Quantifying syntax similarity with a polynomial representation of dependency trees

Pengyu Liu[1,2], Tinghao Feng[3], Rui Liu[4,5*]

[1] Department of Microbiology and Molecular Genetics, University of California, Davis
[2] Department of Mathematics, Simon Fraser University
[3] Department of Computer Science, Appalachian State University
[4] Department of Chinese Language and Literature, Beijing Normal University, Zhuhai
[5] Center for Linguistic Sciences, Beijing Normal University, Zhuhai
[*] Corresponding author's email: liu_rui@bnu.edu.cn

DOI: https://doi.org/10.53482/2022_53_402

ABSTRACT

We introduce a graph polynomial that distinguishes tree structures to represent dependency grammar and a measure based on the polynomial representation to quantify syntax similarity. The polynomial encodes accurate and comprehensive information about the dependency structure and dependency relations of words in a sentence, which enables in-depth analysis of dependency trees with data analysis tools. We apply the polynomial-based methods to analyze sentences in the Parallel Universal Dependencies treebanks. Specifically, we compare the syntax of sentences and their translations in different languages, and we perform a syntactic typology study of available languages in the Parallel Universal Dependencies treebanks. We also demonstrate and discuss the potential of the methods in measuring syntax diversity of corpora.

Keywords: dependency tree, polynomial representation, syntax similarity, Parallel Universal Dependencies (PUD), syntactic typology

1 Introduction

Dependency grammar is an important framework for syntactic analysis (Imrényi and Mazziotta, 2020). Dependency focuses on the proximity of words in a sentence, and the hierarchical relations between words in the sentence are represented by a tree structure called the dependency tree of the sentence. Recently, an international collaboration project called Universal Dependency (UD) has created a standard annotation scheme for constructing dependency trees from sentences, and hundreds of UD treebanks of various languages have been made publicly available (de Marneffe et al., 2021). These datasets form key materials for syntax analysis, providing new opportunities for automated text processing and syntactic typology studies to name a few. Parallel Universal Dependency (PUD) treebanks are a class of UD treebanks consisting of dependency trees of 1,000 sentences and their translations to other languages (Zeman et al., 2017). The 1,000 sentences are randomly selected from the news domain and Wikipedia

and are originally written in English, French, German, Italian or Spanish. At the time of writing, there are 20 PUD treebanks containing the dependency trees of the 1,000 sentences in 20 languages respectively. These UD treebanks have stimulated novel computational methods for syntax analysis and the development of quantitative measures for syntax similarity (H. Liu and Xu, 2012; Vulić et al., 2020; Wong et al., 2017). However, current methods describing dependency trees mainly focus on partial syntactic information recorded in the structures such as the order of words and the dependency distance (Chen and Gerdes, 2017, 2022; Gerdes et al., 2021; Lei and Wen, 2020). In this work, we introduce a comprehensive representation of dependency trees based on a tree distinguishing polynomial. The polynomial takes into account all syntactic information recorded in a dependency tree, and two sentences have the same dependency structure if and only if the polynomials of their dependency trees are identical.

Structural polynomials are well studied objects in mathematical areas such as knot theory and graph theory, and they have natural applications in characterizing topological and discrete structures. In the theory of knots and links, Jones polynomial (Jones, 1985) and HOMFLY polynomial (Freyd et al., 1985) have been used to characterize properties of knots and links such as crossing number (Kauffman, 1987; Thistlethwaite, 1987) and braid index (Diao et al., 2020; Murasugi, 1991). In the study of graphs, the Tutte polynomial (Tutte, 1954) contains the information about graphs including the number of spanning trees of the graph and the number of graph colorings. Recently, a structural polynomial that distinguishes unlabeled trees has been defined and studied in (P. Liu, 2021). This builds an one-to-one correspondence between unlabeled trees and a class of bivariate polynomials, that is, two unlabeled trees are isomorphic if and only if they have the same polynomial. This tree distinguishing polynomial has been applied to study phylogenetic trees and pathogen evolution (P. Liu et al., 2022) and generalized to represent some classes of phylogenetic networks (Janssen and Liu, 2021; Pons et al., 2022; van Iersel et al., 2022). It has been shown that the polynomial-based methods for tree comparison have better accuracy and computational efficiency, when compared to other tree comparison and representation methods such as sequence-based representations, Laplacian spectrum of trees and summary statistics (P. Liu et al., 2022). Current methods to compare dependency trees are mainly based on summary statistics including tree kernels and their generalizations (Culotta and Sorensen, 2004; Luo and Xi, 2005) or tree edit distances (Reis et al., 2004) which only take into account local structures rather than the global structure of trees. Here, we generalize the tree distinguishing polynomial for representing dependency trees and define a distance between the polynomials to measure syntax similarity. We apply the polynomial-based methods to the dependency trees in the PUD treebanks, and we compare the syntax of sentences with small and large distances. We also perform a syntactic typology study for currently available languages in the PUD treebanks. Furthermore, we show that the pairwise distances between sentences can be used to measure syntax diversity of a corpus and discuss its potential applications.

2 Material

2.1 Dependency trees

A *dependency tree* of a sentence is a rooted node-labeled tree representing grammatical relations between words in the sentence. Each node in a dependency tree corresponds to a word in the sentence. An edge in a dependency tree connects two nodes and represents a grammatical connection between the two corresponding words: The node closer to the root is the *head* of the edge and the other node is a *dependent* of the head. A head can have multiple dependents, while every dependent has only one head. The label of a dependent indicates the grammatical relation to its head. In a dependency tree of a sentence, the root node representing the head of the entire sentence is not a dependent, so its label only shows that it is the root. Furthermore, a sentence can contain words with the same grammatical relation, so dependents in a dependency tree can have identical labels. In Figure 1, we display the dependency tree of an English sentence and the dependency tree of a Chinese translation of the sentence. In these examples of dependency trees, the numbers in parentheses after each word are the node labels representing head-dependent grammatical relations listed in Table 1. All dependency trees used in the paper are constructed by crosslinguistically consistent morphosyntactic annotation under the Universal Dependencies (UD) framework (de Marneffe et al., 2021).

Table 1: The indices of head-dependent relations of the Universal Dependencies (UD) framework.

Index	Relation	Index	Relation
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		

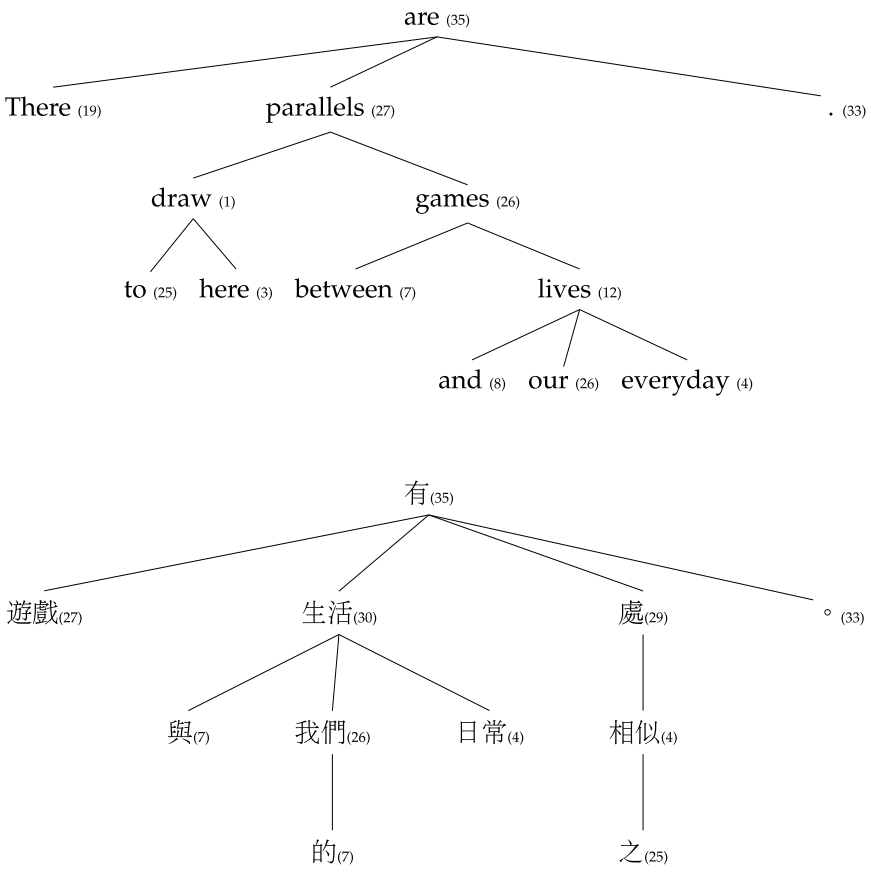

Figure 1: Examples of dependency trees. Top: the dependency tree of an English sentence: "There are parallels to draw here between games and our everyday lives." Bottom: the dependency tree of a Chinese translation of the sentence. The numbers in parentheses after each word are labels representing head-dependent relations listed in Table 1.

