Language Resources and Evaluation (2025) 59:3445–3485
https://doi.org/10.1007/s10579-025-09826-3

ORIGINAL PAPER

Parallel Trees: a novel resource with aligned dependencyand constituency syntactic representations

Chiara Alzetta[1] · Alessio Miaschi[1] · Felice Dell'Orletta[1] · Giulia Venturi[1] · Simonetta Montemagni[1]

Accepted: 2 April 2025 / Published online: 28 June 2025

Abstract

The paper introduces Parallel Trees, a novel multilingual treebank collection that includes 20 treebanks for 10 languages. The distinguishing property of this resource is that the sentences of each language are annotated using two syntactic representation paradigms (SRPs), respectively based on the notions of dependency and constituency. By aligning the annotations of existing resources, Parallel Trees represents an example of exploiting pre-existing treebanks to adapt them to novel applications. To illustrate its potential, we present a case study where the resource is employed as a benchmark to investigate whether and how BERT, one of the first prominent neural language models (NLMs), is sensitive to the dependency- and constituency-based approaches for representing the syntactic structure of a sentence. The case study results indicate that the model's sensitivity fluctuates across languages and experimental settings. The unique nature of the Parallel Trees resource creates the prerequisites for innovative studies comparing dependency and phrase-structure trees, allowing for more focused investigations without the interference of lexical variation.

Keywords Parallel treebanks · Syntactic representation · Diagnostic probing paradigm · Neural language model

✉ Chiara Alzetta
chiara.alzetta@ilc.cnr.it

✉ Alessio Miaschi
alessio.miaschi@ilc.cnr.it

✉ Giulia Venturi
giulia.venturi@ilc.cnr.it

Felice Dell'Orletta
felice.dellorletta@ilc.cnr.it

Simonetta Montemagni
simonetta.montemagni@ilc.cnr.it

[1] Istituto di Linguistica Computazionale "A. Zampolli", CNR-ILC, ItaliaNLP Lab, via G. Moruzzi 1, Pisa, Italy

1 Introduction

In the last few years, the advent of neural language models (NLMs) based on Transformer architectures has revolutionized the natural language processing (NLP) field by achieving unprecedented performance across a wide range of natural language understanding and generation tasks (Wang et al., 2018). Despite their remarkable success, these models pose an open challenge due to their inherent opacity, making it difficult to decipher the mechanisms underlying their operations (Belinkov, 2022) and to holistically assess their multifacet abilities (Liang et al., 2023). For this reason, the NLP community has witnessed a surge in studies aimed at interpreting and evaluating these models, seeking to unveil their inner workings as well as potential limits.

Of particular interest to the present study is the considerable body of research that has exploited either pre-existing or newly built resources to assess the linguistic abilities of NLMs, with a specific focus on evaluating their syntactic competencies (Waldis et al., 2024). These studies select multiple linguistic abilities to be assessed, whether they involve linguistic structures (e.g., semantic or syntactic relation types) or targeted linguistic phenomena (e.g., subject-verb agreement).

To the best of our knowledge, although a wide range of syntactic phenomena has been tested, a less-explored research area focuses on the models' sensitivity to alternative ways to represent the syntactic structure of sentences, for instance in terms of dependencies or constituents. This topic has been discussed, among others, by Kulmizev et al. (2020); Kogkalidis and Wijnholds (2022); Arps et al. (2022). However, it seems that the research community has not yet reached a consensus on whether different syntactic formalisms have an impact on NLMs' abilities to master specific syntactic structures (Muñoz-Ortiz et al., 2023). This highlights the need for resources that enable direct comparisons across syntactic representation paradigms, making their development both timely and relevant.

To fill this gap, this paper introduces *Parallel Trees*, a newly developed multilingual resource designed to serve as a benchmark for exploring how different syntactic representation paradigms (SRPs) represent linguistic information. The resource comprises a collection of 20 treebanks for 10 different languages belonging to multiple language families. We named it Parallel Trees to highlight its primary novel characteristic: for each language, the resource includes sentences acquired through an articulated alignment process from pre-existing treebanks, each of which is assigned two types of syntactic representations, respectively based on a constituency (hereafter C-SRP) and a dependency (D-SRP) model of syntax. Note that here the parallelism, rather than being over parallel corpora (i.e. translationally equivalent texts in different languages) as in the case of parallel treebanks, is referred to the syntactic representation. For both considered SRPs, we resorted to *de facto* representation standards. Concerning D-SRP, we choose the treebanks developed in the framework of the Universal Dependencies (UD) project, since it offers the most comprehensive collection of multilingual resources annotated with a shared dependency-based annotation scheme (Nivre et al., 2016, 2020; de Marneffe et al., 2021). For C-SRP, we selected the annotation scheme of the Penn Treebank (PTB)

(Marcus et al., 1994) whenever possible, as it has a long-standing history as a training and evaluation corpus in natural language processing, also for languages beyond English.

To the best of our knowledge, Parallel Trees represents a novel and original resource among syntactically annotated treebanks with the potential to avoid interference from other linguistic elements - particularly the semantic content of sentences - thereby facilitating comparative studies based on contrastive analysis of dependency and constituency trees. Methodologically, the resource also offers an example of re-using and adapting pre-existing resources to make them suitable to address novel challenges and research questions.

To demonstrate one of the possible applications of the Parallel Trees resource, the paper presents a case study in which the resource is used to investigate the sensitivity of neural language models to the two considered syntactic representation paradigms. In line with the literature on assessing the linguistic competencies of NLMs, we adopted one of the mostly used evaluation methodologies: the *diagnostic probing approach* (Conneau et al., 2018; Warstadt et al., 2019a). To this aim, we designed four probing tasks that capture key properties of a sentence. In these tasks, the probing model uses, as input, sentence embeddings from one of the first prominent NLMs, BERT (Devlin et al., 2019). Specifically, we investigated whether BERT encodes dependency and constituency information in distinct ways, thus revealing its sensitivity to different syntactic representation formalisms and whether this sensitivity varied across the observed linguistic properties and languages.

The remainder of the paper is organised as follows. We present the related work and background in Sect. 2. Section 3 introduces the novel Parallel Trees resource. In particular, we describe the alignment methodology that led to the construction of the novel resource (Sect. 3.1), introduce the languages and source treebanks covered (Sect. 3.2), and illustrate the distribution of a selection of linguistic features across the considered treebanks (Sect. 3.3). Section 4 presents the case study. The models employed and the experimental settings are detailed in Sect. 4.1, and the experimental results are reported in Sects. 4.2 and 4.3. Section 4.4 summarises the main results of this work and Sect. 5 presents the conclusions and possible future developments.

2 Background

The background of the work reported in this paper is provided from different perspectives. Section 2.1 focuses on the syntactic representation paradigms covered by the Parallel Trees resource: we start from the notions of constituency and dependency for the syntactic modelling of sentences and briefly survey the debate around them from different perspectives. Section 2.2 discusses the emerging roles that annotated linguistic resources play in assessing NLM competencies. Finally, Sect. 2.3 presents the background of the case study, illustrating prior research in which the diagnostic probing paradigm, i.e. the approach we employed in our case study, has been used to investigate whether and how different syntactic formalisms impact probing results.

2.1 Syntactic representation paradigms

Constituency and dependency representations are based on totally different assumptions. In C-SRP, a given sentence is divided into phrasal constituents, which include terminal nodes (labelled with vocabulary items) and non-terminal nodes (labelled with syntactic categories, e.g. Noun Phrase, Verb Phrase, etc.). This representation type is highly hierarchical and divides the sentences into recursively embedded phrasal constituents. As opposed to C-SRP, dependency-based representations directly encode word-to-word grammatical relationships, without making use of phrasal constituents. Dependency links are established between a head (or governor) and a dependent: essentially, dependencies are binary, asymmetric governance relations holding between words. These relations are usually labelled with functional labels such as Subject, Direct_Object, Modifier, etc.

Given the different assumptions underlying C-SRP and D-SRP, let us consider some of the main differences in modelling the syntactic structure of a sentence. Constituency trees, as already mentioned, contain non-terminal nodes, while dependency representations do not. From this it follows that dependency trees look flatter compared to the hierarchical syntactic structures in which nested constituency plays a key role. Osborne (2014) reports that the number of nodes in dependency-based structures tends to be approximately half that of constituency-based structures (p. 624). These differences lead to different treatments of non-local phenomena, such as ellipsis, and discontinuity. C-SRP representations often address these constructions using a trace-filler notation, which incorporates empty nodes to represent missing or distant constituents. Conversely, dependency trees typically do not represent empty categories (Rambow & Joshi, 1997), unless explicit mechanisms are introduced, such as zero wordforms (Mel'čuk et al., 1988; Mel'čuk & Polguère, 2009).

In both C-SRP and D-SRP, the syntactic structure of a sentence is represented as a tree, as exemplified in Fig. 1 which shows the constituency- and dependency-based

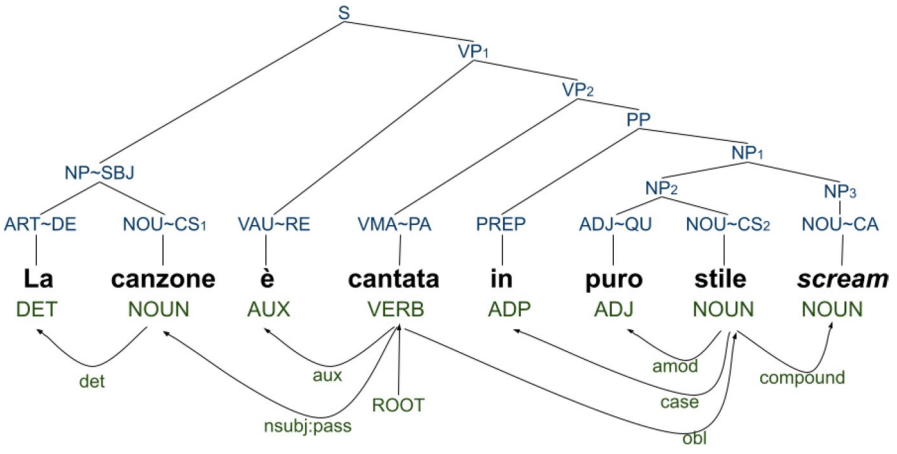

Fig. 1 Constituency (top) and dependency (bottom) trees representing an Italian sentence [transl. "*The song is sung in pure scream style*"] extracted from the Turin University Treebank

representations for the following Italian sentence extracted from the Turin University Treebank (Bosco et al., 2012):

(1) *La canzone è cantata in puro stile scream.* [transl. 'The song is sung in pure scream style.']

Looking at the two representations,[1] it can be noticed that while the dependency tree (bottom) is represented in terms of head-dependent relations between words, the phrase structure tree (top) includes both non-terminal (*phrases*) and terminal nodes (*words*). As a consequence, for example, in the phrase structure tree, the verb *cantata* ('sung') is linked to its subject *canzone* ('song') and to the prepositional phrasal modifier *in puro stile scream* ('in pure scream style') via multiple or one non-terminal nodes, respectively. In contrast, the dependency tree explicitly encodes predicate-argument relationships through labeled dependencies between words, without the mediation of intermediate phrasal nodes.

Given the distinct principles of syntactic organization underlying these two representations, their relationship has been investigated in the literature from different perspectives. In fact, it has been a long-standing problem in theoretical linguistics. The debate revolved around different topics, ranging from their formal equivalence (see, among others, Hays (1964); Gaifman (1965); Hudson (1980)), the representational economy and simplicity (e.g., Osborne (2014)), to the empirical adequacy to be understood as cross-linguistic data coverage (de Marneffe et al., 2021). Despite a lack of consensus on formal aspects of the correspondence between D-SRP and C-SRP, linguists often still look at them as equivalent notational variants. Despite a lack of consensus on formal aspects of the correspondence between D-SRP and C-SRP, linguists often still look at them as notational variants (Nefdt & Baggio, 2023) or weakly equivalent (Matthews, 1981) (because establishing a one-to-one correspondence between them is not always feasible).

In natural language processing, dependency-based methods for syntactic parsing have gained increasing popularity in recent years with respect to constituency parsing (Kübler et al., 2009). Among the practical and technical reasons for their increasing popularity, it is worth mentioning here: the direct applicability to NLP tasks, including information extraction, semantic parsing, machine translation and question answering; the lower computational complexity with respect to constituency parsing, especially for large-scale applications; the adaptability across languages, even those with flexible or free word order; last but not least, the availability of dependency treebanks for different languages, reinforcing dependency parsing's popularity.

In psycholinguistics and cognitive studies, both constituency-based and dependency-based syntactic representations offer distinct insights into how humans process language, and each framework highlights different aspects of cognitive language processing. Constituency-based representations are particularly valuable for

[1] Note that the constituency tree reflects the Italian PTB schema, while the dependency tree follows the UD scheme.

examining hierarchical aspects of language comprehension, while dependency-based representations resonate with theories of immediate, word-to-word processing, reflecting distinct but complementary cognitive processes. As widely discussed in Nefdt and Baggio (2023), it appears that neither C-SRP nor D-SRP alone are sufficient to model human syntactic competence and to explain all cross-linguistic, psycholinguistic, and neurolinguistic observations reported in the literature (see Oota et al. (2023); Lopopolo et al. (2021) to mention only a few): they both seem necessary (and may be jointly sufficient) to achieve full explanatory adequacy.

2.2 Linguistic resources as benchmarks for language models

The reuse of pre-existing linguistic resources and the development of new ones that include targeted linguistic phenomena as benchmarks have gained renewed importance in recent years. These benchmarks serve to test whether and to what extent NLMs can master diverse linguistic phenomena, as well as their abilities to resolve specific NLP tasks. While the literature on this topic is extensive and an in-depth survey is beyond the scope of this study, this section aims to provide a broad overview of the fundamental role of both new and existing linguistic resources in the novel research scenarios of natural language processing.

