Dominique Brunato* and Giulia Venturi*

Why is this language complex? Cherry-pick the optimal set of features in multilingual treebanks

Abstract: This paper investigates linguistic complexity across natural languages from a corpus-based perspective and relies on the assumptions of linguistic profiling as a methodological framework. We focus in particular on the domain of syntactic complexity and analyze the distribution of a set of features taken as proxies of complexity phenomena at sentence level, which were extracted from 63 treebanks annotated according to the Universal Dependencies formalism. This dataset guarantees that the features considered are modeling the same linguistic phenomena in different treebanks, allowing reliable comparison among languages. We show that our approach is able to identify tendencies of structural proximity between languages not necessarily in line with typologically-supported classification, thus shedding light on new corpus-based findings.

Keywords: Linguistic Complexity; Linguistic Profiling; Universal Dependencies; Syntactic Domain

1 Introduction

Linguistic complexity, along with its detection, evaluation and processing, is a topic that has long attracted researchers embracing different perspectives ranging from typological linguistics (Miestamo, Sinnemaki & Karlsson 2008), first and second language acquisition (Kortmann & Szmrecsanyi 2012), computational linguistics and related fields (Brunato et al. 2016). Despite the debated and multidimensional nature of the notion, a quite established theoretical distinction identifies an "absolute complexity", that refers to the formal properties of linguistic systems, and a "relative complexity", that defines complexity in relation to the language user (e.g. speaker, listener or learner) thus considering complexity in terms of processing difficulty (Miestamo 2008). The absolute viewpoint encounters itself a main methodological obstacle that, in Miestamo's words, can be summarized as follows: "even if this

***Corresponding author: Dominique Brunato,** Institute for Computational Linguistics "A. Zampolli" (ILC-CNR) - ItaliaNLP Lab, Pisa (Italy)
***Corresponding author: Giulia Venturi,** Institute for Computational Linguistics "A. Zampolli" (ILC-CNR) - ItaliaNLP Lab, Pisa (Italy)

were theoretically possible, it would be beyond the capacities of the mortal linguist to exhaustively count all grammatical details of the languages studied, especially in a large-scale cross-linguistic study". Accordingly, if studying the *global* complexity of a language is perceived as a very ambitious and probably hopeless endeavor, the dominant and more feasible approach addresses *local* complexity, i.e. the complexity in the different sub–domains of the language (Kortmann & Szmrecsanyi 2012).

In this scenario, the growing availability of linguistically annotated corpora for many languages has promoted the exploitation of data-driven approaches focused on detecting and measuring a large variety of complexity phenomena across corpora representative of different languages and language varieties, with a particular emphasis on syntactic-related peculiarities showing to be consistent across many languages. It is the case, for example, of dependency length – the distance between syntactically related words in a sentence –, which is considered as a reliable measure of sentence complexity according to both experimental and theoretical language research, see e.g. Gibson (1998, 2000), Futrell, Mahowald & Gibson (2015) and Liu (2017).

With this respect, the benefits of acquiring information about linguistic complexity from multilingual treebanks have been recently promoted by the Universal Dependencies (UD) project,[1] an international initiative with over 300 contributors producing nearly 200 treebanks in over 100 languages. The project allowed the definition of a framework for cross-linguistically consistent treebank annotation aiming to capture similarities as well as idiosyncrasies among typologically different languages (Nivre 2015). The first related initiative took place in 2018 as a satellite event of the "Evolution of Language International Conferences" (EVOLANG), named "Workshop on Measuring Language Complexity (MLC)".[2] By relying on the morpho-syntactic and syntactic formalism of the UD treebanks, seven teams of researchers designed 34 different measures of linguistic complexity for 37 language varieties belonging to seven families (Berdicevskis et al. 2018). The 2019 "Interactive Workshop on Measuring Language Complexity (IWMLC)"[3] allowed a continuation of the debate about cross-linguistic complexity research prompted by the use of UD treebanks as source corpora.

Our contribution: The present contribution stems from our participation in the 2019 IWMLC workshop, where we originally presented our approach based on linguistic profiling to measure and compare languages according to their absolute complexity. In this paper we illustrate the fundamentals of this approach and

1 https://universaldependencies.org/

2 http://www.christianbentz.de/MLC_index.html

3 http://christianbentz.de/MLC2019_index.html

extend the preliminary findings presented in that context. As described in the following section, the core of this approach is the extraction from multilingual treebanks of a same set of features modeling phenomena of sentence complexity in different sub-domains of language, with a main focus on the syntactic one. From this perspective, our study lies in the framework of linguistic research aiming at acquiring quantitative evidence about linguistic complexity from large-scale data representative of real language usage, and in particularly dependency annotated corpora. The rich variety of features here considered aims to empirically prove that the notion of syntactic complexity is not monolithic. As previously observed, there is a wide consensus in considering it as a multifaceted notion covering several aspects also within the same domain. Thus, with our perspective we would like to underline the need for 'cherry-picking' which feature is more reliable to model a specific aspect of complexity. The approach has been tested on 63 UD treebanks, presented in Section 2.1, while the linguistic features are illustrated in Section 2.2. The choice of considering multi-lingual treebanks possibly containing different textual genres is also motivated by our intention of showing that treebanks may be only partially representative of a given language and that, as a consequence, any quantitative evidence about the complexity of a language cannot be generalized to the whole system but instead should be related to the text typologies of its representative corpus. In Section 3.1 our set of features is first analyzed separately, that is considering each feature as a distinct complexity metric. In Section 3.2, we inspect the results of a cluster analysis based on the combination of all features showing that our approach is able to identify tendencies of structural proximity between languages not necessarily in line with typologically-supported classifications.

