Quasy 2021

Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)

Proceedings

To be held as part of SyntaxFest 2021
21–25 March, 2022
Sofia, Bulgaria

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 978-1-955917-15-5

Preface

Following its first edition that was held in Paris in 2019, the Second Workshop on Quantitative Syntax (Quasy 2021) takes place again as part of SyntaxFest, which co-locates four related but independent events:

- The Sixth International Conference on Dependency Linguistics (Depling 2021)
- The Second Workshop on Quantitative Syntax (Quasy 2021)
- The 20th International Workshop on Treebanks and Linguistic Theories (TLT 2021)
- The Fifth Workshop on Universal Dependencies (UDW 2021)

The reasons that suggested bringing these four events together in 2019 still hold in 2021. There is a continuing, strong interest in corpora and dependency treebanks for empirically validating syntactic theories, studying syntax from quantitative and theoretical points of view, and for training machine learning models for natural language processing. Much of this research is increasingly multilingual and cross-lingual, made in no small part possible by the Universal Dependencies project, which continues to grow at currently nearly 200 treebanks in over 100 languages.

For these reasons and encouraged by the success of the first SyntaxFest, which was held in 2019 in Paris, we – the chairs of the four events – decided to bring them together again in 2021. Due to the vagaries of the COVID-19 pandemic, it was eventually decided to push the actual SyntaxFest 2021 back to March 2022. In order not to delay the publication of new research and not to conflict with other events, we decided however to publish the proceedings that you are now reading in advance, in December 2021.

As in 2019, we organized a single reviewing process for the whole SyntaxFest, with identical paper formats for all four events. Authors could indicate (multiple) venue preferences, but the assignment of papers to events for accepted papers was made by the program chairs.

38 long papers were submitted, 25 to Depling, 11 to Quasy, 17 to TLT, and 24 to UDW. The program chairs accepted 30 (79%) and assigned 8 to Depling, 5 to Quasy, 7 to TLT, and 10 to UDW. 22 short papers were submitted, 6 to Depling, 7 to Quasy, 9 to TLT, and 9 to UDW. The program chairs accepted 14 (64%) and assigned 3 to Depling, 3 to Quasy, 3 to TLT, and 5 to UDW.

At the time of this writing, we do not yet know whether SyntaxFest will be a hybrid or purely online event. We regret this uncertainty but are nevertheless looking forward to it very much. Our sincere thanks go to everyone who is making this event possible, including everybody who submitted their papers, and of course the reviewers for their time and their valuable comments and suggestions. We would like to thank Djamé Seddah, whose assistance and expertise in organizing SyntaxFests was invaluable. Finally, we would also like to thank ACL SIGPARSE for its endorsement and the ACL Anthology for publishing the proceedings.

Radek Čech, Xinying Chen, Daniel Dakota, Miryam de Lhoneux, Kilian Evang, Sandra Kübler, Nicolas Mazziotta, Simon Mille, Reut Tsarfaty (co-chairs)

Petya Osenova, Kiril Simov (local organizers and co-chairs)

December 2021

Program co-chairs

The chairs for each event (and co-chairs for the single SyntaxFest reviewing process) are:

- Depling:
  - Nicolas Mazziotta (Université de Liège)
  - Simon Mille (Universitat Pompeu Fabra)
- Quasy:
  - Radek Čech (University of Ostrava)
  - Xinying Chen (Xi'an Jiaotong University)
- TLT:
  - Daniel Dakota (Indiana University)
  - Kilian Evang (Heinrich Heine University Düsseldorf)
  - Sandra Kübler (Indiana University)
- UDW:
  - Miryam de Lhoneux (Uppsala University / KU Leuven / University of Copenhagen)
  - Reut Tsarfaty (Bar-Ilan University / AI2)

Local Organizing Committee of the SyntaxFest

- Petya Osenova (Bulgarian Academy of Sciences)
- Kiril Simov (Bulgarian Academy of Sciences)

Program Committee for the Whole SyntaxFest

Patricia Amaral (Indiana University Bloomington)
Valerio Basile (University of Turin)
David Beck (University of Alberta)
Ann Bies (Linguistic Data Consortium, University of Pennsylvania)
Xavier Blanco (UAB)
Igor Boguslavsky (Universidad Politécnica de Madrid)
Bernd Bohnet (Google)
Cristina Bosco (University of Turin)
Gosse Bouma (Rijksuniversiteit Groningen)
Miriam Butt (Universität Konstanz)
Marie Candito (Universtité Paris 7 / INRIA)
Radek Cech (University of Ostrava)
Giuseppe Giovanni Antonio Celano (University of Pavia)
Xinying Chen (Xi'an Jiaotong University)
Silvie Cinková (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics)
Cagri Coltekin (University of Tuebingen)
Benoit Crabbé (Université Paris 7 / Institut national de recherche en informatique et en automatique, Paris)
Daniel Dakota (Indiana University)
Eric De La Clergerie (Institut national de recherche en informatique et en automatique, Paris)
Felice Dell'Orletta (Institute for Computational Linguistics, National Research Council, Pisa)
Kaja Dobrovoljc (Jožef Stefan Institute)
Kilian Evang (Heinrich Heine University Düsseldorf)
Thiago Ferreira (University of São Paulo)
Ramon Ferrer-I-Cancho (Universitat Politècnica de Catalunya)
Kim Gerdes (Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay)
Carlos Gómez-Rodríguez (Universidade da Coruña)
Jan Hajic (Institute of Formal and Applied Linguistics, Charles University, Prague)
Eva Hajicova (Institute of Formal and Applied Linguistics, Charles University, Prague)
Dag Haug (University of Oslo)
Richard Hudson (University College London)
András Imrényi (Eszterházy Károly Egyetem)
Leonid Iomdin (Institute for Information Transmission Problems, Russian Academy of Sciences)
Jingyang Jiang (Zhejiang University)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre / CNRS)
Vaclava Kettnerova (Institute of Formal and Applied Linguistics)
Sandra Kübler (Indiana University Bloomington)
Guy Lapalme (University of Montreal)
François Lareau (Observatoire de linguistique Sens-Texte, Université de Montréal)
Alessandro Lenci (University of Pisa)
Nicholas Lester (University of Zurich)
Lori Levin (Carnegie Mellon University)
Haitao Liu (Zhejiang University)

Marketa Lopatkova (Institute of Formal and Applied Linguistics, Charles University, Prague)
Olga Lyashevskaya (National Research University Higher School of Economics)
Teresa Lynn (Dublin City University)
Jan Macutek (Mathematical Institute of the Slovak Academy of Sciences / Constantine the Philosopher University in Nitra)
Robert Malouf (San Diego State University)
Alessandro Mazzei (Dipartimento di Informatica, Università di Torino)
Nicolas Mazziotta (Université de Liège)
Alexander Mehler (Text Technology Group, Goethe-University Frankfurt am Main)
Wolfgang Menzel (Department of Informatics, Hamburg University)
Jasmina Milicevic (Dalhousie University)
Simon Mille (Pompeu Fabra University)
Yusuke Miyao (The University of Tokyo)
Simonetta Montemagni (Institute for Computational Linguistics, National Research Council, Pisa)
Kaili Müürisep (University of Tartu)
Alexis Nasr (Laboratoire d'Informatique Fondamentale, Université de la Méditerranée, Aix-Marseille II)
Sven Naumann (University of Trier)
Anat Ninio (The Hebrew University of Jerusalem)
Joakim Nivre (Uppsala University)
Pierre Nugues (Lund University, Department of Computer Science Lund, Sweden)
Kemal Oflazer (Carnegie Mellon University-Qatar)
Timothy Osborne (Zhejiang University)
Petya Osenova (Sofia University / Institute of Information and Communication Technologies, Sofia)
Robert Östling (Department of Linguistics, Stockholm University)
Agnieszka Patejuk (Polish Academy of Sciences / University of Oxford)
Alain Polguère (Université de Lorraine)
Prokopis Prokopidis (Institute for Language and Speech Processing/Athena RC)
Laura Pérez Mayos (Pompeu Fabra University)
Owen Rambow (Stony Brook University)
Rudolf Rosa (Institute of Formal and Applied Linguistics, Charles University, Prague)
Tanja Samardzic (University of Zurich)
Giorgio Satta (University of Padua)
Nathan Schneider (Georgetown University)
Olga Scrivner (Indiana University Bloomington)
Djamé Seddah (Alpage, Université Paris la Sorbonne)
Alexander Shvets (Institute for Systems Analysis of Russian Academy of Sciences)
Maria Simi (Università di Pisa)
Achim Stein (University of Stuttgart)
Reut Tsarfaty (Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot)
Francis M. Tyers (Indiana University Bloomington)
Zdenka Uresova (Institute of Formal and Applied Linguistics, Charles University, Prague)
Gertjan Van Noord (University of Groningen)
Giulia Venturi (Institute for Computational Linguistics, National Research Council, Pisa)
Veronika Vincze (Hungarian Academy of Sciences, Research Group on Artificial Intelligence)
Relja Vulanovic (Kent State University at Stark)

Chunshan Xu (anhui jianzhu university)
Xiang Yu (University of Stuttgart)
Zdenek Zabokrtsky (Institute of Formal and Applied Linguistics, Charles University, Prague)
Amir Zeldes (Georgetown University)
Daniel Zeman (Institute of Formal and Applied Linguistics, Charles University, Prague)
Hongxin Zhang (Zhejiang University)
Yiyi Zhao (Institute of Applied Linguistics, Communication University of China, Beijing)
Heike Zinsmeister (University of Hamburg)
Miryam de Lhoneux (University of Copenhagen)
Marie-Catherine de Marneffe (The Ohio State University)
Valeria de Paiva (Samsung Research America and University of Birmingham)

Additional Reviewers

Chiara Alzetta
Aditya Bhargava
Lauren Cassidy
Simon Petitjean
Xenia Petukhova
Daniel Swanson
He Zhou
Yulia Zinova

Table of Contents

Successes and failures of Menzerath's law at the syntactic level

Aleksandrs Berdicevskis
Språkbanken Text, Department of Swedish, University of Gothenburg
aleksandrs.berdicevskis@gu.se

Abstract

Menzerath's law is a quantitative generalization which predicts a negative correlation between the mean size of parts of a unit and the number of parts in the unit. In this paper, I use Universal Dependencies to perform a cross-linguistic test of Menzerath's law at two syntactic levels: whether the number of clauses in a sentence negatively correlates with mean clause length in this sentence and whether the number of words in a clause negatively correlates with mean word length in this clause. Menzerath's largely holds at the former level and largely does not at the latter. I discuss other interesting patterns observed in the data and propose some tentative partial explanations.

1 Introduction

Quantitative laws such as, for instance, Zipfian rank-frequency law (Piantadosi, 2014) or abbreviation law (Bentz and Ferrer-i-Cancho, 2016) are perhaps ones of the most universal generalizations that can be made about language. *Universal* here can be understood as both 'true for all / most languages' and 'true for various domains / levels of language'.

Another oft-cited generalization is Menzerath's law (Altmann, 1980; Stave et al., 2021), also called Menzerath-Altmann law. Menzerath's law predicts a negative correlation between the mean size of parts of a unit and the number of parts in the unit. Thus, the more sub-units (constituents) a linguistic unit (carrier unit, or construct) has, the shorter these units are expected to be on average. For instance, the more clauses a sentence contains, the shorter the mean length of these clauses (in words) is expected to be (Altmann, 1980).

Menzerath's law has been tested for various types of units in various languages (and also beyond language) and mostly (though not universally) found to be true (see an overview in Section 2). Most studies, however, used relatively small corpora (or even dictionaries), often of just one language, often not open-access, often shallowly annotated for the specific study. I use the Universal Dependencies (UD) collection to perform the largest-scale (to date) study of Menzerath's law at two syntactic levels: sentence–clause–word and clause–word–grapheme (see Section 3). I demonstrate that Menzerath's law works quite well at the former level, but not at the latter (see Sections 4, 5 and 6).

There is currently no consensus on *why* Menzerath's law emerges (or why it does not), and thus I cannot fully explain the observed results. In Section 7, however, I discuss which insights can be gleaned from the UD analysis and which hypotheses deserve further testing.

2 Background

2.1 Defining Menzerath's law

Any particular application of Menzerath's law has to be described at three levels: the length of a *unit* (for instance, clause), measured in *sub-units* (for instance, words), is supposed to negatively correlate with the mean length of sub-units, measured either in *sub-sub-units* (for instance, phonemes or graphemes) or using a suitable continuous measure (for instance, seconds). In this paper, two triples will be analyzed: sentence–clause–word and clause–word–grapheme.

Menzerath's law has been shown to hold at different levels in different languages, but it is sometimes overlooked that there are at least two ways to interpret the claim *Menzerath's law holds*. One interpretation (which will be used in this paper) is 'the mean size of a sub-unit and the number of sub-units in the unit are negatively correlated' (Stave et al., 2021). Another interpretation is 'the relation between the mean size of a sub-unit (y) and the number of sub-units (x) can be approximated by a specific function'. The function is typically assumed to be $y(x) = ax^b e^{-cx}$ (Altmann, 1980), often simplified to $y(x) = ax^b$, though other variants have also been proposed (Milička, 2014). Sometimes the first interpretation is labelled as Menzerath's law, while the second one as Menzerath–Altmann's law (Ferrer-i-Cancho et al., 2014). Both interpretations rest on the assumption that number of sub-units and the mean size of sub-units are related (Mačutek et al., 2019).

While it is possible that there is a negative correlation *and* the relation can be approximated by Menzerath–Altmann's function (a power law with an exponential cutoff), it may also be that the latter is true, while the former is not. In Chinese, for instance, Menzerath–Altmann's function works well for sentence–clause–word ($R^2 = 0.85$) and clause–word–component ($R^2 = 0.77$; *component* is a constructing unit of a logogram) (Chen and Liu, 2019). Visual inspection of Chen and Liu's data, however, shows that the relation is clearly non-monotonic (down-up), and measuring Spearman's correlation coefficient shows there is no negative correlation for clause–word–component ($r = 0.42, p = 0.016$), while for sentence–clause–word the results are somewhat ambivalent ($r = -0.51, p = 0.052$). In a similar vein, Buk and Rovenchak (2008) report fitting Menzerath–Altmann's function for sentence–clause–word in Ukrainian, but the visualization of the data shows a non-monotonic (up-down) pattern, and Spearman's coefficient (calculated only for those sentence lengths which Buk and Rovenchak consider "reliable", that is, those for which at least 20 datapoints are available) does not show a negative correlation ($r = -0.13, p = 0.748$).

It can be argued that the latter, fit-a-model approach offers a more exact description of the reality. Following this logic, many studies (Cramer, 2005; Kelih, 2010; Baixeries et al., 2013; Milička, 2014) focus on finding the most appropriate formula and fine-tuning the parameters. On the other hand, if the model is complex enough, virtually any curve can be approximated reasonably well. To avoid overfitting, a clear theoretical explanation of the model is desirable. The existing explanations of Menzerath's law (see Section 2.3) mostly address the negative correlation, though attempts at explaining the Menzerath–Altmann's function and even interpreting its parameters have also been made (Köhler, 1984). I am not, however, aware of any explanation that would have successfully addressed the non-monotonic patterns observed above. Thus, in this study, Menzerath's law is understood as the negative correlation, without an attempt to describe the exact mathematical relation. The purpose of the study is to find out whether the law holds at the syntactic level.

2.2 Existing evidence

Altmann (1980, p. 129) predicts that Menzerath's law will hold for sentence–clause–word (otherwise the sentence presumably loses clarity). He also considers sentence–word–subword unit (word length can be measured in different ways: phonemes, graphemes, syllables, morphemes), but does not make a specific prediction, noting that "a monotonic decrease of word length can hardly be expected".

The first hypothesis (sentence–clause–word) has been tested before. Apart from the references mentioned in Section 2.1, Teupenhayn and Altmann (1984) report that Menzerath's law holds for German. Hou et al. (2017) find that in Chinese, it holds in formal written texts, but not in other registers. Xu and He (2020), however, demonstrate that in English, it holds for different registers. Roukk (2011), analyzing parallel texts in Russian and German and Russian and English, reports poor fitting results. Her data are too small for a correlation test to yield reliable results, but from a visual inspection it is obvious that there is no clear downward trend.

The second hypothesis (clause–word–subword unit) has received much less attention, but see the aforementioned study by Buk and Rovenchak (2008) and a relevant discussion by Altmann (1983)

Mačutek et al. (2017) look at clause–phrase–word in Czech, where *phrase* is defined as a subtree consisting of a node which is directly dependent of the clause predicate and all nodes that are (directly or indirectly) dependent on this node. They report good fitting results, and applying Spearman's test to their data (following their approach, only to those clause lengths that have more than 10 datapoints) yields a strong negative correlation ($r = -0.92, p = 0.001$).

Note that all these studies have at least one (often more) of the following limitations: they were performed for one language only; small corpora were used; those corpora had shallow annotation (for instance, number of clauses estimated by simply counting the number of finite verbs), often created specifically for the study; the data are not openly available.

In fact, the only large cross-linguistic study on Menzerath's law that I am aware of was performed by Stave et al. (2021), but it deals with the word–morpheme–grapheme level.

2.3 Explanations

There is no ubiquitously accepted explanation of why Menzerath's law is expected to hold. It has been argued to be mathematically trivial, but Ferrer-i-Cancho et al. (2014) provide evidence against this view.

It is typically assumed that Menzerath's law, similarly to Zipf's abbreviation law, emerges from efficiency pressures, but what exactly those pressures are is not fully understood. Köhler (1984) hypothesizes that the sub-units and the "structural information" about the connections between them must be stored at the same "register" in the brain (Vulanovic and Köhler, 2005). As the number of sub-units increases, so does the amount of structural information, and the only way to free up the necessary storage space is to use shorter sub-units. Milička (2014) develops this hypothesis further, but in both accounts the notion of structural information remains very vague.

Gustison et al. (2016) claim that Menzerath's law is caused by pressure for compression. They propose a unified formal mathematical framework for the explanation of Menzerath's law and Zipf's law of abbreviation.

It can actually be asked whether Menzerath's law cannot (at least in some cases) be reduced to Zipf's law of abbreviation. Consider, for instance, the word–morpheme–phoneme level. The more morphemes in a word, the higher the chance that many of them will be affixes rather than roots, have higher frequency and be on average shorter. Stave et al. (2021), however, show that for word–morpheme–grapheme, both Zipf's and Menzerath's law are at work, and removing one of them results in a poorer fit of a model.

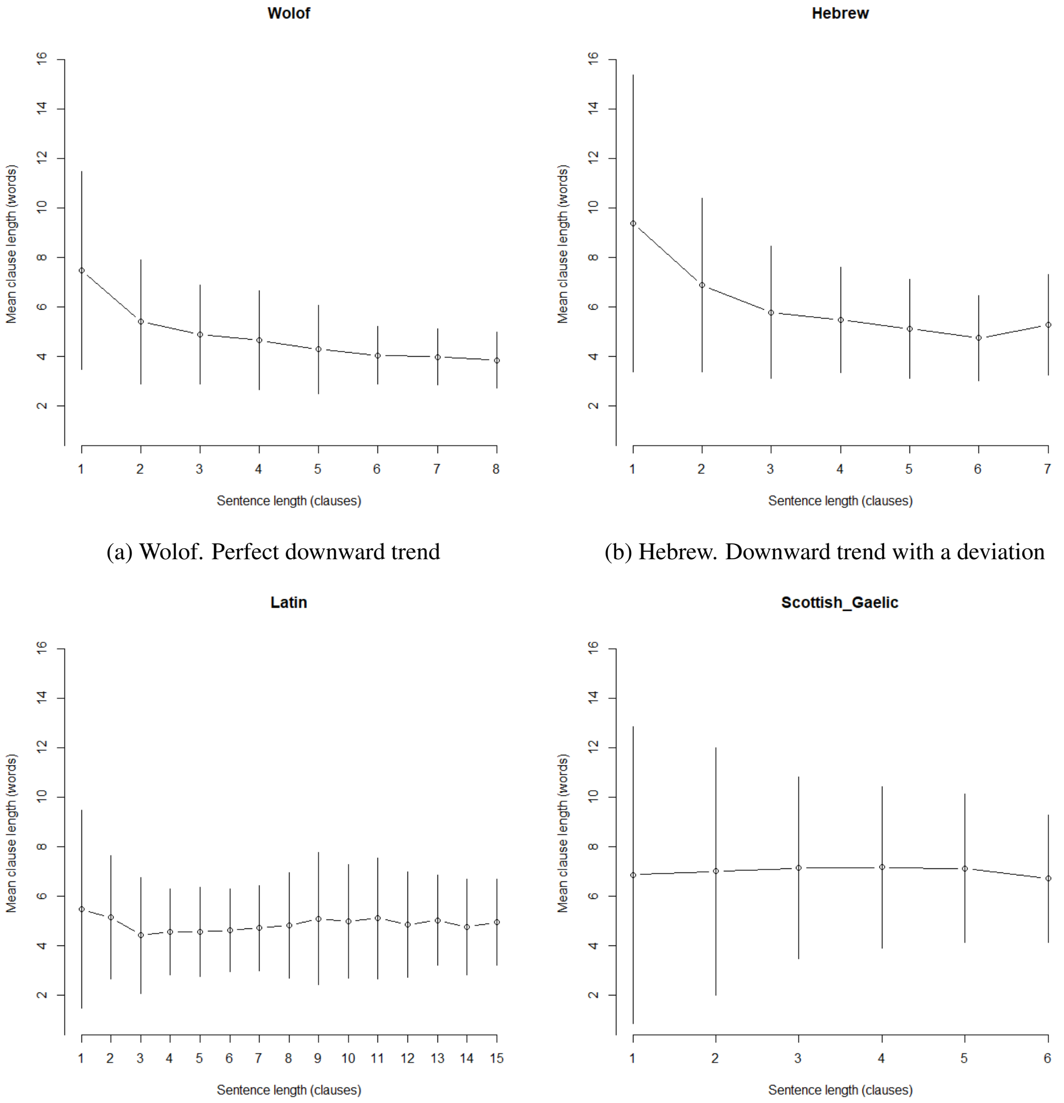
Coming back to syntax, the following level-specific explanation can be proposed for the sentence–clause–word level. Clauses often share certain elements. Open clausal component (raising and control structures, `xcomp` in UD), for instance, by definition does not have an internal subject, but the main clause may contain an element that functions as an (external) subject (cf. *Mary wants to buy a book*, where *Mary* is the subject of *wants*, but also the (external) subject of *buy*). Coordinated clauses can have shared arguments (*Mary is singing and dancing*, where *Mary* is the subject of both verbs), while repeated verbs can be omitted (gapping: *I like tea, and you coffee*). It can be expected that the number of clauses may correlate with the number of shared elements, thus reducing the average clause length. In a similar vein, clauses can act as elements of another clause. The length of the main clause *per se* can decrease, if dependent clauses fulfill the roles that would otherwise have been played by non-clausal dependents (see a test of this hypothesis in Section 4.2).