One of the first objectives in this line of research has been to explore how resources can be used to assess NLMs' abilities to implicitly encode a wide range of linguistic knowledge within their sentence representations. This knowledge spans from surface-level properties of sentences, as explored in studies like those by Conneau et al. (2018), to more complex syntactic (Conneau et al., 2018; Hewitt & Manning, 2019) and semantic or discourse-level structures (Ettinger, 2020). A significant number of studies focus on datasets explicitly built to test targeted linguistic phenomena. Notable examples include the CoLA dataset (Warstadt et al., 2019b), which contains 10,000 sentences covering a variety of linguistic phenomena drawn from linguistics papers and books, and the Benchmark of Linguistic Minimal Pairs (BLiMP), comprising 67 minimal pair paradigms, each with 1,000 sentence pairs in American English grouped into 12 categories (Warstadt et al., 2020). Another prominent example is SyntaxGym (Gauthier et al., 2020), a platform implementing the targeted syntactic evaluation paradigm designed to serve users without advanced technical skills or computational resources to evaluate models' syntactic knowledge on many specific phenomena such as subject-verb agreement. More recently, Waldis et al. (2024) introduced the Holmes benchmark, which features 208 datasets addressing 66 distinct phenomena across morphology, syntax, semantics, reasoning, and discourse. With the goal of disentangling NLMs' syntactic and semantic knowledge, Arps et al. (2024) proposed SPUD (Semantically Perturbated Universal Dependencies), a framework for creating nonce treebanks for multilingual Universal Dependencies corpora.

Beyond the evaluation of specific linguistic phenomena, linguistic resources have been extensively developed and employed as benchmarks for testing NLMs' capabilities across a wide range of natural language understanding and generation tasks. Notable examples include GLUE (Wang et al., 2018) and SuperGLUE (Wang et al.,

2019), BIG-bench (Srivastava et al., 2023). The impact of these evaluation benchmarks yielded the development of several platforms such as the OpenLLM Leaderboard (Beeching et al., 2023). These resources allow testing models on applied competencies, such as commonsense reasoning, sentence similarity and domain-specific problem-solving.

To the best of our knowledge, no existing resource offers a systematic framework for testing NLMs' abilities to compare the two SRPs using treebanks aligned at the token level with constituency and dependency annotation. The only notable exception is the dataset developed for the 2013 Statistical Parsing of Morphologically Rich Languages (SPMRL) Shared Task (Seddah et al., 2013), covering nine languages and with parallel constituency and dependency annotations. The dataset, however, is no longer available. Although in the SPMRL dataset constituency- and dependency-annotated sentences were aligned at the token level, for each language distinct annotation schemes and criteria were used for the different levels (morpho-syntax, constituency and dependency syntactic representations). Furthermore, most annotations were automatically generated for the SPMRL shared task and used without manual revision.

2.3 Diagnostic probing for syntactic representational structure

The *diagnostic probing paradigm* (Conneau et al., 2018; Warstadt et al., 2019a) is one of the most widely adopted interpretability approaches for NLMs in the literature. Specifically, it relies on the rather simple idea that if a lightweight classifier (the *probing model*) successfully predicts a given language property (the *probing task*) by using the sentence embeddings of a pre-trained NLM as input, we can assume that the model somehow encoded the property. To this end, several methods have been implemented also taking inspiration from human language experiments (Ettinger, 2020) and demonstrated that sentence-level representations encode linguistic knowledge in a hierarchical manner (Belinkov et al., 2017; Blevins et al., 2018; Tenney et al., 2019), and can even support the extraction of either undirected (Hewitt & Manning, 2019) or labelled and directed dependency parse trees (Müller-Eberstein et al., 2022).

Despite the substantial body of work dedicated to probing the inherent abilities of NLMs, there are still several open questions concerning their use (Hewitt & Liang, 2019; Maudslay & Cotterell, 2021; Belinkov, 2022). Among the many open methodological aspects, it is worth mentioning a cross-cutting direction of research that concerns the potential influence of syntactic formalisms on the probing outcomes. This line of research investigates how the choice of syntactic representation paradigms used for annotating benchmark resources might affect the results of the probing tasks.

As highlighted by Belinkov (2022), the majority of probing studies have relied on dependency-based syntactic representations, with the Universal Dependencies (UD) formalism being particularly prominent. However, few studies have examined the impact of alternative syntactic paradigms on evaluating NLMs' syntactic competencies. The most commonly considered alternative is the constituency-based approach

even if few studies are focused on probing NLMs solely for constituency structure. A notable exception is the study by Arps et al. (2022), who used probing techniques to assess the accuracy of representing constituents of different categories within the neuron activations of a LM such as RoBERTa. On the contrary, a more prominent line of research is dedicated to the comparative study of the models' sensitivity to the two syntactic representation paradigms. Among the first studies, Tenney et al. (2019) conducted a comparative analysis of four contextualized representation models, probing their abilities across a suite of eight core natural language processing tasks. These tasks encompassed both syntactic and semantic phenomena, including both constituent and dependency labelling. The authors used distinct benchmarks for each task and did not directly compare the two syntactic labelling paradigms. However, their results indicated higher probing accuracies in the dependency labelling task, suggesting that the models may have a preference for implicitly encoding this paradigm.

An investigation more focused on a direct comparison between two different formalisms of the same dependency representation paradigm is the main focus of Kulmizev et al. (2020). The authors conducted a study on 13 languages and probed two NLMs (BERT and ELMO) for two formalisms, i.e. UD and Surface Universal Dependencies (SUD), a representation scheme more oriented towards surface structure (Gerdes et al., 2018). Although the experiments show different sensitivities of the models to the two formalisms, the results seem mainly influenced by the typological properties of the considered languages rather than the syntactic formalisms. It's essential to note that SUD and UD are very similar formalisms, both based on D-SRP, diverging primarily with respect to the inventory of labels and the use of functional words as heads.

BERT and ELMO were also used by Vilares et al. (2020) to compare the effectiveness of precomputed embeddings for dependency and constituency parsing. Results suggest that the embeddings of both models are particularly effective in constituency parsing. However, relying on different treebanks for each parsing task makes the results obtained not directly comparable. Furthermore, this work is constrained by focusing solely on the English language. Divergent findings are reported by Muñoz-Ortiz et al. (2023), who did not find clear evidence indicating a preference for encoding dependency-based syntactic information over constituency-based information in pre-trained word vectors of three NLMs. They performed a multilingual analysis by selecting 13 UD treebanks and 8 constituency treebanks. However, they did not rely on parallel resources.

Based on the current state of the art, there is no clear consensus about which syntactic representation paradigm a language model is more sensitive to.

3 The Parallel Trees resource

Parallel Trees is a unique multilingual resource specifically conceived to compare the linguistic structure of a sentence formalized by constituency- vs. dependency-based representations. The resource originates from a set of pre-existing treebanks for 10 languages which were annotated from both perspectives. Note that we

deliberately selected treebanks openly and freely accessible for research purposes to ensure transparency and replicability of results.

As introduced in Sect. 1, for the D-SRP we choose the treebanks developed in the framework of the Universal Dependencies (UD) project, offering a wide collection of multilingual resources annotated with a shared syntactic annotation scheme (de Marneffe et al., 2021). For what concerns C-SRP, we considered the annotation scheme of the Penn Treebank (PTB) originally developed for English (Marcus et al., 1994). Over the years, many treebanks have been developed for different languages using PTB or PTB-inspired annotations. Differently from the UD resources sharing the same annotation scheme, PTB annotation for different languages might feature language-specific adaptations, in particular for what concerns not directly observable data, handled in PTB in terms of empty categories. Nevertheless, cross-language comparability is guaranteed by the fact that all languages share the same core phrase-based annotation principles.

As detailed in Table 1, the newly created multilingual resource consists of 87,376 sentences with parallel annotations, which represent a subset of the original treebanks. This subset is the result of a comprehensive alignment strategy (see Sect. 3.1) aimed at ensuring consistency between the source constituency and dependency treebanks. The selection of the 10 languages included in Parallel Trees was guided by the availability of treebanks annotated according to the two annotation schemas considered, namely UD and PBT.

3.1 The alignment approach

In Parallel Trees, each sentence is assigned both a dependency- and a constituency-based representation. As a first step, we identified languages with the same treebank annotated according to both dependency- and constituency-based syntactic representations. For treebanks where the PTB format was unavailable, we checked whether the resource was distributed in other constituency-based formats comparable to the Penn Treebank. If so, we included these resources in our dataset and the subset of sentences annotated according to both SRPs was extracted. Specifically, sentence trees were aligned by matching words or assigned IDs between sentences in the source dependency and constituency treebanks. Furthermore, and most importantly, we verified that the sentences paired using IDs were equally segmented and tokenized. In this way, we ensured that the constituency and dependency parallel trees refer to exactly the same sentences, and differ only at the level of the assigned syntactic annotation.

The number of parallel trees, reported in the last column of Table 1, reflects a reduction in the original dataset sizes due to our alignment strategy. For instance, some sentences were excluded due to different splitting strategies used in the constituency and dependency versions (e.g., whether semicolons trigger sentence split or not), or were omitted by curators during formalism conversion. Despite the reduction, this approach ensures consistent alignment of trees.

Furthermore, it is worth mentioning that the *Parallel Trees* collection was challenged also by treebank license restrictions, particularly in the case of constituency

Table 1 Source Treebanks details

Lang (code)	Genre	PST				DT					Parallel
		Name	Sents	Annotation	License	Name	Sents	Annotation	1st vers.	License	
Catalan (CAT)	News	AnCora-CA	16,591	Semi-auto	GNU-GPL	AnCora	16,678	Semi-auto	1.3	CC BY 4.0	14,861
Dutch (DUT)	News	Alpino	7,138	Automatic	GNU-GPL	Alpino	13,603	Semi-auto	1.2	CC BY-SA 4.0	7,120
French (FRE)	News	French Treebank	21,550	Automatic	CC-BY-NC-ND	FTB	18,535	Automatic	2.0*	LGPL-LR	18,509
Icelandic (ICE)	Non fiction	IcePaHC	4,671	Manual	CC BY 4.0	Modern	6,928	Semi-auto	2.7	CC BY-SA 4.0	2,410
Indonesian (IND)	News, non fiction	Kethu	1,031	Manual	GNU AGPL	CSUI	1,030	Automatic	2.7	CC BY-SA 4.0	923
Italian (ITA)	News, non fiction	TUT	2,859	Semi-auto	CC BY-NC-SA 2.5	ISDT	14,167	Semi-auto	1.0	CC BY-NC-SA 3.0	1,418
Portuguese (POR)	News	Bosque	9,368	Semi-auto	CC BY-NC-SA 3.0	Bosque	9,357	Semi-auto	1.2	CC BY-SA 4.0	8,862
Spanish (SPA)	News	AnCora-ES	17,376	Semi-auto	GNU-GPL	AnCora	17,662	Semi-auto	1.3	CC BY 4.0	15,704
Turkish (TUR)	News, non fiction	Turkish-Penn	9,561	Manual	CC BY	Penn	16,396	Manual	2.8	CC BY-SA 4.0	9,494
Vietnamese (VIE)	News	VietTreebank	10,165	Manual	CC BY-NC-SA 4.0	VTB	8,120	Automatic	1.4	CC BY-SA 4.0	8,075
Grand Total			100,310				122,476				87,376

*FTB was removed from UD since version 2.11

For each language, we report the ISO-639-2 language code and the genre of the texts

For both constituency treebanks (PST section of the table) and dependency treebanks (DT section of the table), we report the treebank name, the sentence count (Sents), the annotation approach, and the treebank distribution license

For DTs we also report the first UD release of the treebank. Concerning the annotation approach, note that 'Manual' signifies that all levels of the annotation have been carried out manually, 'Automatic' indicates that the annotation was performed fully automatically, 'Semi-auto' indicates that the treebank was converted automatically and then fully or partially revised manually

The last column (Parallel) reports the number of sentences annotated according to both syntactic representation paradigms and included in the Parallel Trees resource

treebanks. These resources are not always freely accessible, easily available, or actively maintained, and the treebanks selected for the study do not encompass all existing treebanks annotated according to both SRPs. Rather, these were specifically chosen because they were curated and maintained by their respective authors, who provided the necessary support during the data collection process thus facilitating the reproducibility of the experiments (see Acknowledgments).

To further support the research community, the Parallel Trees resource is available in the Open Collection of the CLARIN repository and the following webpage.

3.2 Languages and source treebanks

The set of ten languages included in the Parallel Trees resource spans across languages showing different typological properties. Specifically, the set covers four language families, internally distinguished in genera according to the World Atlas of Language Structures (WALS) (Dryer & Haspelmath, 2013).[2] Most of the languages fall within the Indo-European family, which includes two major groups: Romance, represented in this work by Catalan, French, Italian, Portuguese, and Spanish, and Germanic languages, here represented by Dutch and Icelandic. The remaining three languages belong to different families: Indonesian belongs to the Austronesian family, while Turkish and Vietnamese belong to the Turkic and Austro-Asiatic families, respectively.

The languages included are somewhat limited in terms of typological variety, having 5 out of 10 languages belonging to the Romance genus, a branch of the Indo-European language family. However, we selected these languages because they are the only ones with available treebanks annotated using both dependency- and constituency-based syntactic representations. Additionally, they are diverse enough to allow investigations into whether differences and similarities between syntactic representation paradigms relate to typological properties.

Below, we present the source treebanks for each language, along with details regarding the construction process of their constituency- and dependency-based versions. Links to the source treebanks and the indication of the versions we used are provided in Appendix A, while the key properties specific to each treebank are summarised in Table 1.

Catalan and Spanish

These two languages are covered by the AnCora project (Taulé et al., 2008), which provides treebanks for Catalan and Spanish (Castillian) annotated according to multiple SRPs.

Constituency treebank: The AnCora phrase-structure annotation is based on a theory-neutral annotation scheme, with language-specific tagset and annotation criteria. Among its main features, it is worth mentioning that: no formal distinction between arguments and adjuncts is made in the syntactic tree; the verb node is represented as a sister node of its complements (either arguments or adjuncts); elliptical subjects are encoded as empty nodes; nominal, adjectival and adverbial phrases

[2] http://wals.info.

are assigned a multilevel representation modelling their internal structure, possibly highly nested. These properties, as we will discuss further, have consequences at the level of tree depth.

Dependency treebank: The first conversion of AnCora treebanks into UD dependency trees was carried out automatically from the original constituency-based versions (Civit et al., 2006); the result was released for the UD version 1.3. Subsequent releases underwent manual revision to remove errors and inconsistencies (Alonso & Zeman, 2016).

Dutch

The Alpino Treebank of Dutch originates from a combination of samples from various treebanks annotated using the Alpino annotation tools and guidelines. In its native format, distributed in XML, Alpino employs a hybrid annotation scheme, encoding both dependency and constituency information (Van der Beek et al., 2002).

Constituency treebank: The constituent annotation of Alpino employs a theory-independent annotation format, where constituents are marked through bracketing and assigned a syntactic category. Note that functional information is annotated as node attributes, without affecting the overall tree structure.