2 Linguistic profiling of multilingual treebanks

The approach presented here has been inspired by research on "linguistic profiling" which is grounded on two main ingredients: i) large-scale (automatically or manually) annotated corpora representative of a given language variety and ii) counts of linguistic features, extracted from different levels of annotation, which all together model properties related to the form of a text (van Halteren 2004). Although it was originally developed for authorship recognition or verification purposes, this methodology proved to be effective in multiple scenarios, for example to study variations related to genre and register (Argamon et al. 2003) or to the social dimension of language (Nguyen et al. 2016), or also to model stylometric characteristics (Daelemans 2013). It is worth mentioning here that many of the linguistic features used for profiling purposes include fine-grained predictors of

linguistic complexity. Accordingly, similar sets of features have been used to assess the readability level of texts (Collins-Thompson 2014), to predict human judgments on sentence complexity (Brunato et al. 2018) or to study diachronic variation in syntactic complexity (Lei & Wen 2020).

In this study we relied on the linguistic profiling methodology described in Brunato et al. (2020) and implemented in Profiling-UD,[4] the first web-based tool conceived to linguistically profile multilingual texts by relying on the UD formalism. This tool computes a very large set of linguistic features either extracted from a document or a single sentence. The application of profiling at sentence-level allows focusing on specific instances of phenomena which might be flattened when computed at document level. This is precisely the case of the corpora we are analyzing, since UD treebanks are not homogeneous with respect to textual genres (Plank 2016) and thus linguistic features are unevenly distributed across each corpus. Moreover, for our investigation, we selected only the features particularly relevant for operationalizing sentence complexity in the syntactic sub-domain and we computed their value for each sentence of the considered UD treebanks. The final value for each language corresponds to the average value that the feature has in all sentences of the reference treebank(s) for that language. Finally, following the outcome of the literature on sentence complexity from different perspectives (cognitive, corpus-based, computational), we assumed that the higher this value, the more complex the language usage observed in the treebank with respect to each feature.

2.1 Universal Dependencies Treebanks

As aforementioned, our investigation was based on a subset of UD treebanks released in version 2.3. The UD project is aimed not only at promoting the development and comparative evaluation of multilingual Natural Language Processing systems but also at enabling comparative linguistic studies (Nivre 2015). In fact, corpora annotated with the same inventory of morpho-syntactic categories and dependency relations are paving the way toward methods able to track and quantify linguistic variation across languages avoiding possible interference due to multiple annotation schemata.

Table 1 reports the languages considered, together with the corresponding language family and genus according to the *World Atlas of Language Structures* (WALS) (Dryer & Haspelmath 2013),[5] the most commonly-used and broadest

4 http://www.italianlp.it/demo/profiling-UD/

5 http://wals.info

Language	Family	Genus	TB	Tokens
Arabic (ARA)	Afroasiatic	Semitic	1	282k
Hebrew (HEB)	Afroasiatic	Semitic	1	161k
Turkish (TUR)	Altaic	Turkic	1	57k
Uyghur (UIG)	Altaic	Turkic	1	40k
Vietnamese (VIE)	Austroasiatic	Viet-Muong	1	43k
Indonesian (IND)	Austronesian	Malayo-Sumbawan	1	121k
Basque (BAQ)	Basque	Basque	1	121k
Latvian (LAV)	Indo-European	Baltic	1	152k
Afrikaans (AFR)	Indo-European	Germanic	1	49k
Danish (DAN)	Indo-European	Germanic	1	100k
German (GER)	Indo-European	Germanic	1	292k
English (ENG)	Indo-European	Germanic	4	465k
Dutch (DUT)	Indo-European	Germanic	2	326k
Norwegian (NOR)	Indo-European	Germanic	2	301k
Swedish (SWE)	Indo-European	Germanic	2	175k
Greek (GRE)	Indo-European	Greek	1	63k
Hindi (HIN)	Indo-European	Hindi	1	351k
Urdu (URD)	Indo-European	Indic	1	138k
Persian (PER)	Indo-European	Iranian	1	152k
Catalan (CAT)	Indo-European	Romance	1	531k
French (FRE)	Indo-European	Romance	2	470k
Italian (ITA)	Indo-European	Romance	3	477k
Portuguese (POR)	Indo-European	Romance	1	227k
Romanian (RUM)	Indo-European	Romance	2	413k
Spanish (SPA)	Indo-European	Romance	2	980k
Bulgarian (BUL)	Indo-European	Slavic	1	156k
Czech (CZE)	Indo-European	Slavic	3	2167k
Croatian (HRV)	Indo-European	Slavic	1	197k
Polish (POL)	Indo-European	Slavic	2	213k
Russian (RUS)	Indo-European	Slavic	2	99k
Slovak (SLO)	Indo-European	Slavic	1	106k
Slovenian (SLV)	Indo-European	Slavic	1	140k
Serbian (SRP)	Indo-European	Slavic	1	86k
Ukrainian (UKR)	Indo-European	Slavic	1	116k
Japanese (JPN)	Japanese	Japanese	1	184k
Korean (KOR)	Korean	Korean	2	430k
Chinese (CHI)	Sino-Tibetan	–	1	123k
Estonian (EST)	Uralic	Finnic	1	434k
Finnish (FIN)	Uralic	Finnic	2	361k
Hungarian (HUN)	Uralic	Ugric	1	42k
Ancient languages				
Gothic (GOT)	Indo-European	Germanic	1	55k
Ancient Greek (GRC)	Indo-European	Greek	2	416k
Old Church Slavonic (CHU)	Indo-European	Slavic	1	57k
Latin (LAT)	Indo-European	–	2	552k

Tab. 1: Overview of languages (with their ISO-639-2 code), corresponding WALS language family and genus, number of treebanks per language (*TB*) and treebank size in k of tokens.