The present study is thus exploratory rather then confirmatory. It seeks to test whether Menzerath's law holds for sentence–clause–word and clause–word–grapheme across languages, and whether the cross-linguistic data lend support to any tentative explanations.

3 Materials and methods

I use corpus data from Universal Dependencies (UD) 2.8.1 (Zeman et al., 2021). All treebanks that do not have surface forms or that have less than 10,000 tokens are excluded from consideration. Naija-NSC treebank is also excluded, since it has an unusually high proportion (29%) of `dep` relation (which should normally be avoided).

If a language has more than one treebank that fit the requirements, they are all concatenated. The



(a) Wolof. Perfect downward trend

(b) Hebrew. Downward trend with a deviation

(c) Latin. No clear downward trend

(d) Scottish Gaelic. No clear downward trend

Figure 1: Examples of correlation between sentence length (in clauses) and mean clause length (in words). Error bars show interquartile range. Sentence lengths with fewer than 50 datapoints were excluded

final dataset has 78 languages from 15 families (Indo-European, Afro-Asiatic, Mande, Basque, Mongolic, Sino-Tibetan, Uralic, Austronesian, Turkic, Mayan, Korean, Dravidian, Tai-Kadai, Austro-Asiatic, Niger-Congo). Note that how different genres are represented varies strongly across languages and treebanks. Genre is likely to affect the distribution of lengths of all units (sentences, clauses, words) and thus may potentially be a relevant factor. Nonetheless, since many treebanks do not have explicit detailed metadata about which sentence belongs to which genre, I do not attempt to control for genre.

The key notions ("sentence", "word", "clause") are operationalized as follows. "Sentences" are equivalent to UD sentences. Note, however, that sentence segmentation may not be a trivial task, for instance, for oral speech, ancient languages and social media, and thus even at the sentence level some inconsistencies across treebanks are possible.

"Words" are equivalent to UD tokens with minor exceptions. Punctuation marks (PUNCT) are excluded. Symbols (SYM) and unclassifiable tokens (X) must also be excluded (these labels can be used, for instance, for very long tokens like URLs, which can skew the results). However, unlike PUNCTs, SYMs and Xs can potentially have their own dependents and thus be important elements of the syntactic structure: it is not clear whether in such cases it is legitimate to remove them, but leave the rest of the sentence. For this reason, all sentences containing at least one SYM or X are excluded completely. Empty nodes (nodes with IDs like 1.1) are excluded, since they do not exist (should not inflate clause length) and do not have their own length. For multiword tokens, the token denoted by the range ID (e.g. 1–3), i.e. the surface token, is included in the analysis, the corresponding syntactic tokens (1, 2, 3) are excluded.

"Clauses" are most problematic, since there is no straightforward way to demarcate clauses in UD (as in most dependency grammars). Here, a clause consists of a node which has an incoming "clausal" relation (clausal root) and all descendants (both direct and indirect: children, grandchildren and so on) of the clausal root that do not have an incoming "clausal" relation.

Clausal relations are `root`, `csubj` (clausal subject), `ccomp` (clausal complement), `advcl` (adverbial clause modifier), `acl` (relative clause modifier), `parataxis`, some cases of `xcomp` (open clausal complement), some cases of `conj` (coordination). Relation subtypes are not distinguished (i.e. everything after a colon is ignored: `csubj:pass` is treated as `csubj`, `acl:relcl` as `acl`). `xcomp` is considered a clausal relation if its child is a verb, i.e. *great* in *You look great* does not start a new clause, while *work* in *I started to work there yesterday* does. `conj` is considered a clausal relation if either the parent or the child is a verb. The idea is to distinguish between clausal and non-clausal coordination, but the problem is that in cases of ellipsis, the head of a clause is not necessarily a verb. The rule "at least one of the conjuncts has to be a verb" covers some of such cases, but not the ones like *Jack a teacher, Jill a doctor*, which are possible and frequent in some languages.

Overall, the operationalization is necessarily crude and will certainly to some extent err on both sides: make false clause splits and fail to split when it should. Apart from the coordination problem, the following issues can be mentioned. Words (e.g. participles) can have `verb` as the POS label, but not actually behave like verbs. The `parataxis` relation can be argued to not always introduce a new clause. `dep`, `reparation` and `discourse` may deserve a special treatment, the former two should possibly be excluded, while the latter can be argued to introduce a new clause, at least in some cases. All these questions cannot be properly addressed without a thorough language- and treebank-specific linguistically-informed manual analysis.

Removing punctuation marks and empty tokens may in some cases lead to clauses or sentences consisting of zero words (e.g. if a clause/sentence consisted of an exclamation mark). All sentences where at least one clause has zero length are excluded.

Note that the language sample is not balanced (either genetically, areally or typologically). Excluding overrepresented language groups (mostly Indo-European) would lead to an undesirable data loss, since these languages also tend to have larger treebanks, which can be assumed to yield more robust and reliable results. To this end, I avoid averaging across languages (with the exception of clause type analysis in Section 4.2). Readers are encouraged to keep in mind that certain biases can emerge from the sample properties.

	negative	**positive**	**none**
sentence–clause–word	68	0	10
clause–word–grapheme	12	29	37
sentence–word–grapheme	26	19	30
sentence–clause–phrase	38	5	35
clause–phrase–word	58	2	18
phrase–word–grapheme	11	22	45

Table 1: Summary of the correlation tests across languages at different levels. The total amount of languages may vary across levels, since languages which do not have enough datapoints are excluded from the analysis

Section 4.2 describes an additional analysis that seeks to explore whether Menzerath's at the sentence–clause–word level can be explained by the fact that clauses share elements. Section 6 describes additional robustness analyses.

The code that was used to run the analyses and its detailed output are available at `https://github.com/AleksandrsBerdicevskis/menzerath`.

4 Results: Sentence–clause–word

4.1 General results

For every sentence in every language, I measured (according to operationalizations outlined in Section 3) how many clauses it contains and how many words the clauses in this sentence on average contain. I visually inspected the relation between the two variables for all languages. To prevent the results being skewed by outliers (usually a very small number of very long sentences), only those sentence lengths which had at least 50 datapoints were included. Note that languages vary greatly in terms of how many sentence lengths are represented in the data. After the 50-sentence filter is applied, Kiche, for instance, only has sentences which contain one or two clauses, while Icelandic covers the whole range from one to fifteen clauses.

Most typical patterns are represented by examples in Figure 1. In the vast majority of languages, the average clause length decreases monotonically according to what seems to be a power law (see, for instance, Wolof in Figure 1a). Sometimes, minor deviations from monotonicity are observed, often at large sentence length values (see Hebrew in Figure 1b). Nonetheless, even with the deviations most languages still exhibit a clear general downward trend. Those few for which the downward trend is not observed include, for instance, Latin (Figure 1c) and Scottish Gaelic (Figure 1d). In Latin, there is a decrease, but only in the beginning, while in Scottish Gaelic, there is rather a very small upward trend.

To do a formal test, I calculated Spearman's correlation coefficients between sentence length and clause length for all languages. They are reported in Table 3 in Appendix A (together with corresponding p-values) and summarized in Table 1. When summarizing the results, p-values, however, should be treated with caution, since they strongly depend on sample size, that is, how many different sentence lengths are represented in the data. For languages with small range of lengths p-values will never be small, even if perfect correlation is observed. Komi-Zyrian, for instance, demonstrates a perfect negative correlation, but only four different sentence lengths are represented (1–4 clauses), and the p-value is a theoretical minimum of 0.083. For this reason, the following criteria were applied. If the absolute value of the correlation coefficient was equal to or larger than 0.70, the language was labelled as demonstrating as either negative or positive correlation, regardless of the p-value. The same was done if the absolute value of the coefficient was larger than or equal to 0.30 and smaller than 0.70 and the p-value was smaller than or equal to 0.05. In other cases the correlation was assumed to be absent.

Under this interpretation, there are no cases of (anti-Menzerathian) positive correlation and 10 cases where the correlation could not be detected. It is difficult to tell whether it happens because it is truly absent or whether the sample is too small, and thus not clear whether the datapoints should be interpreted as ‘Menzerath’s law does not hold’ or ‘Unknown whether Menzerath’s law holds’. If we concentrate on languages with ranges 1–6 or larger (on the assumption that they yield more reliable samples), then seven languages out of 52 do not clearly conform to Menzerath’s law: Latin, Scottish Gaelic, Icelandic, Old East Slavic, Old French, Finnish and Turkish. The other three (“small”) languages that do not have a correlation are Manx, Breton and Sanskrit.

4.2 Is sharing elements across clauses the answer?

As a preliminary test of the hypothesis that Menzerath’s law at the sentence–clause–word level can be explained by the fact that certain words are syntactically shared between clauses, I performed the following analyses.

First, I compared average lengths of various types of clauses. For every language, I extract all sentences that have exactly two clauses, one main (matrix) clause, one dependent (though in cases of coordination, the main–dependent contraposition is actually somewhat artificial). For dependent clauses, “type” is equivalent to the incoming relation of the clause root (that is, `ccomp`, `conj` etc.). For every clause type, its average length within language is calculated (types with less than 50 datapoints within a language were excluded). Note that if sentences which contain more than two clauses were included, the comparison would have to become much more nuanced. Main clauses could have different number of dependent clauses, while dependent clauses could have double roles: act as main clauses for their own dependents. These factors can potentially affect length distribution, and would have to be taken into account. For simplicity, the analysis is limited to two-clause sentences.

Clause lengths vary greatly across languages and treebanks. To correct for that and focus on the comparison across clause types within language, I normalized the average length of every type by average length of a simple sentence within the same language (that is, a sentence consisting of one and only one clause). These normalized lengths are then averaged across languages. The results are presented in Table 2. Keep in mind that such averaging may yield heavily skewed results, since the language sample is not balanced (and interquartile ranges suggest large variation for all types). Note also that not every clause type is represented in every language.

`xcomp`, as expected, tends to be short. The same is true for `parataxis`, probably because this relation is often used for short interjected clauses like parenthetical constructions (for example, *for example* or *of course*), tag questions etc. Interestingly, the longest type is not `main`, but `ccomp`.

As mentioned in Section 2.3, dependent clauses may perform functions of non-clausal dependents. `csubj` functions as a subject and is a clausal equivalent of `nsubj`, `advcl` is a clausal equivalent of `advmod`, `ccomp` and `xcomp` can be said to be clausal equivalents of `obj`, though note that this last correspondence is less clear. Consider now a main clause which has one of these clausal dependents, for instance, `csubj`. According to the operationalization used in this paper, the words contained by the dependent clause are not included into the main clause. In other words, there is a subject, but it is “outside” of a clause. If, however, the dependent was non-clausal (`nsubj`), it (and all its dependents) would have been inside the main clause and contributed to its length. It is no surprise then that main clauses are shorter than simple sentences. It is, however, interesting whether this is the only reason. To test that, I measure the decrease in length caused by having a clausal dependent (e.g. `csubj`) is approximately equal to the average length of a corresponding non-clausal dependent (`nsubj`).

I label every main clause in the two-clause-sentence sample described above by the type of dependent clause it has (`xcomp` and `ccomp` are merged together and labelled `comp`). The mean length of every “main-clause type” (normalized by the length of the simple sentence) across languages is reported in Table 2 in the column “main length”.

Using the simple-sentence sample, I calculated the mean length (in words) of `nsubj`, `advmod` and `obj`. The column “diff” in Table 2 shows the normalized difference between the simple sentence length and the sum of two lengths: that of main-clause type (e.g. `csubj`) and the corresponding non-clausal dependent (`nsubj`). As all other numbers in the table, the difference is normalized by the simple-sentence length. If the hypothesis is correct, the difference should be close to zero, and indeed it is for `comp` and `advcl` (though note large interquartile ranges), but not for `csubj`.

Type	Length	IQR	Main length	IQR	Diff	IQR
ccomp	0.93	0.20	0.66	0.19	-	-
main	0.86	0.13	-	-	-	-
csubj	0.84	0.17	0.54	0.19	0.18	0.22
conj	0.78	0.17	0.90	0.16	-	-
advcl	0.74	0.16	0.89	0.17	-0.06	0.19
acl	0.72	0.19	1.06	0.21	-	-
xcomp	0.65	0.13	0.66	0.19	-	-
parataxis	0.62	0.24	0.82	0.19	-	-
comp	-	-	0.61	0.14	0.02	0.16

Table 2: Mean lengths across languages. The "main length" column should be read as 'mean length of a main clause having a dependent clause of the specified type'. "Diff" is a difference between the simple sentence length and the sum of two lengths: that of main-clause type ("main length") and the corresponding non-clausal dependent (e.g. textttnsubj for texttcsubj). All numbers are normalized by the mean length of a simple sentence in the same language. IQR = interquartile range.

No other clear patterns are observed. There does not seem to be any strong correlation between the length of the dependent clause of a certain type and corresponding main clause type. Interestingly, main clauses that have an `acl` clause are slightly longer than simple sentences.

5 Results: Clause–word–grapheme

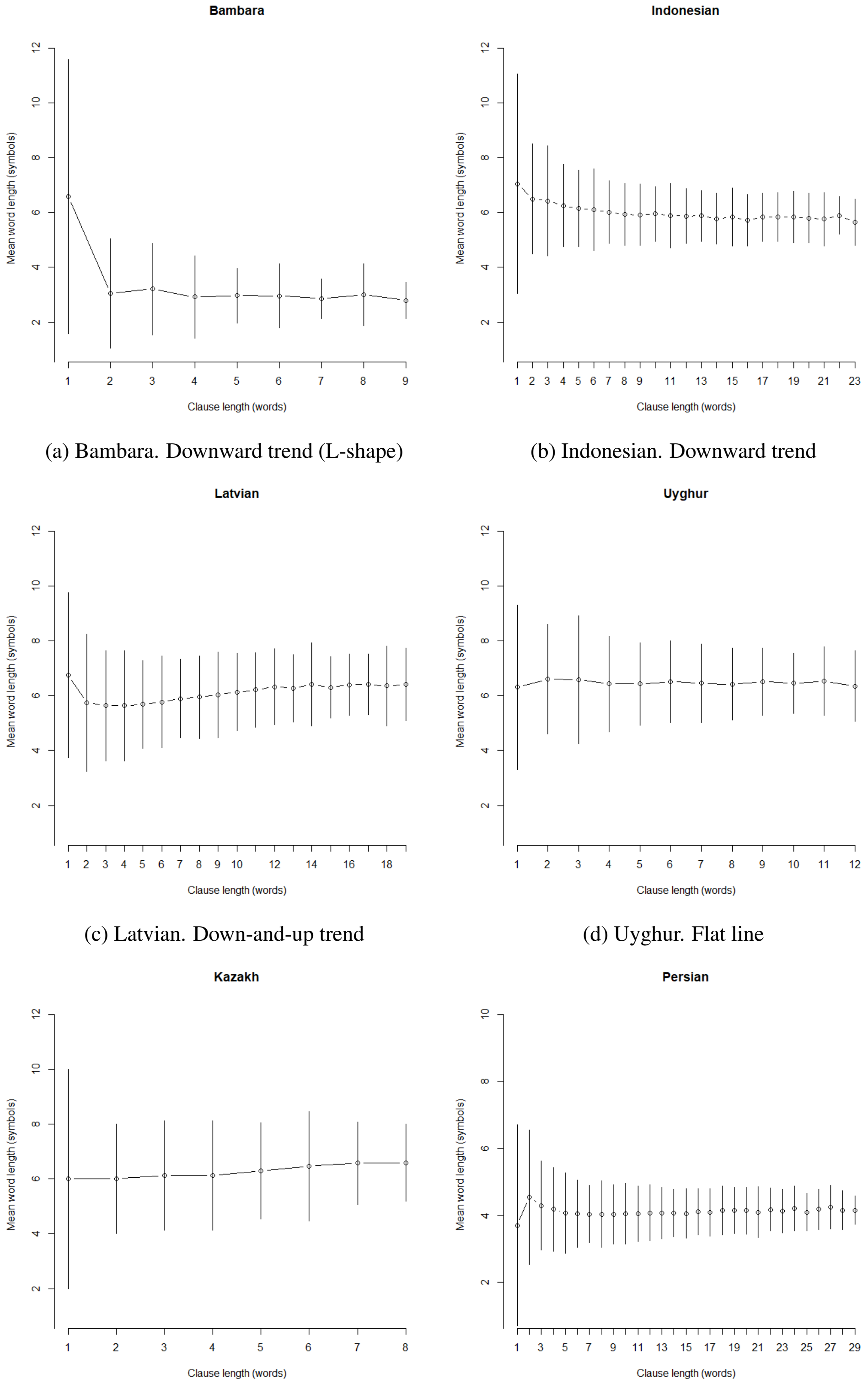
Exactly as with the sentence–clause–word analysis, I measured for every clause in every language how many words it contains and how many graphemes the words in this clause on average contain. I visually inspected the relation between the two variables for all languages. Again, only those lengths that had at least 50 datapoints were included.

Most typical patterns are represented by examples in Figure 2. Overall, the results were more variable than for the sentence–clause–word analysis, where one dominant pattern was observed. For clause–word–grapheme, 29 languages also exhibit a downward trend. Most often, it is L-shaped: a very steep decrease in the beginning, followed by a nearly flat line (see, for instance, Bambara in Figure 2a). In a few cases, the decrease is more gradually spread over the curve (Indonesian in Figure 2b). For 42 languages, an U-curve is observed, first a decrease and then a comparable increase (Latvian in Figure 2c). For four languages, the differences are so small that the pattern is best described as a flat line (see, for instance, Uyghur in Figure 2d). For four languages, there is an upward trend (Kazakh in Figure 2e). Finally, Persian (Figure 2f) exhibits a unique pattern: an inverted U-curve, an increase followed by a decrease.

Spearman's correlation coefficients were calculated in the same way as for the sentence–clause–word level and summarized in Table 1. As can be seen from the summary, the adherence to Menzerath's law at the clause–word–grapheme level is much weaker. Note, however, that correlation coefficients are not really informative for the languages with clear non-monotonic patterns. Since I can propose no explicit hypothesis to explain the observed data, an inferential test is not appropriate: it is unclear *what* it can infer. Technically, some kind of non-linear regression model could of course be fitted to the data, but in the absence of a specific theory to test, the model would end up having many researcher degrees of freedom (Tong, 2019; Simmons et al., 2011), which is undesirable. I limit myself to labelling the observed curves as DOWN, UP, DOWN-UP, FLAT or UP-DOWN. The formalized procedure to determine the shape of the curve is described in Appendix B. The results are summarized above and reported in detail in Table 3 in Appendix A.

6 Robustness analyses

In order to test whether the results are robust, I reran the analyses with various thresholds instead of 50 datapoints per sentence/clause length (0, 20, 100). There were no qualitative changes of the overall



(a) Bambara. Downward trend (L-shape)

(b) Indonesian. Downward trend

(c) Latvian. Down-and-up trend

(d) Uyghur. Flat line

(e) Kazakh. Upward trend

(f) Persian. Up-and-down trend

Figure 2: Examples of correlation between clause length (in words) and average word length (in graphemes). Error bars show interquartile range. Clause lengths with fewer than 50 datapoints were excluded

picture.

Since clause can be argued to be a problematic and / or imperfectly operationalized construct, I ran an analysis for the sentence–word–grapheme level (ignoring the clause level). The results resemble the ones for clause–word–grapheme (see Table 1).