Dependency treebank: The Alpino treebank was automatically converted to UD and included in the training set of UD Alpino Treebank since the UD v1.2 release (Bouma & van Noord, 2017). For the UD version 2.1 release, the treebank underwent manual revision.

French

The French Treebank (Abeillé et al., 2000) is the outcome of a long-term project aimed at building a theory-neutral, surface-oriented treebank with both constituency- and dependency-based annotation.

Constituency treebank: The phrase-structure representation of the French Treebank is based on the PTB bracketed formalism, but the schema was slightly adjusted to produce surface and shallow annotations with major phrases and little internal structure (e.g. in the noun phrase, determiners and modifying adjectives are sister nodes). For verbal phrases, only the minimal verbal nucleus (clitics, auxiliaries, negation and verb) is marked, excluding complements.

Dependency treebank: UD-FTB (Seddah et al., 2018) is the UD treebank resulting from the automatic conversion of the constituency trees of the French Treebank. The conversion of the French Treebank to UD started with the v2.0 release, and was performed automatically (Bonfante et al., 2018).

Icelandic

The Icelandic Parsed Historical Corpus (IcePaHC) is a diachronic corpus with samples of written Icelandic spanning from the 12th century to modern times (Rögnvaldsson et al., 2012). For the specific concerns of this study, we focused on the 21st-century texts from the Modern Treebank. This subset includes parliamentary speeches delivered by four Icelandic members of parliament between 2011 and 2015, as well as newswire articles about sport.

Constituency treebank: The phrase-structure annotation of the Icelandic Treebank follows the annotation guidelines of The Penn Parsed Corpora of Historical English

(PPCHE),[3] the version of the Penn annotation scheme adapted for Old English, with minor modifications tailored to Icelandic. The annotation, carried out automatically, was then manually revised.

Dependency treebank: The Icelandic Modern treebank became part of UD since the v2.7 release thanks to an automatic conversion process from the Penn Treebank format to Universal Dependencies (Arnardóttir et al., 2020).

Indonesian

The Kethu treebank, derived from the Universitas Indonesia Constituency Treebank (UI-CTB) (Dinakaramani et al., 2014), is a resource focusing on formal Indonesian newswire articles (Arwidarasti et al., 2019).

Constituency treebank: The Kethu constituency treebank results from the conversion of UI-CTB to the Penn Treebank format, which was modified for what concerns tokenization and compound annotation. The conversion was carried out automatically, followed by manual revision.

Dependency treebank: An automatic process converted Kethu to UD, resulting in the CSUI-UD treebank (Alfina et al., 2020), a quite small treebank built at the Faculty of Computer Science, Universitas Indonesia, which is part of the Universal Dependencies project since the v2.7 release.

Italian.

The Turin University Treebank (TUT) (Bosco et al., 2000) is an Italian resource that features parallel syntactic annotations in various formats. The treebank comprises miscellaneous texts, spanning from newswire articles, legal texts, and Wikipedia pages.

Constituency treebank: Although the original TUT annotation follows a language-specific dependency-based schema, a constituency-based representation of TUT sentences has been made available in the Penn Treebank phrase structure format (slightly adapted to Italian) (Bosco, 2007). The constituency PTB format results from an automatic conversion applied to the language-specific dependency format in which the resource was originally annotated.

Dependency treebank: TUT is part of UD since v1.0. It was automatically converted to the UD annotation scheme and merged with other existing Italian resources to create the Italian UD Treebank (Bosco et al., 2013). In subsequent releases, the treebank annotation underwent multiple steps of manual revision (Alzetta et al., 2017).

Portuguese.

For Portuguese, we focused on the European variant represented by the Bosque treebank (Afonso et al., 2002), CETEMPúblico portion.

Constituency treebank: Bosque is a constituency-annotated resource, part of the Floresta Sintá(c)tica treebank, originally created using a Constraint Grammar parser. The treebank has been converted to multiple formalisms, including the PTB format.

Dependency treebank: Bosque was automatically converted to the UD format by applying a context-sensitive set of Constraint Grammar rules (Bick, 2016) with additional manual corrections (Rademaker et al., 2017) and became part of the UD resource collection since v1.2.

[3] https://www.ling.upenn.edu/ppche/ppche-release-2016/annotation

Turkish

For Turkish, we started from the Turkish Penn Constituency Treebank, consisting of a collection of 10,000 sentences, each with a maximum length of 15 tokens (Kuzgun et al., 2021).

Constituency treebank: The Turkish Penn Constituency Treebank includes sentences translated from the original version of the Penn Treebank (Kara et al., 2020). For their annotation, the PTB scheme was adapted to provide a more accurate syntactic representation of the Turkish language. Annotation was carried out manually in parallel with the translation process.

Dependency treebank: The UD Turkish Penn treebank is the result of semi-automatic morphological parsing combined with manual annotation of UD dependencies. During the dependency annotation process, annotators were able to view the original sentences from the Penn Treebank, allowing them to verify and correct the sentences based on the original data. This treebank has been part of Universal Dependencies since the UD v2.8 release.

Vietnamese

The Vietnamese Language and Speech Processing (VLSP) project is aimed at building resources and tools for processing Vietnamese. Among these resources, the VietTreebank (Nguyen et al., 2009) is the monolingual treebank of Vietnamese which features syntactic trees with both constituent and functional information. Note that Vietnamese is an isolating language and has no word delimiter, thus words here correspond to the smallest unit which is syntactically independent.

Constituency treebank: The original annotation of the VietTreebank includes both constituency and functional information. Among the supported representation formats there is also the bracketing representation of the Penn Treebank. The annotation was obtained semi-automatically: a first portion of the treebank was manually annotated, while the subsequent portions were annotated automatically and then manually revised.

Dependency treebank: The Vietnamese UD treebank is a conversion of the constituent treebank created in the VLSP project. The treebank, among the UD treebanks since the v1.4 release, was included in the training set of the VLSP shared task on Vietnamese universal dependencies parsing (Linh et al., 2020).

This overview indicates that, for each language, the selected treebanks were originally conceived in various annotation formalisms. Six of the considered resources have been originally developed as constituency treebanks (namely, Catalan, Icelandic, Indonesian, Portuguese, Spanish and Turkish). Three of them (Dutch, French and Vietnamese) were hybrid resources combining constituency and functional information, whereas the Italian treebank is the only one originally conceived as a dependency treebank.

For what concerns the annotation schema that we consider in our resource, all the Universal Dependencies (UD) annotations (version 2.9, Zeman et al., 2021) were obtained via conversion, carried out automatically, followed by manual revision in many cases. The situation of constituency-based treebanks is more intricate. Two languages (Icelandic and Turkish) were natively annotated using the Penn Treebank (PTB) annotation scheme. French, Indonesian, Italian, Portuguese and Vietnamese treebanks are distributed across various formats and schemas, including the

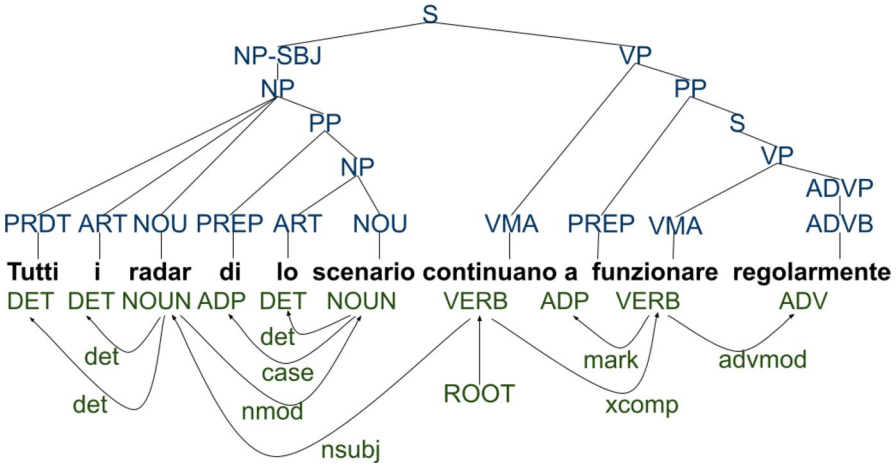

Fig. 2 Constituency (top) and dependency (bottom) tree representing the example sentence (2) extracted from the Turin University Treebank

Penn Treebank one, possibly with minimal language-specific adaptations to handle discontinuous constituents and traces. The remaining three, Catalan, Spanish and Dutch, are annotated following a constituency-based schema inspired by PTB, but with major modifications tailored to handle specific constructions.

Besides the differences at the level of adopted annotation scheme, it is worth mentioning that the treebanks analyzed primarily consist of newspaper articles, with the exceptions being the Icelandic, Italian, and Indonesian treebanks, which also include other text types. This consistency in genre across most of the treebanks is considered important for minimizing genre-specific biases or interference in linguistic studies and applications based on the Parallel Trees resource.

Since this information is crucial for preservation and distribution, Table 1 provides the licensing details for each treebank. It is important to note that while all UD treebanks are freely available (at least for the morpho-syntactic and dependency annotations), some of the constituency-based treebanks may require users to accept a license agreement with the source treebank curators in order to access the data.

3.3 Profiling the parallel trees treebanks

To illustrate the degree of equivalence or divergence between dependency-based and constituency-based representations in the Parallel Trees resource, we analyzed the trees by computing the distribution of a set of selected features. As detailed in Sect. 3.3.1, these features capture structural properties related to two fundamental dimensions of a syntactic tree: i.e. tree width and depth. These properties are independent of specific constituency or dependency labels and can thus be used to contrastively characterize the dependency vs constituency representations for each language.

3.3.1 Linguistic features

In what follows, we introduce the linguistic features and describe how they were computed using the following sample sentence taken from the Italian Turin University Treebank, whose PTB and UD representations are reported in Fig. 2:

(2) *Tutti i radar dello scenario continuano a funzionare regolarmente.* [transl. 'All the radars in the scenario continue to operate normally.']

- **Maximum depth of the tree** (*max-depth*)**:** It represents a global property of the sentence structure. For dependency trees (DTs), the length is computed as the path, in terms of dependency links, from the root to the furthest leaf node. For phrase structure trees (PSTs), it is determined by the longest path from the root to a non-terminal node reporting the part of speech (PoS) of a word. We do not include the path connecting the PoS to its corresponding word (which functions as one of the tree's leaves) as they are positioned at the same level according to the annotation schema. As in both cases, we consider the longest path of the tree, the maximum depth can be assimilated to the tree height. In the UD representation of (2) the value of this feature is equal to 3, corresponding to the three intermediate dependency links that are crossed in the path going from the root of the sentence (*continuano*, *'continue'*) to each of the more distant leaf nodes, represented by two words: di ('in') and lo ('the').[4] When computed on the PST, the feature value is 6, corresponding to the six intermediate nodes between the root node (S) to the adverbial node (ADVB), which is the most distant one.
- **Average length of phrases/dependency links** (*avg-len*)**:** it refers to a local property of a sentence and is computed on the basis of the linear order of elements. Specifically, in DTs it represents the average number of words occurring between a syntactic head and its dependent, excluding punctuation. For PSTs, we followed Nenkova et al. (2009) who computed this feature as "the number of words comprising a given type of phrase, divided by the number of phrases of this type". However, we computed the average length by considering the lengths of all phrase types in a sentence. The average dependency length of example (2) is 1.89, while the average phrase length is 3.55.
- **Maximum phrase/link length** (*max-len*)**:** it complements the *avg-len* feature by referring to the longest dependency link and largest phrase in a DT and PST representation respectively. In (2), the longest link connects 'radar' and *'continuano'* and it is 4-token long, while the largest phrase is 6-token long.[5]
- **Average clause length** (*avg-cl-len*)**:** it is aimed at modeling a local structure of the sentence and it is computed as the average number of tokens per clause. Due to the fact that DTs do not explicitly mark clause boundaries, the feature is calculated in a UD tree as the ratio between the number of tokens in a sentence and

[4] The syntactic head is always marked in italics.

[5] Note that we do not consider the length of the phrase introduced by the root node, labelled as 'S' in example (2), as this would result in the longest phrase always corresponding to the length of the sentence.

the number of verbal and nominal predicates, identified by either a verbal head or a copula. Conversely, the feature is computed on a PST by relying on the non-terminal nodes which are indicated as introducing a clause in the annotation documentation of each treebank.[6] Thus, in the 10-token long sample sentence (2), the average value is 5 when computed on the UD tree since there are two verbal heads ('*continuano*' and '*funzionare*'). When computed on the PST, the average clause length is also 5, although this time the value is obtained by averaging the length of the two phrases introduced by 'S': one is the 10-token long sentence, and the other is the 2-token long phrase '*funzionare regolarmente*'.

3.3.2 Feature distribution in parallel trees

Table 2 and Table 3 report multiple statistics acquired from the constituency and dependency sections of each treebank. For each treebank section, reported statistics refer to *All* sentences, whereas the values in the *Min* and *Max* rows refer respectively to the top 10% of shortest and longest sentences of each treebank. We decided to focus separately on these length-based sentence subsets since the linguistic features considered here are strongly correlated with sentence length. This makes it possible to explore the variation between the sentences with the highest and lowest lengths and the full set. In particular, we computed the average length of the sentences (column *len* in both tables), the average feature values,[7] and the variation coefficient (*cv*), corresponding to the internal variability within each set of sentences in relation to their mean values. We also computed the absolute effect size (r) of the Wilcoxon Signed-Rank Test[8] between paired dependency (DT) and constituency (PST) treebanks to quantify the extent of variation between the two series.

The significance test revealed highly significant differences ($p < 0.001$) across the two formalisms for most languages and features. The absolute effect size scores show that the feature consistently exhibiting the highest average r scores across the ten treebanks is *max-depth*. This indicates a substantial distinction in the values representing the depth of syntactic trees between dependency and constituency representation formalisms. Indeed, the feature distributions in the tables demonstrate that, as expected (see Sect. 2.1), in all languages DTs exhibit lower feature values compared to the corresponding constituency trees. In contrast, *avg-cl-len* displayed the

[6] The complete list of non-terminal node labels introducing a clause is provided for each language below. Additional details about the types of clauses associated with each label, where applicable, can be found in the guidelines specific to each language. Catalan and Spanish: 'sentence', 'S'; Dutch: 'top', 'smain', 'ssub'; French: 'SENT', 'VPinf', 'Sint', 'VPpart', 'Ssub', 'Srel', 'COORD'; Icelanding: 'IP-MAT', 'CP-THT', 'CP-CAR', 'CP-CLF', 'CP-CMP', 'CP-DEG', 'CP-FRL', 'CP-REL', 'CP-QUE', 'CP-ADV', 'CP-EOP', 'CP-TMC'; Indonesian: 'S'; Italian: 'S'; Portuguese: '+fcl', '+icl', '+acl', '+cu'; Turkish: 'S'; Vietnamese: 'S'.

