SIGTYP 2021

The 3rd Workshop on Research in Computational Typology and Multilingual NLP

Proceedings of the Workshop

June 10, 2021

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-34-3

SIGTYP 2021 is the third edition of the workshop for typology-related research and its integration into multilingual Natural Language Processing (NLP). The workshop is co-located with the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021), which takes place virtually this year. Our workshop includes a shared task on robust language identification from speech.

The final program of SIGTYP contains 4 keynote talks, 3 shared task papers, 10 archival papers, and 14 extended abstracts. This workshop would not have been possible without the contribution of its program committee, to whom we would like to express our gratitude. We should also thank Claire Bowern, Miryam de Lhoneux, Johannes Bjerva, and David Yarowsky for kindly accepting our invitation as invited speakers. The workshop is generously sponsored by Google.

Please find more details on the SIGTYP 2021 website: `https://sigtyp.github.io/ws2021.html`

Organizing Committee:

Ekaterina Vylomova, University of Melbourne
Elizabeth Salesky, Johns Hopkins University
Sabrina Mielke, Johns Hopkins University
Gabriella Lapesa, University of Stuttgart
Ritesh Kumar, Bhim Rao Ambedkar University
Harald Hammarström, Uppsala University
Ivan Vulić, University of Cambridge
Anna Korhonen, University of Cambridge
Roi Reichart, Technion – Israel Institute of Technology
Edoardo M. Ponti, Mila Montreal and University of Cambridge
Ryan Cotterell, ETH Zurich

Program Committee:

Željko Agić, Corti
Emily Ahn, University of Washington
Isabelle Augenstein, University of Copenhagen
Emily Bender, University of Washington
Johannes Bjerva, University of Copenhagen
Claire Bowern, Yale University
Miriam Butt, University of Konstanz
Giuseppe Celano, Leipzig University
Agnieszka Falenska, University of Stuttgart
Richard Futrell, University of California, Irvine
Elisabetta Ježek, University of Pavia
Gerhard Jäger, University of Tubingen
John Mansfield, University of Melbourne
Paola Merlo, University of Geneva
Joakim Nivre, Uppsala University
Robert Östling, Stockholm University
Thomas Proisl, FAU Erlangen-Nurnberg
Michael Regan, University of New Mexico
Ella Rabinovich, University of Toronto
Tanja Samardžić, University of Zurich
Richard Sproat, Google Japan
Sabine Stoll, University of Zurich
Daan van Esch, Google AI
Giulia Venturi, ILC “Antonio Zampolli”
Nidhi Vyas, Apple
Ada Wan, University of Zurich
Eleanor Chodroff, University of York
Elizabeth Salesky, Johns Hopkins University
Sabrina Mielke, Johns Hopkins University
Edoardo M. Ponti, University of Cambridge
Damián Blasi, Harvard University
Adina Williams, Facebook
Ivan Vulić, University of Cambridge
Arturo Oncevay, University of Edinburgh
Koel Dutta Chowdhury, Saarland University

Elena Klyachko, National Research University Higher School of Economics
Alexey Sorokin, Moscow State University
Sylvain Kahane, Université Paris Nanterre
Taraka Rama, University of North Texas
Harald Hammarström, Max Planck Institute for the Science of Human History
Olga Lyashevskaya, National Research University Higher School of Economics
Kaushal Kumar Maurya, IIT Hyderabad
Johann-Mattis List, Max Planck Institute for the Science of Human History
Garrett Nicolai, University of British Columbia
Yevgeni Berzak, Technion – Israel Institute of Technology
Olga Zamaraeva, University of Washington
Zoey Liu, Boston College
Jeff Good, University at Buffalo
Priya Rani, National University of Ireland
Silvia Luraghi, University of Pavia
Beata Trawinski, University of Vienna
Miryam de Lhoneux, University of Copenhagen
Kemal Kurniawan, University of Melbourne
Andreas Shcerbakov, University of Melbourne
Ritesh Kumar, Bhim Rao Ambedkar University

Invited Speakers:

Claire Bowern, Yale University
Miryam de Lhoneux, Uppsala University / KU Leuven / University of Copenhagen
Johannes Bjerva, Aalborg University
David Yarowsky, Johns Hopkins University

Table of Contents

Inferring morphological complexity from syntactic dependency networks: a test

Luca Brigada Villa (Università degli Studi di Pavia) & Guglielmo Inglese (KU Leuven – FWO)

Abstract

Research in linguistic typology has shown that languages do not fall into the neat morphological types (synthetic vs. analytic) postulated in the 19th century. Instead, analytic and synthetic must be viewed as two poles of a continuum and languages may show a mix analytic and synthetic strategies to different degrees. Unfortunately, empirical studies that offer a more fine-grained morphological classification of languages based on these parameters remain few. In this paper, we build upon previous research by Liu & Xu (2011) and investigate the possibility of inferring information on morphological complexity from syntactic dependency networks.

1 Introduction

Language classification based on morphological profiles has prominently featured in the linguistic typology research agenda since the earliest days of the discipline.