2.2 Parallel Universal Dependencies

We analyze dependency trees in the Parallel Universal Dependencies (PUD) treebanks (version 2.10), which were created in a shared task of the Conference on Computational Natural Language Learning (CoNLL 2017) (Zeman et al., 2017). To construct the PUD treebanks, 1,000 sentences were randomly selected from online news or Wikipedia articles, and there were 750 of the sentences originally in English, 100 in German, 50 in French, 50 in Italian and 50 in Spanish. Then, the 1,000 sentences were translated by professional translators to other languages. A PUD treebank contains 1,000 dependency trees of the translated or original sentences in a language. Currently, there are 20 PUD treebanks available, containing dependency trees of the 1,000 translated or original sentences in 20 languages including Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai and Turkish.

2.3 Datasets

We divide the 1,000 sentences of the PUD treebanks into 5 datasets based on their original languages and name the 5 datasets using the capital ISO 639-2/B codes of the sentences' original languages. Throughout the paper, the capital ISO 639-2/B codes of languages only refers to the 5 datasets, and results based on different datasets are visualized in different colors. Table 2 shows the corresponding color of each dataset and the number of sentences in each dataset. Note that every sentence has 20 dependency trees corresponding to its translations in the 20 languages, so the number of dependency trees in each dataset is the number of sentences multiplied by 20 which is also displayed in Table 2.

Table 2: Information of datasets constructed from the Parallel Universal Dependencies (PUD) treebanks.

Dataset	Original language	Number of sentences	Number of dependency trees	Color
ENG	English	750	15000	Blue
GER	German	100	2000	Yellow
FRE	French	50	1000	Purple
ITA	Italian	50	1000	Green
SPA	Spanish	50	1000	Red

3 Methodology

3.1 Tree distinguishing polynomial

We review the graph polynomial that distinguishes unlabeled trees introduced in (P. Liu, 2021). Every rooted unlabeled tree T corresponds to a unique bivariate polynomial $P(T, x, y)$. To compute the polynomial $P(T, x, y)$ for the unlabeled tree T, we recursively assign a polynomial to each node in T from the leaf nodes to the root, and the polynomial at the root is $P(T, x, y)$. Let $P(n, x, y)$ denote the polynomial at node n. If node n is a leaf node, then we assign the polynomial $P(n, x, y) = x$ to node n. Let node m be an internal (non-leaf) node with k child nodes $n_1, n_2, \ldots, n_k$. The polynomial at node m is $P(m, x, y) = y + \Pi_{i=1}^{k} P(n_i, x, y)$. We say that the *topology* of a dependency tree is the tree structure without any labels. In Figure 2, we show the recursive process for computing the polynomials representing the topologies of the dependency trees displayed in Figure 1. It is proved that two unlabeled trees are isomorphic if and only if they have the same polynomial. Furthermore, each term in the polynomial of an unlabeled tree is interpretable and corresponds to a specific subtree of the unlabeled tree. See (P. Liu, 2021) for more details about the tree distinguishing polynomial, and see (P. Liu et al., 2022) for distances and methods based on the polynomial to analyze tree structures.

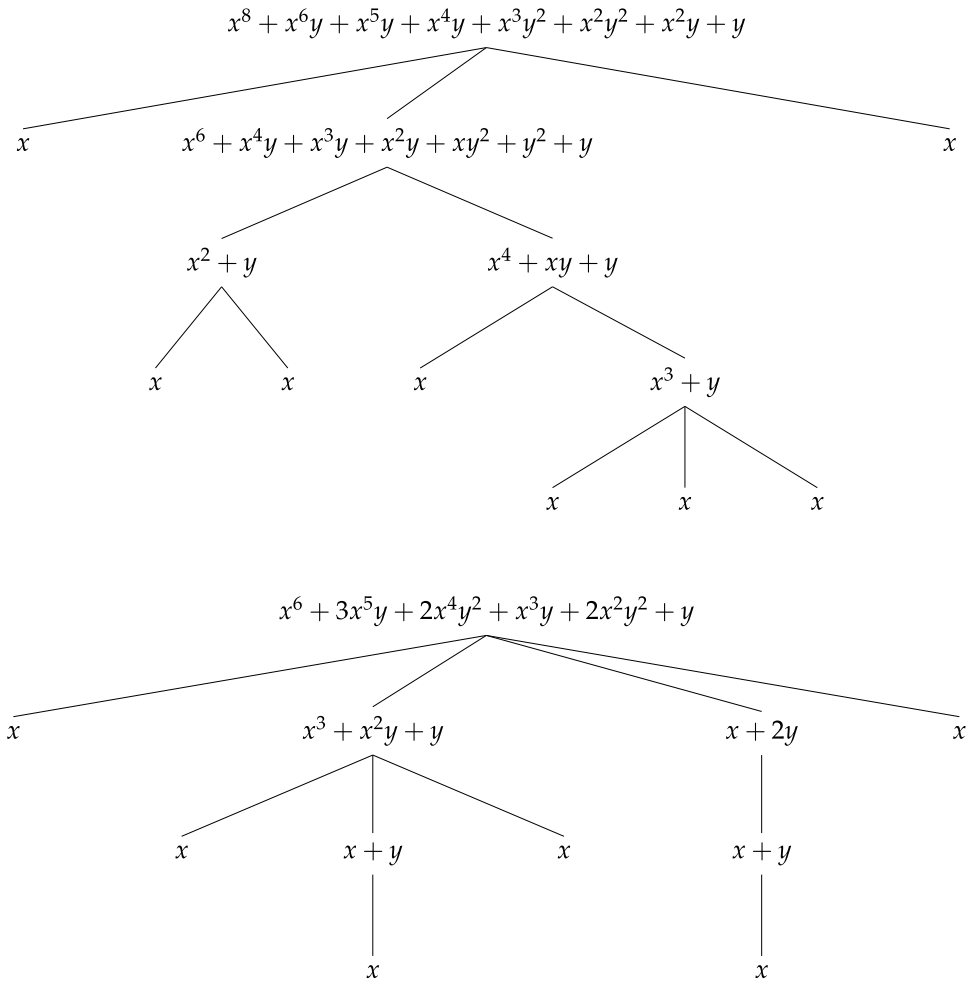

Figure 2: Examples of polynomials of unlabeled trees. The recursive process of computing the polynomials of topologies of dependency trees displayed in Figure 1 from the leaf nodes to the roots. The polynomials at the roots represent the two unlabeled trees.