Mačutek et al. (2017) reported that Menzerath's law holds for clause–phrase–word in Czech (see Section 2.2). It can be questioned whether their operationalization of phrase is theoretically adequate (in general, *phrase* is a less theory-neutral notion than *clause*), but I used it to run the analyses for clause–phrase–word and sentence–clause–phrase. I reproduced their findings for Czech, but overall, compared to sentence–clause–word, the adherence to Menzerath's law was slightly lower for clause–phrase–word, much lower for sentence–clause–phrase and even lower for phrase–word–grapheme (see Table 1).

7 Discussion

At the sentence–clause–word level, Menzerath's law largely holds, regardless of corpus size and typological or genealogical properties of language. It is not clear what is special about ten (or seven, depending on how one counts) languages that do not demonstrate an expected correlation. It can be noticed that four (or three) of them are ancient languages: Latin, Old East Slavic, Old French (and Sanskrit), but there are other ancient languages (e.g. Old Church Slavonic or Classical Chinese) that conform to Menzerath's law.

Clink and Lau (2020), analyzing primate communication, reach a somewhat similar conclusion: Menzerath's law holds in some cases, but not always (though in their study the adherence rate is much lower). They hypothesize that while the pressure for efficiency may facilitate compliance with Menzerath's law, other pressures may affect communication, sometimes to the extent that the law no longer holds. It is not, however, clear, which pressures could affect, for instance, non-Menzerathian Finnish and Icelandic, but not Menzerathian Estonian and Norwegian.

One potential confound is register, or genre (Hou et al., 2017). It is a question for future research to what extent Menzerath's law is robust to genre (and if it is not, why).

The explanation of why Menzerath's law (largely) holds is still wanting. The shared-element account that I propose seems to explain some cases, but not all, and it is not clear whether it is the sole reason. This hypothesis can potentially be further explored by using enhanced dependencies available in some UD treebanks (e.g. by measuring whether the shorter length of coordinated clauses can be "compensated" by taking into account shared dependents and elided verbs, or whether `xcomp` is shorter solely because it does not have an internal subject). Overall, it may be useful to consider whether Menzerath's law should be explained by level-specific factors, general optimization principles, or both.

At the clause–word–grapheme level, Menzerath's law generally does not hold. The observation by Altmann (1980) that the relationship is probably not monotonic turns out to be at least partly true. In the vast majority of cases, the mean word length as a function of number of words follows one of the two patterns: either L-shaped (steep decrease and then an almost flat line) or U-shaped (decrease and increase). L-shaped cases can be said to adhere to Menzerath's law, but first, they are less frequent than U-shaped ones, second, not all of them demonstrate a strong negative correlation.

Again, there does not seem to be any obvious way to explain the observed variance by different properties of languages or treebanks. Writing system may potentially be a confound. Apart from alphabets, the writing systems represented in the sample include (impure) abjads (vowels are omitted or partly omitted; e.g. Arabic), abugidas (consonant-vowel units are based on a consonant letter; vowel notation is secondary; e.g. Hindi) and logographic scripts (e.g. Mandarin). Japanese is a special case, using a mixture of a syllabary (kana) and a logographic script (kanji). However, if the writing system plays a role, its contribution is inconsistent: (Mandarin) Chinese and Classical Chinese do not conform to Menzerath's law, while Cantonese does; Hindi (Devanagari, abugida) does not, while Amharic (Ge'ez script, abugida) does.

It can be argued that graphemic word length is not the most adequate measure, and that phonemic length should be preferred. These measures, however, tend to be strongly correlated (Piantadosi et al., 2011). Moreover, should Menzerath's law hold for phonemes, it would probably mean that there is some kind of optimization pressure due to which it emerges in oral speech. But then it is very likely that the same pressure would also affect written language (and most of the analyzed corpora contain predominantly written texts) and the law should hold for graphemes, too.

An anonymous reviewer raises two more important concerns. First, it can be questioned whether Menzerath's law should actually hold for *corpora* and not *texts* (cf. a discussion about inter- and intratextual laws by Grzybek and Stadlober (2011)). Given that the law is formulated as a relationship between the length of a unit and a sub-unit, and that it is hypothesized to emerge due to some kind of optimization pressure, I do not see any reason to assume that it should be valid only for texts and not for any sample of units, provided that the sample is large and representative. For any corpus, it can of course be questioned whether it is large and representative enough, but usually corpora tend to do better on these two scales than single texts.

The second concern is that Menzerath's law may not be valid if the unit, the sub-unit and the sub-sub-unit are not at the adjacent levels of the hierarchy. It can be argued that by testing the law on clause–word–grapheme, I am hopping over a level, since grapheme is not an immediate constituent of a word, and instead syllables or morphemes should be used. It is, however, unclear, which is the more appropriate unit, syllable or morpheme (or whether the law should work equally well for both). Furthermore, it is likely that graphemic (and phonemic) length is highly correlated with both syllabic and morphemic length. (To give an example: I measured the Spearman's correlation coefficient between the graphemic and the morphemic length of Swedish words, using the CoDeRooMor dataset (Volodina et al., 2021): $r = 0.83, p < 0.001$.) Note also that the robustness analyses described in Section 6 suggest that while adding or removing hierarchical levels (e.g. removing clause or adding phrase) affects the results, it does not change the overall picture. Nonetheless, this is a reasonable concern, and it would of course be beneficial to reproduce this study with syllable or morpheme as a sub-sub-unit. The problem is that the necessary resources are lacking.

Unlike Stave et al., I do not test for the role of Zipf's abbreviation law. For sentence–clause–word, this hardly is possible, since clauses are not repeated in languages often enough to enable frequency estimates. For clause–word–grapheme, I cannot propose an explicit prediction for the role of clause length that could have been tested by a regression model (see Section 5).

To conclude, Menzerath's law does not seem to be universal. It does not hold at some levels of analysis, and even at those where it does, some languages (or at least corpora) are exceptions. The reasons for that (both compliance and non-compliance) are not fully clear. Further studies should focus on explanatory approaches and on reproducing the existing results on larger and better samples.[1]

Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

References

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10.

Gabriel Altmann. 1983. H. Arens'«Verborgene Ordnung» und das Menzerathsche Gesetz. In Manfred Faust, Roland Harweg, Werner Lehfeldt, and Götz Wienold, editors, *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik*, pages 31–39. Gunter Narr, Tübingen.

Jaume Baixeries, Antoni Hernández-Fernández, Núria Forns, and Ramon Ferrer-i-Cancho. 2013. The parameters of the Menzerath-Altmann law in genomes. *Journal of Quantitative Linguistics*, 20(2):94–104.

Chris Bentz and Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, pages 1–4. University of Tübingen.

[1] Supplementary materials are available at `https://github.com/AleksandrsBerdicevskis/menzerath`.

Solomija Buk and Andrij Rovenchak. 2008. Menzerath–Altmann law for syntactic structures in Ukrainian. *Glottotheory*, 1(1):10–17.

Heng Chen and Haitao Liu. 2019. A quantitative probe into the hierarchical structure of written Chinese. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 25–32, Paris, France, August. Association for Computational Linguistics.

Dena J. Clink and Allison R. Lau. 2020. Adherence to Menzerath's law is the exception (not the rule) in three duetting primate species. *Royal Society Open Science*, 7(11):201557.

Irene Cramer. 2005. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics*, 12(1):41–52.

Ramon Ferrer-i-Cancho, Antoni Hernández-Fernández, Jaume Baixeries, Łukasz Dębowski, and Ján Mačutek. 2014. When is Menzerath-Altmann law mathematically trivial? A new approach. *Statistical Applications in Genetics and Molecular Biology*, 13(6):633–644.

Peter Grzybek and Ernst Stadlober. 2011. Do we have problems with Arens' law? A new look at the sentence-word relation. In Peter Grzybek and Reinhard Köhler, editors, *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, pages 203–215. De Gruyter Mouton.

Morgan L. Gustison, Stuart Semple, Ramon Ferrer-i-Cancho, and Thore J. Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, 113(19):E2750–E2758.

Renkui Hou, Chu-Ren Huang, Hue San Do, and Hongchao Liu. 2017. A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 24(4):350–366.

Emmerich Kelih. 2010. Parameter interpretation of the Menzerath law: evidence from Serbian. In Peter Grzybek, Emmerich Kelih, and Ján Mačutek, editors, *Text and Language*, pages 71–80, Wien. Presens Verlag.

Reinhard Köhler. 1984. Zur Interpretation des Menzerathschen Gesetzes [On the interpretation of the Menzerath's law]. *Glottometrika*, 6:177–183.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100–107, Pisa, Italy, September. Linköping University Electronic Press.

Ján Mačutek, Jan Chromý, and Michaela Koščová. 2019. Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics*, 26(1):66–80.

Jiří Milička. 2014. Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, 21(2):85–99.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Maria Roukk. 2011. The Menzerath-Altmann law in translated texts as compared to the original texts. In Peter Grzybek and Reinhard Köhler, editors, *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, pages 605–610. De Gruyter Mouton.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366. PMID: 22006061.

Matthew Stave, Ludger Paschen, François Pellegrino, and Frank Seifart. 2021. Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard*, 7(s3):20190076.

Regina Teupenhayn and Gabriel Altmann. 1984. Clause length and Menzerath's law. *Glottometrika*, 6:127–138.

Christopher Tong. 2019. Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician*, 73(sup1):246–261.

Elena Volodina, Yousuf Ali Mohammed, and Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 178–189, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Relja Vulanovic and Reinhard Köhler. 2005. Syntactic units and structures. In Reinhard Köhler, Gabriel Altmann, and Rajmund Piotrowski, editors, *Quantitative linguistics: An international handbook*, pages 274–291. Walter de Gruyter, Berlin.

Lirong Xu and Lianzhen He. 2020. Is the Menzerath-Altmann law specific to certain languages in certain registers? *Journal of Quantitative Linguistics*, 27(3):187–203.

Daniel Zeman, Joakim Nivre, et al. 2021. Universal Dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix A. Detailed results across languages

	sentence–clause–word			**clause–word-grapheme**			**general info**	
language	*r*	*p*	range	trend	min	range	corpus size	family
Afrikaans	-1.00	0.083	4	down	4	18	49	ine
Akkadian	-0.80	0.333	4	down*	10	11	25	afa
Amharic	-1.00	0.333	3	down*	5	5	10	afa
Ancient Greek	-0.75	0.018	10	down*	15	22	417	ine
Arabic	-0.96	0.003	7	up*	5	22	303	afa
Armenian	-0.96	0.003	7	down-up	3	14	53	ine
Bambara	-0.90	0.083	5	down*	9	9	14	dmn
Basque	-1.00	<0.001	7	down-up	6	15	121	eus
Belarusian	-0.96	0.003	7	down-up	5	21	305	ine
Breton	-0.50	1.000	3	down	5	9	10	ine
Bulgarian	-1.00	0.003	6	down-up	3	22	156	ine
Buryat	-1.00	0.083	4	down-up	7	9	10	xgn
Cantonese	-0.90	0.083	5	down-up	2	10	14	sit
Catalan	-1.00	<0.001	9	down	8	36	547	ine
Chinese	-0.96	<0.001	11	flat	1	20	285	sit
Clas. Chinese	-1.00	0.017	5	up*	1	13	269	sit
Coptic	-1.00	<0.001	7	down	4	8	49	afa
Croatian	-1.00	<0.001	8	down-up	4	23	199	ine
Czech	-0.99	<0.001	10	down-up	3	36	2223	ine
Danish	-1.00	0.003	6	down-up	5	20	101	ine
Dutch	-1.00	0.003	6	down-up	5	25	307	ine
English	-0.97	<0.001	9	down-up	3	26	556	ine
Erzya	-1.00	0.017	5	down*	6	8	17	urj
Estonian	-0.93	0.001	9	down-up	4	19	507	urj
Faroese	-0.79	0.048	7	down	4	13	50	ine
Finnish	-0.94	0.017	6	down-up	4	15	397	urj
French	-0.98	<0.001	9	down	7	32	583	ine
Galician	-1.00	0.003	6	down	8	28	164	ine
German	-1.00	<0.001	7	down	2	35	3754	ine
Gothic	-0.94	0.017	6	down-up	8	14	55	ine
Greek	-0.94	0.017	6	down	13	19	63	ine
Hebrew	-0.89	0.012	7	down	7	19	161	afa
Hindi	-1.00	0.017	5	down-up	6	34	376	ine
Hungarian	-1.00	0.017	5	down-up	4	19	42	urj
Icelandic	-0.06	0.822	15	down-up	3	28	1162	ine
Indonesian	-0.89	0.012	7	down*	23	23	168	map
Irish	-0.94	0.017	6	down-up	3	23	131	ine
Italian	-1.00	<0.001	9	down-up	5	34	819	ine
Japanese	-0.98	<0.001	9	down	20	25	237	jpx
Kazakh	-1.00	0.333	3	up*	1	8	11	trk
Kiche	-1.00	1.000	2	down*	7	8	10	myn
Komi Zyrian	-1.00	0.083	4	down*	8	8	10	urj
Korean	-0.99	<0.001	10	down*	16	17	447	(kor)
Kurmanji	-1.00	0.333	3	down-up	5	11	10	ine
Latin	0.12	0.676	15	down-up	15	31	978	ine
Latvian	-1.00	<0.001	8	down-up	4	19	252	ine

Continued on next page

Table 3 – continued from previous page

	sentence–clause–word			**clause–word-grapheme**			**general info**	
language	*r*	*p*	range	trend	min	range	corpus size	family
Lithuanian	-1.00	<0.001	7	down-up	5	14	75	ine
Maltese	-1.00	<0.001	8	down	6	15	44	afa
Manx	0.50	1.000	3	down-up	5	10	21	ine
North Sami	-0.80	0.333	4	down-up	4	10	27	urj
Norwegian	-0.93	0.002	8	down-up	3	26	667	ine
OCS	-0.82	0.034	7	down*	10	13	58	ine
OES	-0.18	0.713	7	down-up	6	20	180	ine
Old French	-0.29	0.556	7	down-up	8	17	171	ine
Persian	-0.88	0.003	9	up-down	1	29	655	ine
Polish	-1.00	<0.001	9	down-up	4	22	499	ine
Portuguese	-0.93	0.001	9	down	21	34	571	ine
Romanian	-0.86	0.001	11	down-up	5	32	938	ine
Russian	-0.87	0.001	11	down-up	5	28	1421	ine
Sanskrit	-0.70	0.233	5	down	8	10	29	ine
Scottish Gaelic	-0.03	1.000	6	down-up	3	19	72	ine
Serbian	-1.00	<0.001	7	down-up	3	20	98	ine
Slovak	-0.90	0.083	5	down-up	4	16	106	ine
Slovenian	-1.00	<0.001	7	down-up	3	20	170	ine
Spanish	-0.99	<0.001	11	down	22	37	1015	ine
Swedish	-1.00	0.083	4	down-up	4	14	207	ine
Tamil	-1.00	0.083	4	down	5	10	12	dra
Thai	-1.00	<0.001	7	down-up	4	14	22	(tai)
Turkish	-0.45	0.267	8	down	4	20	592	trk
Turkish German	-1.00	0.003	6	down*	14	14	37	ine
Ukrainian	-1.00	<0.001	7	down-up	3	19	122	ine
Upper Sorbian	-1.00	0.333	3	up	5	11	11	ine
Urdu	-1.00	0.017	5	down-up	6	30	138	ine
Uyghur	-1.00	0.017	5	flat	1	12	40	trk
Vietnamese	-1.00	0.017	5	down	7	8	44	aav
Welsh	-1.00	0.017	5	down-up	6	17	37	ine
West. Armenian	-0.86	0.024	7	down-up	6	14	36	ine
Wolof	-1.00	<0.001	8	down-up	2	13	44	nic

Table 3: Results across **languages** (OCS = Old Church Slavonic, OES = Old East Slavic). For **sentence–clause–word** analysis: r = Spearman's correlation coefficient, p = corresponding p-value, range = maximum sentence length (in clauses) for which 50 datapoints are available. For **clause-word-grapheme** analysis: trend = the shape of the curve, min = clause length for which the shortest mean word length is observed, range = maximum clause length (in words) for which 50 datapoints are available. For languages with DOWN, UP or FLAT trend, asterisk marks those where $|r| \geq 0.70$ or $|r| \geq 0.30$ and $p \leq 0.05$. Corpus size is given in K words, families are denoted by ISO-639 codes. There are no ISO-639 codes for Koreanic (the code for Korean is used) and Tai-Kadai (the code for the Tai branch is used).

Appendix B. The procedure for determining the shape of the curve

The formal procedure of determining the shape of the curve ("trend") for the clause–word–grapheme (reported in Table 3 in Appendix A) was as follows. The extrema (maximum and minimum) of the curve were identified. Then four points (first point, the smallest clause length; maximum; minimum; last point, the largest clause length) were compared by means of *t*-tests between adjacent pairs of points. In many cases, there were actually only three or even two points, because either maximum or minimum (or both) coincided with either first or last point (or both). Thus, the number of *t*-tests varied from one to three (Bonferroni correction for multiple comparisons was applied). If *p*-value was smaller than 0.05 and the absolute value of Cohen's *d* (effect size) was larger than 0.20, then the difference was considered to be large enough to label the corresponding part of the curve as going either DOWN or UP, otherwise it was ignored. If there were no differences at all, the whole curve was labelled as FLAT.

Bear in mind that the procedure is descriptive rather than inferential (even though it uses inferential statistics as a technique). It is approximately equivalent to manually classifying the patterns, but relies on formalized criteria and thus is more reproducible. See main text for the reasons why more sophisticated inferential tests were not applied.

Corpus-based Language Universals Analysis using Universal Dependencies

Hee-Soo Choi
Inria, LORIA & ATILF,
Université de Lorraine, CNRS,
F-54000 Nancy, France
hee-soo.choi@loria.fr

Bruno Guillaume
Université de Lorraine, CNRS,
Inria, LORIA,
F-54000 Nancy, France
bruno.guillaume@inria.fr

Karën Fort
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
Sorbonne Université, F-75006 Paris, France
karen.fort@loria.fr

Abstract

This paper presents experiments aiming at verifying Greenberg's universals based on Universal Dependencies (UD) corpora (de Marneffe et al., 2021). We adopt a corpus-based approach that allows us to highlight inconsistencies between corpora of the same language and to explore the causes of these inconsistencies. In addition to intra-language inconsistency, our analysis on 141 corpora, i.e. 74 languages, also shows cross-language inconsistency and questions the adaptability of UD annotations for some linguistic concepts.

1 Introduction and Related Work

Despite the evident diversity of languages, similarities between them have allowed linguists to extract common properties called language universals. The notion itself is difficult to define with the fine line between an absolute universal, a property valid over all languages, and a strong tendency. Linguists generally opposed two approaches: the typological approach and the generative approach (Comrie, 1989; Croft, 2003). While the latter insists on the common genetic character of languages (Chomsky, 1982), the typological approach relies more on empirical data from different languages, represented in Greenberg's work (Greenberg, 1966). Considered as one of the pioneers of modern typology, Greenberg defined 45 universals dealing with basic word order, morphology and syntax, based on 30 languages. In particular, he established a classification of the 30 languages according to three basic factors: the existence of prepositions as against postpositions, the order of subject, object and verb in declarative sentences with nominal subject and object and the order of the qualifying adjective in relation to the noun.

The development of natural language processing (NLP) tools and digital resources, especially treebanks, allowed for the multiplication of multilingual research work aiming at testing typological features on a large number of languages (Dryer, 1992; Liu, 2010; Östling, 2015; Futrell et al., 2015; Levshina, 2019). More recently, UD treebanks have been used in several typological and language universal studies: Sharma et al. (2019) showed that language networks constructed from UD treebanks cluster correctly languages based on Greenberg's word order typology, allowing for a representation of linguistic generalizations, while Gerdes et al. (2019) explored Greenberg's universals in a quantitative way on treebanks annotated in dependency syntax using an annotation scheme derived from UD (SUD) and identified quantitative universals using diagrams, which they call "typometric diagrams" (Gerdes et al., 2021). Moreover, Dönicke et al. (2020) proposed a framework to investigate Greenberg's typological universals on UD by using real-valued logics.

Our study aims at examining Greenberg's observations on the basis of the large amount of data provided by the UD corpora.[1] Our results represent empirical information that can confirm or even com-

[1] We decided not to use SUD for our study for two reasons. First, many occurrences of verbal forms with a subject and an object would require a much more complex set of patterns (a specific pattern is needed when there is one auxiliary which is the governor of the subject in SUD; another one when there are two auxiliaries…). Second, most of SUD corpora are produced automatically from UD data and we prefer to work on original data.

plete existing typological databases. Although our experiments are in line with the work of Gerdes et al. (2021), we decided not to group corpora of the same language together to preserve their specificities and evaluate their influence. Our paper is structured as follows: in Section 2, we present our sample of 141 UD corpora as well as the tool and metrics we used for our experiments; in Section 3, we describe our experiments on determining three word orders and verifying four universals; finally, in Section 4, we discuss issues in UD annotations that our analysis brought to light.