database of structural (phonological, grammatical, lexical) properties of languages. The rationale behind the choice of these languages lies in our participation in the shared task organized in conjunction with the 2019 "Interactive Workshop on Measuring Language Complexity (IWMLC)", where these languages were considered as a reasonable test-bed to compare different measures of linguistic complexity. For each language we also specify the number of available treebanks and their size in number of tokens. As it can be seen, the majority of languages (31 out of 44, i.e. 70%) belongs to the Indo-European family, which is internally distinguished into eight genera with three major groups, i.e. Slavic, Germanic and Romance. Concerning the number of treebanks per language, 66% of the languages (29 languages) is represented by one treebank and 29% (13 languages) by two. As it will be discussed in the following sections, it is not always the case that different treebanks of the same language have similar linguistic features. This has a well-known impact on cross-linguistic studies grounded on corpora which may be biased by corpora variations (Chen & Gerdes 2017) mostly due to the multiple genres and domains contained in the different treebanks available for each language (Plank 2016).

2.2 Linguistic Features

The set of features here considered is a subset of the ones described by Brunato et al. (2020) and has been chosen to be representative of different macro-areas of language complexity phenomena. In what follows, we will describe how they were computed using the following sample sentence taken from the English treebank (EWT). (Figure 1 shows the tree graphical representation):

(1) You wonder if he was manipulating the market with his bombing targets.

Basic text properties

- Sentence length (*sent_length*): it is calculated as the average number of words per sentence. Sentence length is typically used as an approximation of syntactic complexity, for example in traditional formulas developed for the automatic assessment of text readability (Kincaid et al. 1975). (1) is 13 tokens long.
- Word length (*word_length*): it is calculated as the average number of characters per word (excluded punctuation). It is a basic indicator of word complexity and, similarly to sentence length, it is used by traditional readability formulas as an approximation of lexical complexity. (1) contains words that are 4.83 characters long on average.

Parse-tree structure

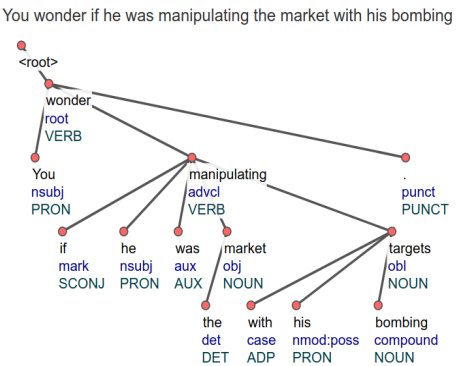

Fig. 1: Linguistic annotation of the example sentence.

– Average length of clauses (*clause_length*): it is measured as the number of tokens per clause, which is calculated as the ratio between the number of tokens in a sentence and the number of either verbal or copular heads. Syntactic metrics relying on clause length, such as T-Unit (Hunt 1966), are widely used in studies on human production and comprehension of complex sentences, as well as in first and second language acquisition to assess the development of syntactic competence. In the sample sentence the average clause length is 6.5 tokens, since there are two verbal heads ('wonder' and 'manipulating').
– Average length of dependency links (*dep_links_len*): this is calculated as the average number of words occurring between the syntactic head and the dependent. As mentioned in Section 1, longer dependencies represent a source of greater processing difficulties for both humans, see (Gibson 1998, 2000), and Demberg & Keller (2008), and statistical parsers, see McDonald & Nivre (2007), Rimell, Clark & Steedman (2009), Nivre et al. (2010), and Gulordava & Merlo (2015). This measure is also considered as a universal property by typological studies, which demonstrate that dependency length is actually minimized in real utterances across many languages and language families, even if with some differences due to language-specific grammatical constraints, syntactic choices (Temperley & Gildea 2018) or diachronic changes (Gulordava & Merlo 2015). The average value in (1) is 2.36: four links have a distance of one from their syntactic head ('You' '*wonder*'; 'was' '*manipulating*'; 'the' '*market*'; 'bombing' '*targets*'),[6] three links have a distance of two ('he' '*manipulating*'; 'market' '*manipulating*'; 'his' '*targets*'), two have a distance of three ('if' '*manipulating*';

6 The syntactic head is always marked in italic.

'with' '*targets*'), one has a distance of four ('manipulating' '*wonder*') and the longest one, i.e. a distance of six, links 'targets' to *manipulating*.

– Depth of the whole parse tree (*tree_depth*): it corresponds to the longest path from the root of the dependency tree to some leaf nodes. The measure originates from studies on "relative complexity" showing that deeper syntactic trees hamper human sentence processing (Frazier 1985). In (1), 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 (*'wonder'*) to each of the more distant leaf nodes, represented by the words 'the', 'with', 'with', 'his' and 'bombing'.

Subordination

– Percentage distribution of subordinate clauses (*subord_dist*): it is calculated as the percentage distribution of main vs subordinate clauses, where the latter are identified on the basis of the UD guidelines that distinguish four types.[7] We included this and the following feature as the use of subordination is a broadly studied marker of structural complexity, for example for text simplification purposes (Bott & Saggion 2014). (1) is articulated into a main ('wonder') and a subordinate clause ('manipulating'), headed by the verbal root 'wonder' and marked as an adverbial clause modifier (*advcl*). Thus, the percentage distribution of this features is 50%.
– Average depth of 'chains' of embedded subordinate clauses (*subord_chain_len*): once the sub-tree of the subordinate clause is identified, a subordinate 'chain' is calculated as the number of subordinate clauses recursively embedded in the top subordinate clause. In the sample sentence, the value of this feature is equal to one, since it contains only one single subordinate clause.