3.2 Polynomial of dependency trees

Here, we generalize the tree distinguishing polynomial for representing dependency trees. Compared with tree topologies, dependency trees have node labels. In this study, there are 37 labels representing head-dependent relations listed in Table 1. These labels may appear in both leaf nodes and internal nodes of dependency trees. So, we represent dependency trees using a generalized tree distinguishing polynomial with 74 variables classified into two sets: $X = \{x_1, x_2, \ldots, x_{37}\}$ and $Y = \{y_1, y_2, \ldots, y_{37}\}$. We denote the generalized polynomial for a dependency tree T by $P(T, X, Y)$. Similarly, we compute the polynomial $P(T, X, Y)$ recursively from the leaf nodes to the root for the dependency tree T. Suppose that node n^ℓ is a leaf node with label ℓ, then we assign the polynomial $P(n^\ell, X, Y) = x_\ell$ to the leaf node. Let node m^ℓ be an internal node with label ℓ which has k child nodes $n_1, n_2, \ldots, n_k$, then the polynomial at node m^ℓ is $P(m^\ell, x, y) = y_\ell + \Pi_{i=1}^{k} P(n_i, x, y)$. Figure 3 shows the process of recursively

computing the generalized polynomials representing the two dependency trees displayed in Figure 1. Since this is a generalization of the polynomial that distinguishes unlabeled trees, two dependency trees have the same generalized polynomial if and only if they are isomorphic and corresponding nodes have the same labels. Therefore, two sentences have exactly the same dependency structure if and only if the generalized polynomials of the dependency trees of the sentences are identical. For simplicity, we call the generalized polynomial of the dependency tree of a sentence the *dependency tree polynomial* of the sentence.

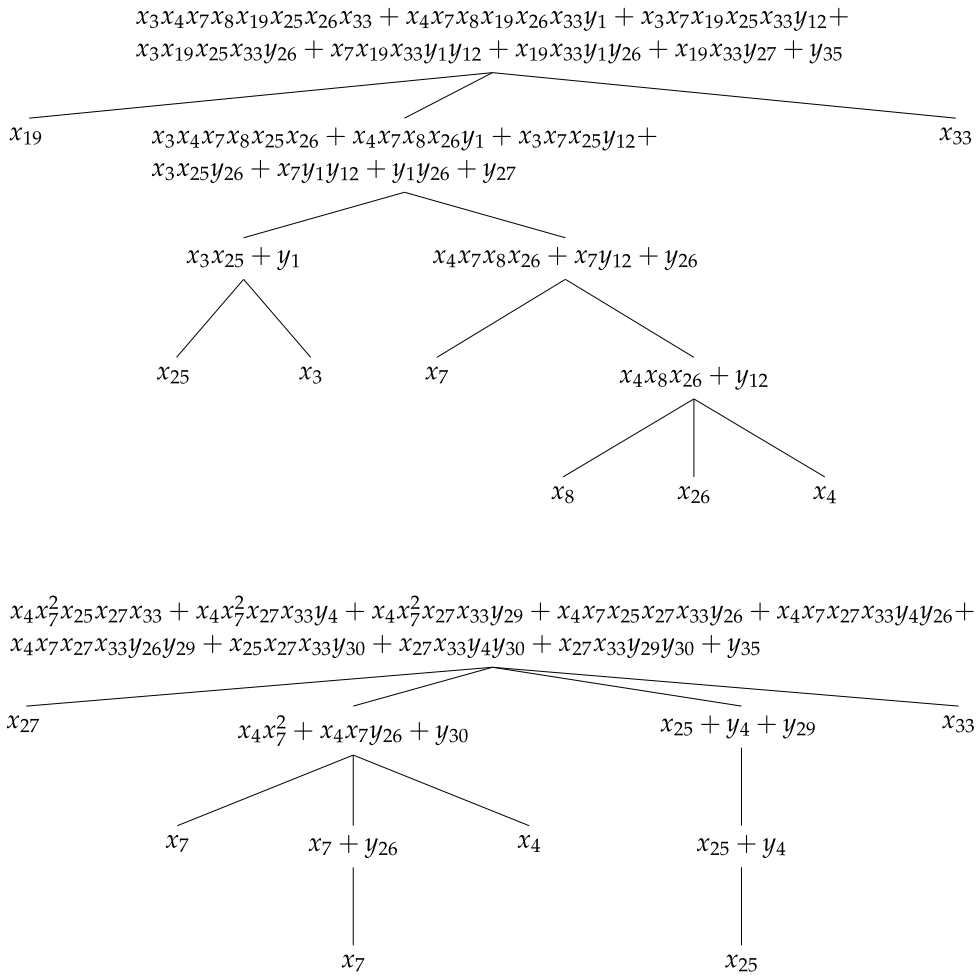

Figure 3: Polynomials of the dependency trees. The recursive process of computing the polynomials representing the dependency trees displayed in Figure 1 from the leaf nodes to the roots. The polynomials at the roots represent the two dependency trees.

3.3 Polynomial distance of dependency trees

In the polynomial representing an unlabeled tree, the information about the hierarchical structure is encoded in the coefficient and exponents of each term; see Figure 2. In the polynomial representing a dependency tree, the syntactic information is encoded mainly in the exponents of each term due to the introduction of additional variables; see Figure 3. We develop a new measure to compare dependency tree polynomials, hence the dependency trees. The polynomial $P(T, X, Y)$ representing a dependency tree T can be described term by term. We write each term of the polynomial as a vector with 75 entries $t = [e_{x_1}, e_{x_2}, \ldots, e_{x_{37}}, e_{y_1}, e_{y_2}, \ldots, e_{y_{37}}, c]$, where the exponent of variable x_i is e_{x_i}, the exponent of variable y_i is e_{y_i} and the coefficient of the term is c. We call such a vector a *term vector* of the polynomial $P(T, X, Y)$. Let P and Q be two dependency tree polynomials and $\mathcal{V}_P$ and $\mathcal{V}_Q$ be the corresponding sets of term vectors of P and Q. We denote the number of term vectors in $\mathcal{V}_P$ (or $\mathcal{V}_Q$) by $|\mathcal{V}_P|$ (or $|\mathcal{V}_Q|$). Let s and t be two term vectors. We denote the Manhattan distance (Craw, 2010) between s and t by $\|s - t\|_1$ and define the *polynomial distance* for the pair of dependency tree polynomials P and Q using Formula (1).

$$d(P, Q) = \frac{\sum_{s \in \mathcal{V}_P} \min_{t \in \mathcal{V}_Q} \|s - t\|_1 + \sum_{t \in \mathcal{V}_Q} \min_{s \in \mathcal{V}_P} \|s - t\|_1}{|\mathcal{V}_P| + |\mathcal{V}_Q|} \tag{1}$$

Since polynomials and dependency trees are in one-to-one correspondence, the defined distance for dependency tree polynomials is also for dependency trees. Without ambiguity, the polynomial distance between dependency trees refers to the distance between dependency tree polynomials throughout the paper. Furthermore, each sentence in the PUD treebanks also has a unique dependency tree constructed under the UD framework, so, without ambiguity, the polynomial distance between sentences refers to the distance between their dependency tree polynomials.

3.4 Experiments

Note that every sentence in the 5 datasets is written in 20 languages, and 20 dependency trees are constructed for each sentence based on the original sentence and its 19 translations. So, for each of the 5 datasets, a dependency tree can be identified by the original sentence and the language to which the original sentence is translated. For each dataset, we compute the polynomials of all the dependency trees, and calculate the pairwise polynomial distances between the 20 dependency trees for every sentence. We analyze the syntax of the sentences whose polynomial distances between a pair of translations are the smallest and the largest. For each sentence, the pairwise distances between the 20 dependency trees form a 20×20 distance matrix, which we call the *translation distance matrix* of the sentence. We say that the

distance stored in each entry of the translation distance matrix of a sentence is the *translation distance* of the sentence between the corresponding languages of the entry. We take the mean value of each entry in the translation distance matrices over all sentences in a dataset and call the resulting matrix the *language distance matrix* of the dataset. We say that an entry in the language distance matrix of a dataset is the *pairwise language distance* between the corresponding pair of languages in the dataset. The numeric value at each entry of the language distance matrix of a dataset indicates syntax similarity of a pair of languages based on the sentences in the dataset. We summarize the language distance matrices of the 5 datasets by showing the mean and median of all pairwise language distances and pairs of the nearest and farthest languages in the pairwise language distance. We also take the mean value of the pairwise language distances between a language and other 19 languages and call the mean value the *average language distance* of the language. We show the languages with smallest and largest average language distances in the 5 datasets. We use the language distance matrices of the 5 datasets to perform a syntactic typology study of the 20 available languages in the PUD treebanks. We visualize the language distance matrices using multidimensional scaling (MDS) (Cox and Cox, 2001), and we construct dendrograms by applying the unweighted pair group method with arithmetic mean (UPGMA) method to the language distance matrices (Sokal and Michener, 1958). These visualizations provide different perspectives for analyzing syntax similarity of languages based on the sentences in the PUD treebanks. Lastly, for each dataset, we consider the translations of all sentences in a language as a corpus of the language, and there are 20 corpora for each dataset. We calculate all pairwise distances between translated sentences in each of the 20 corpora, and we call such a pairwise distance a *pairwise sentence distance* in the corpus. We show the distribution of pairwise sentence distances for each corpus, and we call the maximum pairwise sentence distance in a corpus the *diameter* of the corpus. The diameter is a simple measure of diversity (Bryant and Tupper, 2012), and we discuss the potential of the polynomial methods in measuring syntax diversity.