Earlier 19th century classifications essentially focused on morphological complexity in terms of the number of morphemes per word and the number of meanings per morpheme, and proposed that languages may be typologized into neatly discrete type, e.g. 'isolating', 'agglutinative', 'inflectional' (see Schwegler 1990).[1] However, it soon became clear that such a holistic approach does not adequately capture the variation of natural languages (already Sapir 1921). Instead, morphological complexity should be viewed as an empirically measurable "multidimensional typological space" (Arkadiev & Klamer 2018: 444), in which languages can be arranged based on a number of parameters.[2]

Based on this line of reasoning, scholars have variously tried to measure morphological complexity by means of quantitative methods and classify languages accordingly. In this paper, we build upon a proposal by Liu & Xu (2011) and investigate whether syntactic dependency networks can be effectively used as tools for measuring (at least some aspects of) morphological complexity.

The paper is structured as follows: in Section 2 we review previous research on quantitative approaches to morphological typology. Section 3 briefly introduces syntactic dependency networks and network analysis. Section 4 is devoted to our own analysis. We first illustrate our data and methods (Section 4.1 and 4.2), and then present and discuss our results (Section 4.3 and 4.4). Section 5 contains a summary of our findings.

2 Quantitative morphological typology: previous research

Scholars generally agree that a more accurate and realistic morphological typology can only be achieved through empirical investigations of naturalistic (corpus) data, but how this measurement is to be carried out remains a matter of debate. To our knowledge, there exist two main approaches that have so far been pursued in the quantitative study of morphological typology.[3]

[1] We use the term *morphological complexity* in the narrow sense of *enumerative complexity*, that is, "the number of elements of which a given morphological entity consists, mainly inventory size and string length" (Arkadiev & Gardani 2020: 8).

[2] For the large scale cross-linguistic investigation of some of these parameters see e.g. Bickel & Nichols (2013a; 2013b; 2013c).

[3] By quantitative study, we intend here typological studies based on corpus data, that is, what Levshina (2019) refers to

Proceedings of the Third Workshop on Computational Typology and Multilingual NLP, pages 10–22

The first approach stems from Greenberg (1960). Greenberg proposes that morphological complexity be decomposed in a few easily measurable indexes, e.g. the number of morphemes per word and the number of meanings expressed by each morpheme. To test this approach, Greenberg calculated each index by looking at 100-word stretches of texts in 8 different languages.

Siegel et al. (2014) follow a similar approach and focus on two morphological indexes, that is, the analyticity and the syntheticity indexes. They measure these by taking into account several parameters, including e.g. number of morphemes per words, in randomized samples of 1000 manually annotated token for 19 languages (4 languages plus 13 varieties of English and two English-based creoles).

The main advantage of the approach pursued by Greenberg (1960) and Siegel et al. (2014) is that they employ indexes that are theoretically well-grounded and offer an accurate morphological typology of the languages investigated. However, previous studies of this type present two major shortcomings. The first one concerns the data: both studies focus on a relatively narrow set of languages. The second one concerns the methodology: the indexes must be calculated by manually annotating (a sample of) tokens in each of the languages under investigation. While this methodology undoubtedly results in high quality and reliable data, it is a labor-intensive and time-consuming task, less suitable to investigate morphological complexity on a large cross-linguistic scale.

As an alternative, Liu & Xu (2011) propose to use syntactic dependency networks to explore morphological typology. The main assumption behind this approach is that network structure can be used as a proxy of morphological complexity, which can thus be measured by means of topological indexes of networks (see Section 3). The main advantage of this approach is that it allows to compare a potentially large number of languages for which annotated corpora are available, without the need to manually code each token for its morphological features.

Liu & Xu (2011) results suggest that networks can indeed be a useful tool to explore morphological typology, but their work may be improved in a number of respects. First, the methodology needs to be tested on a wider set of languages (Liu and Xu's sample includes only 15 languages, with a significant overrepresentation of Indo-European languages). Secondly, the authors partly leave open the question of which network measure best captures morphological complexity.

3 Syntactic dependency networks

In this section, we describe syntactic dependency networks and their properties (Section 3.1), and we illustrate various indexes that can be used to interpret network structure (Section 3.2), with a focus on those indexes that we use in our own analysis in Section 4.

3.1 Defining syntactic dependency networks

A network is a structure consisting of a set of objects, called vertices or nodes, and a set of links, called edges. Edges connect two nodes and may be directed, if two nodes are involved in a hierarchical structure, or undirected. Directed and undirected networks differ based on whether they feature directed or undirected edges, respectively.[4]

Networks have been shown to be a suitable tool to represent syntactic relations (Liu 2008; Čech & Mačutek 2009; Čech, Mačutek & Žabokrtský 2011; Passarotti 2014; Čech, Mačutek & Liu 2016). This holds particularly true for dependency grammars, which view syntactic structures as binary and hierarchical relations between lexical nodes (Robinson 1970), thereby allowing the representation of sentences as rooted trees.[5] In Figure 1, we illustrate the representation of the sentences 'John calls Mary', 'John eats an apple', 'The apple is red' and 'Mary buys some apples' as dependency trees.

as *token-based typology* and Gerdes et al. (2021) as *typometrics*. This contrasts with e.g. the classifications proposed by Bickel & Nichols (2013a; 2013b; 2013c), which are based on a sample of few formatives per language (Bickel & Nichols 2013d) and thus fall within the more traditional *type-based* typology (Levshina 2019).

[4] For the purpose of this work, we treat dependencies as undirected.

[5] A tree is a graph in which no cycle can be found. A rooted tree is a tree in which one node is designated as the root of the tree.