2 Material and Methods

2.1 From UD 2.7 to UD 2.7_{1K}

In our work, we used UD version 2.7, which contains 104 languages and 183 corpora. Since we consider each corpus separately, we decided not to take into account the corpora containing less than 1,000 sentences, which we consider too small to be representative. We thus obtained a set of 74 languages and 141 corpora, which constitutes our experimental data, UD 2.7_{1K}. There is a strong bias in the languages represented in UD in general, therefore in our data: 76% of the sentences and 65% of the corpora in UD 2.7_{1K} belong to the Indo-European family. Of the 30 languages in Greenberg's sample, only 14 are present in UD 2.7_{1K}, including all the European languages used by Greenberg.

2.2 GREW: a tool for fine-grained observation on corpora

In our experiments, we need to count the number of occurrences of specific patterns in each corpus. We use the graph matching mechanism available in the GREW tool[2] (Guillaume, 2021). It allows to write complex patterns, with combination of constraints on dependency relations, on node features; it is also possible to add negative constraints to refine the queries. Figure 1 shows a query with constraints on relations, on words, on word order (<< symbol) and with negative constraints (*without* part).

2.3 Quantifying qualitative concepts

We used the same method as in (Choi et al., 2021) and defined the notion of dominant word order quantitatively as follows. We computed the ratio between the two most frequent orders: if it is greater than or equal to 2, the most frequent order is the dominant order, if it is strictly inferior to 2, the corpus is considered to have no dominant order (NDO). In the case where only two orders are possible (for example, adjective-noun / noun-adjective), if the ratio is greater than 2, the frequency of the most frequent order is greater than $\frac{2}{3}$.

In order to compare if two corpora show the same distribution of some observations, we compute the cosine value between the two vectors representing the proportion of each observation. For instance with the orders between Subject, Object and Verb, we have vectors with six dimensions for the proportion of the six possible relative orders. We expect two corpora of the same language to have the same distribution and thus, the cosine to be closed to 1. In our experiments, a lower cosine value indicates a greater inconsistency between the two corpora.

3 Experiments and Results

3.1 Order of Subject, Object and Verb

To be consistent with Greenberg's observations, we determined a dominant order of Subject (S), Object (O) and Verb (V) in declarative sentences with nominal subject and object. Inspired by Choi et al. (2021), we decided to refine the GREW pattern by setting the POS tags to filter the nominal subjects and objects (see Figure 1). Indeed, even if the `nsubj` relation inherently involves nominal dependents, we noticed that this is not the case in all corpora (see Section 4). Furthermore, UD annotations do not allow us to filter declarative sentences precisely, we therefore relied on punctuation by eliminating all sentences whose verb is linked to a question mark or an exclamation mark (which corresponds to the *without* part of the pattern). However, this method remains fragile since corpora without punctuation exist and we

[2] https://grew.fr

have no certainty that the punctuation system is the same in all the concerned languages. Moreover, we could not filter out imperative sentences because some corpora do not indicate the mood of the verbs.

```
pattern {
  V [upos=VERB];
  V -[1=nsubj]-> S; S[upos=PROPN|PRON|NOUN];
  V -[1=obj]-> O; O[upos=PROPN|PRON|NOUN];
  S << V; V << O;
}
without {
  V -[punct]-> P; P [lemma="?"|"!"];
}
```

Figure 1: GREW pattern for SVO order.

Using criterion described in Section 2.3 on 141 corpora, we observed that 91 are SVO, 24 are SOV, 4 are VSO and 22 have no dominant order (NDO). Of the 29 multi-corpora languages, all corpora of the same language show the same dominant order, except for six languages: German, Arabic, Ancient Greek, Latin, Dutch and Romanian. Our analyses of the inconsistency in these six languages agree with those of Choi et al. (2021), however we gained in consistency with seven languages with a minimum cosine value of less than 0.95 against ten languages for them.

3.2 Prepositions/Postpositions

In his work, Greenberg uses the terms "prepositional language" or "language with prepositions" but does not provide a precise definition of what this means. Inspired by WALS (Dryer, 2013b), we decided to extract occurrences of adpositions linked to a noun phrase, which corresponds to a noun, a pronoun or a proper noun in UD. Figure 2 shows a GREW pattern for the order adposition - noun phrase (preposition). We fixed a node A labeled as an adposition (ADP), related to a noun phrase N by a `case` relation. Moreover, we decided to exclude sentences where the adposition is part of a multi-words expression, marked with a `fixed` or a `flat` relation in UD.

```
pattern {
  A [upos=ADP];
  N [upos=NOUN|PRON|PROPN];
  N -[1=case]-> A;
  A << N;
}
without {
  A -[1=fixed|flat]-> X
}
```

Figure 2: GREW pattern for adposition - noun phrase order.

The results show a marked tendency for one of the two types of adposition. Of the 141 corpora, 108 have prepositions (Pr), 30 have postpositions (Post), one corpus has no dominant order (`Chinese-PUD`) and two corpora show no occurrence of either type (`Korean-PUD` and `Sanskrit-Vedic`). For the 29 multi-corpora languages, all corpora of the same language present the same type of adposition, except for Chinese and Korean due to the two atypical corpora. To measure the degree of consistency, we computed the cosine values and extracted the minimum cosine between all possible pairs. 27 languages show a high consistency between corpora with a minimum cosine above 0.99. Only two languages have minimum cosine values below 0.99: Chinese (0.8771) and Persian (0.9658)[3].

Chinese The `Chinese-PUD` is the only Chinese corpus without a dominant order. The results on other Chinese corpora in Table 1 show that they generally use prepositions[4]. The minimum cosine value of 0.8771 is observed between the `Chinese-PUD` and the `Chinese-GSD`.

[3] For Korean, we have excluded the corpus `Korean-PUD` which present no occurrence of either adposition.

[4] There is a `Chinese-GSDSimp` which is a simplified Chinese version of the corpus of the `Chinese-GSD`. We consider only the second one here.

Corpus	Pr	Post
Chinese-GSD	99.92%	0.08%
Chinese-HK	87.40%	12.60%
Chinese-PUD	**64.57%**	**35.43%**

Table 1: Proportions of prepositions and postpositions in the Chinese corpora.

Corpus	`acl`	`appos`	`case`	`case:loc`	`conj`	`mark`
Chinese-GSD	**98.41%**	0.10%	0.20%	0.00%	0.20%	1.09%
Chinese-HK	0.00%	0.00%	9.68%	**90.32%**	0.00%	0.00%
Chinese-PUD	0.00%	0.00%	0.00%	**99.39%**	0.00%	0.61%

Table 2: Syntactic relation types between the noun phrase and the postposition in the Chinese corpora.

Table 2 presents syntactic relation types between postpositions and noun phrases. In the `Chinese-GSD`, postpositions are not annotated with a `case` relation but with a `acl` relation, they are therefore not covered by the pattern we used. In contrast, the `Chinese-PUD` has 99% of postpositions identified with a `case:loc` relation, a subtype of the `case` relation covered by the pattern. The same goes for the `Chinese-HK` with 90.32% of `case:loc` relation and 9.68% `case` relation. One could imagine that the relations `case` and `acl` involve postpositions of a different nature, but all Chinese corpora use postpositions that generally correspond to location-indicating postpositions such as: 上 (shàng = above/on), 中 (zhōng = between, in the middle), 下 (xià = below/under). In this case, the annotators simply made a different choice of annotation. Taking into account the postpositions related by the relation `acl` in `Chinese-GSD`, we obtain a distribution of 72.87% of prepositions and 27.13% of postpositions, which is consistent with the distributions of the `Chinese-PUD` and the `Chinese-HK`.

The UD guidelines describe the `acl` relation as a clausal modifier of noun. In Chinese, clausal modifiers may precede the noun and may be formed with the particle 的. They may also follow the noun, in which case they will be juxtaposed after the noun[5]. The `acl` relation must link a noun and the head of clause. It is quite rare to find an adposition depending on a relation `acl`, which would be contrary to the UD guidelines. We therefore assume that there is an inconsistency in annotations in these corpora[6].

Conflict between ADP and PART tags The distinction between a particle and an adposition is difficult to define. Adpositions, coordinating conjunctions and subordinating conjunctions are considered particles in UD, but the guidelines specify that the most precise label should be used.

Korean is an agglutinating language using postpositions at the end of words to designate their functions. These postpositions are sometimes referred to as particles. In UD, it is stated that the PART tag designates a functional word and should only be used when the word does not fit the definitions of other functional words such as adpositions and conjunctions[7]. However, in the `Korean-PUD` corpus, the ADP tag is not used and all adpositions are annotated with the PART tag. Moreover, in version 2.7 of the corpus, adpositions are not linked with a relation `case` but with a relation `dep:prt`[8].

As in `Korean-PUD`, our pattern did not allow to find any occurrences of adpositions for the `Sanskrit-VEDIC` because they are annotated as a particle. With the PART annotation, we obtain 79.76% of postpositions and thus 20.24% of prepositions. Sanskrit being a dead language, it is not listed in WALS and we cannot offer a more in-depth analysis without a specialist of the language.

Finally, in Persian, the cosine between the two corpora is 0.9658. While both corpora present mostly prepositions, the `Persian-PerDT` presents 21.18% of postpositions and the `Persian-Seraji` only 0.01% because postpositions are annotated as PART and not ADP in the corpus.

[5] `https://universaldependencies.org/zh/dep/acl`, August 2021.

[6] This inconsistency was corrected after we entered an issue on the GitHub.

[7] `https://universaldependencies.org/u/pos/PART`, August 2021.

[8] In version 2.8, the relation `dep:prt` has been replaced by the relation `case`.

Comparison with WALS WALS presents the order of adposition and noun phrase under the feature 85A (Dryer, 2013b). On the 74 languages of our study: 50 languages have the same order as WALS, 21 languages are not in WALS (mainly dead languages) or do not present the feature 85A, three languages do not have the same order: Chinese, Cantonese and Amharic which are considered NDO by WALS.

The `Cantonese-HK` has 87.84% of prepositions. Since Cantonese and Chinese have relatively similar syntax, we can assume that Cantonese can indeed have prepositions and postpositions but that this corpus either has few constructions with postpositions or the annotations do not allow us to cover all cases with our pattern. As for the Amharic corpus (`Amharic-ATT`), it has 83.81% of prepositions. Greenberg also considers Amharic to be a prepositional language.

3.3 Order of the Adjective and the Noun

Determining a dominant order of Adjective (Adj) and Noun (N) turns out to be simpler, since the concepts are present in all languages and are precisely annotated. We used the pattern presented in Figure 3, where we define a noun N and an adjective Adj linked by the relation `amod`.

```
pattern {
  N [upos=NOUN];
  Adj [upos=ADJ];
  N -[1=amod]-> Adj;
  Adj << N
}
```

Figure 3: GREW pattern for the Adj-N order.

Of the 141 corpora, 84 present the Adj-N order, 43 present the N-Adj order and 14 do not have a dominant order: one French corpus, three Italian corpora, two Polish corpora, the two Ancient Greek corpora, the four Latin corpora, the Old Russian corpus and the Gothic corpus.

All corpora of the same language have the same order, except for French, Italian and Polish. 24 multi-corpora languages have a high consistency in their corpora with minimum cosine values above 0.99. Five languages are below 0.99: Italian (0.9146), French (0.9171), Romanian (0.9771), Polish (0.9779) and Latin (0.9836). In French, Italian and Polish, the adjective can be placed either before the noun or after the noun. Generally, the place of the adjective does not change the meaning of the sentence, except in special cases[9]. However, there are rules such as placing short adjectives before the noun or placing adjectives of color after the noun in French and Italian. It is also possible to change the place of the adjective to emphasize the quality carried by the adjective.

French The `French-Spoken` is the only corpus of French without a dominant order. This corpus is the only spoken French corpus among the seven French corpora. As shown in Table 3, the written French corpora have very similar proportions with about 70% N-Adj, while the `French-Spoken` has both orders relatively homogeneously. The minimun cosine value is 0.9171 between the `French-Spoken` and the `French-Sequoia`.

Corpus	Adj-N	N-Adj
French-FQB	29.31%	70.69%
French-FTB	30.86%	69.14%
French-GSD	29.95%	70.05%
French-ParTUT	27.10%	72.90%
French-PUD	31.13%	68.87%
French-Sequoia	24.24%	75.76%
French-Spoken	**46.71%**	**53.29%**

Table 3: Proportions of Adj-N and N-Adj orders in the French corpora.

[9] For example, in French "une ancienne maison" (what used to be a house) has a different meaning than "une maison ancienne" (an old house).

The written corpora are mostly from newspaper articles and Wikipedia, therefore the sentences present adjectives from more scientific domains which tend to be placed after the noun. On the other hand, the oral French corpus shows an over-representation of common adjectives such as “petit” (small) and “grand” (big) which are placed before the noun.

Italian In Italian, three corpora have a dominant N-Adj order and three have no dominant order. Of the three corpora without dominant order, two are tweets corpora, the `Italian-PoSTWITA` and the `Italian-TWITTIRO`. The third corpus is the `Italian-ParTUT`, but this result is certainly due to a threshold effect, the distribution being 33.99% and 66.01%.

As in French, the three N-Adj dominant order corpora as well as the `Italian-ParTUT` have texts extracted from newspaper articles and Wikipedia. The distribution of the orders is similar in these corpora with a proportion around 70% for the N-Adj order. The results show that the genre of the corpora has an influence on the type of adjectives used and thus on the place of the adjective in relation to the noun.

Corpus	Adj-N	N-Adj
Italian-ISDT	29.89%	70.11%
Italian-ParTUT	**33.99%**	**66.01%**
Italian-PoSTWITA	**41.31%**	**58.69%**
Italian-PUD	31.04%	68.96%
Italian-TWITTIRO	**51.69%**	**48.31%**
Italian-VIT	31.76%	68.24%

Table 4: Proportions of Adj-N and N-Adj orders in the Italian corpora.

Polish In Polish, adjectives can be placed before or after the noun, except for short adjectives which are always placed before. Placing an adjective after the noun is possible to emphasize it. Unlike in French and Italian, we cannot explain the differences in distribution by the genre of the texts. The corpora are all composed of various genres: newspaper articles, fiction, Wikipedia.

Corpus	Adj-N	N-Adj
Polish-LFG	71.30%	28.69%
Polish-PDB	**64.48%**	**35.52%**
Polish-PUD	**59.73%**	**40.27%**

Table 5: Proportions of Adj-N and N-Adj orders in the Polish corpora.

Dead languages Latin, Ancient Greek, Old Russian and Gothic corpora do not have a dominant order of Adjective and Noun. The proportions between the two orders are homogeneous, with ratios very close to 1 for these languages. We therefore can assume that these languages allow for both orders, as they are considered free word order languages (Levshina, 2019).

Comparison with WALS WALS presents the order of Adjective and Noun in the feature 87A (Dryer, 2013a). Of the 74 languages: 54 languages show the same order as WALS, 17 languages are not in WALS or do not have the feature 87A, three languages have some corpora with no dominant order in our results: Italian, French and Polish. However, the corpora where order is defined for these languages are all consistent with WALS’ results.

3.4 Greenberg’s Universal 1

Universal 1 *In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.*

The three orders where the subject precedes the object are: SVO, SOV and VSO. The results obtained show that on 141 corpora 91 are SVO, 24 are SOV, four are VSO and 22 are NDO. Our results therefore confirm Greenberg's Universal 1 for 119 corpora and 59 languages. For NDO corpora, without taking into account the ratio, the most frequent order is either SVO, SOV or VSO, except for two corpora: the Amharic-ATT and the Latin-LLCT (see Table 6).

Corpus	**SVO**	**SOV**	**VSO**	**VOS**	**OSV**	**OVS**
Amharic-ATT	4.70%	**28.86%**	8.95%	0.45%	12.97%	**44.07%**
Latin-LLCT	**30.18%**	**29.00%**	4.31%	0.93%	**32.65%**	2.91%

Table 6: Distribution of Subject, Object and Verb orders in Amharic-ATT and Latin-LLCT.

For Amharic, the most frequent order is OVS at 44.07%, followed by SOV at 28.86%. The OVS order remains a rather rare case of dominant order. According to WALS and Greenberg, Amharic is a SOV language, which is the second most frequent order in our results. Without a speaker of Amharic, we can only assume that there are either: i) annotation errors in the corpus, ii) the possibility of using the OVS order in some sentences, iii) an influence of the genre of the texts of the corpus, which is very heterogeneous since the sentences can be examples of grammars, extracted from texts of fiction, from the bible and from newspapers among others.

The Latin-LLCT has a relatively homogeneous distribution between three orders SVO, SOV and OSV which are frequent at about 30%. Latin being a free word order language, this may explain the absence of a dominant order. Moreover, the texts come from different centuries, which has some influence on the way sentences are constructed.

3.5 Greenberg's Universals 3 and 17

Universals 3 and 17 concern VSO languages, so we treat them together in this section.

Universal 3 *Languages with dominant VSO order are always prepositional.*

Universal 17 *With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.*

Table 7 shows that our results confirm Greenberg's Universals 3 and 17, but only on four corpora. For Arabic, two other corpora are available but do not present a dominant order, with a conflict between VSO (31.20% and 49.45%) and SVO orders. These corpora are largely prepositional and have almost 100% N-Adj order. It is interesting to note that Arabic-PADT exhibits these characteristics while having a higher frequency of SVO than VSO (48.15%). This result is consistent with Greenberg's observation that SVO languages are more correlated with prepositionnaly and N-Adj order than postpositionnaly and Adj-N order.

Corpus	**VSO**	**Pr**	**N-Adj**
Arabic-NYUAD	54.56%	99.97%	99.69%
Irish-IDT	99.14%	99.78%	98.91%
Scottish_Gaelic-ARCOSG	97.49%	100%	84.82%
Welsh-CCG	78.57%	100%	82.54%

Table 7: Proportions of prepositions (Pr) and N-Adj order in the VSO corpora.

3.6 Greenberg's Universal 4

Universal 4 *With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.*

According to our results, 24 corpora have a dominant SOV order, which corresponds to 15 languages. To visualize this universal, we used the typometric graph of Gerdes et al. (2021) in Figure 4.

Corpora at the top right of the figure correspond to very strongly SOV and postpositional corpora. The languages represented are: Bambara, Hindi, Japanese, Kazakh, Korean, Telugu, Turkish,

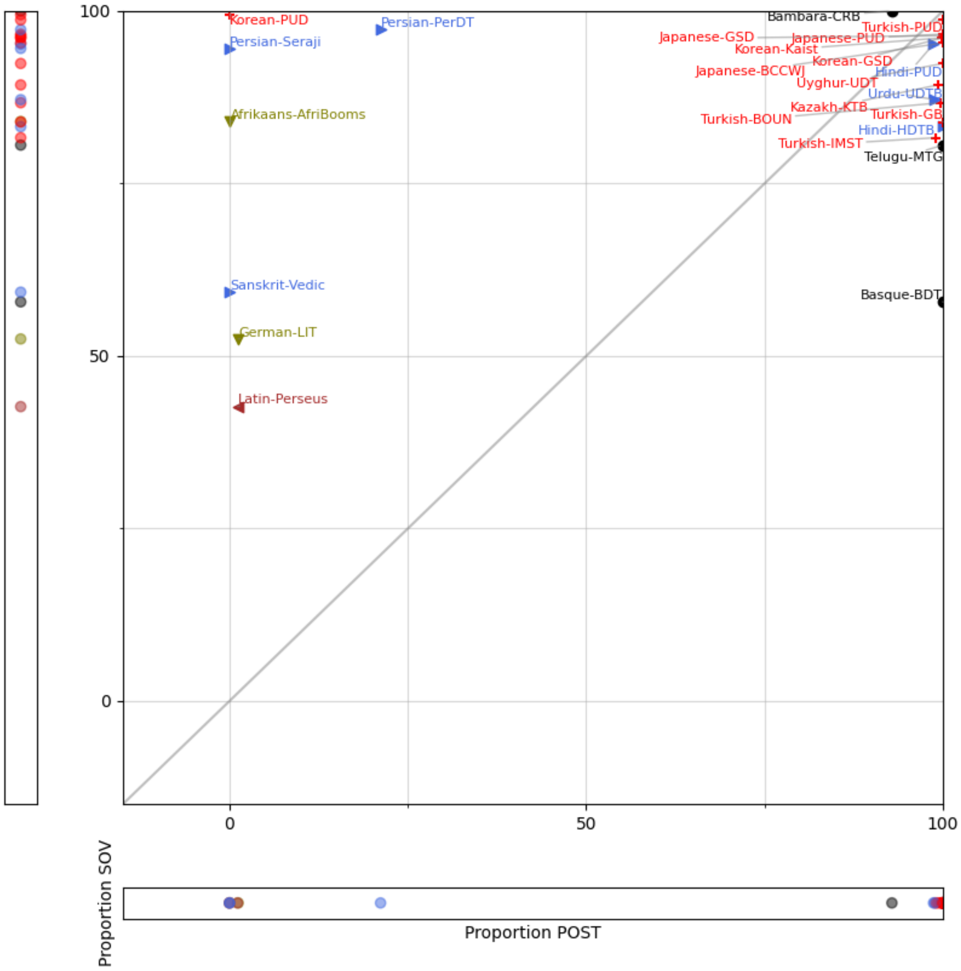


Figure 4: SOV corpora according to their proportions of postpositions.