4 Results

4.1 Syntax comparison of sentences

The newly defined distance of dependency tree polynomials provides a quantitative measure of sentences' syntax similarity. If two sentences have isomorphic dependency structures with identical labels for corresponding words, then the distance between the dependency tree polynomials of the sentences is zero. A smaller distance between a pair of sentences suggests that they are similar in syntax, and a larger distance between a pair of sentences suggests the syntax being more different. The distance between the dependency trees in Figure 1 is 5.06. See Supplementary Figure 1 and 3 for pairs of dependency trees with polynomial distance zero.

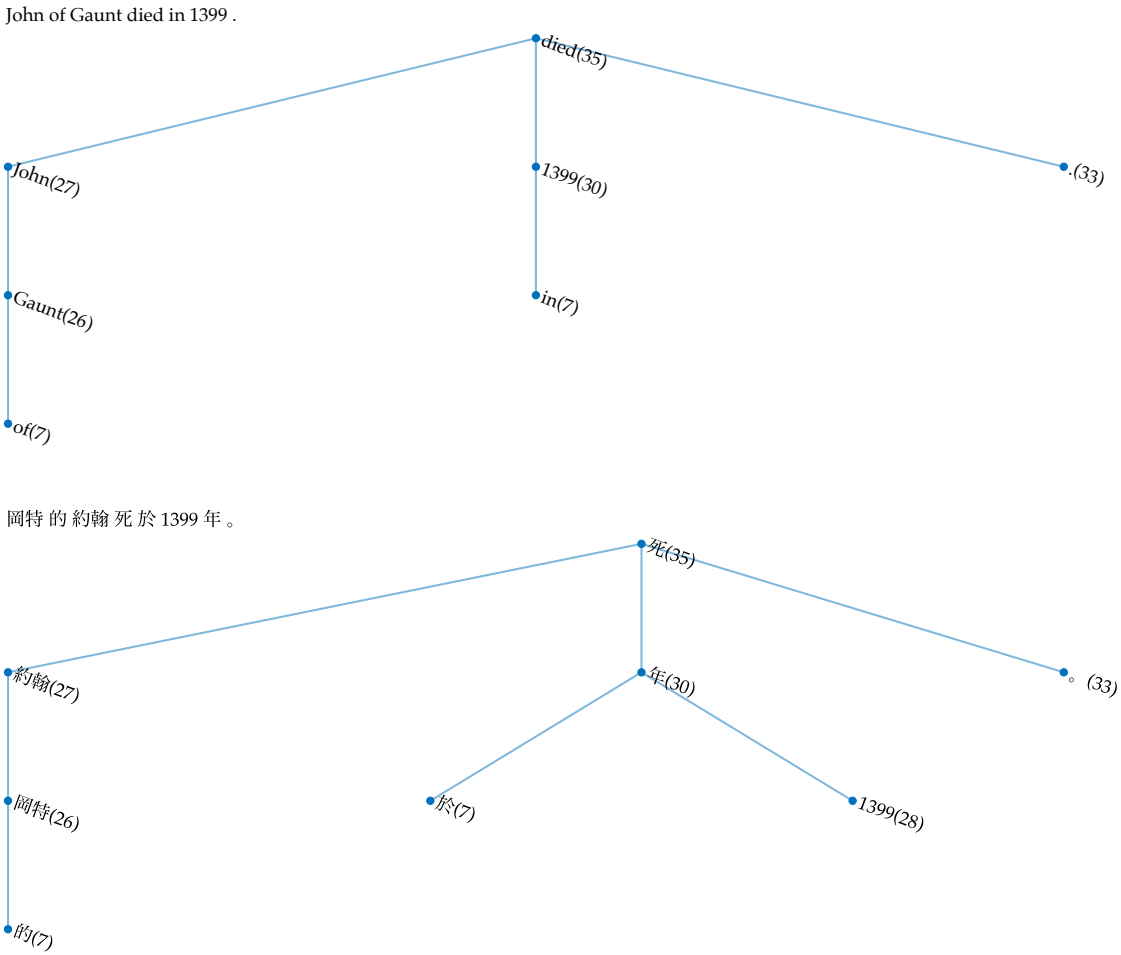

Figure 4: The sentence in the ENG dataset with minimum polynomial distance between its English and Chinese translations. Top: the dependency tree of the sentence's English translation (the original sentence since it is in the ENG dataset). Bottom: the dependency tree of the sentence's Chinese translation. The polynomial distance between the dependency trees is 0.43.

In Figure 4, we display the dependency tree of an English sentence in the ENG dataset and the dependency tree of its Chinese translation in the dataset. The sentence's English and Chinese translations have a polynomial distance 0.43, which is the minimum distance over all sentences in the ENG dataset when comparing the distance between their English and Chinese translations. The English sentence and the Chinese translation are syntactically similar. The stems of both sentences are in subject-predicate form, and the subjects of both sentences are complex noun phrases. The only difference between the sentences is at time adverbials, where the Chinese sentence has a word after "1399" to indicate that the numeral represents a year.

In Figure 5, we display the dependency tree of an English sentence in the ENG dataset and the dependency