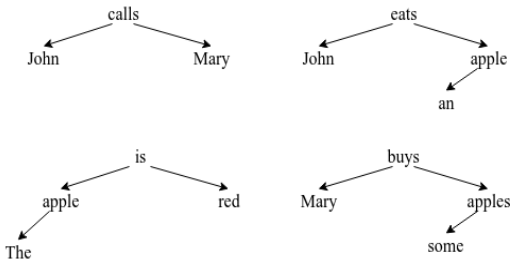

Figure 1: Dependency trees

A syntactic dependency network is a network representing dependency relations. We follow the definition of syntactic dependency network given by Ferrer i Cancho et al. (2004), that is, a set of words *V*, consisting of the vocabulary of a language, and an adjacency matrix *A*. If it happens in at least one sentence that two elements of *V*, let us call them *x* and *y*, are syntactically related, then the value in *A*, corresponding to column *x* and row *y*, will be equal to 1, otherwise it will be 0. The network is then induced from the matrix. This means that syntactic dependency networks built from treebanks actually consist of the combination of all networks that can be drawn from individual dependency trees. Taking the trees in Figure 1 as representing our treebank, the corresponding network has the structure shown in Figure 2.

Dependency networks can be further differentiated into word-based and lemma-based networks (see Čech & Mačutek 2009). The former feature words occurring in sentences as nodes, while in the latter the nodes consist of lemmas. The difference between word- and lemma-based networks is shown in Figure 2 and 3.

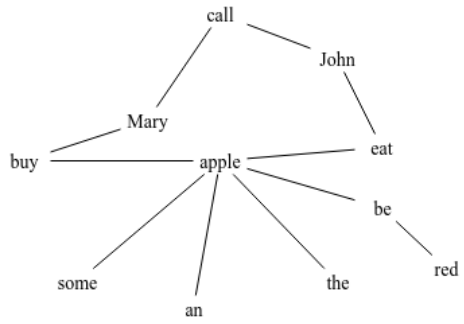

Figure 3: Lemma-based dependency network

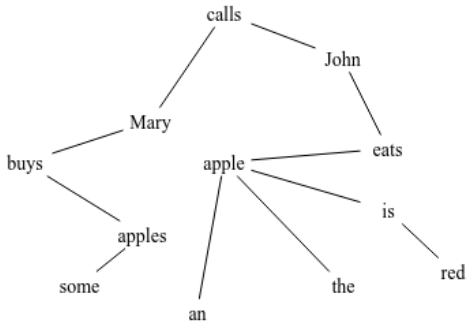

Figure 2: Word-based dependency network

3.2 Network indexes

The structure of networks can be analyzed by taking into account a number of parameters, or indexes. Here, we briefly illustrate the network topological indexes that we employ in our analysis (we refer to Albert & Barabasi 2002; Liu & Xu 2011 for extensive discussion on how the indexes are measured).

Number of edges and nodes: this is the total count of all nodes and edges featured in a network.

Average degree: the count of the links in which a node is involved is called *degree*. The average of the degrees of a network is the simplest measure that can be calculated.

Average path length: in a connected network, it is always possible to find a path between two given nodes. If two nodes are connected, the path length between them is 1, if they are not directly connected, then the path length is computed 'jumping' from one node to another starting from the source node until the target node is reached. The distance is calculated by considering the shortest possible path. The average path length refers to the average of the distances between each pair of nodes in the network.

Clustering coefficient: syntactic dependency networks have the tendency to form clusters in which groups of three elements are completely connected. Clustering coefficient measures the proportion of fully connected triplets of nodes over the number of all the possible groups of three nodes in the network.

Diameter: the diameter of a network is the maximal distance between any pair of its nodes.

Network centralization (Horvath & Dong 2008): network centralization (NC) is a measure to find the most central nodes in a network.

Gamma: according to Albert & Barabási (2002), in so-called real networks the degree distribution follows a power-law. It has been shown that syntactic dependency networks are real networks and likewise follow a power-law $P(k) \sim k^{-\gamma}$ (thus Ferrer i Cancho et al. 2004).

In particular, based on data discussed by (Ferrer i Cancho 2005), it seems that syntactic dependency networks share a common behavior: their degree distributions follow a power-law, their average path length is similar to average path length in random graphs (Erdös-Rényi graphs) and their clustering coefficient is significantly higher than clustering coefficient in random graphs. These features allow us to consider syntactic dependency networks as *small-world* and *scale-free* networks (see further Albert & Barabási 2002; Ferrer i Cancho et al. 2004; Liu & Xu 2011 for discussion).

4 Using networks to measure morphological complexity

Studies by Liu & Xu (2011) and Čech & Mačutek (2009) make a strong case that dependency networks may be used to infer morphological complexity. In this paper, we focus on the networks' potential to explore one component of morphological complexity, that is, the analyticity/syntheticity index. This index reflects the prevalence of synthetic vs. analytic strategies in individual languages. Based on Greenberg's (1960) insights, our assumption is that the index is a gradient, and languages may vary from highly synthetic (prevalence of synthesis) to highly analytic (prevalence of analysis), with several intermediate types.

Following Siegel et al. (2014: 52–53), we distinguish analytic vs. synthetic strategies based on how they convey grammatical information: analytic strategies use free markers, whereas synthetic strategies use bound markers (see also Bickel & Nichols 2013a for discussion).

Dependency treebanks are well suited to explore analyticity/syntheticity for a number of reasons. First, treebanks are already tokenized, which makes it straightforward to single out free vs. bound markers. [6] Moreover, the number of dependencies in a sentence can be indirectly taken as a sign of higher/lower analyticity.