Urdu and Uyghur. At the top left, some corpora are strongly SOV but with very few postpositions: `Afrikaans-AfriBooms`, `Persian-Seraji`, `Persian-PADT` and `Korean-PUD`. Afrikaans and Persian are both SOV and prepositional languages. The `Korean-PUD` does not present postpositions due to annotation issues detailed in section 3.2. We can also note that the two other corpora of Korean largely present postpositions.

Moreover, three corpora stand out with a percentage of SOV around 50%: the `German-LIT`, the `Latin-Perseus` and the `Sanskrit-Vedic`. The German and Latin corpora have a dominant SOV order but their other corpora do not have a dominant order. These two languages are also generally considered to have no dominant order. We assume that Greenberg's formulation "normal SOV order" allows us not to take into account these languages in the universal. For Sanskrit, we are in the same situation as the `Korean-PUD`: our pattern did not allow us to find any occurrences of pre or postpositions.

Finally, the `Basque` is isolated because of a relatively low percentage of SOV at 57%, but sufficient to be considered the dominant order, the second order being SVO at 19%. WALS also considers Basque to be an SOV and postpositional language.

4 Discussing UD annotations

4.1 Nominal dependent relations

To determine the order of Subject, Object and Verb, we had to consider only nominal subjects and objects. In UD, these functions are annotated with the relations `nsubj` and `obj` respectively. We ran a first experiment without fixing the POS tags of the subject and object as noun, pronoun or proper noun, to avoid redundancy in the patterns. We obtained heterogeneous results, especially between corpora of the same language. On closer inspection of some corpora, we observed several times and in corpora of different languages that the dependent `nsubj` and `obj` relations were not nominal, which is inconsistent with the definitions of these relations.

With GREW, we computed the proportion of non-nominal dependents of the relations `nsubj` and `obj` linked to a verb. Out of the 141 corpora, six corpora present more than 20% of non-nominal `nsubj` dependents and four present more than 20% of non-nominal `obj` dependents. Tables 8 and 9 detail the relative proportions for each corpus as well as the most represented POS tags for these dependents. The non-nominal `nsubj` dependents of the three `PROIEL` corpora are mostly adjectives between 72% and 78%. The `Thai-PUD` and `Slovenian-SST` have a high proportion of determiners and out of the 25% of non-nominal `nsubj` dependents of the `Arabic-PADT`, 50% are annotated X, a label used when the other POS tags are not adapted.

As for the four corpora with more than 20% non-nominal `obj` dependents, the most frequent POS is verb for three corpora and adjective for the last one.

Corpus	**Non-nominal** `nsubj` **dependents**	**VERB**	**ADJ**	**DET**	**X**
Old_Church_Slavonic-PROIEL	27.06%	20.94%	**73.89%**	0.00%	0.00%
Thai-PUD	25.86%	14.40%	0.28%	**83.10%**	0.00%
Arabic-PADT	25.15%	0.34%	10.72%	31.88%	**50.34%**
Ancient_Greek-PROIEL	23.67%	17.55%	**78.55%**	0.00%	0.00%
Gothic-PROIEL	21.82%	22.42%	**72.17%**	0.00%	0.00%
Slovenian-SST	20.24%	0.00%	9.46%	**79.28%**	1.35%

Table 8: Corpora with the most non-nominal `nsubj` dependents and proportions of POS tags for these.

Corpus	**Non-nominal** `obj` **dependents**	**VERB**	**ADJ**
Turkish-IMST	32.80%	**65.72%**	26.03%
Hindi-HDTB	26.43%	**89.00%**	9.55%
Urdu-UDTB	24.87%	**88.83%**	8.14%
Ancient_Greek-PROIEL	20.96%	19.12%	**76.95%**

Table 9: Corpora with the most non-nominal `obj` dependents and proportions of POS tags for these.

4.2 The `case` relation

According to UD, the relation `case` allows to annotate any case-marking element which is treated as a separate syntactic word (including prepositions, postpositions, and clitic case markers). These elements are dependents on the nouns to which they are attached[10]. The relation `case` thus involves nominal governors. In the same way as for the relations `nsubj` and `obj`, we calculated the proportion of `case` relations with non-nominal governors in the 141 corpora. Table 10 shows the seven corpora which have more than 20% non-nominal governors involved in the `case` relation.

We explored the possible reasons for these results in Turkish, Chinese and Korean as we can compare the results between different corpora. We could not add the two mono-corpus languages, Basque and

[10] `https://universaldependencies.org/u/dep/case`, September 2021.

Corpus	**Non-nominal** `case` **governors**	**VERB**	**NUM**	**ADJ**	**DET**	**PART**
Turkish-BOUN	45.95%	**54.73%**	0.96%	17.46%	1.33%	0.00%
Turkish-IMST	37.49%	**55.40%**	6.89%	29.46%	3.51%	0.00%
Chinese-GSD	28.77%	**60.76%**	1.02%	6.31%	0.09%	27.27%
Basque-BDT	28.27%	0.00%	19.85%	24.95%	**38.94%**	0.00%
Amharic-ATT	22.47%	**75.41%**	1.64%	7.38%	5.74%	0.00%
Korean-Kaist	20.59%	19.51%	**56.50%**	5.69%	0.00%	0.00%

Table 10: Proportions and POS tags of non-nominal `case` governors linked to an adposition.

Amharic in our analysis, as there was no other corpus to compare them to.

Turkish The percentages of non-nominal governors are high for two Turkish corpora (45.95% and 37.49%) and more than half of them are verbs. In comparison, the other two Turkish corpora have low proportions of non-nominal governors: the `Turkish-GB` is at 6.23% and the `Turkish-PUD` is at 4.97%. Syntactic relations of the `Turkish-BOUN` were manually annotated in the UD scheme, while the `Turkish-IMST` is the result of a semi-automatic conversion of the IMST Treebank (Sulubacak et al., 2016). As the `Turkish-BOUN` is manually annotated by native speakers, we can assume that they made an annotation choice that is contradictory to the UD instructions. For the `Turkish-IMST`, this may be due to the semi-automatic conversion. In version 2.8 of UD, four new corpora have been added and it is indicated that updates have been made to gain consistency across all corpora in Turkish[11]. However, our results remain the same for the `Turkish-BOUN` and the `Turkish-IMST` in UD 2.8.

Chinese The `Chinese-GSD` have 28.77% of non-nominal governors in the relation `case`, mostly verbs (more then 60% of cases). Due to the Chinese language structure, the `case` relation definition is slightly different from the universal definition. `case` is used on particles marking relations such as genitive, prepositions including coverbs and valence markers[12].

The coverbs correspond to particles or adpositions linked to verbs such as: 在+ 落在 (fall on), 向+ 奔向 (run to). In `Chinese-GSD`, they are annotated as adpositions and are thus linked to verbs with the relation `case`. In the other two Chinese corpora, the `Chinese-PUD` and the `Chinese-HK`, the percentages of non-nominal governors are 2.29% and 5.22% respectively. In these corpora, adpositions and verbs are linked with the `mark` relation. In Chinese, it is used on a functional word marking a clause as subordinate to another clause[13]. In the universal definition, this relation involves a subordinating conjunction SCONJ rather than an adposition but Chinese corpora use both tags. For Chinese, we face two difficulties: i) the use of `case` and `mark` relations to annotate coverbs, ii) the conflict between the POS tags ADP, PART and SCONJ. As previously observed in Section 3.2, the `Chinese-GSD` differs from the other two corpora because of some inconsistent annotations with the UD guidelines.

Korean The `Korean-Kaist` has a percentage of non-nominal governors in the relation `case` at 20.59% and 56.50% of these governors are numbers. To annotate numbers followed by symbols, there is no real consensus between languages on the entity that should carry the syntactic relation. For example, in French corpora, the numeral and the symbol are linked by a relation `nummod` and the symbol is linked to the noun to which it is attached by a relation `nmod`. Although the UD guidelines state that the `case` relation involves nominal governors, other languages have `case` relations between an adposition and a numeral. In French, for example, `French-Sequoia` includes this construction in the form “en 1980”.

5 Conclusion

Our results are mostly consistent with Greenberg’s observations, but also with the information from WALS concerning the three orders. Our results constitute new typological information based on large

[11] `https://github.com/UniversalDependencies/UD_Turkish-BOUN`, September 2021.

[12] `https://universaldependencies.org/zh/dep/case`, September 2021.

[13] `https://universaldependencies.org/zh/dep/mark`, September 2021.

amounts of data that can fill in gaps in the existing databases. In particular, we treated seven languages for which values of the three orders are not provided in WALS: Afrikaans, Faroese, Galician, Kazakh, Maltese, Naija and Slovak. The corpus-based approach allows us to evaluate the consistency between corpora of the same language and show a great variation according to the corpus types: oral language, written language in newspapers, tweets, poetry, novels, grammars, etc.

Moreover, our study raises several issues related to UD, especially to the universality of its annotation scheme. UD being initially based on an annotation scheme created for English (de Marneffe et al., 2014), for languages with a different structure than English, adapting the annotations leads the creators of the corpora to make annotation choices that they have to justify and that are not necessarily consistent with other corpora in the language. We therefore end up with some inconsistencies between languages but also between corpora of the same language, inconsistencies for which we provided analyses either by examining the corpus documentation or by asking native speakers. The collaborative aspect of UD project allowed us to share our observations and thus contribute to UD corpora improvement.

It is worth noting that our experiments can be replicated and extended given that the tool GREW is available online[14] along with the UD corpora[15]. Similarly, the scripts and patterns can be found on a Gitlab repository[16].

References

Hee-Soo Choi, Bruno Guillaume, Karën Fort, and Guy Perrier. 2021. Investigating Dominant Word Order on Universal Dependencies with Graph Rewriting. In *Recent Advances in Natural Language Processing (RANLP2021)*, en ligne, Bulgarie, September.

Noam. Chomsky. 1982. *Some concepts and consequences of the theory of government and binding / Noam Chomsky*. MIT Press Cambridge, Mass.

Bernard Comrie. 1989. *Language universals and Typology: Syntax and Morphology*. University of Chicago Press.

William Croft. 2003. *Typology and Universals*. Cambridge University Press, New York.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Islande, May. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Tillmann Dönicke, Xiang Yu, and Jonas Kuhn. 2020. Real-valued logics for typological universals: Framework and application. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3990–4003, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

M. Dryer. 1992. The greenbergian word order correlations. *Language*, 68:138 – 81.

Matthew S. Dryer. 2013a. Order of adjective and noun. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013b. Order of adposition and noun phrase. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Suède, August. Uppsala University, Uppsala, Suède.

[14] `https://grew.fr/`

[15] `https://universaldependencies.org/`

[16] `https://gitlab.inria.fr/ud-greenberg/udworkshop-2021`

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2019. Rediscovering Greenberg's word order universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131, Paris, France, August. Association for Computational Linguistics.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics from implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, 6, 02.

Joseph H. Greenberg. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online, April. Association for Computational Linguistics.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578. Contrast as an information-structural notion in grammar.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Pékin, Chine, July. Association for Computational Linguistics.

Kartik Sharma, Kaivalya Swami, Aditya Shete, and Samar Husain. 2019. Can Greenbergian universals be induced from language networks? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 25–37, Paris, France, August. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japon, December. The COLING 2016 Organizing Committee.

A Quantitative Approach towards German Experiencer-Object Verbs

Johanna M. Poppek
Linguistic Data Science Lab
Ruhr-Universität Bochum
johanna.poppek@rub.de

Simon Masloch
Linguistic Data Science Lab
Ruhr-Universität Bochum
simon.masloch@rub.de

Amelie Robrecht
Faculty of Technology
Universität Bielefeld
arobrecht@techfak.uni-bielefeld.de

Tibor Kiss
Linguistic Data Science Lab
Ruhr-Universität Bochum
tibor.kiss@rub.de

Abstract

Despite being studied for several decades, the properties of experiencer-object (EO) verbs remain under discussion. This holds especially for the question whether they display distinctive properties that distinguish them from "regular" transitive verbs. We performed a large-scale annotation study of German EO verb syntactic distribution patterns which shows that EO verbs differ largely in how frequently they occur in the eponymous pattern, that verbs taken to belong to the same subclass can differ largely in their pattern distribution, and that the negative correlation between the number of occurrences on the reflexive pattern and the passive patterns is smaller than previously assumed. This means that a number of verbs considered "typical" for this verbal class appear to have stronger associations with syntactic patterns other than the prototypical one, which is of special importance for experimental work.

1 Introduction

Experiencer-object (EO) verbs are usually defined as psychological predicates whose experiencer is (normally) realised as the object. In most publications, the term psych verb is referring to verbs where one argument expresses the experiencer of some psychological state, cf. e.g. (Landau, 2010). The idea of a further classification in the domain of psych verbs dates back at least to Belletti and Rizzi (1988), whose division leads to the following three classes:[1]

(1) Belletti and Rizzi (1988)'s classes:

I. *nominative experiencer, accusative stimulus*
Mary fears John/the noise.

II. *nominative stimulus, accusative experiencer*
John/The noise frightens Mary.

III. *nominative stimulus, dative experiencer*
John/The book appeals to Mary.

The ability of the verbs belonging to class II or III to realise the experiencer in the object position ignited vivid discussions among researchers about its peculiar position in the verbal domain. While the above classification has been very influential, it resulted in a presumably premature focus on classes that may be too internally heterogeneous to allow a productive analysis of their members' properties. Particularly, these verbs are known for their variation in argument realisation patterns that call for a more fine-grained classification than currently presented. This requires an overview on the general distribution of syntactic occurrence profiles besides the prototypical experiencer-object pattern. These unresolved issues appear particularly problematic since many of these verbs are used in experimental syntax research without

[1] We depart from their terminology here. One may note that Belletti and Rizzi propose this distinction only for Italian and explicitly state that they do not take into account derivational processes.

further reflection on the classes that are presupposed or with a focus only on selected differences (cf. (Haupt et al., 2008; Verhoeven, 2015; Ellsiepen and Bader, 2018; Scheepers et al., 2000), among many others). It is also subject to debate if the larger category of psych verbs does in fact have distinctive properties compared to other verbs or if the psych-based constructions do not fundamentally differ from other transitive constructions (for the latter view, cf. e.g. (Grafmiller, 2013; Żychliński, 2016)). Psych verbs in general and EO psych verbs in particular have usually been approached with a specific syntactic pattern of realisation in mind. In the case of EO psych verbs, this pattern would be a subtype of a transitive construction where the role of the stimulus (STM) is assigned to the subject, and the role of the experiencer (EXP) is assigned to the object. But even superficial scrutiny soon reveals that *forms* of the respective verbs occur in other syntactic patterns. The semantics assigned to these patterns is sometimes transparently linked to the semantics of the "prototypical" pattern. In other cases, a link is much harder to detect, and in some cases, it appears quite opaque, even to the effect that a description of the verb as EO verb seems hard to defend. A large corpus-based resource on EO verb distribution patterns and their general frequency and distribution contributes to a quantitative perspective towards the phenomenon.

1.1 Why We Need a Large-Scaled Approach

While the possibility of the verbs' appearance in certain syntactic environments has played a prominent role in the literature on experiencer-object verbs for several decades (particularly their ability to passivise, cf. Belletti and Rizzi (1988), Pesetsky (1995), Landau (2010) among many others) and alternations between some realisation patterns have gained some prominence in the literature recently (cf. i.a. (Alexiadou and Iordăchioaia, 2014; Pijpops and Speelman, 2017; Hirsch, 2018; Rott et al., 2020)), we are not aware of a large-scaled corpus study investigating the syntactic distributional patterns of the posited class of experiencer-object psych verbs in German: Engelberg (2018) and Cosma and Engelberg (2014) look at 11 (German) verbs only, Becker and Guzmán Naranjo (2020) annotate 30 sentences for 12 "psych concepts" (in 7 languages), and Verhoeven (2015) performs annotations for 30 EO verbs, but she is interested in word order differences and does not aim to capture the whole syntactic distribution. Given the large amount of candidate verbs in German, the need for a comprehensive data-driven approach and the fruitfulness of its ultimate results for theoretical work is evident, but it may also be used to improve experimental work (which was the original motivation for the annotation effort at hand): The observed large differences between verbs within broad classes like "accusative EO" suggest that one should not assume all its members to behave alike and that insights gained from testing a small number of verbs might not generalise to the whole assumed class. If one is interested in a specific pattern, a Reliance analysis (cf. Section 2.1) will help to find verbs that typically occur in it. Also, the annotations enable the search for sentences fulfilling specific criteria, which can – in a modified form – be used in experiments in turn, cf. the methodology of *modified stimulus composition* as described in (Börner et al., 2019).

Our findings based on a large-scaled corpus-based analysis strengthen the hypothesis that the often (implicitly) assumed homogeneity of a category like "accusative/dative EO verb" is not reflected in actual empirical behaviour. While some existing works argue to provide a subclassification of the psych verb domain (e.g. (Hirsch, 2018)), these works did not include strong corpus-linguistic aspects. Other corpus-based works (e.g. (Möller, 2015) on the past participle of German psych verbs) focused on a small number of syntactic phenomena and did not pursue a broader perspective.

1.2 Resource and Annotation Process

The basis of our annotation study are randomly extracted sentences from a corpus of the NZZ (*Neue Zürcher Zeitung*, volumes 1993-1999, cf. (Keßelmeier et al., 2009; NZZ, 1995 to 1999)). We chose 64 German EO verbs based on previous experimental and corpus studies (among others: (Rääts, 2011; Temme and Verhoeven, 2017; Hirsch, 2018; Engelberg, 2018)). To be included, the verb should be grammatically possible within a transitive EO-construction. We operationalised that by including verbs that are cited frequently as EO verbs in relevant publications or are clearly possible in such a construction by the intuition of all three (German native speaker) annotators. Semantically, the verb should display psych-predicate properties by clearly denoting an emotional or mental state or event. Both aspects are commonly referred to as distinctive features of psych EO verbs in the literature (cf. (Landau, 2010) among many others). Further constraints on the data set were imposed by balancing on overall frequency, case preference, morphological variety, and perfect tense auxiliary selection preference. For each of the candidate verbs, up to 200 samples were randomly extracted from the NZZ corpus. Roughly one third of these 64 verbs did not yield complete samples of 200 sentences due to their low corpus frequencies (for all sample sizes, cf. Appendix A). The samples were divided among and annotated by three native speakers of German with respect to a variety of syntactic patterns, the animacy of the stimulus (if present), an eventual stimulus-indicating PP or other kind of stimulus adjunct, syntactic aspects like control and a number of other potentially relevant factors.

The annotated syntactic patterns include: prototypical EO-transitive (X-STM_V_Y-EXP), intransitive (X-STM_V) without a syntactically realised EXP, and Acc/Dat-EXP_V without a phoric subject, but with a dative or accusative EXP. We further annotated both the stative (sein_V-PII) and the eventive/verbal (werden_V-PII) passive, and reflexive variants, where the experiencer is the subject (X_V_refl). Other patterns include constructions based on the past/perfect participle[2] like refl_V-PII_zeigen ("to show oneself/feel V-ed"), wirken/scheinen_V-PII ("seem V-ed"), NoAux_V-PII (without an auxiliary), where the status as a verb is rather doubtful (the same applies to the stative passive), a kind of causative pattern (X-CAUS_V_Y-EXP_PP), the reflexive pattern + an additional genitive NP (X-EXP_V_refl_Gen-STM)[3], *let*-constructions with a reflexive (X_lassen_refl_V), and a pattern that looks similar to the reflexive one but uses ablaut instead of reflexivisation (Nom-EXP_V), as well as modal infinitives and embedding into the *tough*-construction.

After the first annotation stage was completed, each of the samples was revised by at least one further annotator in a subsequent adjudication step to decide on problematic cases. We did not consider a classic inter-annotator agreement calculation as fruitful in this case, but verified every annotation by at least a simple majority decision among the annotators. This resulted in a data set with a total of 10,290 annotated examples. All analyses and visualisations were performed in R (R Core Team, 2020) using the *tidyverse* (Wickham et al., 2019). The data and accompanying material is publicly available via `https://github.com/Linguistic-Data-Science-Lab/German_EO_verbs`. A comprehensive table containing the frequency data for all verbs and patterns can be found in Appendix A.

2 Results and Implications

As our data shows, the vast majority of the given verbs display a large variety of syntactic patterns. This syntactically promiscuous behaviour has been considered typical for psych verbs (cf. e.g. (Hirsch, 2018)). The present corpus-annotation proved the domain of these verbs even more syntactically heterogeneous than expected. We are aware that our approach is, by all means, a frequentist one, which entails that we have to face the well-known issues that come along with frequentist methods, e.g. that non-occurrence in a corpus does not prove ungrammaticality *per se*, and the actual distribution might be to some extent corpus-dependent. However, we assume that higher frequencies of a specific verb in a specific syntactic configuration do reflect some characteristics of a verb. Additionally, it *is* possible to find occurrences of specific verbs in specific patterns that they were hitherto thought to disallow (cf. Section 2.3).

