Ocneanu, A. A new polynomial invariant of knots and links. *Bulletin of the American Mathematical Society* **1985**, *12*, 239–246.
13. H. Kauffman, L. State models and the jones polynomial. *Topology* **1987**, *26*, 395–407.
14. Fei, H.; Wu, S.; Ren, Y.; Zhang, M. Matching Structure for Dual Learning. Proceedings of the International Conference on Machine Learning, ICML, 2022, pp. 6373–6391.
15. Li, J.; Xu, K.; Li, F.; Fei, H.; Ren, Y.; Ji, D. MRN: A Locally and Globally Mention-Based Reasoning Network for Document-Level Relation Extraction. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1359–1370.
16. Thistlethwaite, M.B. A spanning tree expansion of the jones polynomial. *Topology* **1987**, *26*, 297–309.
17. Diao, Y.; Hetyei, G.; Liu, P. The braid index of reduced alternating links. *Mathematical Proceedings of the Cambridge Philosophical Society* **2020**, *168*, 415–434.
18. Li, J.; Fei, H.; Liu, J.; Wu, S.; Zhang, M.; Teng, C.; Ji, D.; Li, F. Unified Named Entity Recognition as Word-Word Relation Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 10965–10973.
19. Murasugi, K. On the Braid Index of Alternating Links. *Transactions of the American Mathematical Society* **1991**, *326*, 237–260.
20. Fei, H.; Wu, S.; Ren, Y.; Li, F.; Ji, D. Better Combine Them Together! Integrating Syntactic Constituency and Dependency Representations for Semantic Role Labeling. Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, 2021, pp. 549–559.
21. Tutte, W.T. A Contribution to the Theory of Chromatic Polynomials. *Canadian Journal of Mathematics* **1954**, *6*, 80–91.
22. Liu, P. A tree distinguishing polynomial. *Discrete Applied Mathematics* **2021**, *288*, 1–8.
23. Wu, S.; Fei, H.; Li, F.; Zhang, M.; Liu, Y.; Teng, C.; Ji, D. Mastering the Explicit Opinion-Role Interaction: Syntax-Aided Neural Transition System for Unified Opinion Role Labeling. Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, 2022, pp. 11513–11521.
24. Shi, W.; Li, F.; Li, J.; Fei, H.; Ji, D. Effective Token Graph Modeling using a Novel Labeling Strategy for Structured Sentiment Analysis. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4232–4241.
25. Fei, H.; Zhang, Y.; Ren, Y.; Ji, D. Latent Emotion Memory for Multi-Label Emotion Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 7692–7699.
26. Wang, F.; Li, F.; Fei, H.; Li, J.; Wu, S.; Su, F.; Shi, W.; Ji, D.; Cai, B. Entity-centered Cross-document Relation Extraction. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9871–9881.
27. Liu, P.; Biller, P.; Gould, M.; Colijn, C. Analyzing Phylogenetic Trees with a Tree Lattice Coordinate System and a Graph Polynomial. *Systematic Biology* **2022**, *71*, 1378–1390.
28. Janssen, R.; Liu, P. Comparing the topology of phylogenetic network generators. *Journal of bioinformatics and computational biology* **2021**, *19*, 2140012.
29. Pons, J.C.; Coronado, T.M.; Hendriksen, M.; Francis, A. A polynomial invariant for a new class of phylogenetic networks. *PLOS ONE* **2022**, *17*, 1–22.
30. van Iersel, L.; Moulton, V.; Murakami, Y. Polynomial invariants for cactuses. *Preprint* **2022**. doi:10.48550/arxiv.2209.12525.
31. Shang, M.; Li, P.; Fu, Z.; Bing, L.; Zhao, D.; Shi, S.; Yan, R. Semi-supervised Text Style Transfer: Cross Projection in Latent Space. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 4939–4948.

32. Wu, S.; Fei, H.; Ren, Y.; Ji, D.; Li, J. Learn from Syntax: Improving Pair-wise Aspect and Opinion Terms Extraction with Rich Syntactic Knowledge. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 3957–3963.
33. Fei, H.; Li, F.; Li, B.; Ji, D. Encoder-Decoder Based Unified Semantic Role Labeling with Label-Aware Syntax. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 12794–12802.
34. Fei, H.; Wu, S.; Li, J.; Li, B.; Li, F.; Qin, L.; Zhang, M.; Zhang, M.; Chua, T.S. LasUIE: Unifying Information Extraction with Latent Adaptive Structure-aware Generative Language Model. Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022, 2022, pp. 15460–15475.
35. Wu, S.; Fei, H.; Ji, W.; Chua, T.S. Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 2593–2608.
36. Culotta, A.; Sorensen, J. Dependency Tree Kernels for Relation Extraction. Proceedings of the 42nd annual meeting on association for computational linguistics, 2004, p. 423.
37. Luo, Q.; Xi, J. A novel similarity measure for dependency trees [query answer system example]. Proceedings. 2005 International Conference on Communications, Circuits and Systems, 2005, p. 785.
38. Reis, D.C.; Golgher, P.B.; Silva, A.S.; Laender, A.F. Automatic Web News Extraction Using Tree Edit Distance. Proceedings of the 13th International Conference on World Wide Web; Association for Computing Machinery: New York, NY, USA, 2004; WWW '04, p. 502–511.
39. Cox, T.F.; Cox, M.A. *Multidimensional scaling*, 2nd ed.; Monographs on statistics and applied probability; 88, Chapman & Hall, 2001.
40. Sokal, R.R.; Michener, C.D. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin* **1958**, *38*, 1409–1438.
41. Bryant, D.; Tupper, P.F. Hyperconvexity and tight-span theory for diversities. *Advances in Mathematics* **2012**, *231*, 3172–3198.
42. Forkel, R.; Hammarström, H. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web* **2022**, *13*, 917–924.
43. Robbeets, M.; Bouckaert, R.; Conte, M.; Savelyev, A.; Li, T.; An, D.I.; Shinoda, K.i.; Cui, Y.; Kawashima, T.; Kim, G.; Uchiyama, J.; Dolińska, J.; Oskolskaya, S.; Yamano, K.Y.; Seguchi, N.; Tomita, H.; Takamiya, H.; Kanzawa-Kiriyama, H.; Oota, H.; Ishida, H.; Kimura, R.; Sato, T.; Kim, J.H.; Deng, B.; Bjørn, R.; Rhee, S.; Ahn, K.D.; Gruntov, I.; Mazo, O.; Bentley, J.R.; Fernandes, R.; Roberts, P.; Bausch, I.R.; Gilaizeau, L.; Yoneda, M.; Kugai, M.; Bianco, R.A.; Zhang, F.; Himmel, M.; Hudson, M.J.; Ning, C. Triangulation supports agricultural spread of the Transeurasian languages. *Nature* **2021**, *599*, 616–621.