[7] Refer to Appendix B.1 for values plus standard deviations.

[8] The Wilcoxon effect size is a measure of the magnitude of difference between the values of two series. It provides a standardized way to interpret the practical significance of the results, complementing the p-value by quantifying the strength of the observed difference, typically interpreted as $r < 0.3$ small effect, $0.3 < r < 0.5$ moderate effect, and $r >= 0.5$ large effect.

Table 2 For each language and their experimental setting (*Min*, *Max* and *All*) the table reports the average sentence length (*len*), and the average value and variation coefficient (*cv*) of the max-depth and avg-len features computed over the parallel UD dependency (*DT*) and constituency trees (*PST*). We also report the absolute effect size (*r*) of the Wilcoxon Signed-Rank Test ($p < 0.001$). – indicates that the test was non-significant

			Max-depth			Avg-len		
Lang	Set	Len	DT (cv)	PST (cv)	r	DT (cv)	PST (cv)	r
CAT	All	33.79	5.45 (0.37)	15.25 (0.29)	0.86	2.69 (0.25)	3.23 (0.36)	0.37
	Max	69.5	8.03 (0.24)	17.6 (0.23)	0.87	3.43 (0.26)	4.65 (0.26)	0.72
	Min	9.53	2.58 (0.37)	13.55 (0.35)	0.87	1.88 (0.27)	1.74 (0.27)	0.29
DUT	All	19.75	3.83 (0.42)	5.91 (0.49)	0.86	2.98 (0.33)	4.79 (0.45)	0.79
	Max	42.05	6.11 (0.23)	10.07 (0.27)	0.87	3.94 (0.24)	7.76 (0.25)	0.85
	Min	5.44	1.56 (0.45)	1.83 (0.53)	0.54	1.54 (0.42)	1.6 (0.74)	–
FRE	All	30.97	4.91 (0.42)	6.09 (0.47)	0.33	2.78 (0.31)	4.47 (0.4)	0.7
	Max	68.54	7.57 (0.27)	9.69 (0.29)	0.63	3.7 (0.31)	6.69 (0.3)	0.84
	Min	6.42	1.89 (0.46)	2.13 (0.49)	0.56	1.65 (0.42)	2.33 (0.37)	0.68
ICE	All	20.4	4.35 (0.53)	7.41 (0.56)	0.57	2.4 (0.31)	4.83 (0.6)	0.68
	Max	50.47	8.07 (0.27)	12.93 (0.41)	0.7	3.28 (0.25)	8.55 (0.44)	0.82
	Min	4.92	1.45 (0.41)	3.51 (0.69)	0.86	1.54 (0.27)	2.47 (0.72)	0.53
IND	All	26.72	5.16 (0.34)	9.09 (0.41)	0.74	2.64 (0.26)	5.56 (0.33)	0.84
	Max	48.24	6.83 (0.27)	12.2 (0.31)	0.84	3.37 (0.24)	7.48 (0.26)	0.87
	Min	9.38	2.77 (0.35)	3.82 (0.54)	0.72	1.87 (0.23)	3.03 (0.26)	0.85
ITA	All	22.41	4.57 (0.44)	8.65 (0.49)	0.87	2.28 (0.23)	4.82 (0.47)	0.84
	Max	51.01	7.5 (0.22)	14.65 (0.25)	0.87	2.9 (0.15)	7.95 (0.25)	0.87
	Min	6.08	1.82 (0.32)	3.04 (0.28)	0.87	1.67 (0.19)	1.86 (0.36)	–
POR	All	25.44	4.92 (0.44)	5.5 (0.58)	0.14	2.5 (0.28)	5.22 (0.45)	0.82
	Max	60.24	8.21 (0.27)	10.17 (0.32)	0.6	3.3 (0.28)	8.58 (0.29)	0.86
	Min	5.87	2.2 (0.29)	1.67 (0.47)	0.73	1.68 (0.22)	2.52 (0.24)	0.83
SPA	All	32.81	5.47 (0.4)	15.48 (0.3)	0.86	2.67 (0.27)	3.26 (0.39)	0.37
	Max	65.2	8.09 (0.24)	18.02 (0.24)	0.87	3.33 (0.27)	4.76 (0.27)	0.78
	Min	7.53	2.14 (0.42)	13.42 (0.39)	0.87	1.69 (0.34)	1.59 (0.29)	0.26
TUR	All	9.15	3.01 (0.42)	4.68 (0.3)	0.71	2.1 (0.32)	1.81 (0.23)	0.35
	Max	15.05	4.37 (0.26)	5.77 (0.2)	0.74	2.74 (0.21)	2.16 (0.15)	0.69
	Min	3.31	1.13 (0.57)	2.53 (0.28)	0.86	1.04 (0.47)	1.15 (0.19)	0.6
VIE	All	20.62	4.31 (0.44)	5.98 (0.5)	0.79	2.32 (0.3)	4.7 (0.44)	0.85
	Max	45.8	6.89 (0.27)	9.92 (0.3)	0.83	3.17 (0.26)	7.45 (0.3)	0.86
	Min	5.53	1.78 (0.4)	2.23 (0.48)	0.53	1.45 (0.3)	2.25 (0.34)	0.78

lowest effect size scores, or in some cases, non-significant variations. This suggests that there is no strong trend towards higher values in either of the two formalisms.

To normalise the feature values and allow a cross-feature comparison, we computed the differences between the two SRPs as a percentage increase from the lowest value (which in almost all cases corresponds to that of the DTs). These values are reported in Appendix B (see Table 4) and are aligned with the results above. Indeed, the feature

Table 3 Continuing Table 2 for the max-len and avg-cl-len features

			Max-len			Avg-cl-len		
Lang	Set	len	DT (cv)	PST (cv)	r	DT (cv)	PST (cv)	r
CAT	All	33.79	14.72 (0.69)	20.08 (0.66)	0.32	14.25 (0.69)	13.06 (0.56)	0.07
	Max	69.5	31.26 (0.51)	44.16 (0.34)	0.62	19.53 (0.9)	17.85 (0.66)	–
	Min	9.53	4.17 (0.47)	4.64 (0.51)	0.41	7.08 (0.6)	7.45 (0.4)	0.42
DUT	All	19.75	10.85 (0.64)	12.45 (0.69)	0.25	9.9 (0.57)	12.77 (0.52)	0.44
	Max	42.05	22.16 (0.38)	27.84 (0.32)	0.46	12.9 (0.58)	21.54 (0.34)	0.74
	Min	5.44	2.51 (0.54)	2.0 (0.83)	0.45	3.98 (0.73)	4.08 (0.5)	-
FRE	All	30.97	14.79 (0.74)	17.38 (0.74)	0.15	12.28 (0.69)	12.75 (0.61)	0.06
	Max	68.54	33.67 (0.45)	39.68 (0.43)	0.33	16.55 (0.68)	19.45 (0.59)	0.25
	Min	6.42	3.02 (0.57)	3.59 (0.5)	0.52	3.47 (1.04)	4.76 (0.46)	0.43
ICE	All	20.4	8.9 (0.81)	15.04 (0.88)	0.39	7.6 (0.5)	13.47 (0.52)	0.64
	Max	50.47	22.02 (0.48)	36.65 (0.55)	0.59	7.91 (0.31)	21.84 (0.41)	0.84
	Min	4.92	2.31 (0.39)	5.04 (1.31)	0.59	4.09 (0.51)	6.24 (0.8)	0.71
IND	All	26.72	12.85 (0.58)	21.29 (0.53)	0.54	10.53 (0.62)	10.53 (0.54)	–
	Max	48.24	23.9 (0.36)	40.11 (0.29)	0.76	13.67 (0.62)	13.91 (0.51)	–
	Min	9.38	4.4 (0.45)	6.2 (0.43)	0.77	6.27 (0.7)	7.06 (0.54)	–
ITA	All	22.41	9.4 (0.77)	15.55 (0.73)	0.73	9.41 (0.72)	11.37 (0.76)	0.44
	Max	51.01	21.39 (0.41)	37.27 (0.28)	0.81	12.64 (0.47)	19.84 (0.48)	0.76
	Min	6.08	2.42 (0.29)	2.91 (0.36)	0.68	1.4 (1.83)	1.35 (1.83)	-
POR	All	25.44	11.37 (0.78)	13.56 (0.76)	0.17	10.63 (0.69)	9.35 (0.55)	0.1
	Max	60.24	26.76 (0.49)	33.12 (0.37)	0.46	14.51 (0.65)	14.25 (0.39)	0.09
	Min	5.87	2.86 (0.36)	2.97 (0.34)	0.51	3.88 (0.74)	3.44 (0.68)	0.38
SPA	All	32.81	14.36 (0.67)	20.24 (0.64)	0.35	12.21 (0.63)	12.03 (0.54)	–
	Max	65.2	28.5 (0.43)	41.93 (0.31)	0.69	14.88 (0.6)	16.11 (0.49)	0.16
	Min	7.53	3.24 (0.54)	3.78 (0.59)	0.52	5.09 (0.75)	6.15 (0.41)	0.53
TUR	All	9.15	5.68 (0.55)	5.38 (0.49)	0.06	7.12 (0.67)	7.54 (0.5)	0.07
	Max	15.05	10.24 (0.27)	8.35 (0.28)	0.55	12.15 (0.42)	11.14 (0.34)	0.26
	Min	3.31	1.2 (0.58)	1.57 (0.52)	0.76	1.4 (1.32)	1.87 (1.04)	0.7
VIE	All	20.62	9.8 (0.74)	13.8 (0.68)	0.57	5.85 (0.65)	13.58 (0.6)	0.81
	Max	45.8	22.91 (0.44)	31.56 (0.38)	0.55	6.35 (0.62)	22.01 (0.59)	0.85
	Min	5.53	2.3 (0.45)	3.47 (0.39)	0.71	3.19 (0.81)	5.53 (0.25)	0.85

showing the higher percentage differences between the two formalisms is *max-depth* while *avg-cl-len* tends to be the characteristic showing similar trends across the DTs and PSTs.[9] Again, the inherent differences between the two paradigms affect these results: constituency trees are typically deeper due to their nested phrase structures, while dependency trees represent linear word-to-word relationships, resulting in flatter structures. Nested structures in dependency trees can create long dependency links and

[9] The average percentage difference between languages of the features is 78.29 for max-depth, 68.45 for avg-len, 36.45 for max-len, and 29.07 for avg-cl-len.

non-projective structures but do not achieve the same depth as constituency trees. On the other hand, the length of a clause is highly related to the presence of verbs in a sentence. This was also exemplified in Sect. 3.3.1, where the average number of tokens per clause and the ratio between the sentence length (measured in tokens) and the number of verbal heads or copula is the same.

These trends can be observed for the entire treebanks (*All*), but not when we focus on the subsets of shortest (*Min*) and longest (*Max*) sentences. Specifically, the percentage increase of depth (*max-depth*) of PSTs compared to those of DTs is more pronounced in the set of short sentences, while in the case of the longest sentences, the key distinguishing characteristic is *avg-len*. This outcome seems to suggest that both subsets include sentences with structural syntactic properties that markedly differ from those found in the entire treebanks.

Note that these trends differ across the ten languages considered. Catalan and Spanish are the languages with the most marked difference between the depth of the constituent and dependency trees. In the case of Catalan and Spanish, this substantial difference may be attributed to the intermediate empty nodes that represent the internal nested structure of phrases (see Sect. 3.2). However, for most languages, the most distinctive feature is *avg-len*, while Vietnamese is the only language for which DTs and PSTs differ mainly for the average clause length.

The latter observation is supported by an analysis of the coefficient of variation (*cv*), which measures the degree of variation in feature values among sentences within each group. As shown in Tables 2 and 3, even if the coefficient exhibits differences across the *All*, *Min*, and *Max* groups, the set of shortest sentences tend to display higher internal variation in all features and languages with respect to the *cv* values in the *Max* group, which exhibits greater internal consistency.

In conclusion, we observe statistically significant differences between the two SRPs, particularly with respect to specific sentence properties. While this result aligns with expectations, the outcome of our linguistic profiling provides multilingual evidence contributing to the discussion on the topic (see Sect. 2.1). In particular, we empirically demonstrated that, although these differences hold across the ten languages examined, their strength varies from language to language. These findings suggest that, beyond the inherent differences between the two SRPs, the history of the construction of each treebank (i.e. language) plays a role. Indeed, while multilingual representation schemes potentially mask local language differences, thus resulting in a greater similarity across languages, they might still incorporate adjustments to the schema to accommodate language-specific cases. These facts underscore the importance of accounting for both the general differences across paradigms and the specificities of each original treebank when using the Parallel Trees resource for further research or applications.

4 Parallel trees for comparing syntactic representation paradigms

In order to show possible usages of the Parallel Trees resource, we present here the results of a case study carried out within the broader field of interpretability research. The goal is to demonstrate how the newly developed resource can serve as a benchmark for assessing the linguistic abilities of BERT, one of the most widely

influential neural language models. In particular, the case study shows how the Parallel Trees resource can be used to address the following research questions: does BERT encode dependency- and constituency-based syntactic representations in distinct ways, revealing sensitivity to the two syntactic representation paradigms? Moreover, are specific linguistic phenomena represented differently across SRPs also reflected in BERT's sentence embeddings?

4.1 Experimental setting

4.1.1 Probing tasks

We adopted the approach of linguistic probing tasks defined in Conneau et al. (2018) and subsequently tailored in Miaschi et al. (2022). Note that the latter approach has proven to be robust for these types of assessments since it does not rely on spurious signals unrelated to the linguistic properties under consideration to solve the probing tasks. Consequently, there was no necessity to introduce control tasks aimed at determining whether probing tasks might inadvertently extract information about the NLM representation solely through the probe's ability to discern surface patterns in the data, as introduced by Hewitt and Liang (2019).

For the specific purpose of our case study, we designed four probing tasks, each focused on predicting the values of one of the four linguistic properties described in Sect. 3.3.1. This selection is motivated by the work of Miaschi et al. (2020), which demonstrated that these tasks are reliable for assessing BERT's multi-level linguistic abilities. Additionally, as discussed in Sect. 3.3.1, these properties capture significant differences between dependency- and constituency-based sentence representations, making them an ideal testbed for investigating the model's sensitivity to diverse syntactic representation paradigms. Given these variations, we can hypothesize that the model's performance may vary depending on its sensitivity to these distinct SRPs.