2.1 The Prototypical EO-transitive Pattern

As the corpus study on the syntactic pattern distribution of EO verbs has shown, their behaviour is notably heterogeneous (cf. Section 2.2). If a number of verbs that are frequently cited as examples for a particular verb class defined with a certain argument pattern in mind and researched (using introspective or experimental methods) as exemplars of that class turn out to occur in this pattern only comparatively rarely, then this can be crucial for experimental work as well as theoretical reflections. This may happen regardless of the general grammaticality in the respective pattern, which is, in our case, the transitive EO

[2] PII in the pattern names (after its traditional German name *Partizip II* "participle II").

[3] Most of these constructions appear limited to a small number of the candidate verbs, constructions like EXP_V_refl_Gen-STM are arguably no longer productive in Modern Standard German, cf. (Hirsch, 2018).

pattern.

To quantify the relation between the verbal lemma and the target construction mostly associated with the psych-EO class, we calculated the overall Reliance measure (introduced by Schmid (2000), defined in equation 1) for each verb (excluding examples where the lemma clearly does not appear in its psych reading[4]) for the transitive EO pattern (with an overt object experiencer, as in (2)) as well as the "object drop" intransitive construction, where the experiencer argument is not represented syntactically but is semantically present (a kind of arbitrary experiencer), as in (3).

(2) (NZZ_1995_11_28_a97_seg3_s1)

Die	Aura	der	Stararchitekten	bezaubert	neuerdings	die	Welt.
the.NOM	aura.NOM	the.GEN	star.architects.GEN	charms	recently	the.ACC	world.ACC

The aura of star architects recently charms the whole world.

(3) (NZZ_1994_08_24_a85_seg3_s10)

Die	musikalischen	Leistungen	imponierten	fast	durchweg.
the.NOM	musical.NOM	achievements.NOM	impressed	almost	the.entire.time

The musical achievements impressed almost consistently.

The association measure Reliance mirrors to what extent a certain lexeme (dis-)prefers a certain syntactic slot, calculated by the number of occurrences in a given construction (l_c) divided by the number of all occurrences of the lexeme, i.e. the sum of l_c and the number of observed occurrences in other constructions, $l_{\neg c}$.

$$R = \frac{l_c}{l_c + l_{\neg c}} \tag{1}$$

Figure 1 displays the Reliance for all verbs on the transitive (in black) and the intransitive (in grey) pattern, higher scores entailing a higher preference for the given construction. It strikes us as surprising that – while some verbs display a strong preference for the transitive EO pattern – particularly a number of verbs frequently used and analysed in works about the syntactic characteristics of EO verbs (e.g. (Verhoeven, 2014; Temme, 2018)) display a relatively low (or mediocre) Reliance score regarding the transitive pattern (with an overt experiencer object). This particularly holds for *verwundern* "to astonish", *verängstigen* "to frighten", *deprimieren* "to depress", *begeistern* "to thrill, enthuse", and *ausreichen* "to suffice", as well as for a number of other verbs, namely *interessieren* "to interest", *freuen* "to be glad", *empören* "to outrage", *amüsieren* "to amuse", *ekeln* "to disgust", *erfreuen* "to enjoy, delight", *langweilen* "to bore", where we find a pattern alternation with the reflexive construction (cf. Figure 2).

2.2 The Reflexive Pattern and its Relation to the Passive

This reflexive pattern is employed by some accusative[5] EO verbs. Here, the experiencer is realised in the subject position and the verb is reflexivised. The stimulus may be dropped entirely or be realised in a PP (the factors determining which preposition is used are still unclear although most verbs (heavily) favour one preposition). In the case of clausal stimuli, a pronominal adverb is used frequently (as in (5)).

(4) (NZZ_1995_08_15_a122_seg6_s3)

Man	amüsiert	sich	über	Tinguelys	Sinn	für	Satire	und	Ironie.
one.NOM	amuses	REFL	about	Tinguely's	sense.ACC	for	satire.ACC	and	irony.ACC

One is amused by Tinguely's sense of satire and irony.

[4] We have also identified a number of occurrences that can be considered "psych ambiguous" due to an ambiguity or vagueness of the verb between a mental state and a non-mental state meaning (an example is *schwerfallen* "to be/feel difficult"). For the Reliance analysis, both unambiguously psych and psych ambiguous examples were included, while unambiguously non-psych occurrences were not considered.

[5] *Gefallen* "to like" is a dative EO verb that has a reflexive variant (and is listed in Figure 2 accordingly) but its meaning in the reflexive variant is somewhat different from the one on the EO pattern and we probably have to deal with a different phenomenon here.

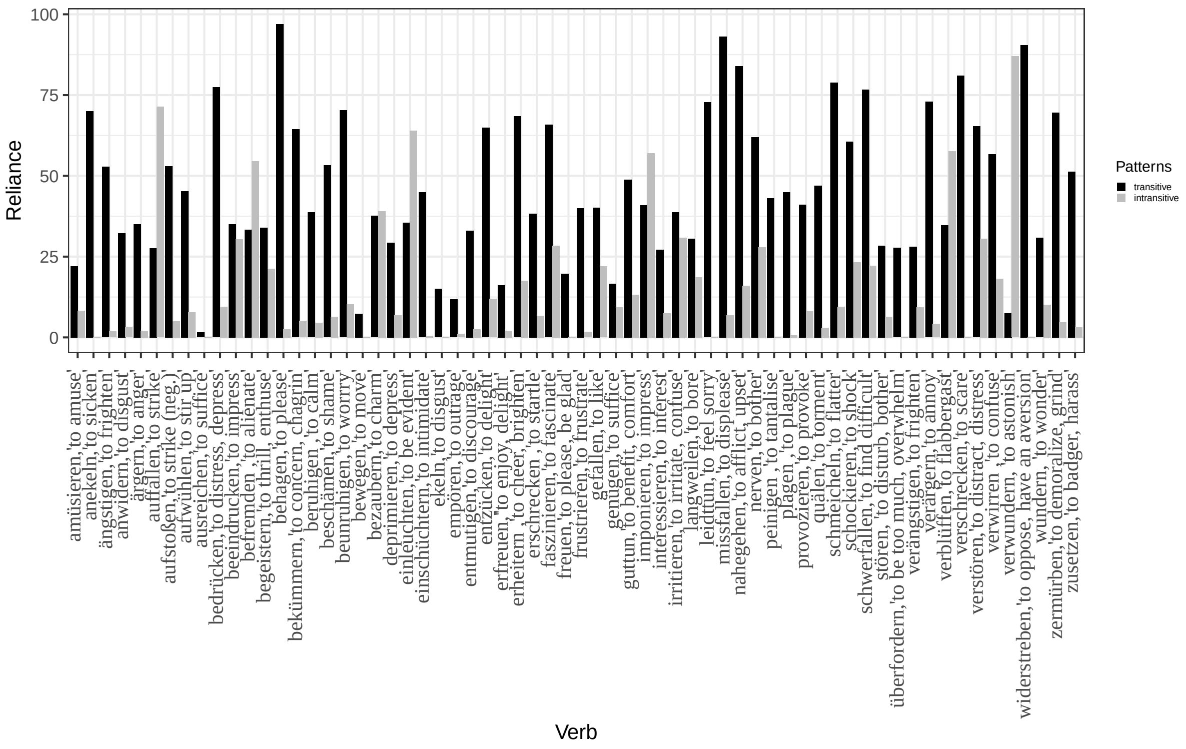


Figure 1: Reliance measure transitive and intransitive patterns

(5) (NZZ_1994_12_16_a188_seg29_s25)

23.00 Ich freue mich, dass du geboren bist.
23.00 I.NOM am.glad REFL that you.NOM born are

23.00 I'm glad you were born.

Engelberg (2018, pp. 61–65) observes that the experiencer subject variant is used far more often than the experiencer-object variant in verbs that allow for it, and asserts a negative correlation between the number of examples in the experiencer subject variant and the number of (eventive or stative) passive sentences (with a correlation coefficient of -0.64, compared to a correlation coefficient of +0.59 between passive sentences and EO sentences) (Engelberg, 2018, p. 64). He speculates that both patterns compete because they serve the same function with regard to information structure, namely allowing the experiencer to be realised in the position typically occupied by topics.[6] While we observe a correlation in our data, it is not nearly as strong as in Engelberg's. This is due to the fact that the verbs from his study (*amüsieren*, "to amuse", *ärgern* "to anger", *aufregen* "to upset" (not in our data), *freuen* "to please, be glad", *interessieren* "to interest", and *wundern* "to wonder") are among the ones employing the reflexive pattern most frequently (cf. Figure 2). This shifted perspective sheds light on the need for larger groups of test verbs.

As illustrated by Figure 2, some verbs, although allowing the pattern, occur in the reflexive pattern much less frequently than others. Furthermore, some verbs occur in the passive as well as in the reflexive pattern, and we observe differences between the eventive (*werden*) and the stative (*sein*) passive. If we consider only the patterns on psych usages of the verbs, we find a negative correlation of -0.12 between the reflexive pattern and both passive variants combined, -0.14 between reflexive and *werden* passive, and -0.06 between reflexive and *sein* passive. While it is possible that both patterns are employed to achieve certain configurations meeting the speaker's information-structural desires, we would not overestimate

[6] One should note that one would expect a third competitor due to the flexible German word order, namely simply putting the experiencer in topic position on the EO pattern. Word order with psych verbs in German is a complex topic though that we cannot delve into here.

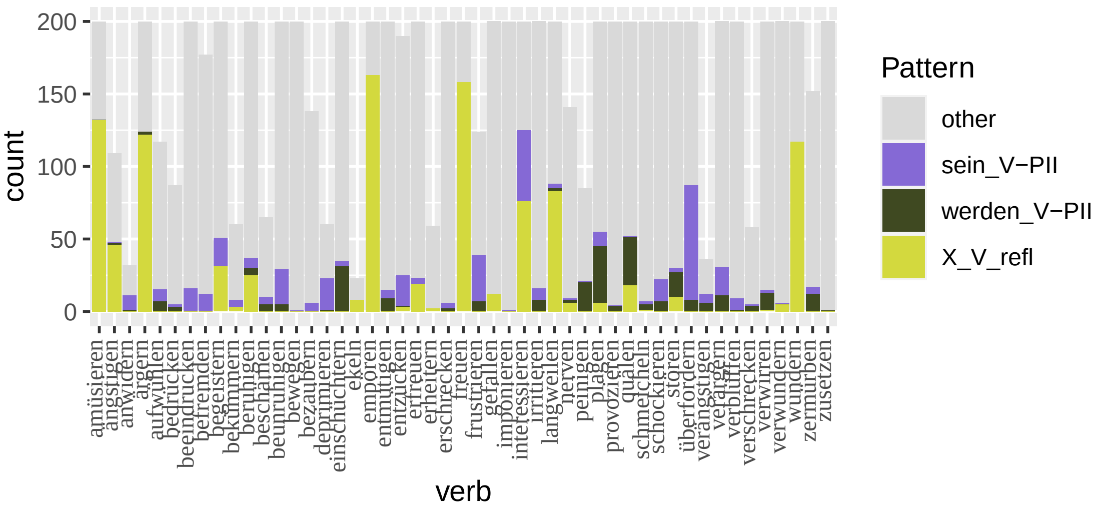


Figure 2: Co-occurrences of passive and reflexive constructions (only verbs occurring at least once in one of them)

this. Rather, we suspect that independent factors are responsible for the (un-)availability of the patterns – which is not to say that there is no (indirect) connection between them.

2.3 Variation among Subclasses

While a discussion of all interesting differences between the verbs would be far beyond the scope of this paper, we illustrate some pattern distribution variation in Figure 3. All of the verbs are accusative EO verbs and could thus naively be considered to fall into the same class. They all occur both within the

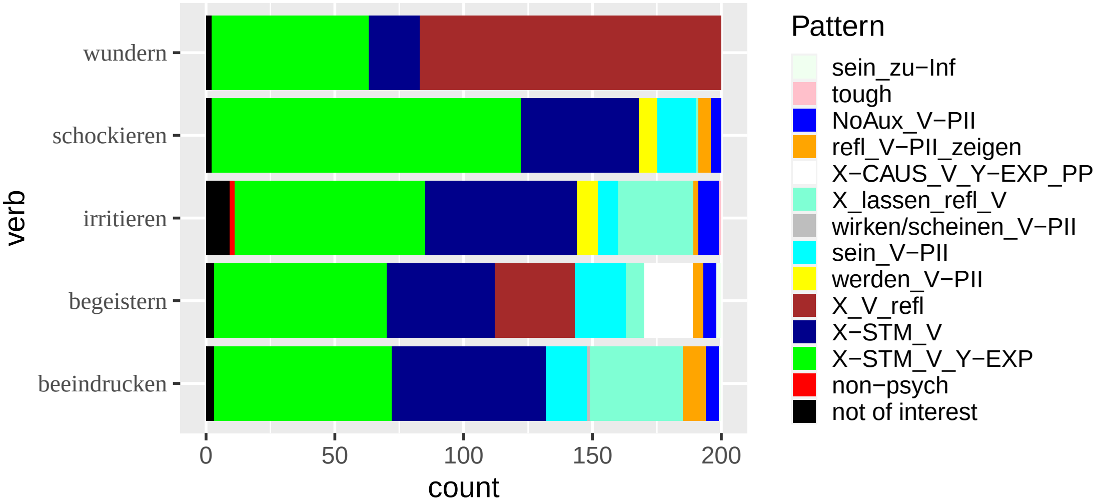


Figure 3: Pattern distribution for selected acc-verbs

transitive and the intransitive pattern. This is intriguing because it poses a challenge for all accounts of accusative EO verbs that do not take the experiencers to be “real” objects since, in German, so-called object-drop (which is what happens with our intransitive pattern) is only considered possible with “real” objects in the literature (cf. (Hirsch, 2018, pp. 163–165)).[7] While *wundern* “to wonder” only occurs in three patterns and is dominated by the reflexive pattern, the other verbs are much more flexible although

[7] Hirsch himself argues based on introspective judgements that a subclass of accusative EO verbs containing *wundern* “to wonder” does in fact not allow it.

only *begeistern* “to thrill, enthuse” also displays the reflexive pattern. It is also one of only two verbs in the data set to showcase the construction we call X-CAUSE_V_Y-EXP_PP, where semantically the subject referent causes the experiencer (realised as the object) to be in the psychological state expressed by the verb towards an object of emotion, which is realised in a PP (cf. (6)).[8]

(6) Der Professor begeisterte seine Studenten für die Linguistik.
the.NOM professor.NOM enthused his.ACC students.ACC for the.ACC linguistics.ACC
The professor made his students get excited about linguistics.

Only *schockieren* “to shock” and *irritieren* “to irritate, confuse” occur in the eventive/verbal passive (werden_V-PII).

3 Conclusion and Further Perspectives

The assumed class of EO verbs and their realisation patterns remain a complex matter. Certain assumptions about verbs considered EO do not appear to hold from a larger-scaled quantitative perspective. This also affects the subclasses proposed on the basis of case preferences. It is also notable that a number of verbs that are considered “typical” for this verbal class (despite its debated heterogeneity and unresolved classification approaches) appear to have a strong association with syntactic patterns other than the prototypical one, e.g. the reflexive construction, which might particularly affect experimental research. We consider both the quantitative perspective as well as a gold-standard annotated resource of sufficient scope as necessary for further research on the issue, particularly in the domain of experimental and theoretical linguistics.

Acknowledgements

This work is part of the project *The grammar of Experiencer-Object verbs: theoretical, computational and experimental approaches towards reflexive binding in German*, which is supported by the Deutsche Forschungsgemeinschaft (DFG) grant KI 759/9-1 to Tibor Kiss. We would like to thank three anonymous SyntaxFest reviewers for their comments.

References

Artemis Alexiadou and Gianina Iordăchioaia. 2014. The psych causative alternation. *Lingua*, 148:53–79.

Laura Becker and Matías Guzmán Naranjo. 2020. Psychpredicates in European languages: A parallel corpus study. *STUF – Language Typology and Universals*, 73(4):483–523.

Adriana Belletti and Luigi Rizzi. 1988. Psych-verbs and θ-theory. *Natural Language & Linguistic Theory*, 6(3):291–352.

Alicia Katharina Börner, Jutta Pieper, and Tibor Kiss. 2019. Corpus data in experimental linguistics. In *Proceedings of Linguistic Evidence 2018: Experimental Data Drives Linguistic Theory*, Tübingen. University of Tübingen.

Ruxandra Cosma and Stefan Engelberg. 2014. Subjektsätze als alternative Valenzen im Deutschen und Rumänischen. In Ruxandra Cosma, Stefan Engelberg, Susan Schlotthauer, Speranța Stanescu, and Gisela Zifonun, editors, *Komplexe Argumentstrukturen: Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen*, pages 339–420. Akademie-Verlag, Berlin.

Emilia Ellsiepen and Markus Bader. 2018. Constraints on argument linearization in German. *Glossa*, 3(1):1–36.

[8] We only subsume examples with the semantics specified above under this label. Examples with e.g. a resultative PP do not fall under it. An anonymous reviewer remarks that other verbs might allow this pattern as well and it might be due to the limited number of occurrences we looked at that we did not find them. This is, of course, true and – as (i) shows – it is possible to construct such examples for *faszinieren* “to fascinate”.

(i) Der Professor faszinierte seine Studenten für die Linguistik.
the.NOM professor.NOM fascinated his.ACC students.ACC for the.ACC linguistics.ACC
The professor made his students get fascinated about linguistics.

Stefan Engelberg. 2018. The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment. In Hans C. Boas and Alexander Ziem, editors, *Constructional Approaches to Syntactic Structure in German*, pages 47–84. De Gruyter Mouton, Berlin, Boston.

Jason Grafmiller. 2013. *The semantics of syntactic choice: An analysis of English emotion verbs*. Ph.D. thesis, Stanford University.

Friederike Haupt, Matthias Schlesewsky, Dieter Roehm, Angela D. Friederici, and Ina Bornkessel-Schlesewsky. 2008. The status of subject object reanalyses in language comprehension architecture. *Journal of Memory and Language*, 56(1):54–96.

Nils Hirsch. 2018. *German psych verbs – insights from a decompositional perspective*. Ph.D. thesis, Humboldt-Universität zu Berlin.

Katja Keßelmeier, Tibor Kiss, Antje Müller, Claudia Roch, Tobias Stadtfeld, and Jan Strunk. 2009. Mining for preposition-noun constructions in German. In Markus Sahlgren and Ola Knutsson, editors, *Mining for Preposition-Noun Constructions in German*, Proceedings of the Workshop on Extracting and Using Constructions in NLP at the 17th Nordic Conference of Computational Linguistics, pages 10–15.

Idan Landau. 2010. *The Locative Syntax of Experiencers*. MIT Press, Cambridge, MA.

Max Möller. 2015. *Das Partizip II von Experiencer-Objekt-Verben: Eine korpuslinguistische Untersuchung*, volume 6 of *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP)*. Narr Francke Attempto.

NZZ. 1995 to 1999. Neue Zürcher Zeitung: NZZ: der komplette Jahrgang 1993–1999.

David Michael Pesetsky. 1995. *Zero syntax: Experiencers and Cascades*. Current Studies in Linguistics. MIT Press, Cambridge, MA.

Dirk Pijpops and Dirk Speelman. 2017. Alternating argument constructions of Dutch psychological verbs: A theory-driven corpus investigation. *Folia Linguistica*, 51(1):207–251.

R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Julian A. Rott, Elisabeth Verhoeven, and Paola Fritz-Huechante. 2020. Valence orientation and psych properties: Towards a typology of the psych alternation. *Open Linguistics*, 6:401–423.

Anni Rääts. 2011. *Semantik und (Morpho-)Syntax der Emotionsverben im Deutschen und im Estnischen*. Ph.D. thesis, University of Tartu.

Christoph Scheepers, Barbara Hemforth, and Lars Konieczny. 2000. Linking Syntactic Functions with Thematic Roles: Psych-Verbs and the Resolution of Subject-Object Ambiguity. In Barbara Hemforth and Lars Konieczny, editors, *German Sentence Processing*, pages 95–135. Springer Netherlands, Dordrecht.

Hans-Jörg Schmid. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Number 34 in Topics in English linguistics. Mouton de Gruyter.

Anne Temme and Elisabeth Verhoeven. 2017. Backward binding as a psych effect: A binding illusion? *Zeitschrift für Sprachwissenschaft*, 36(2):279–308.

Anne Temme. 2018. *The peculiar nature of psych verbs and experiencer object structures*. Ph.D. thesis, Humboldt-Universität zu Berlin.

Elisabeth Verhoeven. 2014. Thematic prominence and animacy asymmetries. evidence from a cross-linguistic production study. *Lingua*, 143:129–161.

Elisabeth Verhoeven. 2015. Thematic asymmetries do matter! a corpus study of German word order. *Journal of Germanic Linguistics*, 27:45–104.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Sylwiusz Żychliński. 2016. *On some aspects of the Syntax of Object Experiencers in Polish and English*. Wydawnictwo Naukowe UAM, Poznań.