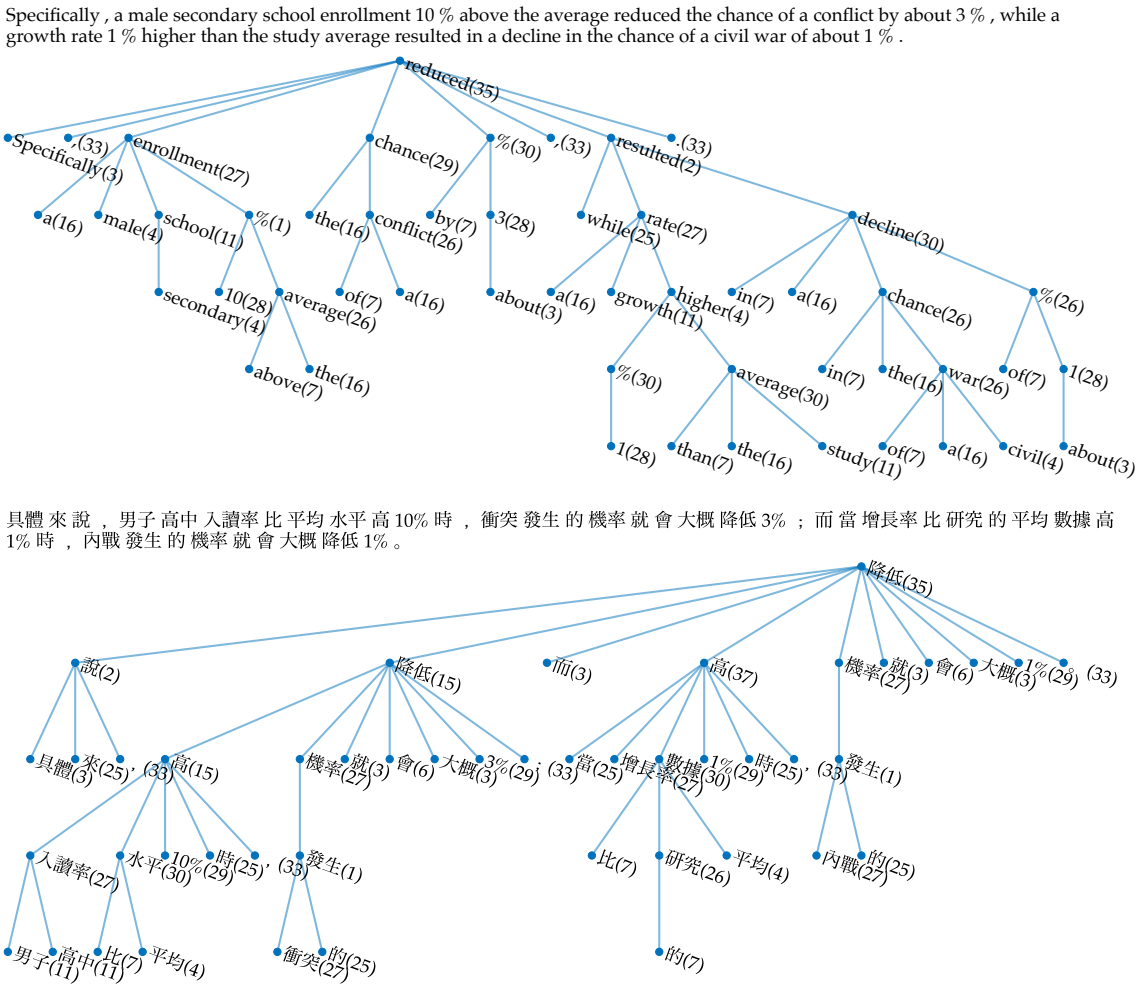

Figure 5: The sentence in the ENG dataset with maximum polynomial distance between its English and Chinese translations. Top: the dependency tree of the sentence's English translation (the original sentence since it is in the ENG dataset). Bottom: the dependency tree of the sentence's Chinese translation. The polynomial distance between the dependency trees is 22.93.

tree of its Chinese translation in the dataset. The sentence's English and Chinese translations have a polynomial distance 22.93, which is the maximum distance over all sentences in the ENG dataset when comparing the distance between their English and Chinese translations. The English sentence and the Chinese translation have more distinct syntax from branches to the stem. The dependency tree of the English sentence is right-branching, that is, there are more modifiers to the right of the root; while the dependency tree of the Chinese translation is left-branching. This difference between English and Chinese is observed in other long sentences in the ENG dataset. In terms of sentence stems, the English sentence has a double-object structure, with "chance" and "'3%" as its objects; the Chinese translation has a single-object structure, with only "1%" as its object. It is worth noting that the UD framework annotates percentages in Chinese and English differently. In Chinese, the the numeral "1" and the

symbol "%" are considered as one word which serves as the object (29) of the sentence; in English, the the numeral "3" and the symbol "%" are treated as two separated words, comprising a numeric modifier (28) and a oblique nominal (30) of the sentence respectively. Furthermore, the Chinese sentence has more adverbials (3) and auxiliaries (6) directly modifying the root of the sentence. These modifiers function in Chinese to make sentences lucid and coherent, but they are not necessary in English. In terms of branches, the complex noun phrase "a male secondary school enrollment 10% above the average" in the English sentence is expressed with a "*bi*"-structure in the Chinese translation, which can be directly translated back to English as "When a male secondary school enrollment is 10% higher than the average". The "*bi*"-structure used to compare the "enrollment" and "the average" form a subject-verb-object clause, which is disparate from the complex noun phrase. In the original English sentence, the word "above" is a preposition bearing a case relationship (7), while its corresponding part in the Chinese translation is the root of the clause.

In general, shorter sentences have fewer options for syntax variation, hence the polynomial distances between shorter sentences are more likely to be small. In contrast, longer sentences have more room for different syntax, so the maximum polynomial distance is more likely between longer sentences. We also display sentences in the ENG dataset with minimum and maximum polynomial distances from the original sentences to their French and Spanish translations. See Supplementary Figure 1-4.

4.2 Syntactic similarity of languages

We measure the syntax similarity between the 20 languages by applying the dependency tree polynomial and the polynomial distance to the 5 datasets. All similarity and closeness are based on the current PUD treebanks and limited to the available languages.

Languages that are related in the genealogical classification of languages based on available historical-comparative research (Glottolog 4.6) (Forkel and Hammarström, 2022) are also clustered with the dependency tree polynomial and polynomial distance; see Figure 6 and Figure 7. Romance languages (French, Italian, Portuguese and Spanish) and Balto-Slavic languages (Czech, Polish and Russian) are close to each other in pairwise language distance (Figure 6). This is also observed in both the MDS plot and the UPGMA dendrogram (Figure 7). The mean pairwise language distance in the ENG dataset is 7.73 (Figure 8). We use mean pairwise language distances as references for syntax similarity between languages: Languages with smaller pairwise language distances are considered similar in syntax, and languages with larger pairwise language distances are considered distinct in syntax. The pairwise language distances between Romance languages are from 4.35 to 5.80, and those between Balto-Slavic

	English	German	Swedish	Icelandic	Italian	French	Portuguese	Spanish	Czech	Polish	Russian	Hindi	Arabic	Chinese	Finnish	Indonesian	Japanese	Korean	Thai	Turkish
English		5.06	4.28	7.64	5.63	6.20	5.08	5.72	6.10	6.08	6.02	7.41	6.69	7.90	6.26	5.24	11.63	7.21	8.72	6.80
German	5.06		5.61	8.15	6.25	6.50	5.91	6.28	6.36	6.51	6.46	8.50	7.42	8.39	6.72	6.36	12.33	7.70	9.23	7.20
Swedish	4.28	5.61		6.95	6.74	7.16	6.33	6.70	5.69	5.79	5.79	7.84	6.37	7.93	5.66	5.19	12.26	6.78	8.44	6.32
Icelandic	7.64	8.15	6.95		9.11	9.24	8.69	8.88	7.32	7.31	7.41	9.37	7.80	8.91	7.10	7.24	12.40	7.65	9.44	7.58
Italian	5.63	6.25	6.74	9.11		5.49	4.61	5.18	7.14	7.08	7.24	8.51	7.76	9.33	8.05	7.01	11.72	8.70	9.67	7.82
French	6.20	6.50	7.16	9.24	5.49		5.16	5.80	7.78	7.64	7.50	8.67	8.18	9.44	8.55	7.52	11.92	9.11	9.83	8.41
Portuguese	5.08	5.91	6.33	8.69	4.61	5.16		4.35	6.99	6.88	6.80	8.11	7.27	9.08	7.84	6.69	11.69	8.42	9.42	7.63
Spanish	5.72	6.28	6.70	8.88	5.18	5.80	4.35		7.28	7.03	7.07	8.21	7.27	9.12	8.03	7.04	11.55	8.39	9.37	7.62
Czech	6.10	6.36	5.69	7.32	7.14	7.78	6.99	7.28		4.60	5.10	8.41	6.29	8.01	5.49	5.84	11.96	6.88	8.92	6.18
Polish	6.08	6.51	5.79	7.31	7.08	7.64	6.88	7.03	4.60		4.79	8.21	6.07	8.08	5.58	5.66	11.75	6.92	8.80	6.13
Russian	6.02	6.46	5.79	7.41	7.24	7.50	6.80	7.07	5.10	4.79		8.19	6.02	8.02	5.77	5.65	11.74	7.17	8.89	6.40
Hindi	7.41	8.50	7.84	9.37	8.51	8.67	8.11	8.21	8.41	8.21	8.19		7.98	9.33	8.85	8.19	10.38	8.96	9.42	8.59
Arabic	6.69	7.42	6.37	7.80	7.76	8.18	7.27	7.27	6.29	6.07	6.02	7.98		8.44	6.92	6.20	11.19	7.58	8.93	6.69
Chinese	7.90	8.39	7.93	8.91	9.33	9.44	9.08	9.12	8.01	8.08	8.02	9.33	8.44		7.83	7.69	11.36	7.21	9.13	8.12
Finnish	6.26	6.72	5.66	7.10	8.05	8.55	7.84	8.03	5.49	5.58	5.77	8.85	6.92	7.83		5.59	12.65	6.13	8.89	5.64
Indonesian	5.24	6.36	5.19	7.24	7.01	7.52	6.69	7.04	5.84	5.66	5.65	8.19	6.20	7.69	5.59		11.77	6.75	8.39	6.03
Japanese	11.63	12.33	12.26	12.40	11.72	11.92	11.69	11.55	11.96	11.75	11.74	10.38	11.19	11.36	12.65	11.77		11.52	12.07	11.90
Korean	7.21	7.70	6.78	7.65	8.70	9.11	8.42	8.39	6.88	6.92	7.17	8.96	7.58	7.21	6.13	6.75	11.52		8.45	6.25
Thai	8.72	9.23	8.44	9.44	9.67	9.83	9.42	9.37	8.92	8.80	8.89	9.42	8.93	9.13	8.89	8.39	12.07	8.45		8.82
Turkish	6.80	7.20	6.32	7.58	7.82	8.41	7.63	7.62	6.18	6.13	6.40	8.59	6.69	8.12	5.64	6.03	11.90	6.25	8.82	