4.1.2 Models

Neural language model All the experiments rely on the pre-trained cased version of Multilingual BERT. This model is trained on the concatenation of monolingual Wikipedia dumps of 104 languages, which includes the 10 languages of the Parallel Trees resource. The pre-trained model used in this experiment is available through the Huggingface library (Wolf et al., 2020).[10] In our experiments, the sentence-level embeddings for each of the 12 layers are obtained using the activation of the first input token (*[CLS]*), which somehow summarizes the information from the actual tokens, as demonstrated among others by Jawahar et al. (2019) and Wagner and Zarrieß (2023).

Probing model We used a linear support vector regression (LinearSVR) as a probing model. The model takes as input layer-wise sentence-level embeddings extracted from the pre-trained version of BERT and it predicts the value of each considered feature in the dependency and constituency treebanks. In addition, we also experimented

[10] https://huggingface.co/bert-base-multilingual-cased

with a probing model that takes as input the sentence representations of a BERT model with randomly initialized weights. Our intuition is that in principle the randomized BERT should not possess any inherent linguistic knowledge, thus it could be less able to predict the considered syntactic characteristics of a sentence. It is intended to serve as a point of comparison, offering insights into how effectively the pre-trained BERT implicitly encodes linguistic competence. However, it should be noted that in this study we do not intend to compare the two probing models in terms of performance, but we are rather interested in inspecting how the pre-trained BERT representations implicitly encode specificities of one of the two considered SRPs.

In all cases, as an evaluation metric, we used the Spearman correlation coefficient between the values of the linguistic features predicted by the model and their actual gold values. In the remainder of the paper, we refer to the evaluation metric as *probing score*.

For each language, we trained and tested the probing models on DTs and PSTs adopting a cross-validation process. Specifically, we split each treebank into five portions containing the same amount of randomly selected sentences; then, we iteratively trained the probing models on four portions and used the remaining fifth as the test set. This way, the models are trained using a representative sample of the treebank at each iteration. We refer to this scenario as *All*, since it includes the full set of sentences from each treebank.

Additionally, we tested the probing models on two conditional settings, focusing on subsets of the entire sentence set. Specifically, we considered the *Min* and *Max* subsets introduced in Sect. 3.3.2, which contain the top 10% of shortest and longest sentences from each treebank. In both settings, the probing models are trained on the remaining 90% of the treebank. Building on the findings of Miaschi et al. (2023), who demonstrated that sentence length can act as a confounding factor that biases the true estimate of BERT's linguistic abilities, we hypothesize that these subsets may be particularly valuable for evaluating the model's robustness when faced with sentences that deviate significantly from the linguistic norms of the language under consideration. It is expected that these short and long sentences are also underrepresented in the pre-training data for BERT, making them particularly insightful.

4.2 Is dependency and constituency information differently encoded?

In this section, we investigate whether information encoded in the dependency- and constituency-based sentence representation within the Parallel Trees resource is differently encoded in BERT's embeddings.

As a first step, we focus on the full set of sentences for each language and we explore the results of the probing tasks performed by the two probing models, namely the one informed with pre-trained BERT representations and the one using BERT representations with randomly initialised weights.

Figure 3 shows the layer-wise average probing scores for each language across the constituency and dependency treebanks.[11] Overall, the models exhibit similar

[11] Refer to Appendix C (see Fig. 6) for the layer-wise probing scores for individual features and languages obtained by the pre-trained and random models.

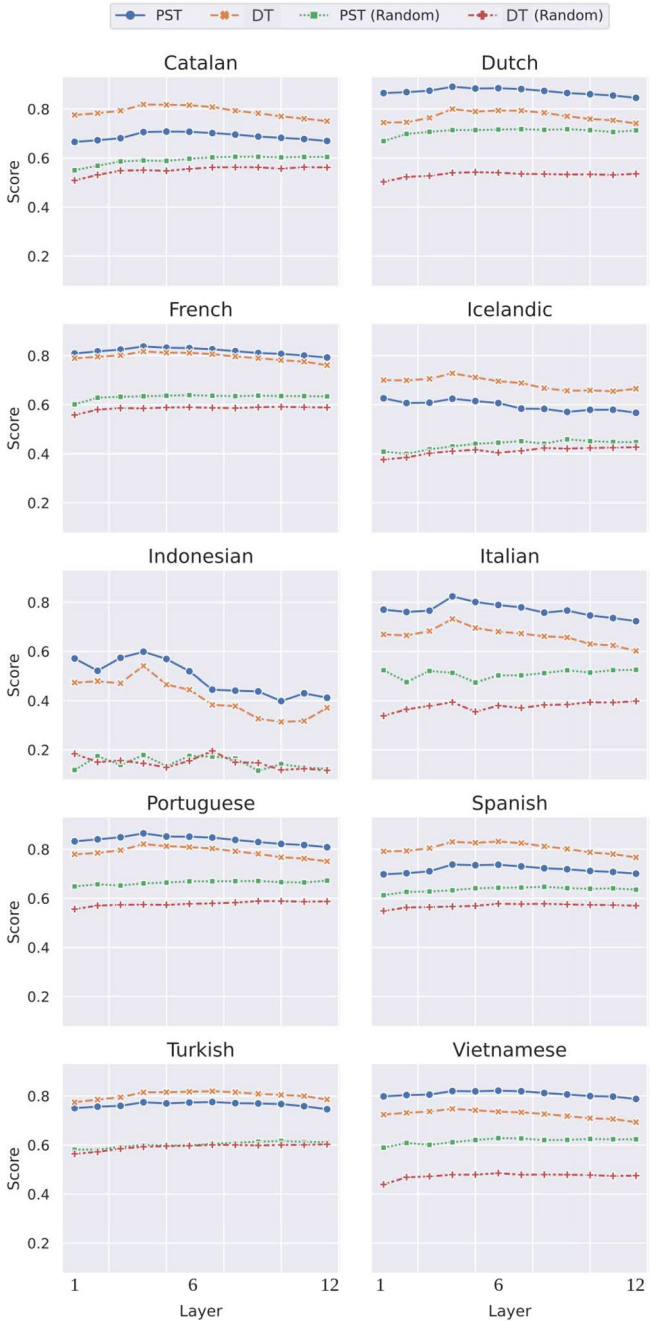

Fig. 3 Layer-wise probing scores (Spearman correlation coefficients) obtained on constituency (PST) and dependency (DT) treebanks of each language using both the pre-trained and randomized BERT sentence representations. Scores are computed for the whole set of sentences

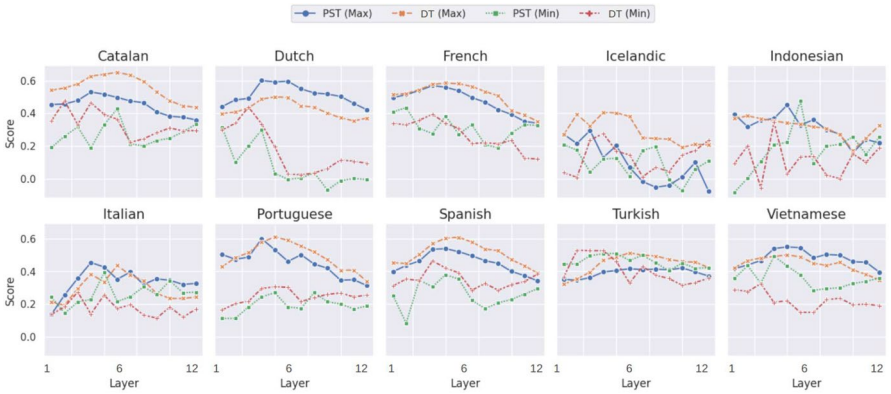

Fig. 4 Layer-wise probing scores on constituency (PST) and dependency (DT) treebanks of each language. Scores are computed for the shortest (*Min*) and longest (*Max*) sets

cross-layer scores, indicating stable syntactic knowledge across layers. However, small differences can be observed in the cases of Italian and Indonesian treebanks that show slightly higher scores in the middle layers. It should be noted that the Indonesian treebanks CSUI and Ketu are the smallest in our set. The limited size of these treebanks possibly contributes to the higher cross-layer variation.

The comparative analysis of the two probing models yields three main results. First, the pretrained model performs better on constituency trees than dependency trees for 6 (out of 10) languages, i.e. Dutch, French, Indonesian, Italian, Portuguese, and Vietnamese. This suggests that, for these languages, BERT better predicted linguistic features in the constituency format. Interestingly, only four languages exhibit higher scores for dependency representations: Turkish, Icelandic, Catalan and Spanish. The latter two were both produced in the context of the AnCora project, thus suggesting a particular challenge in processing the AnCora phrase structure representation. Notably, the difference between the SRPs is minimal, in particular for Turkish, French, and Portuguese, which show the most similar results between the two SRPs. Second, random BERT representations consistently show lower probing scores compared to pre-trained models for both SRPs, although the differences between models are small. The results confirm our intuition that randomized BERT is less able to predict the syntactic characteristics of a sentence, possibly because pre-trained representations encode more linguistic information. However, when we focus on the sensitivity of the random model, we notice that it consistently performs better on constituency trees than dependency trees, though the difference between the SRPs is minimal for Turkish and Indonesian.

Min and Max settings comparison. This analysis focuses only on the probing model scores of pre-trained BERT embeddings. As shown in Fig. 4, the *Min* and *Max* subsets consistently achieve lower scores than the *All* subset across all languages, with variations across layers converging in the output ones. This is likely because these sentences are rarely encountered during BERT's pre-training, possibly impacting the model's ability to encode their linguistic information. However,

we also hypothesize that this variation in performance may be due to the mismatch between the training and test sets in the *All* and the two conditional experimental scenarios. This hypothesis is motivated by the fact that the longest and shortest sentences often display syntactic feature distributions that deviate from the average, as shown in Tables 2 and 3. We recall that the *All* setting used a five-fold cross-validation process, in which the complete treebank was iteratively partitioned so that each portion served once as a test set while the remaining four-fifths were used for training. In contrast, the training sets for the *Min* and *Max* subsets were created by excluding the 10% shortest and longest sentences from the complete treebank.

To investigate this hypothesis, we analyzed the distribution of probed features within the training sets of the *Min* and *Max* subsets. Specifically, we assessed the relative variation scores for each language, which represent the percentage difference between the average feature values of the complete treebanks and those in the conditional training sets. These scores are detailed in Appendix B.1 (see Table 5). The analysis confirmed low variation scores across all languages, indicating that the feature distributions in the *Min* and *Max* training sets largely align with those in the complete treebanks. The differences are more noticeable for PSTs than for DTs, particularly in the *Max* setting, which excludes the longest sentences from the train set. However, these variations are modest, averaging −7.17% for dependency trees and −8.42% for constituency trees.[12] This underscores that the experiment is not significantly biased, and the differences in probing scores across subsets are mainly due to the linguistic information encoded in the NLM embeddings.

Considering again Fig. 4 in light of these results, we note that the probing model shows a stronger ability to predict values for the longest sentences compared to the shortest ones across both SRPs. Additionally, unlike the *All* setting, the constituency-based formalism is not best predicted for most languages in the *Max* setting. In fact, for 7 out of 10 languages (Catalan, French, Icelandic, Indonesian, Portuguese, Spanish, Turkish), the dependency SRP achieves higher results. Notably, for French and Portuguese, the phrase-based formalism performs better on the entire treebank but, on the longest sentences, BERT shows higher performance with the dependency formalism. For what concerns the shortest sentences, the better predicted SRP varies depending on the layer. This result might be expected because, as we noted in Sect. 3.3.2, the set of shortest sentences displays greater internal variation in the values of the linguistic phenomena considered, compared to the set of longest sentences.

4.3 Focusing on individual features

In this section, we focus on individual probing tasks and investigate whether BERT's sentence embeddings show sensitivity to how SRPs represent specific linguistic phenomena. To this aim, we focus on the probing scores obtained relying on the embeddings of the output layer: they are better suited for Masked Language Modeling (MLM) (Jawahar et al., 2019), and we showed in Sect. 4.2 that the scores

[12] Negative variation values indicate that the feature values are lower in the training sets than in the complete treebanks.

do not vary across layers. Results of the score differences between dependency and constituency treebanks are reported in Fig. 5 for each language and feature. Negative differences (blue shades) indicate higher probing scores on constituency treebanks, while a score difference near zero indicates a similar prediction accuracy across the two SRPs.

The most distinctive features. Although the reported differences are generally minimal, the feature with the highest probing accuracy varies depending on the testing scenario, as indicated by the average scores across languages shown in the last row of each heatmap. Specifically, it is the depth of the syntactic tree (*max-depth*) when BERT is tested on the full set of sentences (*Whole treebank* heatmap), with an average difference of 0.1; it is *avg-cl-len*, with an average difference of 0.11, when tested on the longest sentences (*Longest sentences*); and it is *avg-len*, with an average negative difference of −0.11, when focusing on the shortest sentences (*Shortest sentences*). In two of the three testing scenarios, BERT shows a higher ability to encode the dependency-based representation of the feature. Note, however, that these average differences are also influenced by the specific languages. Specifically, the higher scores for the dependency-based representation of **syntactic tree depth are largely driven by the substantial differences observed for Catalan and Spanish**, both of which report higher probing scores for the D-SRP. As discussed in Sect. 3.3.2, these two languages are characterized by the high recurrence of deeper internal structures. BERT shows sensitivity to how this feature is differently represented on the basis of the two SRPs, performing worse when tree depth refers to non-terminal nodes rather than to direct links between words. However, considerable differences are also observed for Icelandic, with a difference of 0.22. Interestingly, for other languages, the differences are smaller and sometimes negative, indicating that BERT achieves higher accuracy when encoding the phrase-based representation of this feature. The **higher probing scores for the D-SRP of average clause length in the longest sentences** are mainly tied to the differences observed for Icelandic (0.45) and Indonesian (0.23). Notably, the D-SRP of this feature achieves higher scores across all languages, though the magnitude of the differences varies. This suggests that BERT finds it more challenging to generalize when non-lexical elements, such

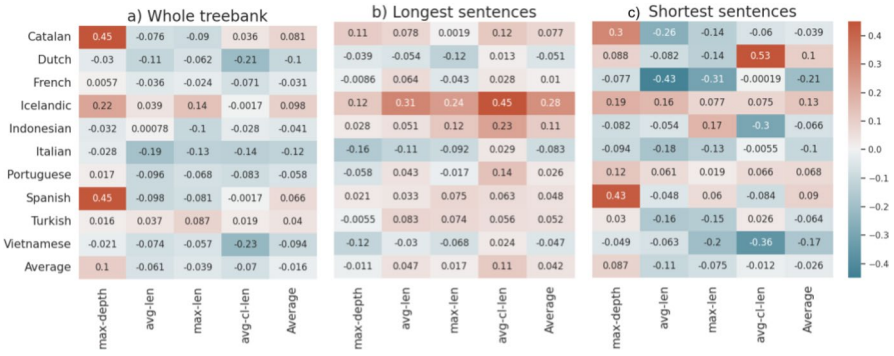

Fig. 5 Differences between the probing scores (12th layer) obtained on DT and PST treebanks computed on each **a** whole treebank, and for the **b** Max and **c** Min subsets

as clause boundaries, are added to the syntactic representation of particularly long sentences. Conversely, the **model's higher ability to encode average length in the phrase-based representation** is mostly due to the higher probing scores achieved for French (with a difference of −0.43 compared to the D-SRP) and Catalan (with a difference of −0.26). This trend is consistent across most languages, with Icelandic and Portuguese being the exceptions.