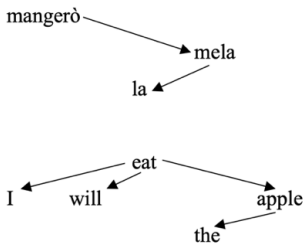

Figure 4: Italian vs. English dependency trees

To illustrate these points, let us compare the dependency trees of the sentence 'I will eat the apple' in Italian and English, as in Figure 4. The main difference between English and Italian is that in Italian grammatical information concerning verbal person/number and TAM is packed by a single form, i.e. *mangerò* 'eat.FUT.1SG', while the same content must be expressed by three free forms *I will eat* in English. In other words, to express future tense, Italian resorts to a more synthetic strategy than English. This is reflected in the number of nodes and links in the trees: the English tree features more nodes and hence more dependencies. This information easily translates into different network structures, in the sense that in principle the more analytic the construction the more edges and nodes the corresponding network will show.

In the reminder of this section, we put Liu & Xu's (2011) intuitions about the connection between analyticity and network structure to a test.

4.1 Data sampling

This study is based on a sample of 42 languages (Appendix A). The sampling procedure has been essentially practical in nature. First, we have only included languages for which treebanks are available in Universal Dependencies (UD) (Nivre et al. 2016; Croft et al. 2017). The reason to work with UD is both practical and theoretical. In the first place, UD allows to easily access already

[6] Clearly, the reliability of tokenization is a potential issue, especially considering problematic items such as clitics. In this study, we work with UD treebanks, which share a uniform tokenization schema. This limits the risk of biases induced by different tokenization styles across treebanks.

annotated data from a variety of languages. From a theoretical viewpoint, the UD annotation schema, which maximizes consistency of annotation across languages, makes UD treebanks particularly well suited for typological studies (see e.g. Levshina 2019; Gerdes et al. 2021).

To maximize diversity among the available UD treebanks, we have picked out one treebank for each language family represented in UD (and one for each branch in each family, where available). Moreover, we have also included historical varieties within the same branch where possible (e.g. Classical Chinese and Mandarin Chinese, Ancient Greek and Modern Greek).

In addition, we have split the treebanks into two groups. The first group features a set of six treebanks that we use to set up our control group. These are languages that can be reasonably taken as instantiating two poles of higher analyticity vs. higher syntheticity. [7] The former include Vietnamese (vie), Mandarine Chinese (zho), and Classical Chinese (lzh). The latter are Russian (rus), Finnish (fin), and Uyghur (uig). The second group includes all the other languages in the sample, whose degree of analyticity/syntheticity we seek to measure.

4.2 Methods

Our study diverges from Liu & Xu (2011) in a number of significant methodological respects. In the first place, Liu & Xu (2011) calculated for each of the 15 languages in their sample several topological indexes and then performed a cluster analysis to classify languages accordingly. In this study, we do not apply clustering techniques. The reason is that clustering analysis may force languages into "hierarchically organized groups" even in absence of a real underlying motivation (Cysouw 2007: 63–64). In our case, we do not in principle expect languages to cluster into neatly defined groups based on their degree of analyticity. Instead, as we have already mentioned, we conceive analyticity/syntheticity as a one-dimension continuum (cf. Gerdes et al. 2021: 13–19).

Abandoning clustering techniques also means that we need to independently single out among the topological indexes those that most likely reflect the difference between the prevalence of analytic vs. synthetic strategies. Moreover, we need take into consideration the different size of the treebanks in our sample (ranging from 955 tokens to 473.881 tokens), as treebank size could lead to potential biases when measuring network indexes.

To overcome these issues, we first established which network indexes perform well in distinguishing analytic vs. synthetic languages irrespective of treebank size. To do so, we set 7 arbitrary sizes (1.000, 5.000, 10.000, 20.000, 30.000, 50.000 and 75.000 tokens) and we extracted one random sub-treebank for each of the above sizes for the languages in the control group.

From each sub-treebank, we induced the corresponding word-based dependency network excluding punctuation marks, symbols and elliptical dependency relations. We calculated the topological indexes described in Section 3 using the python package *igraph* (Csárdi & Nepusz 2006). [8] For the purpose of this paper, we have focused on word-based networks, as these have been claimed to better represent morphological variation than lemma-based networks (Liu & Xu 2011; Čech & Mačutek 2009).

We then carried out a Welch *t*-test (Welch 1947) to establish which indexes are more reliable to separate the two groups, and have picked out only those indexes that perform significantly better across all sub-treebanks' sizes.[9] The Welch *t*-test is used to test the hypothesis that two groups have equal means. The null hypothesis, in our case, was that the two groups means were equal. If a *t*-test performed on a topological index resulted to discard null hypothesis (significance level=0.05), then we consider it as a metric able to separate the two groups, hence possibly reflecting the analytic vs. synthetic distinction.

Once the significant metrics have been singled out, the second step was to measure these indexes for the rest of the languages in our sample and compare them with those of the control group. For

[7] We are aware that the choice of these languages is in part arbitrary, but these are languages (or belong to language families) that have been repeatedly pointed out in the literature as instantiating prototypically analytic vs. synthetic languages.

[8] The code and data used for this study are freely available at https://github.com/bavagliladri/tb2net.