Appendix A: Pattern Distribution

Verb	translation	case EXP	X-STM_V_Y-EXP	X-STM_V	X_V_refl	werden_V-PII	sein_V-PII	wirken/scheinen_V-PII	X_lassen_refl_V	X-CAUS_V_Y-EXP_PP	refl_V-PII_zeigen	Nom-EXP_V	NoAux_V-PII	EXP_V_refl_Gen-STM	Acc/Dat-EXP_V	*tough*	sein_zu-Inf	(non-amb.) non-psych	not of interest	Sample size
amüsieren	amuse	a	43	16	132	0	0	0	0	0	0	0	4	0	0	0	0	0	5	200
anekeln	sicken	a	14	0	0	0	0	0	0	0	1	0	4	0	1	0	0	0	0	20
ängstigen	frighten	a	57	2	46	1	1	0	1	0	0	0	0	0	0	0	0	0	1	109
anwidern	disgust	a	10	1	0	1	10	0	0	0	3	0	6	0	0	0	0	0	1	32
ärgern	anger	a	69	4	122	2	0	0	0	0	0	0	0	0	0	0	0	0	3	200
auffallen	strike	d	55	142	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	200
aufstoßen	strike (neg.)	d	105	10	0	0	0	0	0	0	0	0	0	0	0	0	0	83	2	200
aufwühlen	stir up	a	48	9	0	7	8	1	0	0	2	0	9	0	0	1	0	30	2	117
ausreichen	suffice	d	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	193	5	200
bedrücken	distress, depress	a	62	8	0	3	2	1	0	0	0	0	1	0	0	0	0	7	3	87
beeindrucken	impress	a	69	60	0	0	16	1	36	0	9	0	5	0	0	0	1	0	3	200
befremden	alienate	a	58	95	0	0	12	0	0	0	4	0	5	0	0	0	0	0	3	177
begeistern	thrill, enthuse	a	67	42	31	0	20	0	7	19	4	0	5	0	0	0	2	0	3	200
behagen	please	d	193	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	200
bekümmern	concern, chagrin	a	38	3	3	0	5	0	0	0	1	0	4	0	0	0	0	5	1	60
beruhigen	calm	a	74	9	25	5	7	0	2	0	0	0	5	0	0	0	1	71	1	200
beschämen	shame	a	33	4	0	5	5	0	1	0	1	0	13	0	0	0	0	0	3	65
beunruhigen	worrry	a	137	20	0	5	24	1	3	0	1	0	4	0	0	0	0	0	5	200
bewegen	move	a	10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	125	64	200
bezaubern	charm	a	50	52	0	0	6	0	21	0	0	0	4	0	0	0	0	0	5	138
deprimieren	depress	a	17	4	0	1	22	3	2	0	1	0	7	0	0	1	0	0	2	60
einleuchten	be evident	d	70	126	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	200
einschüchtern	intimide	a	88	1	0	31	4	1	56	0	0	0	8	0	0	2	5	0	4	200
ekeln	disgust	a	3	0	8	0	0	0	0	0	0	0	0	0	4	0	0	5	3	23
empören	outrage	a	22	2	163	0	0	0	0	0	0	0	0	0	0	0	0	0	13	200
entmutigen	discourage	a	65	5	0	9	6	1	91	0	1	0	15	0	0	0	0	4	3	200
entzücken	delight	a	120	22	3	1	21	0	1	0	4	0	13	0	0	0	0	0	5	190
erfreuen	enjoy, delight	a	32	4	19	0	4	0	0	0	3	0	0	133	0	0	0	3	2	200

Verb	translation	case EXP	X-STM_V_Y-EXP	X-STM_V	X_V_refl	werden_V-PII	sein_V-PII	wirken/scheinen_V-PII	X_lassen_refl_V	X-CAUS_V_Y-EXP_PP	refl_V-PII_zeigen	Nom-EXP_V	NoAux_V-PII	EXP_V_refl_Gen-STM	Acc/Dat-EXP_V	*tough*	sein_zu-Inf	(non-amb.) non-psych	not of interest	Sample size
erheitern	cheer, brighten	a	39	10	2	0	0	0	3	0	0	0	2	0	0	1	0	0	2	59
erschrecken	startle	a	74	13	0	2	4	1	4	0	0	91	3	0	0	0	1	0	7	200
faszinieren	fascinate	a	125	54	0	0	0	0	11	0	0	0	0	0	0	0	0	0	10	200
freuen	please, be glad	a	39	0	158	0	0	0	0	0	0	0	0	1	0	0	0	0	2	200
frustrieren	frustrate	a	48	2	0	7	32	1	3	0	8	0	19	0	0	0	0	0	4	124
gefallen	like	d	73	40	12	0	0	0	0	0	0	0	0	0	7	0	0	50	18	200
genügen	suffice	d	23	8	0	0	0	0	0	0	0	0	0	0	0	0	0	161	8	200
guttun	benefit, comfort	d	46	17	0	0	0	0	0	0	0	0	0	0	0	0	0	111	26	200
imponieren	impress	d	79	110	0	0	1	0	3	0	0	0	0	0	0	0	0	0	7	200
interessieren	interest	a	54	15	76	0	49	0	0	1	3	0	1	0	0	0	0	0	1	200
irritieren	irritate, confuse	a	74	59	0	8	8	0	29	0	2	0	8	0	0	1	0	2	9	200
langweilen	bore	a	59	36	83	2	3	1	0	0	1	0	8	0	0	0	0	0	7	200
leidttun	feel sorry	d	139	0	0	0	0	0	0	0	0	0	0	0	52	0	0	0	9	200
missfallen	displease	d	177	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	200
nahegehen	afflict, upset	d	21	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	29
nerven	bother	a	80	36	6	2	1	0	0	0	0	0	4	0	0	0	0	0	12	141
peinigen	tantalise	a	31	0	0	20	1	0	2	0	0	0	18	0	0	0	0	0	13	85
plagen	plague	a	71	1	6	39	10	0	1	0	0	0	28	0	0	0	0	2	42	200
provozieren	provoke	a	77	16	0	4	0	0	3	0	0	0	0	0	0	0	0	97	3	200
quälen	torment	a	93	6	18	33	1	3	1	0	0	0	4	0	0	0	0	39	2	200
schmeicheln	flatter	d	155	17	1	4	2	0	1	0	6	0	2	0	0	0	0	11	1	200
schockieren	shock	a	120	46	0	7	15	0	1	0	5	0	4	0	0	0	0	0	2	200
schwerfallen	find difficult	d	67	14	0	0	0	0	0	0	0	0	0	0	0	0	0	108	11	200
stören	disturb, bother	a	49	12	10	17	3	0	4	0	2	0	0	0	0	0	0	90	13	200
überfordern	be too much, overwhelm	a	53	0	0	8	79	4	0	0	5	0	6	0	0	0	0	40	5	200
verängstigen	frighten	a	9	3	0	6	6	0	0	0	0	0	8	0	0	0	0	0	4	36
verärgern	annoy	a	140	8	0	11	20	0	0	0	3	0	10	0	0	0	0	0	8	200
verblüffen	flabbergast	a	68	113	0	1	8	0	0	0	1	0	5	0	0	0	0	0	4	200

Verb	translation	case EXP	X-STM_V_Y-EXP	X-STM_V	X_V_refl	werden_V-PII	sein_V-PII	wirken/scheinen_V-PII	X_lassen_refl_V	X-CAUS_V_Y-EXP_PP	refl_V-PII_zeigen	Nom-EXP_V	NoAux_V-PII	EXP_V_refl_Gen-STM	Acc/Dat-EXP_V	*tough*	sein_zu-Inf	(non-amb.) non-psych	not of interest	Sample size
verschrecken	scare	a	47	0	0	4	1	0	0	0	1	0	5	0	0	0	0	0	0	58
verstören	distract, distress	a	32	15	0	0	0	0	1	0	0	0	0	0	0	0	0	1	3	52
verwirren	confuse	a	105	34	1	12	2	0	8	0	0	0	4	0	0	1	0	20	13	200
verwundern	astonish	a	15	174	5	0	1	0	0	0	0	0	4	0	0	0	1	0	0	200
widerstreben	oppose, have an aversion	d	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	2	76
wundern	wonder	a	61	20	117	0	0	0	0	0	0	0	0	0	0	0	0	0	2	200
zermürben	demoralize, grind	a	100	7	0	12	5	0	4	0	0	0	17	0	0	0	0	3	4	152
zusetzen	badger, harass	d	86	3	0	1	0	0	0	0	0	0	0	0	0	0	0	101	9	200

Table 1: Pattern distribution and sample size for each verb

The Menzerath-Altmann law in syntactic structure revisited: Combining linearity of language with dependency syntax

Ján Mačutek, Radek Čech, Marine Courtin
Mathematical Institute, Slovak Academy of Sciences, Slovakia & Department of Mathematics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Slovakia
Department of Czech Language, Faculty of Arts, University of Ostrava, Czech Republic
LPP (CNRS) – Sorbonne Nouvelle, France
jmacutek@yahoo.com, cechradek@gmail.com, marine.courtin@sorbonne-nouvelle.fr

Abstract

According to the Menzerath-Altmann law, there is inverse proportionality between sizes of language units and their constituents (i.e., longer language units are composed of shorter constituents, and vice versa). The validity of the law was confirmed many times for the relation between lengths of a word and its syllables. However, the relation between lengths of sentences (measured in clauses) and clauses (measured in words) is problematic. In this paper, a new language unit – linear dependency segment – is introduced with the motivation to avoid some problems connected to the Menzerath-Altmann law on the syntactic level. The new unit is intermediate between clause and word and its definition takes into account both the linearity of language and dependency syntactic structure. It is shown that the relation between sentence length in clauses and clause length measured in linear dependency segments abides by the Menzerath-Altmann law in two Czech dependency treebanks.

1 Introduction

The Menzerath-Altmann law (MAL henceforward) predicts relations between sizes of language units which are neighbours in the language unit hierarchy. According to the law, longer units which are higher in the hierarchy (constructs) consist of shorter lower units (constituents). The formulation of the MAL developed from a verbal one (the longer the word, the shorter on average its syllables; see Menzerath, 1954) to mathematical formula

$$(1) \qquad y(x) = ax^{b}e^{-cx}$$

derived by Altmann (1980). In formula (1), $y(x)$ is the mean size of constituents in the construct of size x; a, b and c are parameters. Very often a simpler formula,

$$(2) \qquad y(x) = ax^{b},$$

is used, which is a special case of function (1) for $c = 0$.

The MAL was first observed as the relation between word length in syllables and either syllable length in phonemes[1] (Menzerath, 1954), or syllable duration in time (Menzerath and de Oleza, 1928;

[1] Sometimes, word length is measured in graphemes instead of phonemes. This approach is applied mainly in languages which have a close phoneme-grapheme correspondence.

Geršić and Altmann, 1980). The validity of the MAL at this lowest level was scrutinized in many languages (see e.g. Cramer, 2005, and references therein; Kelih, 2010, 2012; Mikros and Milička, 2014; Mačutek et al., 2019).

However, two fundamental problems emerge when one goes higher in the hierarchy of language units. First, it was assumed that the upper neighbours of word are clause and sentence. Although several papers in 1980s (Köhler, 1982; Heups, 1983; Schwibbe, 1984; Teupenhayn and Altmann, 1984) claim that the relation between sentence length in clauses and clause length in words abides by the MAL, more recent results are far from clear. Thus, Kułacka (2010), Chen and Liu (2019), and Xu and He (2020) confirm the older results, while data analysed by Kułacka and Mačutek (2007), Benešová and Čech (2015), and Hou et al. (2017) display a Menzerathian tendency, but they cannot be fitted by function (1) sufficiently well.[2] On the other hand, data presented by Buk and Rovenchak (2008) and by Andres and Benešová (2012) do not confirm to the MAL.[3] Curiously enough, Andres and Benešová (2012) and Hou et al. (2019) are, to our best knowledge, the only two papers which focus also on the relation between lengths of clause (in words) and word (in syllables).[4] This relation, again, cannot be modelled by the MAL. To put it mildly, the empirical evidence of the MAL, especially in form of function (2), is doubtful as soon as we move from word to clause and sentence.

Mačutek et al. (2017) tried to measure clause length in syntactic phrases which are directly dependent on the predicate of the clause (with phrase length being measured in words). The MAL in form (2) achieved a very good fit. The phrase thus became a candidate for an intermediate language unit between word and clause. It must be noted that only main clauses were analysed, and only one Czech treebank was used.

Second, although the linguistic interpretation of the parameters of model (1) is still not known, it was suggested that the MAL has something to do with short term memory (Köhler, 1989; Grzybek, 2013; see also Yngve, 1960, 1996).[5] According to the well-known paper by Miller (1956), the capacity of short-term memory is approximately seven. With the exception of polysynthetic languages, words only seldom contain more than seven syllables (or morphemes[6]), and the same is true for sentence length in clauses. However, clauses longer than seven words are not so rare – the mean clause length in the papers cited above is often somewhere near 10, see e.g. Köhler (1982), Heups (1983), and Teupenhayn and Altmann (1984).

The phrase used by Mačutek et al. (2017) faces the same problem, e.g. there are 7,125 clauses (more than 12%) which contain only one phrase, and their mean length in words is 9.47 (which means that are many phrases longer than 9.47). In addition, consider a sentence consisting only of a single predicate (e.g. Czech sentence *Prší "It rains"*). Such a sentence contains only one clause of length zero (because there is nothing directly dependent on the predicate of the clause), and phrase length cannot be determined at all, as there is no phrase in the sense of the phrase definition from Mačutek et al. (2017). If the definition is modified so that phrase includes also the predicate, the question arises how to determine phrase length in clauses consisting of at least two phrases (such as e.g. in Czech sentence *Petr miluje Marii "Peter loves Mary"*). If the predicate is a part of the phrases, it appears more than once in all calculations. Regardless of these methodological difficulties, phrase has also a drawback of neglecting the linearity of language.

[2] See Mačutek and Wimmer (2013) for an overview of goodness-of-fit criteria usually used in quantitative linguistics.

[3] Admittedly, these papers do not follow the same methodology. In most of them, either finite verbs or punctuation marks (comma and semicolon) to determine sentence length in clauses.

[4] Hou et al. (2019) measure word length in characters, but in written Chinese there is almost one-to-one correspondence between characters and syllables.

[5] Torre et al. (2019) present an attempt to explain the origin of the MAL in spoken language at the level of words and syllables as a consequence of human physiology (in particular the necessity to breathe). These two tentative explanations of the MAL do not exclude each other; rather, both factors (pauses caused by breathing and a limited capacity of short-term memory) are likely to contribute to the shortening of constituents in longer constructs.

[6] See Pelegrinová et al. (2021) and references therein for the MAL as the relation between word length in morphemes and morpheme length in phonemes.

To avoid the abovementioned problems, we suggest another approach, namely, a new language unit between word and clause is introduced. Its definition combines both linear and hierarchical dependency structure of sentence. We focus on the question whether this new unit behaves according to the MAL.

The paper is structured as follows. Section 2 introduces the linear dependency segment, a new unit positioned between clause and word. In Section 3, language material used for the analysis is described. Results achieved are presented in Section 4. The paper is concluded by a short discussion which contains also some ideas for future research in this area.

2 Linear dependency segment

We define the linear dependency segment (LDS henceforward) as the longest possible sequence of words (belonging to the same clause[7]) in which all linear neighbours (i.e. words adjacent in a sentence) are also syntactic neighbours (i.e. they are connected by an edge in the syntactic dependency tree which represents the sentence). Figure 1 presents the dependency tree of sentence *"This black book on the table costs twenty euros, which is too much for me"*.

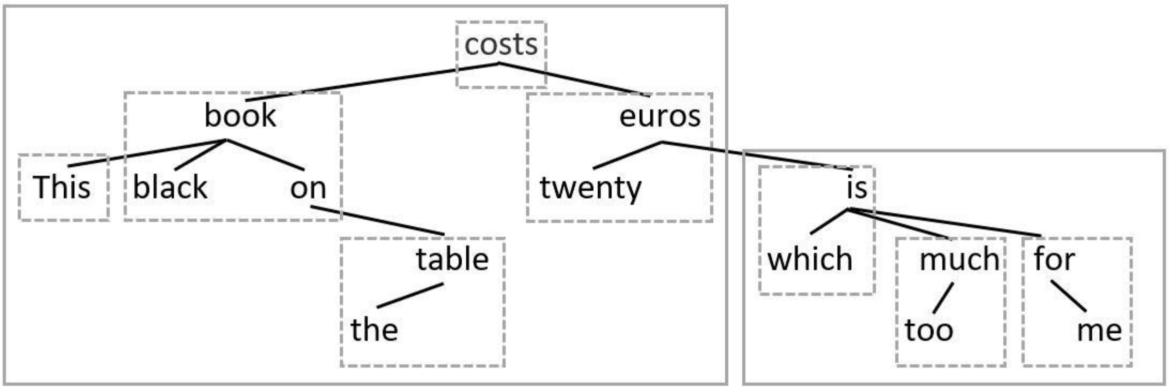


Figure 1. Dependency tree of sentence *"This black book on the table costs twenty euros, which is too much for me"*

Consider the first clause in the sentence. Its first word, "*This*", is syntactically linked with "*book*", but these two words are not linear neighbours. Therefore, the first LDS is [This]. Next, the second word, "*black*", is syntactically linked with "*book*", which is also its linear neighbour, and the third and the fourth words, "*book*" and "*on*", are again both linear and syntactic neighbours. Here the segment is ended, because the next word, "*the*", is not syntactically linked with "*on*". Examining the whole clause we obtain LDSs [This][black book on][the table][costs][twenty euros]. Similarly, the second clause in this sentence has LDSs [which is][too much][for me]. We remind that we define the LDSs as units of which clauses are composed, i.e. a LDS is always ended at the end of a clause.

The definition is good in the sense that every clause can be unambiguously divided into LDSs, and that the intersection of two different LDSs is the empty set (i.e. every word in a clause belongs to one and only one LDS).

From the MAL point of view, clause is a construct and LDS its constituent (which, in turn, is a construct itself, with words being its constituents). We expect that longer sentences (measured in the number of clauses) contain shorter clauses (measured in the number of LDSs), and vice versa. This expectation is based on the fact that dependency links which do not respect the linearity of a sentence are more difficult to process.[8] The same is true for a sentence with many clauses. The MAL does not allow sentences to become too complex, as it "forces" clauses in long sentences (i.e. in ones which

[7] We use the definition of clause from Prague Dependency Treebank 3.0 (https://ufal.mff.cuni.cz/pdt3.0/documentation#__RefHeading__42_1200879062), according to which "[a] clause typically corresponds to a single proposition expressed by a finite verb and all its arguments and modifiers (unless they constitute clauses of their own)".

[8] The idea that dependency distance in language is shorter than a random baseline can be traced back to Liu (2008).

contain many clauses) to become shorter (i.e. to be composed of fewer LDSs). Fewer LDSs mean that there are fewer dependency distances (as defined by Liu, 2008, p. 164) longer than one (as all dependency distances within one LDS are minimal, i.e. equal to one).

Provided that the MAL is valid as a model for the relation between lengths of sentences and clauses, a sentence can be composed either of more clauses which are shorter in terms in LDSs (which means that they are syntactically simpler[9]), or of fewer clauses which are "allowed" to contain more LDSs (and consequently to be syntactically more complex)

3 Language material

For the analysis, we used two Czech treebanks, the Czech-PDT UD[10] and the FicTree (Jelínek, 2017). The treebanks were converted to the Surface Syntactic Universal Dependencies (SUD) annotation scheme (Gerdes et al., 2018). The use of the Universal Dependency annotation scheme (de Marneffe et al., 2021) was also considered. However, we prefer the SUD approach because it is based on surface-syntactic distributional criteria that fit the nature of our analysis better than the Universal Dependency approach which is based on "a mixture of semantic and syntactic motivations" (Osborne and Gerdes, 2019).

The Czech-PDT UD consists of 87,913 Czech sentences from non-abbreviated newspaper, business and popular scientific journal articles published from 1991 to 1995. The FicTree consists of 12,760 sentences from Czech literary works published between 1991 and 2007. The treebanks were also merged and treated as one whole in which different genres are represented. Sentences without a predicate (especially titles of newspaper articles) were removed. We thus analysed altogether 86,266 sentences.

4 Results

As we study the relation between sentence length and the mean clause length, the number of clauses from which the mean is calculated cannot be too low if the result should be robust. We decided to take into account sentence lengths with frequencies which make at least 0.1% of our language material. We thus disregarded sentences containing more than eight clauses (together 76 sentences, i.e. 0.09%). Very complicated structures, such as several clauses placed in brackets, clauses separated by a colon, or citations, are typical for these long sentences. The possibility to check thoroughly sentences which do not conform to the MAL was also the reason why we focus only on Czech treebanks in this paper – one of the coauthors is a native Czech speaker. It is obvious that our choice substantially limits the scope of this paper, but given that it is the first attempt to study the LDS as a language unit, we prefer this more careful approach.

The relation between sentence length in clauses and the mean clause length measured in LDSs is presented in Table 1.

[9] If we consider the extreme case, a clause consisting of only one LDS either contains only one word, or it reaches the minimum of dependency distance (in such a clause all dependency distances are equal to one).