Figure 6: The language distance matrix of the ENG dataset. The languages are ordered based on Glottolog 4.6 classification: Indo-European languages are listed first and grouped according to their subclasses (Germanic, Romance, Balto-Slavic and Indo-Iranian), and other languages are following in the alphabetical order.

languages are from 4.60 to 5.10, all smaller than the mean value 7.73. Germanic languages (English, German, Swedish and Icelandic) are not as clustered as Romance languages or Balto-Slavic languages. The pairwise language distances between English, German and Swedish are from 4.28 to 5.61, smaller than the mean value 7.73, but the pairwise language distances from Icelandic to English and German are 7.63 and 8.15, close to or larger than the mean value. This is in part because Icelandic is one of only two Germanic languages that preserve noun declension, and this makes sentences in Icelandic have fewer prepositions as grammatical cases are expressed by suffixes. Since prepositions are represented as nodes in dependency trees, the dependency trees of Icelandic sentences can be different from other Germanic languages.

Geographic locations play an important role in shaping syntax similarity of languages. Among the 20 available languages, the pairs of languages that are most similar in terms of syntax include Portuguese and Spanish as well as Czech and Polish (Figure 8). The countries and regions with populations speaking each pair of languages are closely located. The pairs of languages that are most distinct in terms of syntax include Japanese and Finnish as well as Japanese and Icelandic (Figure 8), and the countries and

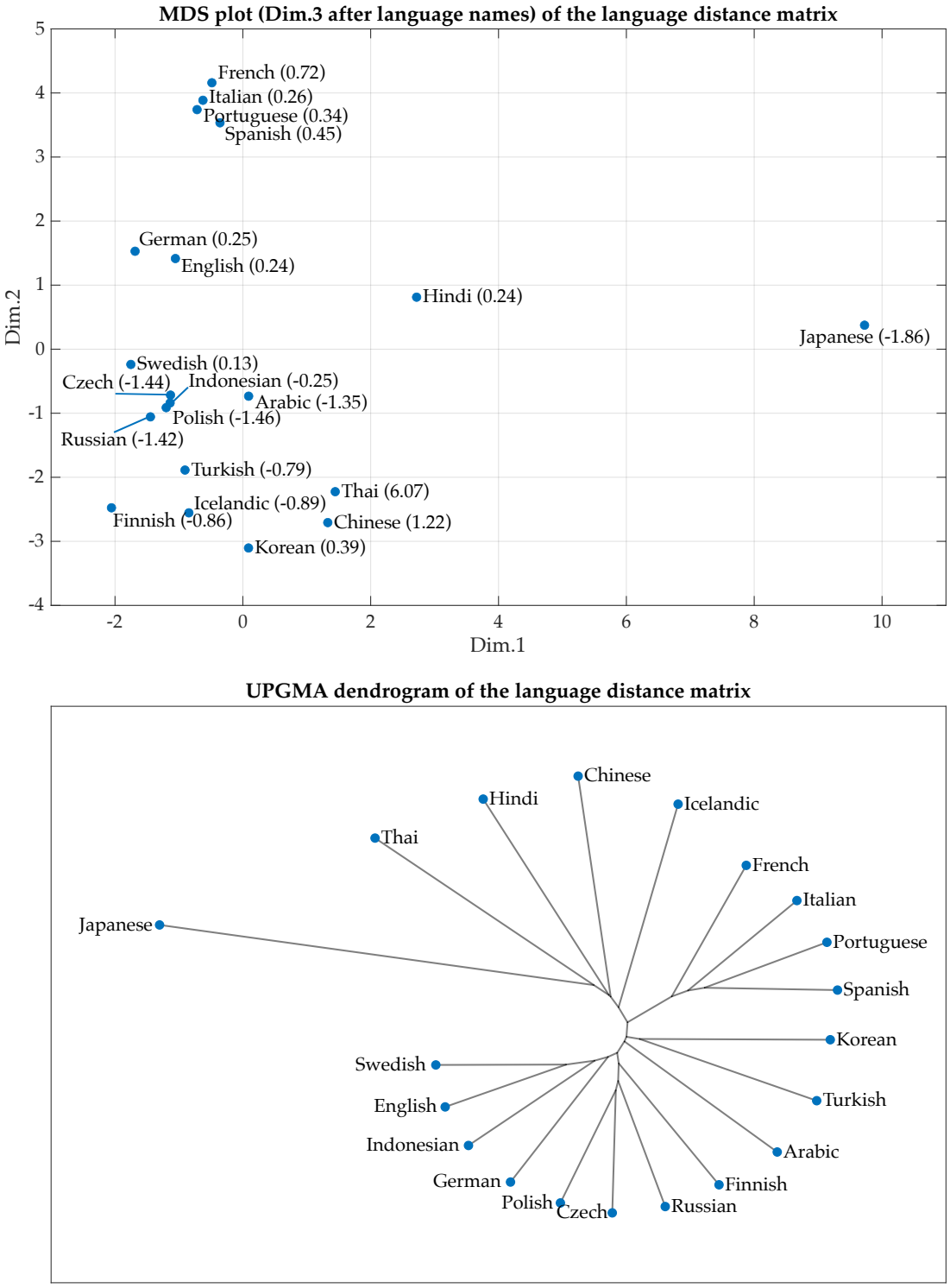

Figure 7: Visualizations the language distance matrix of the ENG dataset. Top: the multidimensional scaling plot of the language distance matrix of the ENG dataset. Bottom: the UPGMA dendrogram constructed based on the language distance matrix of the ENG dataset.

	ENG	GER	FRE	ITA	SPA	
Mean	7.73	6.55	7.40	7.71	7.96	Pairwise language distance
Median	7.55	6.56	7.38	7.39	7.82	
Nearest	4.28 (eng vs swe)	3.24 (por vs spa)	3.60 (por vs spa)	3.99 (por vs spa)	4.13 (ita vs por)	
2nd nearest	4.35 (por vs spa)	3.44 (ita vs spa)	4.14 (pol vs rus)	4.21 (cze vs pol)	4.16 (por vs spa)	
3rd nearest	4.60 (cze vs pol)	3.46 (cze vs pol)	4.39 (ita vs por)	4.39 (pol vs rus)	4.34 (fre vs ita)	
3rd farthest	12.33 (ger vs jpn)	9.80 (ger vs jpn)	11.63 (ger vs jpn)	12.06 (ger vs jpn)	12.51 (ger vs jpn)	
2nd farthest	12.40 (ice vs jpn)	10.08 (fin vs jpn)	11.84 (ice vs jpn)	12.43 (ice vs jpn)	12.78 (ice vs jpn)	
farthest	12.65 (fin vs jpn)	10.18 (ice vs jpn)	12.22 (fin vs jpn)	12.53 (fin vs jpn)	12.96 (fin vs jpn)	
Smallest	6.61 (eng)	5.61 (cze)	6.40 (rus)	6.77 (eng)	6.85 (eng)	Average language distance
2nd smallest	6.73 (swe)	5.70 (pol)	6.47 (eng)	6.78 (por)	6.97 (cze)	
3rd smallest	6.85 (ind)	5.73 (swe)	6.47 (swe)	6.80 (rus)	7.05 (swe)	
3rd largest	8.60 (chi)	7.59 (chi)	8.38 (chi)	8.74 (chi)	9.12 (chi)	
2nd largest	9.20 (tha)	7.93 (tha)	9.07 (tha)	9.49 (tha)	9.67 (tha)	
largest	11.78 (jpn)	9.35 (jpn)	11.00 (jpn)	11.52 (jpn)	11.96 (jpn)	