Verbal predicate structure

– Average number of dependency links of a verbal head (*verb_arity*): this corresponds to the average number of instantiated dependency links (both arguments and modifiers) sharing the same verbal head, excluding auxiliaries bearing the syntactic role of copula according to the UD scheme. This feature reflects the richness of verbal predicates, i.e the higher the score the richer the verbal predicate.[8] Note that this measure might be highly sensitive to language: pro-drop languages, which do not obligatorily require an explicit subject, can

7 https://universaldependencies.org/u/overview/complex-syntax.html#subordination

8 This measure could be refined if corpora had a further level of annotation making explicit the verb argument structure (allowing one to distinguish arguments from adjuncts) or an external sub-categorization lexicon serving as a reference resource.

have fewer dependents since null subjects are not explicitly marked in the UD annotation scheme. In (1) the average arity score is 3, since the main verb 'wonder' has two dependents ('You' and 'manipulating') and the first embedded verb 'manipulating' has four ('if', 'he', 'was' and 'market').

3 Comparison of multilingual treebanks

Based on the analysis of the selected features, our approach to the study of linguistic complexity allowed attaining several outcomes which can be categorized in two main groups. The first one, described in Section 3.1, is meant to investigate how the considered features are able to intercept different aspects of sentence complexity, whether and to what extent their values are stable within each UD treebank, and how they change across languages and also across multiple treebanks available for a language. Through the second group of results (see Section 3.2), we looked at these features from an holistic perspective and used them to cluster all the treebanks considered.

3.1 A feature-based comparison

Figure 2 reports the average distribution of each feature extracted from all sentences of a given treebank. The heatmap provides a direct visualization of which treebank has the higher feature value (darker color) and thus presents a more complex usage with respect to that feature. In each cell we also report in parentheses the position that the treebank occupies in the ranking of all treebanks established by the coefficient of variation for each feature. The coefficient of variation represents a standardized measure of the dispersion of data points around the mean and it is particularly useful for comparing series of data calculated on different scales. Being calculated as the ratio between the standard deviation and the mean, we considered it as particularly appropriate for accounting for the nature of our data as well as for the aim of the analysis. On the one hand, it allows normalizing standard deviation, thus preventing the impact of extreme values, and it turned out to be a reliable index to compare values of linguistic features which can have quite different scales and ranges, such as sentence length (absolute number) and subordinate clauses (distribution). On the other hand, it quantifies the degree of variation within the composition of the considered treebank, on the assumption that the more stable a feature is, the more representative it is for a given language (when we have a

unique treebank per language) or for a language variety (when we have more than one).

Let us start analyzing the treebanks with respect to our basic features of complexity, i.e. sentence and word length. For the former, a distinct result emerges that Romance languages treebanks tend to have longer sentences with respect to the other languages, although the treebank with the longest sentences (∼37 tokens) belongs to the Semitic genus and it represents the Arabic language. The different degree of affixation in word formation clearly affects the resulting length of words. In this respect, our data confirm the 'coarse' distinction into analytical vs synthetic languages as we find Chinese and Finnish in the lowest and highest positions of ranking by word length, respectively (Finnish_TDT: 7.34; Chinese: 1.69). Like Finnish, other typical examples of agglutinative languages like Turkish, Basque and Hungarian are similarly highly ranked, followed by the majority of languages of the Germanic and Romance group which still have a rich inflectional morphology but often realized with fusional suffixes. However, it is generally agreed that the distinction into different morphological types should be considered as more gradient rather than categorical and that the same language can exhibit patterns of a different nature (Haspelmath 2009). Japanese, for instance, is highly synthetic with a complex system of verb inflection, but also highly analytic in not having noun inflection; and this might explain the lower position in our raking.

Interestingly, focusing merely on these raw text features we observe that languages with more than one treebank have different behaviors. Consider for example the case of Italian, for which there are three treebanks, two of them (ISDT and ParTUT) containing miscellaneous textual genres (i.e. legal texts, newspaper articles and Wikipedia pages) and PoSTWITA, a collection of Italian tweets. The limited number of characters for tweets allowed by the Twitter platform necessarily yields shorter sentences in PoSTWITA (18.54) than the other two treebanks (21.06 in ISDT and 26.58 in ParTUT). Such a constraint in terms of length is also reflected by the ranking position established by the coefficient of variation: the first position of PoSTWITA suggests that it is the most stable treebank regarding this feature.

Since it is well-known that sentence length is highly related to features extracted from the syntactic level of annotation, we observe that treebanks can be grouped quite similarly when we consider complexity measures accounting for the parse tree structure. Thus, Arabic, the language with the longest sentences, is also the language with the deepest syntactic trees (7.14). As expected, the genus with the highest *tree_depth* values is the Romance one, with an average depth of 4.59 in the corresponding treebanks, even though Afrikaans, a Germanic language, has the second greatest tree depth (3.1). Also in this case, the Italian Twitter treebank (PoSTWITA) is among the most stable language variety. A slightly different trend can be observed if we focus on the average length of dependency