[9] The test was carried out using the python library SciPy (Virtanen et al. 2020)

Table 1: Results of the *t*-test on the control group per size

Variable	1k	5k	10k	20k	30k	50k	75k
n_edges	0.46654	0.29684	0.23868	0.23032	0.20536	0.36893	0.36173
n_nodes	**0.04801**	**0.02542**	**0.02413**	**0.02402**	**0.02458**	0.17776	0.17076
av_degree	0.00219	0.03466	0.05941	0.06613	0.07789	0.30949	0.31478
avg_path_length	**0.02765**	**0.0017**	**0.0028**	**0.0026**	**0.00441**	0.10372	0.09932
clus_coeff	0.2025	0.0186	0.02495	0.03039	0.03114	0.23311	0.22447
diam	0.04674	0.03303	0.06705	0.0083	0.06032	0.08439	0.21913
nc	0.35234	0.0197	0.01896	0.01642	0.00805	0.04418	0.10778
gamma	0.26583	0.12827	0.03547	0.03731	0.0401	0.26701	0.21379

the other languages we extracted only one treebank for the largest possible size (up to 30k, see Section 4.3), in order to make the best use of the available data. [10] For example, for the UD_Wolof-WTB treebank, whose size is 38.937 tokens, we produced a sub-treebank of 30.000 tokens. From these treebanks, we induced the corresponding dependency networks and calculated the relevant network indexes following the procedure outlined above. The results of our analysis are discussed in the next section.

4.3 Results

Let us first discuss the results of the *t*-test performed on the control group. Table 1 reports the p-value for each index across all treebank sizes (with 3 languages per group in the 1k-30k and 2 languages per group in 50-70k; see Appendix B for the raw data). As the results show, the indexes that consistently give a p-value of less than 0.05 are number of nodes and average path length.

The other indexes give a mixed picture. Number of edges is never significant. However, the other indexes are significant for some specific sub-size(s). For example, unlike Liu & Xu (2011: 4), we do not find network centrality (nc) to be a consistently significant index. This index performs well for treebank size 5k-30k, but not for the smallest size of 1k, and we found a similar result for clustering coefficient. By contrast, average degree gives consistent results only for the smallest sizes 1k and 5k. Nevertheless, since none of these indexes performs consistently well for size 1k-30k, for this preliminary study we have decided to leave these aside and focus only on number of nodes and average path length. More research is needed to fully understand the interplay between treebank size and topological indexes of the corresponding networks, also adopting other statistical tests.

In addition, note that none of the indexes yields significant results when the treebank size is 50k tokens or higher. It may be possible that the significant results obtained from the networks induced from the smaller treebanks are due to chance. However, it must be mentioned that only 4 out of the 6 treebanks of the control group have more than 50k tokens and the reduced size of the control group may have affected the statistical testing. For these reasons, for treebanks more than 30k tokens, we have randomly created 30k size sub-treebanks and have only analyzed the corresponding networks, since beyond this size the indexes appear to be less reliable.

We have then measured number of nodes and average path length for the networks induced from the rest of the languages in our sample. The results are reported in Appendix C. In Figure 5 and 6 we visualize the results for 5k and 30k treebanks respectively. Data is visualized as a one-dimension continuum for each index (see Gerdes et al. 2021: 13–19).

[10] An anonymous reviewer suggests that, as an alternative, one could also place each treebank in the uppermost allowable group and then, for treebanks with more than 5k, sample smaller sub-sets for each of the smaller sizes. While we see the potential for this approach, we have not pursued it in this paper. The reason is that based on the control group, we establish which network indexes perform well irrespective of treebank size. Once treebank size becomes irrelevant, this means that for the rest of the sample we can safely look one treebank of the largest possible size.

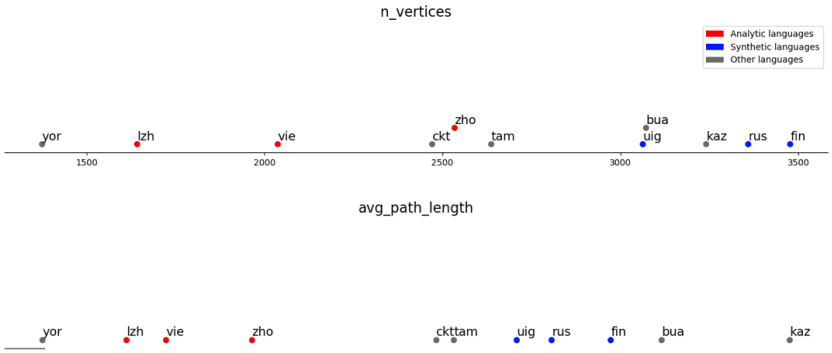

Figure 5: average path length and number of nodes for 5k treebanks

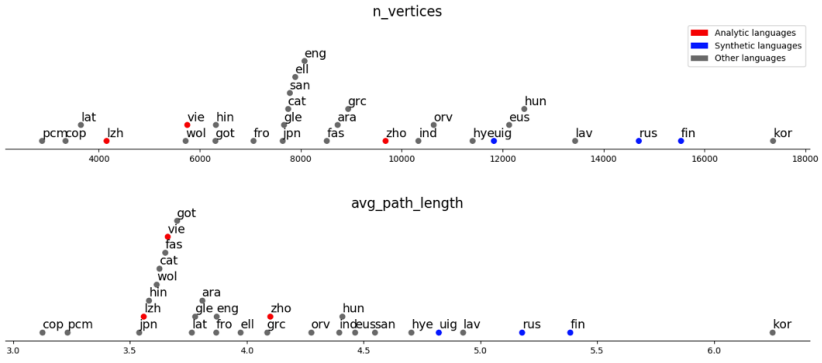

Figure 6: average path length and number of nodes for 30k treebanks

4.4 Discussion

Let us first comment upon the results of the *t*-test on the control group. Our hypothesis that average path length and number of nodes might be taken as proxies for the analyticity index can be linguistically motivated by the nature of networks.