Dataset

Figure 8: Summaries of language distance matrices. Language are represented by their ISO 639-2/B codes.

regions with populations speaking Japanese and those with populations speaking Finnish and Icelandic are located at the opposite ends of Eurasia. Actually, Japanese is consistently with the most distinct syntax (the largest average language distance) from the other 19 languages in the 5 datasets (Figure 8), which is also observed in the visualizations of the language distance matrices; see Figure 6, Figure 7 and Supplementary Figure 5-12 Furthermore, the languages with on average the most similar syntax to other 19 languages Furthermore, the languages that are on average the most similar to other 19 languages (languages with the smallest average language distances) include Balto-Slavic languages and English (Figure 8). This is partly because Balto-Slavic-language-speaking countries and regions are located in the center of Eurasia, and English is the most commonly used international language, which also affects the syntax similarity between languages.

In spite of the influence of geographic locations on syntax similarity of languages, there are also languages that are similar in syntax but distant in location. The most typical example in the 20 languages is the syntax similarity between Finnish, Korean and Turkish. In terms of syntax, Finnish is the most similar language to both Korean and Turkish in all 5 datasets; see Supplementary Figure 5-8 and Supplementary Table 4. The pairwise distance between the three languages are all smaller than the mean value 7.73. The

pairwise language distance between Turkish and Finnish is 5.64 in the ENG dataset, which is as small as the distance between English and Italian (Figure 6). The pairwise language distance between Finnish and Korean is 6.13 in the ENG dataset, which is as small as the distance between English and French (Figure 6). The pairwise language distance between Korean and Turkish is 6.25 in the ENG dataset, which is as small as the distance between German and French (Figure 6). It is also observed in the UPGMA dendrograms that the Korean, Finnish and Turkish are closely related, especially that Korean and Turkish share common ancestry in the dendrograms of ENG and SPA datasets; see Figure 7 and Supplementary Figure 9-12. This coincides with a recent unified study leveraging genetics, archaeology and linguistics to show that Korean and Turkish share common ancestry from northeast Asia (Robbeets et al., 2021). However, the connection between Korean and Finnish is unclear with only initial studies discussing the similarity between the two languages (Hadland, 1989) and studies of ancient genomics revealing the spread of Siberian ancestry in northern Europe (Lamnidis et al., 2018).

4.3 Syntax diversity of corpora

We demonstrate that the polynomial representation of dependency trees together with the distance-based methods can be used to measure syntax diversity, which can be useful in, for example, measuring language acquisition and assessing fidelity of text generated by artificial intelligence.

We consider the translations of all sentences in a language a corpus in a dataset. By comparing the pairwise sentence distances of a corpus, we can describe its syntax diversity. Here, we use two simple measures, the diameter and the mean pairwise sentence distance, to describe the syntax diversity of each dataset's 20 corpora. Each corpus contains the translations of all sentences in the dataset, so the 20 corpora in a dataset express the same content in different languages. The diameters and the mean pairwise sentence distances of the 5 datasets are displayed in Figure 9, and the detailed distributions of the pairwise sentence distances for the corpora of the 5 datasets are displayed in Supplementary Figure 13-17. It is observed that the diameters and the mean pairwise distances for Finnish, Korean and Turkish are consistently smaller than other languages, and the diameters and the mean pairwise distances for Japanese and Hindi are in general larger than other languages. This suggests that to express the same information of the corpora, Finnish, Korean and Turkish use more similar syntax, and Hindi and Japanese use more dissimilar syntax, compared with other languages.

5 Discussion

We have generalized the tree distinguishing polynomial for representing dependency trees and defined a distance between the dependency polynomials for comparing syntax of sentences. Compared to other

	Diameter					Mean pairwise sentence distance				
Arabic	27.73	23.94	16.41	18.48	26.08	8.78	7.17	7.99	7.77	8.78
Chinese	26.41	26.39	20.65	19.11	29.24	9.83	9.19	9.81	9.92	10.53
Czech	24.16	20.40	16.12	20.65	24.57	8.33	7.26	8.15	8.19	8.76
English	30.49	27.69	19.19	18.59	24.31	9.52	8.77	9.04	9.29	10.23
Finnish	18.85	19.65	15.01	17.13	17.75	7.59	6.17	7.50	7.62	7.91
French	33.60	26.88	18.09	19.44	26.07	10.43	9.33	8.82	10.12	10.93
German	33.19	21.46	18.47	18.76	26.24	9.68	8.11	9.25	9.57	10.11
Hindi	32.80	26.67	20.58	25.93	27.52	11.09	9.61	10.32	10.90	11.59
Icelandic	24.75	18.58	17.06	20.09	26.57	8.59	7.91	9.01	8.81	9.49
Indonesian	26.80	22.69	17.11	16.64	22.91	8.54	7.47	7.90	8.29	9.35
Italian	37.33	25.97	18.19	21.44	23.56	9.87	8.28	8.92	8.86	9.85
Japanese	38.92	32.29	24.78	25.36	33.80	10.13	8.59	9.83	10.06	9.71
Korean	20.97	17.55	15.17	13.24	15.31	6.45	6.09	6.29	6.40	7.12
Polish	22.63	20.37	20.43	17.19	20.24	8.24	7.05	7.76	8.07	8.69
Portuguese	30.65	30.68	17.60	16.26	26.11	9.45	8.48	8.60	8.35	9.62
Russian	23.39	19.88	17.60	20.66	19.91	8.45	7.03	7.53	7.63	8.60
Spanish	33.15	28.78	18.25	22.33	22.39	9.51	8.52	8.94	8.97	9.69
Swedish	29.39	24.45	19.24	19.73	27.16	8.92	7.71	8.37	9.57	9.97
Thai	26.22	23.82	18.18	24.24	28.08	10.29	9.21	10.06	10.64	11.00
Turkish	21.29	19.10	16.21	14.66	15.96	7.46	6.45	6.90	7.23	7.98
	ENG	GER	FRE	ITA	SPA	ENG	GER	FRE	ITA	SPA
	Dataset									

Figure 9: The diameters and the mean pairwise sentence distances of the 20 corpora in the 5 datasets. Every row records the diameters and mean pairwise sentence distances of the 5 corpora in the corresponding language of the 5 datasets.

methods for analyzing dependency grammar such as studying order of words (Chen and Gerdes, 2017; Gerdes et al., 2021) and calculating dependency distance (Chen and Gerdes, 2022; Lei and Wen, 2020), the polynomial-based methods analyze dependencies from a more comprehensive perspective, taking into account all structural information and dependency relations. The polynomial representation is in fact a "translation" of dependency trees into a form that can be compared and analyzed by distance-based methods and other data analysis tools.

The polynomial-based methods have been applied to analyze 1,000 sentences in the Parallel Universal Dependency (PUD) treebanks, and each treebank contains the translations of the 1,000 sentences in a language. To analyze their syntax, we divided the sentences into 5 datasets based on their original languages. We have compared the sentences with the minimum and maximum polynomial distances between their English and Chinese, French or Spanish translations. This demonstrates the capability of comparing syntax with polynomial-based methods. With the PUD treebanks, we have computed the average pairwise polynomial distance over all sentences in a dataset for each pair of languages. We have used the pairwise language distance to perform a syntactic typology study of the 20 available languages, and we have conducted the analysis for all 5 datasets. The typological results based on the 5 datasets in general agree the genealogical classification in Glottolog 4.6 (Forkel and Hammarström, 2022), though

there are only 50 to 100 sentences originally written in German, French, Italian and Spanish which form the GER, FRE, ITA and SPA datasets respectively. With the polynomial-based methods, we have also observed less discussed syntactic typology results, for example, Japanese and Finnish being among the pairs of languages with the most distinct syntax, the connection between Finnish and Korean and a recently discussed Korean-Turkish link from a study using genetics, archaeology and linguistics (Robbeets et al., 2021).