Differences across languages. The model's sensitivity, again, turned out to vary depending on the testing scenario. A notable example is the Dutch language. In this case, BERT demonstrates a stronger ability to encode phrase-based representations when tested on the full set of sentences and the longest ones, with average score differences of −0.1 and −0.051, respectively (see the *Average* column in the heatmaps). In contrast, when tested on the shortest sentences, higher probing scores are achieved for the D-SRP, with a positive difference of 0.1. However, these average differences are driven by different features. For instance, the higher scores for the C-SRP are largely attributed to BERT's ability to encode *avg-cl-len* when considering the whole treebank, with a difference of −0.21, but to encode *max-len* when considering the longest sentences, with a difference of −0.12. Conversely, the positive average difference of 0.1 for the shortest sentences is mainly due to BERT's higher ability in representing *avg-cl-len*, with a substantial difference of 0.53. Notably, this is the same feature for which we observed an opposite trend when considering the whole treebank. This confirms a recurring observation in our study: sentence length, and consequently the standard distribution of the linguistic phenomena mostly influenced by length, plays a significant role in testing the multi-level linguistic abilities of a neural language model.

4.4 Discussion

The case study showed how the novel Parallel Trees resource can be employed to investigate BERT sensitivity to SRPs. In particular, the study addressed two research questions: (1) whether BERT encodes different syntactic representation paradigms in distinct ways, and (2) whether the model's varying sensitivity is particularly tied to specific linguistic phenomena.

Regarding the first research question, our findings suggest that BERT is sensitive to different syntactic formalisms, encoding constituency- and dependency-based representations in distinct ways. However, the differences in the variation of probing scores averaged across probing tasks are generally minimal and fluctuate depending on the language. For example, we observe higher probing scores for dependency-based representations in Icelandic, Spanish and Catalan, whereas phrase-based representations led to better results in French, Italian, and Vietnamese. The preference for a specific formalism also varies with sentence length, indicating that BERT's linguistic competence tends to decline when constrained by this raw textual variable. Specifically, while the D-SRP is generally preferred in the longest sentence setting, in the shortest sentence conditional scenario PST is the slightly better predicted SRP, although this preference varies depending on the layer. However, these trends are not uniform across all languages.

These results diverge from previous studies, such as Kulmizev et al. (2020), which found different tendencies. This work, which compared BERT and ELMO on Surface Universal Dependencies (SUD) and Universal Dependencies (UD) representations, found a more consistent preference for the UD formalism across models, layers, and 13 languages, which is surprising given that the theoretical foundations underlying UD and SUD are less divergent than those assumed by dependency and constituency-based representations. The authors also suggest that language typology played a role, with morphologically rich languages favouring shallow dependency structures. Unlike Kulmizev et al. (2020), our study does not show a strong typological effect or that there is a strong preference for one SRP. Furthermore, we have not observed similar trends within closely related languages, like the Romance group: while Catalan and Spanish show a preference for DTs, French, Italian, and Portuguese are generally better represented using phrase structures.

Regarding the second research question, which investigates the model's sensitivity to specific linguistic properties, we observed that BERT's ability to encode the four considered features varies depending on the testing scenarios, with respect to sentence length. Specifically, we found higher probing scores for the dependency-based representation of features where the addition of non-lexical nodes, characteristic of the phrase-based representation, makes encoding more challenging. This is evident in the prediction of syntactic tree depth (*max-depth*) and average clause length (*avg-cl-len*). However, this trend varies according to sentence length. For instance, the difference between probing scores for dependency and constituency trees is more pronounced when BERT is tested on the whole treebank, while predicting clause length becomes more challenging when the model is tested on the longest sentences. We also observed that, for certain languages, this trend aligns with the distribution of feature values in the Parallel Trees resource, as seen with *max-depth* for Catalan and Spanish. However, this pattern is not consistent across all languages and testing scenarios.

At first sight, the complex and articulated evidence emerging from this case study for both investigated research questions might appear inconclusive for what concerns BERT sensitivity to SRPs. On the contrary, it is worth noting that these results may be read from two different perspectives. On the one hand, they show that the linguistic generalization abilities of the tested NLM vary depending on the SRP, thus suggesting that the model is sensitive to the specificities of the two paradigms. Nevertheless, they highlight the need to further investigate the conditions under which BERT implicitly encodes dependency-based or constituency-based syntactic representations more effectively. On the other hand, our findings may contribute to the ongoing debate within the linguistic community about the relationship between the two approaches to encoding the syntactic structure of a sentence. Specifically, our results appear to be in line with the hypothesis put forward by Nefdt and Baggio (2023), who suggest that both constituency-based and dependency-based frameworks contribute to model human syntactic competence and claim that an integrative approach is rather required to achieve full explanatory adequacy.

5 Conclusion

In this paper, we introduced Parallel Trees, a novel resource consisting of 87,376 paired syntactic trees representing identical sentences annotated according to both dependency-based and constituency-based syntactic representation paradigms. The resource covers 10 different languages and results from aligning 20 monolingual treebanks derived from international initiatives such as the Universal Dependencies and the Penn Treebank Project, as well as language-specific projects. We thoroughly identified pairs of identical sentences for each language and retrieved their syntactic trees, providing both dependency-based and phrase-based representations. In essence, Parallel Trees offers a resource that allows for direct comparison of syntactic phenomena as represented in the two SRPs.

The resource is unique in its ability to offer such direct comparisons, providing valuable insights into how syntactic representation paradigms grounded on distinct yet complementary principles encode linguistic information. By profiling the trees according to four linguistic features that capture key properties of syntactic structures, we showed how the two SRPs formalise these features differently. Additionally, the resource exemplifies the effective reuse of pre-existing linguistic resources to create new, valuable datasets. This approach offers several advantages, such as ensuring a high standard of linguistic annotation quality, as the original treebanks have already undergone extensive validation and have been widely used in previous research, and facilitating the comparability of results across different studies and models. Future developments of the resource could focus on expanding Parallel Trees to include more languages. Although identifying further parallel representations might be challenging due to license restrictions and lack of resource curation, such an extension would create the prerequisites for further research on multilingual NLP and cross-lingual syntactic representation.

To illustrate the effectiveness and usefulness of Parallel Trees, we carried out a case study based on the diagnostic probing approach aimed at examining the sensitivity of BERT, one of the first prominent NLMs, to the two syntactic representation paradigms covered by the resource. The results indicate that BERT is sensitive to different approaches to syntactic representation. However, further investigation is needed to determine the conditions under which the model's linguistic generalization abilities vary based on the type of syntactic representation. Further developments of our probing experiments could include additional tasks to test language models' sensitivity to different syntactic paradigms. For instance, an interesting direction would be to investigate whether models can effectively resolve complex linguistic phenomena, such as prepositional phrase attachment, when represented in dependency trees versus phrase structure trees.

Looking forward, resources like Parallel Trees have the potential for a wide range of applications. They can serve as benchmarks for investigating how neural language models like BERT and its successors handle diverse syntactic

formalisms, as shown in our case study. Such resources can also contribute to the broader question of whether there is a linguistic framework that best supports pre-trained NLMs across different downstream tasks, as suggested by studies like Prange et al. (2022). Additionally, Parallel Trees can be valuable for purely linguistic studies focused on comparing syntactic representation strategies for various linguistic phenomena across languages. For example, the resource may facilitate the analysis of structural differences, such as non-projective dependencies in UD versus trace annotations in PSTs. Its parallel nature enables investigations into whether long dependency links align with corresponding phrase-based representations, particularly in cases where traces are either annotated or not. In a multilingual scenario, Parallel Trees can support comparative studies on how phrase-based representations of verbal clauses differ across languages. These studies might focus on identifying which languages distinguish between auxiliaries, modal verbs, and lexical verbs, or investigate whether auxiliaries are treated as heads of verbal projections. Such analyses can shed light on language-specific strategies for representing this typology of clauses and further enhance our understanding of variations in syntactic representation across languages.

Appendix A Links to resources

A.1 Dependency treebanks

- Universal Dependencies Treebanks, version 2.9 (Zeman et al., 2021): AnCora-ca (Catalan), Alpino (Dutch), FTB (French), Modern (Icelandic), CSUI (Indonesian), ISDT (Italian), Bosque (Portuguese), AnCora-es (Spanish), Turkish-Penn (Turkish) and VTB (Vietnamese) treebanks. Download from http://hdl.handle.net/11234/1-4611

A.2 Constituency treebanks

- Catalan and Spanish: AnCora corpus, version 2.0.0. Download from http://clic.ub.edu/corpus/en
- Dutch: Alpino Treebank. Download from https://www.let.rug.nl/~vannoord/trees/

- French: French Treebank FTB, Penn Treebank format. Download from http://ftb.llf-paris.fr/telecharger.php?langue=en
- Icelandic: Icelandic Parsed Historical Corpus (IcePaHC) corpus, version 0.9. Download from https://linguist.is/wiki/index.php?title=Icelandic_Parsed_Historical_Corpus_(IcePaHC)
- Indonesian: Kethu Treebank, version 2.0. Download from https://github.com/ialfina/kethu
- Italian: Turin University Treebank (TUT), TUT-Penn format. Download from http://www.di.unito.it/~tutreeb/treebanks.html
- Portuguese: Bosque Portuguese treebank, version 8.0. Download from https://www.linguateca.pt/Floresta/info_floresta_English.html
- Turkish: Turkish Penn Treebank. Download from https://github.com/olcaytaner/TurkishAnnotatedTreeBank-15
- Vietnamese: Vietnamese treebank. Download from https://vlsp.hpda.vn/demo/?page=resources

Appendix B Profiling treebanks

The treebank analyses reporting the gold values of the structural features (see Table 4) are available in the Open Collection of the CLARIN repository and the following webpage: http://www.italianlp.it/resources/.

Table 4 Average values and standard deviations of 4 structural features computed on the UD dependency trees (DT) and constituency trees (PST) of each language. Values refer to the all treebank ('ALL') and to the subsets comprising the 10% of sentences with maximum ('MAX') and minimum ('MIN') length. The 'Diff' columns report the average differences between the two SRPs as a percentage increase from the lowest value

Lang	Max-depth			Avg-len			Max-len			Avg-cl-len		
ALL												
	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff
CAT	5.45 ±2.02	15.25 ±4.37	179.8	2.69 ±0.67	3.23 ±1.16	20.07	14.72 ±10.18	20.08 ±13.24	36.41	14.25 ±9.88	13.06 ±7.32	8.35
DUT	3.83 ±1.61	5.91 ±2.93	54.31	2.98 ±1	4.79 ±2.15	60.74	10.85 ±6.98	12.45 ±8.57	14.75	9.90 ±5.69	12.77 ±6.59	28.9
FRE	4.91 ±2.08	6.09 ±2.87	24.03	2.78 ±0.86	4.47 ±1.80	60.79	14.79 ±10.98	17.38 ±12.92	17.51	12.28 ±8.46	12.75 ±7.81	3.83
ICE	4.35 ±2.29	7.41 ±4.18	70.34	2.40 ±0.74	4.83 ±2.89	101.3	8.90 ±7.21	15.05 ±13.28	68.99	7.60 ±3.78	13.47 ±6.96	77.24
IND	5.16 ±1.77	9.10 ±3.70	76.16	2.64 ±0.69	5.57 ±1.85	110.6	12.86 ±7.49	21.31 ±11.36	65.68	10.53 ±6.57	10.54 ±5.64	0
ITA	4.56 ±2.02	8.64 ±4.24	89.28	2.28 ±0.54	4.81 ±2.24	111.4	9.39 ±7.19	15.53 ±11.27	65.43	9.41 ±6.78	11.37 ±8.66	20.83
POR	4.92 ±2.19	5.50 ±3.19	11.79	2.50 ±0.69	5.22 ±2.35	108.8	11.37 ±8.84	13.57 ±10.30	19.26	10.63 ±7.30	9.35 ±5.18	12.04
SPA	5.47 ±2.20	15.48 ±4.65	183	2.67 ±0.73	3.26 ±1.26	22.09	14.36 ±9.62	20.24 ±12.96	40.95	12.21 ±7.74	12.03 ±6.45	1.47
TUR	3.01 ±1.27	4.68 ±1.39	55.48	2.10 ±0.68	1.81 ±0.42	13.81	5.68 ±3.11	5.38 ±2.64	5.28	7.12 ±4.74	7.54 ±3.75	5.90
VIE	4.31 ±1.89	5.98 ±2.97	38.75	2.32 ±0.69	4.69 ±2.07	102.6	9.78 ±7.25	13.77 ±9.41	40.82	5.85 ±3.80	13.55 ±8.10	132.1
	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff
MAX												
CAT	8.04 ±1.94	17.6 ±4.12	119.2	3.43 ±0.9	4.65 ±1.23	35.57	31.21 ±15.9	44.2 ±15.22	41.27	19.53 ±17.49	17.82 ±11.72	8.60
DUT	6.11 ±1.43	10.07 ±2.75	64.81	3.94 ±0.96	7.77 ±1.95	96.95	22.12 ±8.41	27.87 ±8.8	25.63	12.87 ±7.52	21.56 ±7.34	66.98
FRE	7.58 ±2.07	9.69 ±2.82	28.01	3.7 ±1.14	6.69 ±2.03	80.81	33.6 ±15.16	39.67 ±16.96	17.85	16.57 ±11.26	19.47 ±11.59	17.52
ICE	8.05 ±2.16	12.97 ±5.27	60.22	3.29 ±0.81	8.56 ±3.74	160.7	22.02 ±10.67	36.58 ±20.15	66.44	7.89 ±2.4	21.87 ±9.02	176.1
IND	6.83 ±1.82	12.2 ±3.78	78.62	3.37 ±0.81	7.48 ±1.95	121.9	23.9 ±8.63	40.11 ±11.74	67.82	13.67 ±8.48	13.91 ±7.16	1.76
ITA	7.48 ±1.63	14.6 ±3.6	95.33	2.9 ±0.41	7.96 ±2.02	174.1	21.55 ±8.6	37.55 ±10.29	74.24	12.64 ±5.92	19.99 ±9.65	56.96
POR	8.21 ±2.25	10.18 ±3.31	23.87	3.3 ±0.93	8.59 ±2.53	160	26.74 ±13.21	33.12 ±12.37	23.77	14.45 ±9.33	14.27 ±5.53	1.79