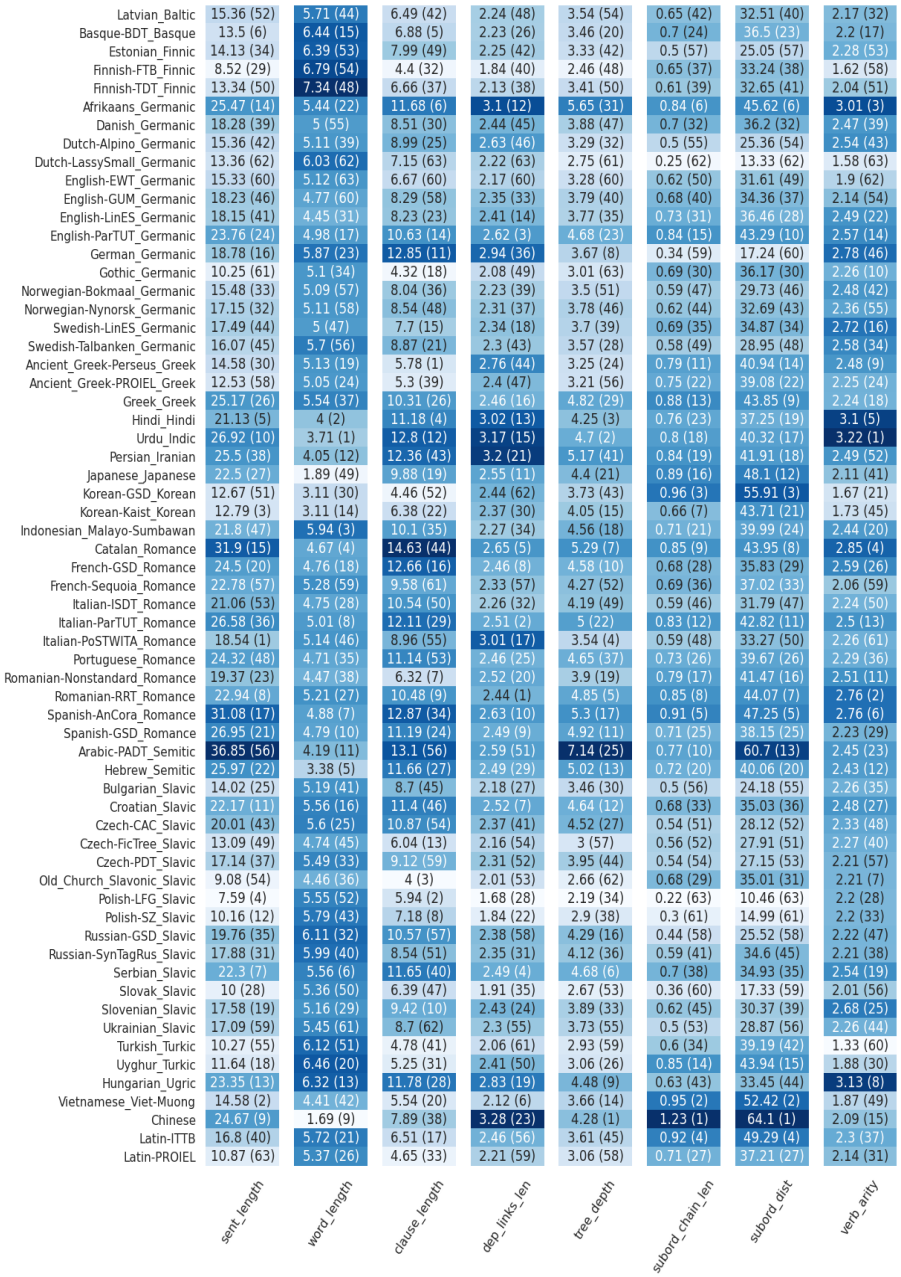

Fig. 2: Distribution of linguistic features for each treebank. In each cell is reported the average value of the feature in the corresponding treebank and the number (in parentheses) indicating the ordinal position that the treebank has in the ranking of all treebanks given by the coefficient of variation for each feature. The lower the number, the more stable the feature in a given treebank.

links, a feature similarly extracted from the syntactic annotation, but accounting for the linear structure. The greatest average dependency lengths occur in Chinese (3.28), Persian (3.2), Urdu (3.17) and Hindi (3.02) sentences. However, because this feature is highly related to the two aforementioned ones, it is to be expected that the treebanks belonging to the Romance genus are still those with longer links (with an average length of 2.52). Romance treebanks are also the most stable with respect to this feature, as we find six Romance languages (Romanian-RRT, Italian-ParTUT, Catalan, French-GSD, Spanish-GSD and Spanish-AnCora) in the top ten ranked treebanks for coefficient of variation.

Treebanks representative of the Romance languages are confirmed to be the most complex ones in terms of sentence structure also when we consider the average clause length. However, the computation of this feature does not allow us to take into account any distinction among the typology of clauses, e.g. subordinate vs coordinate ones. To inspect this aspect we need to examine the values of features explicitly modeling the use of subordination. As it can be seen, this feature does not strictly follow the distribution of the other features. The languages with the most complex use of subordination are Chinese, Korean (GSD), Vietnamese, Latin (ITTB), Spanish (AnCora), Japanese. The Arabic language turns out to be the second most complex one only with respect to the distribution of subordinate clauses, but not when the internal subordinate clause structure is considered (*subord_chain_len*). Interestingly, Chinese, Korean (GSD), Vietnamese, Latin (ITTB) are also the top-four most stable languages for this feature in terms of coefficient of variation.

We conclude this part with some observations about the verbal arity property. As we observed when we explained how it is computed, our intuition is that this measure is highly sensitive to language-specific constraints also related to the obligatory expression of nominal (or pronominal) subject. To verify this hypothesis, we checked in the WALS Online database the feature "Expression of Pronominal Subjects (101A)" and we found that the languages obtaining the highest verbal arity in our analysis, i.e. Urdu (3.22), Hungarian (3.13), Hindi (3.1), Afrikaans (3.01), are not marked for the "Obligatory pronous" value in WALS. This suggests that our feature is able to intercept information not only limited to the nuclear verb structure. Hindi, Urdu and Afrikaans are also among the top-five ranked languages in terms of coefficient of variation.

3.2 A cluster-based comparison

In this last section, we try to understand how languages tend to cluster on the basis of our complexity metrics. To this end, we employ cluster analysis techniques

and specifically we perform a hierarchical clustering using the Ward algorithm on normalized data. We first apply the cluster analysis on all languages of our dataset and then we focus on the most representative language family, i.e. Indo-European.