Average path length represents the average distance between any pair of nodes and therefore reflects connectivity in the network. The more highly connected the nodes are, the easier it will be to reach any node in the network starting from any arbitrary point. In particular, the occurrence of hub nodes, that is, highly connected nodes, will result in a generally lower average path length, because hub nodes frequently serve as bridge between nodes which would otherwise be connected by longer paths. As shown by Passarotti (2014), in the case of syntactic dependency networks, hub nodes are often grammatical words like determiners, adpositions, and auxiliaries. Notably, these are as a general rule preferably used in analytic languages, which by definition tend to express grammatical information by means of independent words as opposed to bound morphology (see Siegel et al. 2014: 52–53). The prediction is thus that analytic languages will have a lower average path length than synthetic languages.

Number of nodes also indirectly reflects morphological complexity. In particular, in word-based networks, languages with inflectional morphology will feature more nodes per lexeme, one for each inflected form, than analytic languages. This can clearly be observed in Figure 1, where *apple* and *apples* are two distinct nodes. The prediction is thus that analytic languages will have a lower number of nodes than synthetic languages.[11]

[11] One anonymous reviewer suggests that the same result, i.e. higher number of nodes correlates with higher synthesis, could also be extracted by simply measuring the ratio of different word forms per lemma in treebanks, without the need to resorting to networks. However, a higher number of word forms per lemma does not necessarily mean that a language is more synthetic, but simply that it has larger inflectional paradigms. To achieve a more fine-grained

Both predictions are fully borne out by data from the control group (see Appendix B): networks induced from synthetic languages have higher average path length and higher number of nodes than those from analytic languages.

Turning to the rest of the languages in the sample, for treebanks with size lower than 30k, in most cases the results seem to match our intuitions about the relationship between the indexes under analysis and the analyticity/syntheticity index. Consider Figure 5. First, languages are indeed placed along a continuum, and do not seem to cluster into neatly defined groups. This matches our assumption that analyticity is a continuum. Languages of the control group indeed seem to occupy different regions of the continuum. The other languages also pattern accordingly. For example, Chukchi (ckt) and Buryat (bua), both rich inflectional language (see Dunn 1999; Skribnik 2003), show an average path length comparable to that of synthetic languages. By contrast, Yoruba, which shows a marked analytic profile (Awobuluyi 1978), shows an average path degree even lower than that of the control group analytic languages.

Unfortunately, the picture is not as neat for the rest of the languages in the sample. This is particularly true for the group of treebanks with 30k size (recall that this group also includes reduced versions of all treebanks with size over 30k in our sample). The results shown in Figure 6 can hardly reflect underlying morphological complexity of the languages under analysis. For example, it is not clear why most languages, even highly inflectional ones such as Latin and Ancient Greek, seem to pattern with the analytic languages in the control group. Further study is needed to understand why we get less reliable results with treebanks of higher size. Note that there seems to be a cluster of languages whose dependency networks have average path length between 3.5 and 4.0. This result has previously not been discussed in the literature, and more research is needed to investigate whether this is accidental or not.

Another limitation of the methodology pursued in this paper is that other indexes of morphological complexity cannot be inferred from network structure alone. For example, syntactic dependency networks do not allow to extrapolate more fine-grained information about the internal structure of words in term of cumulation. This means that distinctions that are crucial to morphological typology, such as the distinction between cumulative vs. agglutinative strategies, cannot be measured with this methodology.

5 Conclusions

In this paper, we have put to an empirical test the proposal advanced by Liu & Xu (2011) that syntactic dependency networks can be exploited to investigate cross-linguistic variation in morphological complexity.

Our findings only partly support the validity of this methodology. While we are sympathetic with the underlying assumptions, we must conclude, against Liu & Xu's (2011) more optimistic view, that when applied to larger cross-linguistic datasets, network indexes do not yet yield consistently interpretable results as to morphological complexity.

This means that more research is needed to fully ascertain the suitability of networks to explore morphological complexity. In particular, more attention needs to be paid to the role of treebank size and to the potential impact of annotation schemas. Another potentially confounding factor is that we have worked on networks directly extracted from treebanks as a whole. It needs to be tested whether better results may be achieved by working with networks that operate a finer-grained distinction for e.g. parts of speech.

Finally, we must stress that even for neat data such as that in Figure 5, the proposed correlation between network indexes and the language's analyticity index must remain at this stage tentative. While there might well be a linguistic motivation to link higher number of nodes and average path length to higher syntheticity, the validity of these assumptions needs to be tested against a finer-grained qualitative assessment such as that

result, one would need to calculate and compare the ratio of word forms per lemma for various lemmas and various parts of speech. This is a more complex procedure than simply exploring the number of nodes in a network, which is therefore in principle a more efficient procedure. Notably, variation in paradigm size in inflectional languages can also be explored with networks, by comparing word-based with corresponding lemma-based networks (see Čech & Mačutek 2009).

proposed by Greenberg (1960) and Siegel et al. (2014).