Table 4 (continued)

MAX	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff
SPA	8.1 ±1.96	18.04 ±4.36	122.7	3.33 ±0.88	4.76 ±1.31	42.94	28.52 ±12.2	41.93 ±13.18	47.12	14.85 ±8.87	16.08 ±7.83	8.27
TUR	4.36 ±1.15	5.75 ±1.17	32.04	2.75 ±0.57	2.16 ±0.33	21.17	10.27 ±2.78	8.34 ±2.36	18.46	12.11 ±5.15	11.12 ±3.75	8.31
VIE	6.89 ±1.88	9.92 ±2.98	43.98	3.17 ±0.82	7.45 ±2.23	135	22.94 ±10.07	31.49 ±11.9	37.76	6.33 ±3.82	21.9 ±13.0	246.6
MIN	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff	DT	PST	Diff
CAT	2.58 ±0.95	13.58 ±4.72	425.2	1.88 ±0.5	1.74 ±0.47	7.45	4.18 ±1.99	4.63 ±2.36	11.27	7.1 ±4.25	7.48 ±3.04	5.23
DUT	1.56 ±0.69	1.84 ±0.98	17.31	1.54 ±0.64	1.6 ±1.18	3.90	2.51 ±1.35	1.99 ±1.66	20.32	3.99 ±2.92	4.09 ±2.05	2.51
FRE	1.89 ±0.87	2.13 ±1.05	12.70	1.65 ±0.7	2.34 ±0.87	41.21	3.01 ±1.72	3.59 ±1.8	18.87	3.46 ±3.58	4.78 ±2.2	37.9
ICE	1.44 ±0.58	3.53 ±2.43	142.1	1.54 ±0.42	2.47 ±1.78	60.39	2.31 ±0.92	5.05 ±6.6	118.2	4.03 ±2.1	6.25 ±4.95	52.57
IND	2.77 ±0.97	3.78 ±2.0	37.91	1.87 ±0.43	3.04 ±0.8	62.03	4.37 ±1.96	6.14 ±2.57	40.91	6.35 ±4.51	7.16 ±3.87	12.60
ITA	1.82 ±0.59	3.03 ±0.82	67.03	1.67 ±0.31	1.86 ±0.65	11.38	2.41 ±0.69	2.9 ±1.04	20.25	1.36 ±2.53	1.36 ±2.5	3.57
POR	2.2 ±0.62	1.66 ±0.79	24.09	1.67 ±0.37	2.51 ±0.6	50	2.86 ±1.03	2.94 ±1.01	3.85	3.83 ±2.86	3.39 ±2.33	11.34
SPA	2.14 ±0.91	13.43 ±5.22	527.1	1.69 ±0.57	1.59 ±0.46	5.92	3.24 ±1.76	3.78 ±2.22	16.67	5.09 ±3.83	6.15 ±2.52	20.83
TUR	1.12 ±0.64	2.52 ±0.7	123.9	1.04 ±0.49	1.15 ±0.21	10.58	1.2 ±0.69	1.57 ±0.8	30.83	1.39 ±1.84	1.87 ±1.94	33.57
VIE	1.78 ±0.7	2.22 ±1.05	25.28	1.45 ±0.43	2.25 ±0.78	55.17	2.31 ±1.04	3.46 ±1.36	50.87	3.2 ±2.58	5.53 ±1.38	73.35

B.1 Relative variations between train sets

Table 5 Relative variation (%) between the average feature values between the whole treebank and the training sets used in the Min (*Train (Min)*) and Max (*Train (Max)*) conditional settings for each language for dependency (DT) and constituency (PST) trees. Note that negative variation scores indicate that the feature value is greater in the full treebank in contrast to the training set. Conversely, positive values indicate that the training set exhibits higher feature values when compared to the entire treebank

		Max-depth		Avg-len		Max-len		Avg-cl-len	
		DT	PST	DT	PST	DT	PST	DT	PST
CAT	Train (Min)	5.55	1.23	3.24	5.00	7.36	7.85	5.25	4.53
	Train (Max)	−5.62	−1.73	−3.07	−5.21	−14.2	−15.4	−4.32	−4.23
DUT	Train (Min)	6.13	7.22	5.1	6.99	7.82	8.52	6.16	7.06
	Train (Max)	−7.28	−8.44	−3.83	−7.4	−13.14	−15.92	−3.56	−8.22
FRE	Train (Min)	6.3	6.74	4.14	4.89	8.14	8.09	7.39	6.52
	Train (Max)	−6.51	−7.03	−4.12	−5.92	−16.55	−16.64	−3.98	−6.16
ICE	Train (Min)	6.85	5.61	3.61	5.11	7.58	6.87	4.88	5.67
	Train (Max)	−10.69	−8.97	−4.35	−9.28	−19.46	−18.99	−0.4	−7.42
IND	Train (Min)	4.8	6.1	2.94	4.96	6.75	7.27	4.27	3.48
	Train (Max)	−3.82	−3.89	−3.53	−3.93	−10.49	−10.83	−3.44	−3.74
ITA	Train (Min)	6.16	6.69	2.98	6.23	7.66	8.26	8.64	8.89
	Train (Max)	−7.78	−8.4	−2.7	−7.83	−16.48	−18.43	−3.98	−9.01
POR	Train (Min)	5.75	7.25	3.47	5.43	7.64	8.01	6.59	6.59
	Train (Max)	−7.89	−10.22	−3.73	−7.63	−17.7	−19.05	−4.22	−6.13
SPA	Train (Min)	6.34	1.4	3.61	5.23	7.95	8.29	6.08	5.13
	Train (Max)	−5.6	−1.91	−3.09	−5.5	−12.28	−13.52	−2.52	−3.89
TUR	Train (Min)	6.52	4.88	5.41	3.72	8.09	7.4	8.13	7.71
	Train (Max)	−5.24	−2.63	−3.45	−2.26	−9.86	−6.53	−8.54	−5.6
VIE	Train (Min)	6.1	6.56	4.13	5.43	7.81	7.69	4.88	6.22
	Train (Max)	−6.95	−7.94	−4.04	−7.06	−17.51	−16.75	−0.86	−7.35

Appendix C Probing scores on individual features

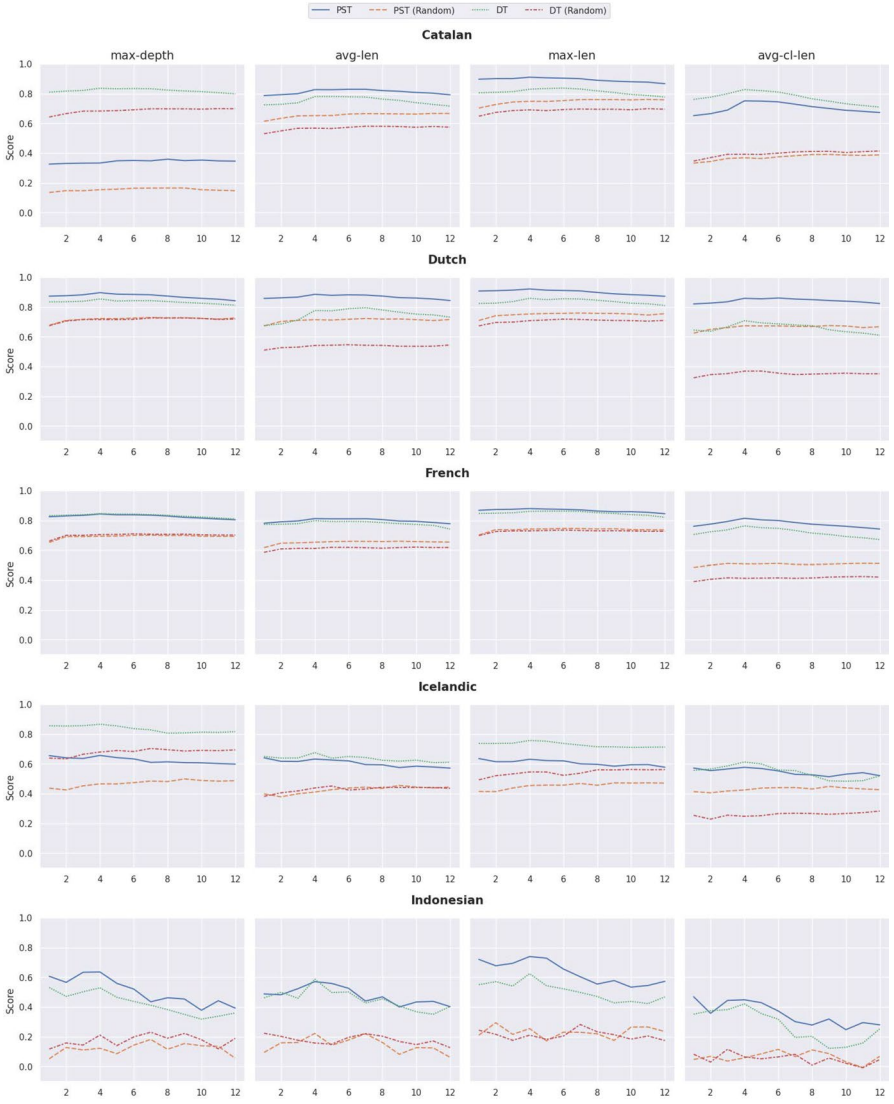

Fig. 6 Layer-wise probing scores (Spearman correlation coefficients) obtained for each probed feature on constituency (PST) and dependency (DT) treebanks of each language using both the pre-trained and randomized BERT sentence representations. Scores are computed for the whole set of sentences. Continues in next page

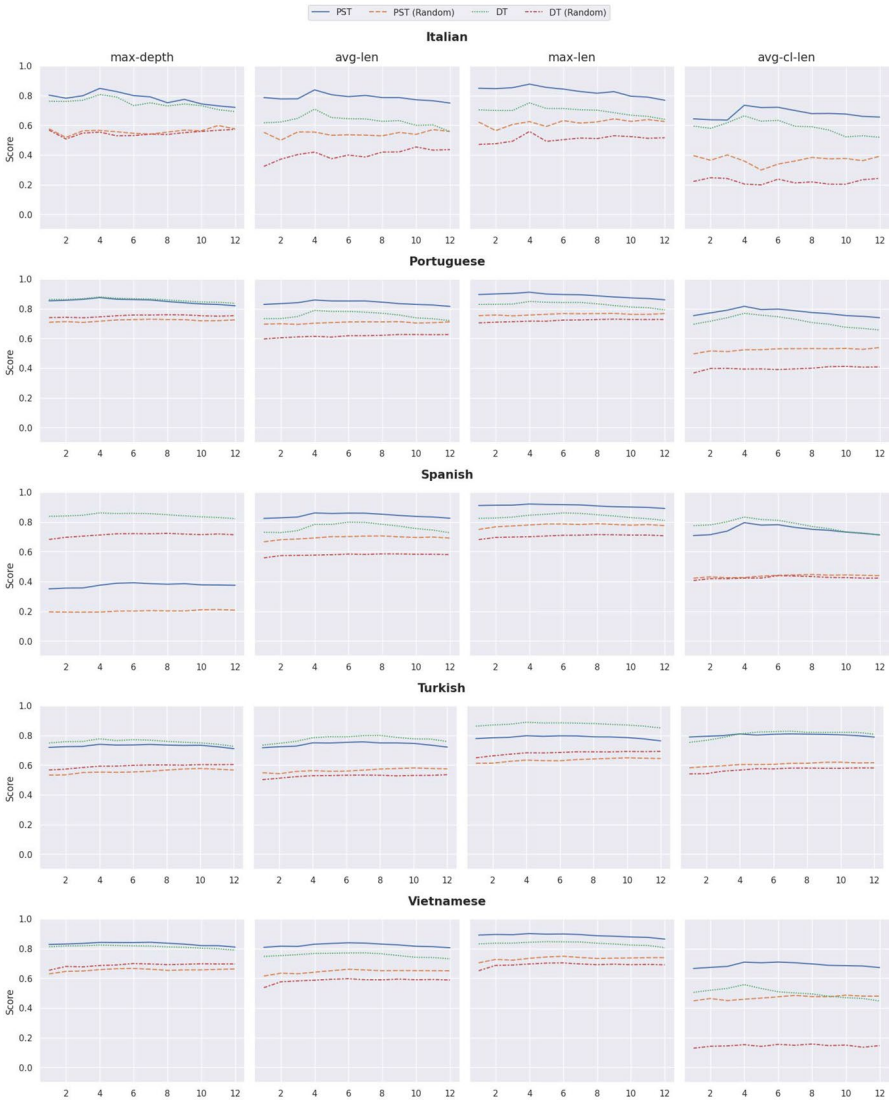

Fig. 6 (continued)

Acknowledgements The authors gratefully acknowledge the support of:the PNRR project FAIR - Future Artificial Intelligence Research (PE0000013); the PNRR project CHANGES - Cultural Heritage Innovation for Next-Gen Sustainable Society (PE00000020); the PRIN 2022 TEAMING-UP - Teaming up with Social Artificial Agents project (20177FX2A7); the PRIN PNRR 2022 Project EKEEL - Empowering Knowledge Extraction to Empower Learners (P20227PEPK); the project "Human in Neural Language Models" (*IsC93_HiNLM*), funded by CINECA (https://www.cineca.it/en) under the ISCRA initiative, for the availability of HPC resources and support. We also thank all the developers and curators of the treebanks we used in this study. In particular, we would like to express our sincere gratitude to Huyen

Nguyen Thi Minh and Ha Linh (VLSP) for their invaluable support and assistance during the data collection phase.