The purpose of a cluster analysis applied to natural languages is to identify coherent groups of languages (i.e. clusters) whose members are more related each other (in some sense) than members in other groups. In many previous works, clustering has been framed in a typological perspective and informed by properties of languages pointing to different aspects of cross-linguistic diversity typically available in descriptive materials. One of the most informative sources used for this purpose is again WALS, which has been used e.g. by Daumé III & Campbell (2007), who proposed a Bayesian approach for automatically uncovering universal implications from sparse data, and by Georgi, Xia & Lewis (2010) to compare phylogenetic groupings to clusters derived from typological features. Other works have studied graph-theoretic properties of dependency trees for language classification. Liu & Li (2010) proposed a method to cluster languages according to parameters derived from complex network analysis. Features derived from labeled dependency parses were also used by Chen & Gerdes (2017) and applied to UD treebanks, which were clustered according to two quantitative measures of syntactic order variation, i.e. dependency direction and head-dependent distance for each order. In line with the authors of this study, we share the assumption that our cluster-based analysis is not expected to find a categorical answer of grouping languages into fixed language groups – as our complexity measures only partially cover the whole spectrum of language variation –, but rather to identify tendencies of structural proximity between treebanks. In this sense, we were inspired by the most recent developments of the Distributional Typology framework for comparative linguistics (see e.g. Bickel (2015) and Gerdes, Kahane & Chen (2021)), which is mainly focused on the use of statistical methods applied to large sets of fine-grained variables in order to identify quantitative trends across languages.

Figure 3 shows the hierarchical similarity tree resulting from clustering all treebanks of the dataset. The horizontal axis corresponds to the distance between each cluster using the Ward method. As it can be seen, starting from the bottom of the hierarchy, Croatian and Serbian, Ukranian and Czech (PDT), and Latin (PROIEL) and Gothic are the first merged pairs, which are clustered together at a distance lower than 0.2, while Chinese and Japanese are similarly paired together but at a higher distance (about 0.7). As we move up the dendrogram at a distance of about 1.1, we see that treebanks belonging to the Romance genus tend to group into a quite homogeneous cluster, even though with some exceptions represented by treebanks not representative of Romance languages (e.g. Greek, Hebrew, Serbian, etc.). It can also be observed that two Italian treebanks (i.e. ISDT and POSTWITA) and the French Sequoia treebank form a bigger and more heterogeneous cluster,

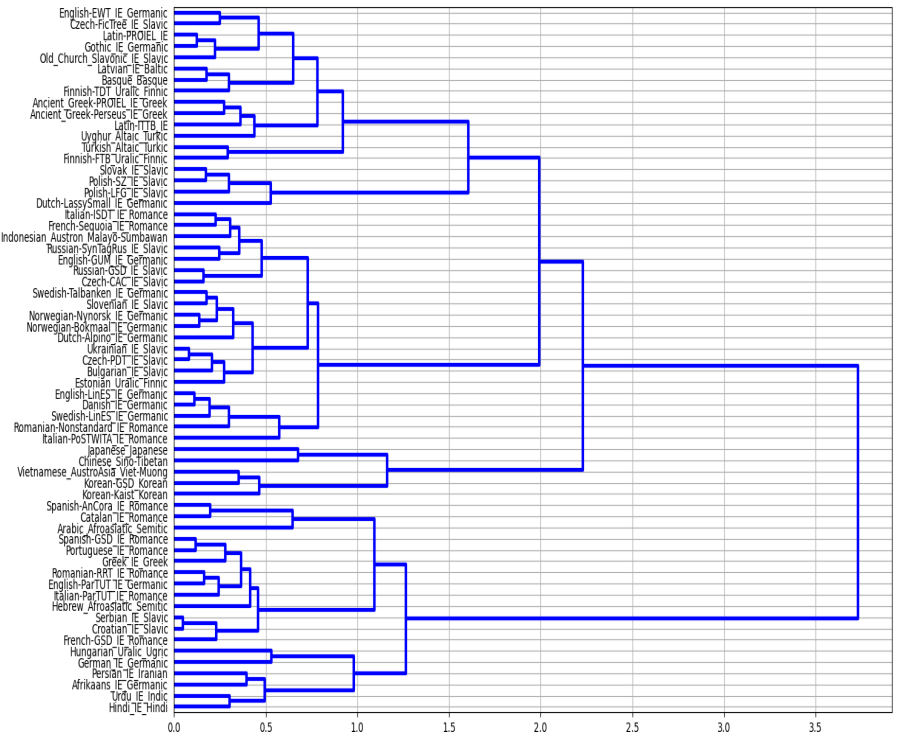

Fig. 3: Hierarchical clustering for all languages.

merged with the former only in a subsequent step (at a distance of about 3.0). We also observe an isolated cluster grouping together Japanese, Chinese, Vietnamese and Korean, which would deserve more in-depth investigation. In fact, despite belonging to different families according to traditional comparative literature, their distance is relatively small with respect to the distribution of the features considered. Among the many possible reasons, a role might be played by the specific annotation criteria defined in the UD project.