Acknowledgments

We would like to express our gratitude to three anonymous reviewers, whose comments have greatly contributed to improve this paper. The remaining shortcomings are our own. Guglielmo Inglese acknowledges the financial support of the FWO – Research Foundation Flanders (grant n. 12T5320N). The paper is the result of close collaboration between the two authors. For academic purposes, Luca Brigada Villa is responsible of Sections 3, 4.1, 4.2, 5 and Guglielmo Inglese is responsible of Sections 1, 2, 4, 4.3 and 4.4.

References

Albert, Reka & Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74(1). 47–97. https://doi.org/10.1103/RevModPhys.74.47.

Arkadiev, Peter & Francesco Gardani. 2020. Introduction: Complexities in morphology. In Peter Arkadiev & Francesco Gardani (eds.), *The Complexities of Morphology*, 1–19. Oxford: Oxford University Press.

Arkadiev, Peter & Marian Klamer. 2018. Morphological theory and typology. In Jenny Audring & Francesca Masini (eds.), *The Oxford Handbook of Morphological Theory*, 436–454. Oxford: Oxford University Press.

Awobuluyi, Oladele. 1978. *Essentials of Yoruba Grammar*. Ibadan: Oxford University Press Nigeria.

Bickel, Balthasar & Johanna Nichols. 2013a. Inflectional Synthesis of the Verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info).

Bickel, Balthasar & Johanna Nichols. 2013b. Fusion of Selected Inflectional Formatives. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Bickel, Balthasar & Johanna Nichols. 2013c. Exponence of Selected Inflectional Formatives. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Bickel, Balthasar & Johanna Nichols. 2013d. Sampling Case and Tense Formatives. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Čech, Radek & Ján Mačutek. 2009. Word form and lemma syntactic dependency networks in Czech: A comparative study. *Glottometrics* 19. 85–98.

Čech, Radek, Ján Mačutek & Haitao Liu. 2016. Syntactic Complex Networks and Their Applications. In Alexander Mehler, Andy Lücking, Sven Banisch, Philippe Blanchard & Barbara Job (eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, 167–186. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-47238-5_8.

Čech, Radek, Ján Mačutek & Zdeněk Žabokrtský. 2011. The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A: Statistical Mechanics and its Applications* 390(20). 3614–3623. https://doi.org/10.1016/j.physa.2011.05.027.

Croft, William, Dawn Nordquist, Katherine Looney & Michael Regan. 2017. Linguistic Typology meets Universal Dependencies. In *TLT*, 63–75.

Csárdi, Gábor & Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* 1695.

Cysouw, Michael. 2007. New approaches to cluster analysis of typological indices. In Peter Grzybek & Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text*, 61–76. Berlin & New York: de Gruyter. https://doi.org/10.1515/9783110894219.61.

Dunn, Michael John. 1999. *A Grammar of Chukchi*. PhD Dissertation, Australian National University.

Ferrer i Cancho, Ramon. 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. *Problems of quantitative linguistics* 60–75.

Ferrer i Cancho, Ramon, Ricard V. Solé & Reinhard Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*. American Physical Society 69(5). 051915. https://doi.org/10.1103/PhysRevE.69.051915.

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). 17. https://doi.org/10.5334/gjgl.764.

Greenberg, Joseph H. 1960. A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics* 26(3). 178–194.

Horvath, Steve & Jun Dong. 2008. Geometric Interpretation of Gene Coexpression Network Analysis. *PLOS Computational Biology* 4(8).

e1000117.
https://doi.org/10.1371/journal.pcbi.1000117.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572. https://doi.org/10.1515/lingty-2019-0025.

Liu, Haitao. 2008. The complexity of Chinese syntactic dependency networks. *Physica A: Statistical Mechanics and its Applications* 387(12). 3048–3058. https://doi.org/10.1016/j.physa.2008.01.069.

Liu, Haitao & Chunshan Xu. 2011. Can syntactic networks indicate morphological complexity of a language? *EPL (Europhysics Letters)* 93(2). 28005. https://doi.org/10.1209/0295-5075/93/28005.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajicˇ, Christopher D Manning, Ryan McDonald, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666.

Passarotti, Marco. 2014. The importance of being sum : network analysis of a Latin dependency treebank. In Roberto Basili, Alessandro Lenci & Bernardo Magnini (eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014: 9-11 December 2014, Pisa*, 291–295. Pisa: Pisa University Press.

Robinson, Jane J. 1970. Dependency Structures and Transformational Rules. *Language* 46(2). 259–285. https://doi.org/10.2307/412278.

Sapir, Edward. 1921. *Language. An Introduction to the Study of Speech*. New York: Harcourt, Brace & World.

Schwegler, Armin. 1990. *Analyticity and Syntheticity: A Diachronic Perspective with Special Reference to Romance Languages*. Berlin & New York: de Gruyter.

Siegel, Jeff, Benedikt Szmrecsanyi & Bernd Kortmann. 2014. Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages* 29(1). 49–85. https://doi.org/10.1075/jpcl.29.1.02sie.

Skribnik, Elena. 2003. Buryat. In Juha Janhunen (ed.), *The Mongolic Languages*, 102–128. London & New York: Routledge.

Welch, B. L. 1947. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika* 34(1/2). 28–35. https://doi.org/10.2307/2332510.