Funding Open access funding provided by ILC - PISA within the CRUI-CARE Agreement. No funding was received for conducting this study.

Declarations

Conflict of interest The authors declare that they do not have any Conflict of interest.

Compliance with Ethical Standards Our work has limited ethical implications since we mainly introduced probing tasks to evaluate a pre-trained neural language model. The treebanks we relied upon were used in compliance with the Terms of Use and the resources and materials produced during this study will be distributed in compliance with the license agreement of each source treebank. The research questions were investigated relying on the treebanks for which we were able to recover parallel sentences (as defined in this work), thus the scalability of our approach is limited to treebanks distributed across multiple annotation formats. We need to point out that our conclusions only concern the sample of treebanks that we analyzed in this study.

References

Abeillé, A., Clément, L., & Kinyon, A. (2000) Building a treebank for French. In Proceedings of the second international conference on language resources and evaluation (LREC'00). European Language Resources Association (ELRA)

Afonso, S., Bick, E., Haber, R. (2002) Floresta sintá(c)tica: a treebank for portuguese. In Manuel González Rodrigues; Carmen Paz Suarez Araujo (ed) Proceedings of the third international conference on language resources and evaluation (LREC 2002)(Las Palmas de Gran Canaria Espanha 29-31 de Maio de 2002) ELRA

Alfina, I., Budi, I., & Suhartanto, H. (2020). Tree rotations for dependency trees: Converting the head-directionality of noun phrases. *Journal of Computer Science, 16*(11), 1585–1597.

Alonso, H. M., & Zeman, D. (2016). Universal dependencies for the ancora treebanks. *Procesamiento del Lenguaje Natural, 57*, 91–98.

Alzetta, C., Dell'Orletta, F., Montemagni, S., et al (2017) Dangerous relations in dependency treebanks. In Proceedings of the 16th international workshop on treebanks and linguistic theories, pp 201–210

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, EF., et al. (2020) A universal dependencies conversion pipeline for a penn-format constituency treebank. In Proceedings of the fourth workshop on universal dependencies (UDW 2020) (pp 16–25). Association for Computational Linguistics

Arps, D., Samih, Y., Kallmeyer, L., et al (2022) Probing for constituency structure in neural language models. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Findings of the association for computational linguistics: EMNLP 2022 (pp 6738–6757). Association for Computational Linguistics, https://doi.org/10.18653/v1/2022.findings-emnlp.502

Arps, D., Kallmeyer, L., Samih, Y., et al. (2024) Multilingual nonce dependency treebanks: Understanding how language models represent and process syntactic structure. In Duh K, Gomez H, Bethard S (eds) Proceedings of the 2024 conference of the North American chapter of the association for

computational linguistics: Human language technologies (Volume 1: Long Papers). Association for Computational Linguistics, pp 7822–7844, https://doi.org/10.18653/v1/2024.naacl-long.433,

Arwidarasti, JN., Alfina, I., & Krisnadhi, AA. (2019) Converting an indonesian constituency treebank to the Penn Treebank format. In 2019 International conference on Asian language processing (IALP) (pp 331–336), IEEE

Beeching, E., Fourrier, C., Habib, N., et al (2023) Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Van der Beek, L., Bouma, G., Malouf, R., & Van Noord, G. (2002). The Alpino dependency treebank. *Computational linguistics in the Netherlands* (pp. 8–22). Leiden: Brill.

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Comput Linguist, 48*(1), 207–219. https://doi.org/10.1162/coli_a_00422

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers). Asian Federation of Natural Language Processing, pp 1–10

Bick, E. (2016) Constraint grammar-based conversion of dependency treebanks. In Proceedings of the 13th international conference on natural language processing. NLP Association of India, pp. 109–114

Blevins, T., Levy, O., & Zettlemoyer, L. (2018) Deep RNNs encode soft hierarchical syntax. In Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 2: Short Papers). Association for Computational Linguistics, pp 14–https://doi.org/10.18653/v1/P18-2003

Bonfante, G., Guillaume, B., & Perrier, G. (2018). *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons.

Bosco, C. (2007) Multiple-step treebank conversion: From dependency to Penn format. In Proceedings of the linguistic annotation workshop. Association for Computational Linguistics, pp. 164–167

Bosco, C., Lombardo, V., Vassallo, D., & Lesmo, L. (2000). Building a treebank for Italian: A data-driven annotation schema. In Proceedings of the second international conference on language resources and evaluation (LREC'00). European Language Resources Association (ELRA)

Bosco, C., Sanguinetti, M., Lesmo, L., et al (2012) The parallel-tut: A multilingual and multiformat treebank. In Proceedings of the eight international conference on language resources and evaluation (LREC'12), European Language Resources Association (ELRA) (pp 1932–1938)

Bosco, C., Montemagni, S., & Simi, M. (2013) Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse. Association for Computational Linguistics (pp 61–69)

Bouma, G., & van Noord, G. (2017) Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017). Association for Computational Linguistics, Gothenburg, Sweden, pp 19–26

Civit, M., Martí, M. A., Bufí, N., et al. (2006). Cat3lb and cast3lb: From constituents to dependencies. In T. Salakoski, F. Ginter, & S. Pyysalo (Eds.), *Advances in Natural Language Processing* (pp. 141–152). Springer.

Conneau, A., Kruszewski, G., Lample, G., et al (2018) What you can cram into a single \$ &!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp 2126–2136, https://doi.org/10.18653/v1/P18-1198

Devlin, J., Chang, MW., Lee, K., et al. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423

Dinakaramani, A., Rashel, F., Luthfi, A., & Manurung, R. (2014). Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In 2014 International conference on Asian language processing (IALP), IEEE, pp 66–69

Dryer, M.S, & Haspelmath, M. (eds) (2013) WALS Online. Max Planck Institute for Evolutionary Anthropology

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics, 8*, 34–48.

Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and control, 8*(3), 304–337.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020) SyntaxGym: An online platform for targeted evaluation of language models. In Celikyilmaz A, Wen TH (eds) Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations. Association for Computational Linguistics, Online, pp 70–76, https://doi.org/10.18653/v1/2020.acl-demos.10,

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In: Proceedings of the second workshop on universal dependencies (UDW 2018). Association for Computational Linguistics, pp 66–74, https://doi.org/10.18653/v1/W18-6008

Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language, 40*(4), 511–525.

Hewitt, J., & Liang, P. (2019) Designing and interpreting probes with control tasks. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, pp 2733–2743, https://doi.org/10.18653/v1/D19-1275

Hewitt, J., & Manning, CD. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics (pp 4129–4138), https://doi.org/10.18653/v1/N19-1419

Hudson, R. A. (1980). *Constituency and dependency*. Walter de Gruyter.

Jawahar, G., Sagot, B., & Seddah, D. (2019) What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 3651–3657, https://doi.org/10.18653/v1/P19-1356

Kara, N., Marşan, B., Özçelik, M., et al (2020) Creating a syntactically felicitous constituency treebank for turkish. In: 2020 Innovations in intelligent systems and applications conference (ASYU) (pp 1–6), IEEE

Kogkalidis, K., & Wijnholds, G. (2022). Discontinuous constituency and BERT: A case study of Dutch. *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 3776–3785). Association for Computational Linguistics.

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Dependency parsing* (pp. 11–20). Cham p: Springer.

Kulmizev, A., Ravishankar, V., Abdou, M., et al (2020) Do neural language models show preferences for syntactic formalisms? In Proceedings of the 58th annual meeting of the association for computational linguistics. association for computational linguistics, Online, pp 4077–4091, https://doi.org/10.18653/v1/2020.acl-main.375

Kuzgun, A., Yıldız OK,, Cesur N., et al. (2021) From constituency to UD-style dependency: Building the first conversion tool of Turkish. In Proceedings of the international conference on recent advances in natural language processing (RANLP 2021). INCOMA Ltd., Held Online, pp 761–769

Liang, P., Bommasani, R., Lee, T., et al. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research https://openreview.net/forum?id=iO4LZibEqW, featured Certification, Expert Certification

Linh, HM., Huyen, NTM., Luong, VX., et al (2020) VLSP 2020 shared task: Universal Dependency parsing for Vietnamese. In Proceedings of the 7th International workshop on Vietnamese language and speech processing. Association for Computational Lingustics (pp. 77–83)

Lopopolo, A., van den Bosch, A., Petersson, K. M., et al. (2021). Distinguishing syntactic operations in the brain: Dependency and phrase-structure parsing. *Neurobiology of Language, 2*(1), 152–175.

Marcus, M., Kim, G., Marcinkiewicz, MA., MacIntyre, R., Bies, A., Ferguson, M., & Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. In Human language technology: Proceedings of a workshop held at plainsboro, March 8-11, 1994

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

Matthews, P. (1981). *Syntax*. Cambridge University Press.

Maudslay, R.H., & Cotterell, R. (2021) Do syntactic probes probe syntax? experiments with jabberwocky probing. In Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 124–1https://doi.org/10.18653/v1/2021.naacl-main.11

Mel'čuk, I., & Polguère, A. (2009) Dependency in linguistic description. Dependency in Linguistic Description pp 1–308

Mel'čuk, I. A., et al. (1988). Dependency syntax: Theory and practice. *SUNY Press, 8*, 303–332.

Miaschi, A., Brunato, D., Dell'Orletta, F., & Venturi, G. (2020). Linguistic profiling of a neural language model. In Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics,(Online), pp 745–756, https://doi.org/10.18653/v1/2020.coling-main.65

Miaschi, A., Sarti, G., Brunato, D., et al. (2022). Probing linguistic knowledge in Italian neural language models across language varieties. *Italian Journal of Computational Linguistics, 8*(1), 25–44. https://doi.org/10.4000/ijcol.965

Miaschi, A., Alzetta, C., Brunato, D., et al. (2023). Testing the effectiveness of the diagnostic probing paradigm on Italian treebanks. *Information*. https://doi.org/10.3390/info14030144

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2022) Probing for labeled dependency trees. In Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp 7711–7726, https://doi.org/10.18653/v1/2022.acl-long.532,

Muñoz-Ortiz, A., Vilares, D., & Gómez-Rodríguez, C. (2023) Assessment of pre-trained models across languages and grammars. arXiv:2309.11165

Nefdt, R. M., & Baggio, G. (2023). Notational variants and cognition: The case of dependency grammar. *Erkenntnis, 89*(7), 2867–2897.

Nenkova, A., Chae, J., Louis, A., & Pitler, E. (2009). Structural features for predicting the linguistic quality of text. In V. John (Ed.), *Empirical methods in natural language generation* (pp. 222–241). Springer.

Nguyen, PT., Vu, XL., Nguyen, TMH., et al (2009) Building a large syntactically-annotated corpus of Vietnamese. In Proceedings of the third linguistic annotation workshop (LAW III). Association for Computational Linguistics, pp 182–185

Nivre, J., de Marneffe, MC., Ginter, F., et al (2016) Universal Dependencies v1: A multilingual treebank collection. In Calzolari N, Choukri K, Declerck T, et al (eds) Proceedings of the tenth international conference on language resources and evaluation (LREC'16). European Language Resources Association (ELRA), pp 1659–1666, https://aclanthology.org/L16-1262

Nivre, J., de Marneffe, MC., Ginter, F., et al (2020) Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari N, Béchet F, Blache P, et al (eds) Proceedings of the twelfth language resources and evaluation conference. European Language Resources Association, pp 4034–4043, https://aclanthology.org/2020.lrec-1.497

Oota, SR., Marreddy, M., Gupta, M., & Bapi, R. (2023). How does the brain process syntactic structure while listening? In Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the association for computational linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 6624–6647, https://doi.org/10.18653/v1/2023.findings-acl.415

Osborne, T. (2014). Dependency grammar. In E. John (Ed.), *The Routledge handbook of syntax* (pp. 604–626). Routledge.

Prange, J., Schneider, N., & Kong, L. (2022). Linguistic frameworks go toe-to-toe at neuro-symbolic language modeling. In Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, pp 4375–4391, https://doi.org/10.18653/v1/2022.naacl-main.325

Rademaker, A., Chalub, F., Real, L., et al. (2017) Universal dependencies for portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling), pp 197–206

Rambow, O., & Joshi, A. (1997). A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena. *Recent trends in meaning-text theory, 39*, 167–190.

Rögnvaldsson, E., Ingason, AK., Sigurðsson, EF., & Wallenberg, J. (2012) The Icelandic parsed historical corpus (IcePaHC). In: Proceedings of the eighth international conference on language resources and evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, pp 1977–1984

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., (2013). Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In Proceedings of the fourth workshop on statistical parsing of morphologically-rich languages, Association for Computational Linguistics, pp 146–182

Seddah, D., de la Clergerie, E., Sagot, B., et al (2018) Cheating a parser to death: Data-driven cross-treebank annotation transfer. In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA)

Srivastava, A., Rastogi, A., Rao, A., et al. (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615

Taulé, M., Martí, MA., & Recasens, M. (2008) AnCora: Multilevel annotated corpora for Catalan and Spanish. In Proceedings of the sixth international conference on language resources and evaluation (LREC'08). European Language Resources Association (ELRA)

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., & Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In Proceedings of the 7th international conference on learning representations (ICLR 2019)

Vilares, D., Strzyz, M., Søgaard, A. (2020). Parsing as pretraining. In Proceedings of the AAAI Conference on Artificial Intelligence, pp 9114–9121

Wagner, J., Zarrieß, S. (2023). Probing BERT's ability to encode sentence modality and modal verb sense across varieties of English. In Amblard M, Breitholtz E (eds) Proceedings of the 15th International Conference on Computational Semantics. Association for Computational Linguistics, pp 28–38, https://aclanthology.org/2023.iwcs-1.3

Waldis, A., Perlitz, Y., Choshen, L., Hou, Y., & Gurevych, I.(2024). Holmes: Benchmark the linguistic competence of language models. arXiv:2404.18923

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, pp 353–355, https://doi.org/10.18653/v1/W18-5446

Wang, A., Pruksachatkun, Y., Nangia, N., et al. (2019). *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc.

Warstadt, A., Cao, Y., Grosu, I., et al. (2019a) Investigating BERT's knowledge of language: Five analysis methods with NPIs. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 2877–2887, https://doi.org/10.18653/v1/D19-1286

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics, 7*, 625–641.

Warstadt, A., Parrish, A., Liu, H., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392. https://doi.org/10.1162/tacl_a_00321

Wolf, T., Debut, L., Sanh, V., et al (2020) Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations. Association for Computational Linguistics, Online, pp 38–https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zeman, D., Nivre, J., & alii, (2021) Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University