The case of Italian and French, whose different treebanks are clustered far away in the tree, also affects other languages for which more than one treebank is available. Note, for instance, the case of English, whose four treebanks are clustered together only at higher levels. Similarly, the Latin PROIEL treebank[9] appears in a small cluster with Gothic and Old Church Slavonic, while the Latin IITB treebank,

9 The Latin PROIEL treebank contains most of the Vulgate New Testament translations plus selections from Caesar's Gallic War, Cicero's Letters to At-

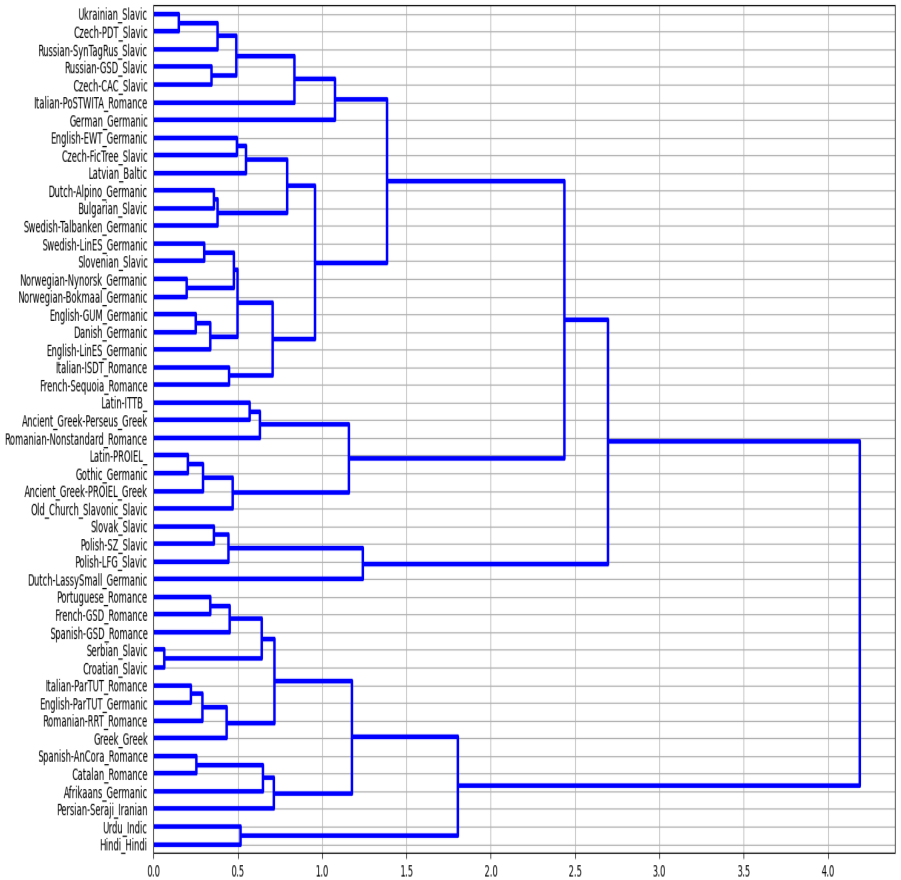

Fig. 4: Hierarchical clustering for Indo-European languages.

which is based on data from the Index Thomisticus corpus,[10] is clustered together with the two treebanks representative of Ancient Greek. These findings suggest that the observed proximity of the considered treebanks may be due not only to language-specific properties, but also to genre-specific features. However, we are aware that a thorough analysis is required to unravel the relationship between genre and complexity starting from an in-depth survey of the textual genres mostly

ticus, Palladius' Opus Agriculturae and the first book of Cicero's De officiis (https://universaldependencies.org/treebanks/la_proiel/index.html).

10 https://universaldependencies.org/treebanks/la_ittb/index.html

represented in each treebank. In fact, there are also languages, such as Norwegian and Korean, which have both their treebanks closely grouped together.

Similar observations hold from inspection of the dendogram resulting from the hierarchical clustering of the Indo-European (Figure 4). Starting from the bottom, languages sharing the same genus and having similar feature values are grouped together. This is the case, for example, of Croatian and Serbian or Catalan and Spanish both represented by the AnCora treebank. We still observe homogeneous groups of languages when we focus on clusters at 0.5 distance. These clusters join together many Slavic, (i.e. Ukrainian, Czech and Russian), and Germanic, (i.e. Swedish, Slovenian, Norwegian, two of the four English treebanks, Danish), languages. In addition, all the treebanks representative of ancient languages are grouped together: the two Latin and Ancient Greek treebanks as well as the Gothic and Old Church Slavonic ones. In this case, the similarity concerns the diachronic variation of language rather than the WALS genus. Interestingly, this cluster also includes one of the two Romanian treebanks, i.e. the NonStandard one, which also contains documents of Old Romanian and folklore.

4 Conclusion

In this study we have proposed a cross-language investigation on linguistic complexity covering more than 60 languages distinguished into different families and genera. We motivated our analysis within the framework of linguistic profiling, a data-driven methodology favored by the availability of large-scale corpora, which assumes that a given language and language variety can be characterized by counting the distribution of a wide set of features representative of phenomena spanning across language domains. We focused here on a rather small subset of features among those that are typically used in linguistic profiling, whose selection has been informed by cognitive, corpus-based and computational linguistics literature on sentence complexity. The availability of multi-lingual treebanks annotated with the same morpho-syntactic and syntactic formalism has guaranteed reliable comparisons since the selected proxies of sentence complexity were computed in the same way across corpora.

We identified tendencies of structural proximity between languages, not always expected in light of typologically-driven classifications. For instance, we observed that languages belonging to the Romance group show a quite homogeneous behavior with respect to several features but also that languages belonging to different language families share a number of characteristics. For example, treebanks representative of the Chinese, Korean, Vietnamese, Japanese, Latin and Spanish

languages contain sentences with the highest use of subordination, and Chinese, Hindi, Urdu and Persian treebanks have the longest dependency links.

The study raises several issues that we believe deserve a thorough analysis. One of these is to establish the effect of textual genre on the assessment of 'general-purpose' language complexity features. We often noticed, in fact, that languages represented by more than one treebank behave quite differently with respect to the same features. Having a better understanding of the relationship between genre and complexity is relevant not only for informing research on genre variation but also from an application perspective: for instance, in the field of readability assessment, to enable the collection of textual resources labeled for genre-specific complexity levels, which can be used as training dataset for machine learning systems. In this respect, a related issue worth investigating concerns the correlation between highest values of the considered features and their variation in a treebank. In our study, we found that in many cases treebanks highly complex for a given feature are also those for which the feature is more stable. This is the case for example for the average length of dependency links or of the use of subordination. Conversely, this does not hold for example for sentence or word length, as treebanks with longer sentences show a high variability.