Appendix A: Language sample

Language*	**ISO code**	**Treebank**	**Token size**
Akkadian	akk	UD_Akkadian-RIAO	21961
Arabic	ara	UD_Arabic-PADT	242383
Bambara	bam	UD_Bambara-CRB	11873
Buryat	bua	UD_Buryat-BDT	8333
Catalan	cat	UD_Catalan-AnCora	473881
Chukchi	ckt	UD_Chukchi-HSE	4740
Coptic	cop	UD_Coptic-Scriptorium	45496
Greek	ell	UD_Greek-GDT	56145
English	eng	UD_English-GUM	97979
Basque	eus	UD_Basque-BDT	101444
Persian	fas	UD_Persian-PerDT	457439
Finnish	fin	UD_Finnish-TDT	171836
Old French	fro	UD_Old_French-SRCMF	170740
Irish	gle	UD_Irish-IDT	104547
Gothic	got	UD_Gothic-PROIEL	55317
Ancient Greek	grc	UD_Ancient_Greek-PROIEL	213980
Mbyá Guaraní	gun	UD_Mbya_Guarani-Thomas	1070
Hindi	hin	UD_Hindi-HDTB	328101
Hungarian	hun	UD_Hungarian-Szeged	36212
Armenian	hye	UD_Armenian-ArmTDP	42213
Indonesian	ind	UD_Indonesian-GSD	103238
Japanese	jpn	UD_Japanese-GSD	172209
Kazakh	kaz	UD_Kazakh-KTB	8316
Korean	kor	UD_Korean-Kaist	310205
Komi Zyrian	kpv	UD_Komi_Zyrian-Lattice	4060
Latin	lat	UD_Latin-LLCT	206859
Latvian	lav	UD_Latvian-LVTB	179744
Classical Chinese	lzh	UD_Classical_Chinese-Kyoto	232188
Erzya	myv	UD_Erzya-JR	13038
Old Russian	orv	UD_Old_Russian-TOROT	149484
Naija	pcm	UD_Naija-NSC	100557
Russian	rus	UD_Russian-GSD	78200
Sanskrit Vedic	san	UD_Sanskrit-Vedic	27117
Nort Sami	sme	UD_North_Sami-Giella	22702
Tamil	tam	UD_Tamil-TTB	8580
Tagalog	tgl	UD_Tagalog-Ugnayan	955
Thai	tha	UD_Thai-PUD	21916
Uyghur	uig	UD_Uyghur-UDT	32401
Vietnamese	vie	UD_Vietnamese-VTB	33887
Wolof	wol	UD_Wolof-WTB	38937
Yoruba	yor	UD_Yoruba-YTB	7119

Chinese	zho	UD_Chinese-GSD	105195

*Languages of the control group are in bold.

Appendix B: number of nodes and average path length for the control group

Size[12]	**Index**	**ISO code**					
		fin	**rus**	**uig**	**lzh**	**vie**	**zho**
1k	avg_path_length	7,0926602	7,1181678	8,3712409	5,1750507	5,7376548	5,729498
	n_nodes	822	783	801	549	650	692
5k	avg_path_length	6,4890935	6,2090146	6,0436406	4,1943988	4,3817425	4,7894008
	n_nodes	3477	3359	3063	1642	2037	2534
10k	avg_path_length	6,0495081	5,8471496	5,5202574	3,8458437	4,0076761	4,4870999
	n_nodes	6309	6125	5306	2362	3050	4300
20k	avg_path_length	5,6139062	5,4378837	5,0921595	3,7027029	3,7923894	4,217152
	n_nodes	11174	10715	8888	3513	4588	7255
30k	avg_path_length	5,3847064	5,1784228	4,8219065	3,5586503	3,6615379	4,1005896
	n_nodes	15535	14699	11828	4154	5753	9682
50k	avg_path_length	5,0969112	4,95705	-	3,428777	-	3,9497258
	n_nodes	23451	22020	-	5273	-	13379
75k	avg_path_length	4,9043123	4,7678429	-	3,3602782	-	3,8341283
	n_nodes	31823	29734	-	6330	-	17393

Appendix C: network indexes for the sample languages

ISO code	**Size**	**avg_path_length**	**n_nodes**
gun	1k	4,3342128	410
kpv	1k	7,5279103	765
tgl	1k	3,8475797	383
bua	5k	6,7303552	3072
ckt	5k	5,6628477	2471
kaz	5k	7,3374322	3241
tam	5k	5,7456242	2637
yor	5k	3,7964219	1375
bam	10k	3,1309446	1063
myv	10k	5,6029887	5137
akk	20k	4,0322794	2802
sme	20k	4,4667009	7750
tha	20k	3,5982284	4076
ara	30k	3,8098087	8732

[12] For reasons of space, in Appendix B and C we only report data on average path lengths and number of nodes. Data on the other indexes can be consulted at https://github.com/bavagliladri/tb2net.

cat	30k	3,6261545	7752
cop	30k	3,1254	3341
ell	30k	3,973389	7892
eng	30k	3,8710204	8076
eus	30k	4,463648	12130
fas	30k	3,6510337	8517
fro	30k	3,8691324	7066
gle	30k	3,7789008	7670
got	30k	3,7018547	6311
grc	30k	4,0874524	8941
hin	30k	3,5812191	6320
hun	30k	4,4091939	12430
hye	30k	4,704363	11406
ind	30k	4,3958122	10332
jpn	30k	3,5393915	7644
kor	30k	6,2509272	17359
lat	30k	3,7646329	3645
lav	30k	4,9259688	13437
orv	30k	4,2763492	10635
pcm	30k	3,2327011	2876
san	30k	4,5486621	7785
wol	30k	3,6142526	5720