We have demonstrated using the polynomial distance to measure the syntax diversity of corpora by showing the distributions of pairwise polynomial distances between all pairs of sentences in the corpora. The diameters and the mean pairwise sentence distances provide simple measures of syntax diversity of the corpora. With proper datasets, the polynomial-based methods can be applied to, for example, measure language acquisition, assess fidelity of artificial intelligence generated text, guide artificial intelligence for generating syntactic diverse content, analyze writing styles and detect languages' syntax change over time.

With more sentences being annotated under the Universal Dependencies framework and more Parallel Universal Dependencies treebanks being constructed, we expect that this method can reveal more information about languages, corpora and their connections and motivate new investigations in linguistics.

Implementation and supplementary material

Code and data for analyses conducted in this paper are available at the repository https://github.com/pliumath/dependencies. Supplementary material the paper including supplementary figures and tables can be found at https://github.com/pliumath/dependencies/blob/main/Supplement.pdf

Acknowledgments

We thank the anonymous reviewer(s) for helpful comments and suggestions. P.L. was partially supported by the grant of the Federal Government of Canada's Canada 150 Research Chair program to Prof. C. Colijn and by the National Science Foundation DMS/NIGMS award #2054347 to Prof. M. Vázquez. R.L. was supported by startup funding 12900-310432118 for scientific research of BNUZ.

References

Bryant, D., Tupper, P. F. (2012). Hyperconvexity and tight-span theory for diversities. *Advances in Mathematics*, *231*(6), pp. 3172–3198. https://doi.org/10.1016/j.aim.2012.08.008

Chen, X., Gerdes, K. (2017). Classifying languages by dependency structure. typologies of delexicalized Universal Dependency treebanks. In:*Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 54–63. https://aclanthology.org/W17-6508

Chen, X., Gerdes, K. (2022). Dependency distances and their frequencies in indo-european language. *Journal of Quantitative Linguistics*, *29*(1), pp. 106–125. https://doi.org/10.1080/09296174.2020.1771135

Cox, T. F., Cox, M. A. (2001). *Multidimensional Scaling* (2nd). Chapman & Hall. https://doi.org/10.1201/9780367801700

Craw, S. (2010). Manhattan distance. In: Sammut, C., Webb, G. I. (Eds.). *Encyclopedia of Machine Learning*, p. 639. Springer US. https://doi.org/10.1007/978-1-4899-7687-1_511

Culotta, A., Sorensen, J. (2004). Dependency tree kernels for relation extraction. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 423–429. https://doi.org/10.3115/1218955. 1219009

de Marneffe, M.-C., Manning, C. D., Nivre, J., Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), pp. 255–308. https://doi.org/10.1162/coli_a_00402

Diao, Y., Hetyei, G., Liu, P. (2020). The braid index of reduced alternating links. *Mathematical Proceedings of the Cambridge Philosophical Society*, *168*(3), pp. 415–434. https://doi.org/10.1017/S0305004118000907

Forkel, R., Hammarström, H. (2022). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, *13*(6), pp. 917–924. https://doi.org/10.3233/SW-212843

Freyd, P., Yetter, D., Hoste, J., Lickorish, W. B. R., Millett, K., Ocneanu, A. (1985). A new polynomial invariant of knots and links. *Bulletin of the American Mathematical Society*, *12*(2), pp. 239–246.

Gerdes, K., Kahane, S., Chen, X. (2021). Typometrics: From implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, *6*(1), 17. https://doi.org/10.5334/gjgl.764

Hadland, J. (1989). The finnish korean connection: An initial analysis. *Language Study (in Korean)*, *25*(3), pp. 689– 703. https://hdl.handle.net/10371/85842

Imrényi, A., Mazziotta, N. (2020). *Chapters of Dependency Grammar: A Historical Survey from Antiquity to Tesnière*. Amsterdam/Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/slcs.212

Janssen, R., Liu, P. (2021). Comparing the topology of phylogenetic network generators. *Journal of bioinformatics and computational biology*, *19*, 2140012. https://doi.org/10.1142/S0219720021400126

Jones, V. F. R. (1985). A polynomial invariant for knots via von neumann algebras. *Bulletin of the American Mathematical Society*, *12*(1), pp. 103–111.

Kauffman, L. (1987). State models and the jones polynomial. *Topology*, *26*(3), pp. 395–407. https://doi.org/10.1016/ 0040-9383(87)90009-7

Lamnidis, T. C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., Khartanovich, V., Balanovsky, O., Ongyerth, M., Weihmann, A., Sajantila, A., Kelso, J., Pääbo, S., Onkamo, P., Haak, W., Krause, J., Schiffels, S. (2018). Ancient fennoscandian genomes reveal origin and spread of siberian ancestry in europe. *Nature Communications*, *9*(1), 5018. https://doi.org/10.1038/s41467-018-07483-5

Lei, L., Wen, J. (2020). Is dependency distance experiencing a process of minimization? a diachronic study based on the state of the union addresses. *Lingua*, *239*, 102762. https://doi.org/10.1016/j.lingua.2019.102762

Liu, H., Xu, C. (2012). Quantitative typological analysis of romance languages. *Poznań Studies in Contemporary Linguistics*, *48*(4), pp. 597–625. https://doi.org/10.1515/psicl-2012-0027

Liu, P. (2021). A tree distinguishing polynomial. *Discrete Applied Mathematics*, *288*, pp. 1–8. https://doi.org/10.1016/j.dam.2020.08.019

Liu, P., Biller, P., Gould, M., Colijn, C. (2022). Analyzing phylogenetic trees with a tree lattice coordinate system and a graph polynomial. *Systematic Biology*, *71*(6), pp. 1378–1390. https://doi.org/10.1093/sysbio/syac008

Luo, Q., Xi, J. (2005). A novel similarity measure for dependency trees [query answer system example]. *Proceedings. 2005 International Conference on Communications, Circuits and Systems*, 785. https://doi.org/10.1109/ICCCAS.2005.1495227

Murasugi, K. (1991). On the braid index of alternating links. *Transactions of the American Mathematical Society*, *326*(1), pp. 237–260. https://doi.org/10.1090/S0002-9947-1991-1000333-3

Pons, J. C., Coronado, T. M., Hendriksen, M., Francis, A. (2022). A polynomial invariant for a new class of phylogenetic networks. *PLOS ONE*, *17*(5), pp. 1–22. https://doi.org/10.1371/journal.pone.0268181

Reis, D. C., Golgher, P. B., Silva, A. S., Laender, A. F. (2004). Automatic web news extraction using tree edit distance. *Proceedings of the 13th International Conference on World Wide Web*, pp. 502–511. https://doi.org/10.1145/ 988672.988740

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., ... Ning, C. (2021). Triangulation supports agricultural spread of the Transeurasian languages. *Nature*, *599*(7886), pp. 616–621. https://doi.org/10.1038/s41586-021-04108-8

Sokal, R. R., Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, pp. 1409–1438.

Thistlethwaite, M. B. (1987). A spanning tree expansion of the jones polynomial. *Topology*, *26*(3), pp. 297–309. https://doi.org/10.1016/0040-9383(87)90003-6

Tutte, W. T. (1954). A contribution to the theory of chromatic polynomials. *Canadian Journal of Mathematics*, *6*, pp. 80–91. https://doi.org/10.4153/CJM-1954-010-9

van Iersel, L., Moulton, V., Murakami, Y. (2022). Polynomial invariants for cactuses. *Preprint*. https://doi.org/10.48550/arxiv.2209.12525

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., Korhonen, A. (2020). Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, *46*(4), pp. 847–897. https://doi.org/10.1162/coli_a_00391

Wong, T.-S., Gerdes, K., Leung, H., Lee, J. (2017). Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 266–275. https://aclanthology.org/W17-6530

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., . . . Li, J. (2017). 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*, pp. 1–19. https://doi.org/10.18653/v1/K17-3001