Our approach has also some limitations which we would like to tackle in the future, starting from the operationalization of some features. For instance, verbal arity as calculated here gives only an approximation of the valency structure of verbal predicates thus not allowing to discriminate obligatory arguments from redundant adjuncts possibly affecting complexity. Similarly, with respect to the use of subordination, it could be also informative to calculate separately the distribution of subordinate clauses of distinct typologies, as well as their relative position with respect to the main clause. Since these properties are known to be related to the interaction between structural and discourse-pragmatic factors (Diessel 2005), they can be relevant also from a language complexity perspective.

References

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3). 321–346.

Berdicevskis, Aleksandrs, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama & Christian Bentz. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the second workshop on universal dependencies (UDW 2018)*, 8–17. Brussels, Belgium: Association for Computational Linguistics.

Bickel, Balthasar. 2015. Distributional typology: statistical inquiries into the dynamics linguistic diversity. In B. Heine & H. Narrog (eds.), *The oxford handbook linguistic analysis*. Oxford University Press.

Bott, Stefan & Horacio Saggion. 2014. Text simplification resources for Spanish. *Language Resources and Evaluation* 48(1). 93–120.

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi & Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. English. In *Proceedings of the 12th language resources and evaluation conference*, 7145–7151. Marseille, France: European Language Resources Association.

Brunato, Dominique, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone & Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2690–2699. Brussels, Belgium: Association for Computational Linguistics.

Brunato, Dominique, Felice Dell'Orletta, Giulia Venturi, Thomas François & Philippe Blache (eds.). 2016. *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*. Osaka, Japan.

Chen, Xinying & Kim Gerdes. 2017. Classifying languages by dependency structure typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 54–63.

Collins-Thompson, Kevin. 2014. Computational assessment of text readability: a survey of current and future research. *ITL - International Journal of Applied Linguistics* 165(1). 97–135.

Daelemans, Walter. 2013. Explanation in computational stylometry. In *Proceedings of the international conference on computational linguistics and intelligent text processing*, 451–462. Springer Berlin Heidelberg.

Daumé III, Hal & Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, 65–72. Prague, Czech Republic: Association for Computational Linguistics. https://www.aclweb.org/anthology/P07-1009.

Demberg, Vera & Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2). 193–210.

Diessel, Holger. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics* 43(3). 449–470.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *Wals online*. Max Planck Institute for Evolutionary Anthropology.

Frazier, Lyn. 1985. Syntactic complexity. In D.R. Dowty, L. Karttunen & A.M. Zwicky (eds.), *Natural language parsing*. Cambridge University Press, Cambridge, UK.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS* 112(33). 10336–10341.

Georgi, R., F. Xia & W. Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Coling*.

Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2021. Typometrics: from implicational to quantitative universals in word order typology. *Glossa: A Journal Of General Linguistics* 6(1). 1–31.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 24(11). 1–76.

Gibson, Edward. 2000. The dependency locality theory: a distance–based theory of linguistic complexity. In W.O.A. Marants & Y. Miyashita (eds.), *Image, language and brain*, 95–126. Cambridge, MA: MIT Press.

Gulordava, Kristina & Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of latin and ancient greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 121–130. Uppsala University, Uppsala, Sweden.

van Halteren, Hans. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the association for computational linguistics*, 200–207.

Haspelmath, Martin. 2009. An empirical test of the agglutination hypothesis. *Scalise, Elisabetta Magni Antonietta Bisetto (eds.), Universals of language today, Dordrecht: Springer*. 13–29.

Hunt, Kellogg W. 1966. Recent measures in syntactic development. *Elementary English* 43(7). 732–739.

Kincaid, J. Peter, Lieutenant Robert P. Fishburne, Richard L. Rogers & Brad S. Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. *Research Branch Report, TN: Chief of Naval Training*. 8–75.

Kortmann, Bernd & Benedikt Szmrecsanyi. 2012. *Linguistic complexity: second language acquisition, indigenization, contact*. De Gruyter.

Lei, Lei & Ju Wen. 2020. Is dependency distance experiencing a process of minimization? A diachronic study based on the state of the union addresses. *Lingua* 239.

Liu, Hai Tao & Wei Wei Li. 2010. Language clusters based on linguistic complex networks. *Chinese Science Bulletin* 55. 3458–3465.

Liu, Haitao. 2017. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2). 159–191.

McDonald, Ryan & Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 122–131. Prague, Czech Republic: Association for Computational Linguistics.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In K. Sinnemaki M. Miestamo & F. Karlsson (eds.), *Language complexity: typology, contact, change*, 23–41. John Benjamins.

Miestamo, Matti, Kaius Sinnemaki & Fred Karlsson (eds.). 2008. *Language complexity: typology, contact, change*. John Benjamins.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé & Franciska de Jong. 2016. Survey: computational sociolinguistics: a Survey. *Computational Linguistics* 42(3). 537–593.

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. In *International conference on intelligent text processing and computational linguistics*, 3–16. Springer.

Nivre, Joakim, Laura Rimell, Ryan McDonald & Carlos Gómez-Rodrıguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)*, 833–841. Beijing, China: Coling 2010 Organizing Committee.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th conference on natural language processing (konvens 2016)*, 13–20.

Rimell, Laura, Stephen Clark & Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, 813–821. Singapore: Association for Computational Linguistics.

Temperley, David & Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics* 4. 67–80.