[10] https://universaldependencies.org/treebanks/cs_pdt/index.html

	merged			PDT			FicTree		
SL	f	rf	MCL	f	rf	MCL	f	rf	MCL
1	36559	0.424	5.02	32002	0.428	5.30	4557	0.396	3.03
2	27735	0.321	3.93	24121	0.323	4.10	3614	0.314	2.82
3	13463	0.156	3.44	11605	0.155	3.54	1858	0.162	2.79
4	5416	0.063	3.17	4537	0.061	3.25	879	0.076	2.77
5	1962	0.023	3.00	1616	0.022	3.07	346	0.030	2.69
6	727	0.008	2.94	580	0.008	3.02	147	0.013	2.64
7	236	0.003	2.84	188	0.003	2.85	48	0.004	2.82
8	92	0.001	2.79	69	0.001	2.93	23	0.002	2.36

Table 1. The MAL in Czech dependency treebanks (SL - sentence length in clauses, f, rf - frequencies and relative frequencies[11] of sentence lengths, MCL – the mean clause length in LDSs).

The MAL in form (2) fits the data from the merged treebanks very well[12], with $R^2 = 0.9836$ ($a = 4.918$, $b = -0.296$).[13] The data and the graph of the function can be seen in Figure 2.

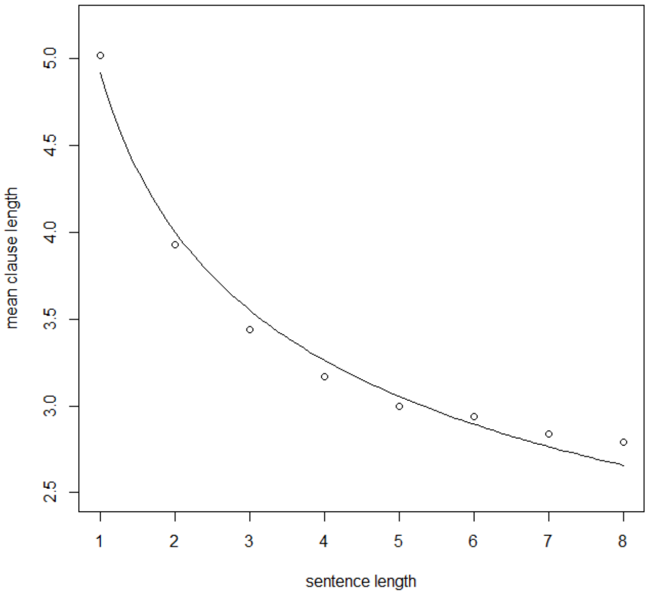


Figure 2. The MAL modelled by function $y(x) = ax^b$ as the relation between sentence length and the mean clause length

The value of parameter a is very close to the mean clause length (measured in the number of LDSs) in sentences consisting of only one clause. If we use this value, i.e. if we set $a = 5.02$ in formula (2),

[11] The relative frequencies do not sum to one, because sentences containing more than eight clauses were disregarded.

[12] The most common rule of thumb in quantitative linguistics is to consider the goodness-of-fit of a model satisfactory if the value of the determination coefficient R^2 is higher than 0.9, see Mačutek and Wimmer (2013).

[13] The fit remains satisfactory also if other options how to deal with low frequency construct length are applied. If all construct lengths with frequency at least 10 are used in the computations (see Mačutek and Rovenchak, 2011), we have $R^2 = 0.9353$, and if we pool low-frequency construct lengths (i.e. sentence which contain more than eight clauses in our case) and compute the weighted mean of clause lengths (see e.g. Pelegrinová et al., 2021), we obtain $R^2 = 0.9649$.

we obtain $b = -0.309$ and $R^2 = 0.9803$, which is still a very good fit. We thus have a very clear interpretation of the parameter a.[14] As for parameter b, its linguistic interpretation remains an open question.

In both PDT and FIC treebanks, the decreasing tendency of the mean clause length can be observed. While the fit of function (2) remains very good ($R^2 = 0.9739$) for PDT, it is much worse ($R^2 = 0.6148$) for the data from the FicTree treebank. However, this is caused by an irregular behaviour of the mean clause length of the two highest values of sentence length, which occur with relatively low frequencies (moreover, the FicTree treebank is much smaller than PDT), and an overall decreasing tendency can be seen also in results from this treebank.

The two treebanks differ also in the mean values of the shortest sentences (i.e. the ones containing only one clause). Most likely, it is a consequence of different sentence length distributions in the treebanks (the mean values are 1.97 for PDT and 2.11 for FicTree; see also relative frequencies of sentence lengths in Table 1). Longer sentences in FicTree are composed of shorter LDSs. We remind that the treebanks consist of journalistic texts (PDT) and fiction (FicTree), and that sentence length depends on genre (see e.g. Kelih et al., 2006; Xu and He, 2020).

5 Conclusion

The achieved results indicate that, at least tentatively, the LDS can be considered a meaningful linguistic unit which allows to model the MAL also on the level of syntax. The LDS avoids the problems frequently encountered when one measures clause length in the number of words the clause contains. From the theoretical point of view, it is important that clause length measured in LDSs correspond with the capacity of short-term memory[15], which is one of theoretical explanations of the MAL. Furthermore, we emphasize that the definition of the LDS takes into account both the linearity of language and the dependency syntactic structure.

Naturally, this paper is only a pilot study, very limited in its scope, and data from many more typologically diverse languages must be analysed before the LDS can establish itself firmly among more traditional language units. Specifically with respect to the MAL, also relations between lengths of clauses (in LDSs) and LDSs (in words) and between lengths of LDSs (in words) and words (in syllables or morphemes) must be investigated. In addition, if the LSD turns out to be a suitable linguistic unit, also its frequencies and its length are supposed to follow distribution laws which are commonly used to model these language properties (i.e. a Zipf-like distribution for LDS frequencies, and a Poisson-like distribution for LSD length, see e.g. Popescu et al., 2009, and Grzybek, 2006, respectively).

Parameter values of the MAL in form of function (2) can probably be used in automatic text classification procedures, as they depend on sentence length, which, in turn, depends on genre.

A possible correspondence between LDSs and dependency distance minimization deserves a closer inspection. While there is a strong evidence that words which are syntactically linked are close to each other also with respect to the linear order of the sentence (see e.g. Liu, 2008; Ferrer-i-Cancho and Liu, 2014; Futrell et al., 2015), short sentences are quite likely not to follow this trend (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). Although sentence length in these studies is expressed in the number of words (as opposed to clauses from our approach) they contain, we can suppose that short sentences mostly contain one or two clauses. The MAL predicts that clauses in short sentences are composed of relatively many LDSs, which means that there must be relatively many dependency distances with values more than one. The findings from Ferrer-i-Cancho and Gómez-Rodríguez (2021) and from this paper thus support each other.

[14] The interpretation of parameter a of the MAL in form (2) as the mean length of constituents of the shortest constructs is not specific to language units analysed in this paper – e.g. Kelih (2010) uses the same approach when investigating the relations between lengths of words and syllables.

[15] Miller (1956) claims that the capacity is roughly seven (although there are also other opinions). Clause length determined in the number of the LDSs only rarely exceeds this value, while clause length in words can be, naturally, (much) higher. Similarly, phrases used by Mačutek et al. (2017) contain more words than LDSs; in addition, the methodology from that paper allows to analyse only main clauses.

Acknowledgements

J Mačutek was supported by research grant VEGA 2/0096/21.

References

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn, editor, *Glottometrika 2*, pages 1–10. Brockmeyer, Bochum.

Jan Andres and Martina Benešová. 2012. Fractal Analysis of Poe's Raven, II. *Journal of Quantitative Linguistics*, 19(4):301–324.

Martina Benešová and Radek Čech. 2015. Menzerath-Altmann law versus random model. In George K. Mikros and Ján Mačutek, editors, *Sequences in Language and Text*, pages 57–69. de Gruyter, Berlin / New York.

Solomija Buk and Andrij Rovenchak. 2008. Menzerath–Altmann law for syntactic structures in Ukrainian. *Glottotheory*, 1(1):10–17.

Heng Chen and Haitao Liu. 2019. A quantitative probe into the hierarchical structure of written Chinese. In Xinying Chen and Ramon Ferrer-i-Cancho, editors, *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 83–88. ACL, Stroudsburg (PA).

Irene M. Cramer. 2005. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, editors, *Quantitative Linguistics. An International Handbook*, pages 659–688. de Gruyter, Berlin / New York.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. (2021).Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez, 2021. Anti dependency distance minimization in short sequences. A graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1):50–76.

Ramon Ferrer-i-Cancho and Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5(2):143–355.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10336–10341.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies Workshop (UDW 2018)*, pages 66–74. ACL, Stroudsburg (PA).

Slavko Geršić and Gabriel Altmann. 1980. Laut – Silbe – Wort und das Menzerathsche Gesetz. In Hans-Walter Wodarz, editor, *Frankfurter phonetische Beiträge 3*, pages 115-123. Buske, Hamburg.

Peter Grzybek. 2006. History and methodology of word length studies. In: Peter Grzybek, editor, *Contributions to the Science of Language and Text. Word Length Studies and Related Issues*, pages 15–90. Dordrecht, Springer.

Peter Grzybek. 2013. Close and distant relatives of the sentence: Some results from Russian. In: Ivan Obradović, Emmerich Kelih, and Reinhard Köhler, editors, *Methods and Applications of Quantitative Linguistics*, pages 59-68. Beograd: Akademska Misao.

Gabriela Heups. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In Reinhard Köhler and Joachim Boy, editors, *Glottometrika 5*, pages 113-133. Brockmeyer, Bochum.

Renkui Hou, Chu-Ren Huang, Hue San Do, and Hongchao Liu. 2017. A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 24(4):350–366.

Renkui Hou, Chu-Ren Huang, Mi Zhou, and Menghan Jiang. 2019. Distance between Chinese registers based on the Menzerath-Altmann law and regression analysis. *Glottometrics*, 45:24–57.

Tomáš Jelínek 2017. FicTree: A manually annotated treebank of Czech fiction. In Jaroslava Hlaváčová, editor, *Proceedings of the 17th Conference on Information Technologies – Applications and Theory (ITAP 2017)*, pages 181–185. http://ceur-ws.org/Vol-1885/181.pdf

Emmerich Kelih. 2010. Parameter interpretation of Menzerath law: Evidence from Serbian. In Peter Grzybek, Emmerich Kelih, and Ján Mačutek, editors, *Text and language. Structures, functions, interrelations, quantitative perspectives*, pages 71–79. Praesens, Wien.

Emmerich Kelih. 2012. Systematic interrelations between grapheme frequencies and words length: Empirical evidence from Slovene. *Journal of Quantitative Linguistics*,19(3):205–231.

Emmerich Kelih, Peter Grzybek, Gordana Antić, and Ernst Stadlober. 2006. Quantitative text typology. The impact of sentence length. In Myra Spiliopoulou, Rudolf Kruse, Christian Borgelt, Andreas Nürnberger, and Wolfgang Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, pages382–389. Springer, Heidelberg / Berlin.

Reinhard Köhler. 1982. Das Menzeratsche Gesetz auf Satzebene. In Werner Lehfeldt and Udo Strauss, editors, *Glottometrika 4*, pages 103–113. Brockmeyer, Bochum.

Reinhard Köhler. 1989. Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In Gabriel Altmann and Michael H. Schwibbe, editors, *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, pages 108–112. Olms, Hildesheim / Zürich / New York.

Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. de Gruyter, Berlin / Boston.

Agnieszka Kułacka. 2010. The coefficients in the formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 17(4):23–32.

Agnieszka Kułacka and Ján Mačutek. 2007. A discrete formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 14(1):257–268.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Ján Mačutek, Jan Chromý, and Michaela Koščová. 2019. Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics*, 26(1):66–80.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni and Joakim Nivre, editors, *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100-107. Linköping University Electronic Press: Linköping.

Ján Mačutek and Andrij Rovenchak. 2011. Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In Emmerich Kelih, Victor Levickij, and Yuliya Matskulyak, editors, *Issues in Quantitative Linguistics 2*, pages 136-147. RAM-Verlag, Lüdenscheid.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3):227–240.

Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*. Dümmler, Bonn.

Paul Menzerath and José M. de Oleza. 1928. *Spanische Lautdauer. Eine experimentelle Untersuchung.* de Gruyter, Berlin / Leipzig.

George Mikros and Jiří Milička. 2014. Distribution of the Menzerath's law on the syllable level in Greek texts. In Gabriel Altmann, Radek Čech, Ján Mačutek, and Ludmila Uhlířová, editors, *Empirical Approaches to Text and Language Analysis*, pages 181–189. RAM-Verlag, Lüdenscheid,

George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Timothy Osborne and Kim Gerdes. 2019. The status of function words independency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, 4(1):17.

Kateřina Pelegrinová, Ján Mačutek, and Radek Čech. 2021. The Menzerath-Altmann law as the relation between lengths of words and morphemes in Czech. *Jazykovedný časopis*, 72 (to appear).

Ioan-Iovitz Popescu, Gabriel Altmann, Petr Grzybek, Bijapur D. Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmila Uhlířová, and Matummal N. Vidya. 2009. *Word Frequency Studies*. de Gruyter, Berlin / New York.

Michael H. Schwibbe. 1984. Text- und wortstatistische Untersuchungen zur Validität der Menzerath'schen Regel. In Joachim Boy and Reinhard Köhler, editors, *Glottometrika 6*, pages 127–138. Brockmeyer, Bochum.

Regina Teupenhayn and Gabriel Altmann. 1984. Clause length and Menzerath's law. In Joachim Boy and Reinhard Köhler, editors, *Glottometrika 6*, pages 152–176. Brockmeyer, Bochum.

Iván G. Torre, Bartolo Luque, Lucas Lacasa, Christopher T. Kello, and Antoni Hernández-Fernández. 2019. On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, 6:191023.

Lirong Xu and Lianzhen He. 2020. Is the Menzerath-Altmann law specific to certain languages in certain registers? *Journal of Quantitative Linguistics*, 27(3):187–203.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444_–446.

Victor H. Yngve. 1996. *From Grammar to Science. New Foundations for General Linguistics*. Benjamins, Amsterdam / Philadelphia.

The Linear Arrangement Library. A new tool for research on syntactic dependency structures.

Lluís Alemany-Puig lluis.alemany.puig@upc.edu
Juan Luis Esteban esteban@cs.upc.edu
Ramon Ferrer-i-Cancho ramon.ferrer@upc.edu
Universitat Politècnica de Catalunya
Jordi Girona 1-3
08034 Barcelona, Catalonia, Spain

Abstract

The new and growing field of Quantitative Dependency Syntax has emerged at the crossroads between Dependency Syntax and Quantitative Linguistics. One of the main concerns in this field is the statistical patterns of syntactic dependency structures. These structures, grouped in treebanks, are the source for statistical analyses in these and related areas; dozens of scores devised over the years are the tools of a new industry to search for patterns and perform other sorts of analyses. The plethora of such metrics and their increasing complexity require sharing the source code of the programs used to perform such analyses. However, such code is not often shared with the scientific community or is tested following unknown standards. Here we present a new open-source tool, the Linear Arrangement Library (LAL), which caters to the needs of, especially, inexperienced programmers. This tool enables the calculation of these metrics on single syntactic dependency structures, treebanks, and collection of treebanks, grounded on ease of use and yet with great flexibility. LAL has been designed to be efficient, easy to use (while satisfying the needs of all levels of programming expertise), reliable (thanks to thorough testing), and to unite research from different traditions, geographic areas, and research fields.

1 Introduction

Quantitative Linguistics is a discipline within Linguistics that aims to unveil linguistic laws and explain their origins (Köhler and Altmann, 2012; Best and Rottmann, 2017). Outstanding examples of these are Zipfian laws, e.g., Zipf's rank-frequency law, Zipf's law of abbreviation (Zipf, 1949), that are defined typically on one of languages' basic units: words. Another discipline in Linguistics is Dependency Syntax, a framework which primarily reduces the syntactic structure of a sentence to word pairwise dependencies. Each of these dependencies has a 'head' word and a 'dependent' word (in fields like Computer Science, these could be called 'parent' and 'child', respectively; the 'head' is also known as 'governor'). The collection of such dependencies in a sentence combined with the linear ordering of the words yields the so-called *syntactic dependency structure* (Mel'čuk, 1988; Kuhlmann and Nivre, 2006; Nivre, 2006; Gómez-Rodríguez et al., 2011) as in Figure 1. Therefore, the underlying structure of a syntactic dependency structure can be seen as a rooted tree (as in Figure 2(b)).

The combination of Quantitative Linguistics with Dependency Syntax has resulted into the emerging field of Quantitative Dependency Syntax `https://quasy-2019.webnode.com/`. The target of this field are syntactic dependency structures and aims to discover and understand statistical patterns in these structures. By linearizing the hierarchical structure "arises the concept of dependency distance or dependency length" (Liu et al., 2017), defined usually as the number of intervening words between the endpoints of the dependency plus one (Ferrer-i-Cancho, 2004) as in Figure 2(a). Another relevant concept is that of *syntactic dependency crossing* (Mel'čuk, 1988). Figure 1 shows two examples of syntactic crossings: two syntactic dependencies cross when the positions of their head and dependent words interleave. Said concept is used to define many formal constraints, such as projective and planar structures (Kuhlmann and Nivre, 2006) and 1-Endpoint-Crossing structures (Satta et al., 2013). See Gómez-Rodríguez et al. (2011) for a review.

Research in Cognitive Science has shown a tendency for languages to reduce dependency distances (Ferrer-i-Cancho, 2004; Liu, 2008; Futrell et al., 2015; Futrell et al., 2020; Ferrer-i-Cancho et al., 2021).

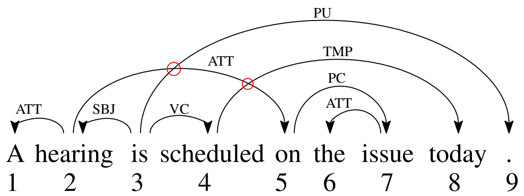


Figure 1: An example of a sentence and the syntactic dependencies among its words (adapted from Pighin (2012)). Relations are labeled with their grammatical category. Numbers below the sentence indicate the positions of the words. In this figure we see two syntactic crossings, marked with small red circles.

According to Liu et al. (2017), Hudson (1995) gave the first definition of dependency distance and presented a cognitive formulation of "the memory burden imposed by dependency distance on language processing". This tendency results from the action of a Dependency Distance Minimization (DDm) principle (Ferrer-i-Cancho, 2003; Ferrer-i-Cancho, 2004), supported by many models and theories (Liu et al., 2017; Temperley and Gildea, 2018) stemming from the more general Principle of Least Effort (Zipf, 1949), hence largely regarded as a linguistic universal. The statistical support for DDm comes from baselines that are used to perform statistical tests on the significance of dependency distances (Ferrer-i-Cancho, 2004; Gildea and Temperley, 2007; Liu, 2008; Park and Levy, 2009; Gildea and Temperley, 2010; Futrell et al., 2015; Yu et al., 2019; Ferrer-i-Cancho et al., 2021). Some of these baselines are defined on extreme conditions (e.g., maximum and minimum sum of dependency distances) which further motivates the study of extremal problems in Computer Science like the Minimum Linear Arrangement problem (Garey and Johnson, 1976; Goldberg and Klipker, 1976; Shiloach, 1979; Chung, 1984) and the Maximum Linear Arrangement Problem (Hassin and Rubinstein, 2000; DeVos and Nurse, 2018) and their variants under formal constraints. Other baselines are defined on 'uniformly random' conditions, typically in uniformly random permutations of the words of a sentence. However, formal constraints from Dependency Grammar, like projectivity and planarity (Sleator and Temperley, 1993; Kuhlmann and Nivre, 2006), have led to defining such random baselines conditioned to those formal constraints (Gildea and Temperley, 2007; Park and Levy, 2009; Futrell et al., 2015; Kramer, 2021; Alemany-Puig and Ferrer-i-Cancho, 2021) although these formal constraints have been argued to be epiphenomena of DDm (Gómez-Rodríguez and Ferrer-i-Cancho, 2017; Gómez-Rodríguez et al., 2020).

In this article we introduce a new tool to support research on the areas and the research problems reviewed above: the Linear Arrangement Library (LAL), which allows researchers to compute easily many of the metrics and algorithms that researchers have been proposing, while simplifying significantly the problem of calculating random or extremal baselines for them. In addition, LAL aims to simplify the process of working with collections of treebanks, one of the most successful recent examples being the Universal Dependencies collection (Zeman et al., 2020) and its variants (Gerdes et al., 2018). LAL is currently available from `https://cqllab.upc.edu/lal`.

In order to grasp the power of LAL we remind the reader that the syntactic dependency structure of a sentence can be defined as a triad composed of (1) a *directed graph structure* in which the vertices of the graph are the words of the sentence, (2) a *linear arrangement* of the vertices of the graph, and (3) *labels* of the edges of the graph which indicate the type of syntactic relationship between the words they relate. Two types of metrics (or scores) can be defined on such structures: word order-dependent, e.g., the sum of dependency distances (Gildea and Temperley, 2007) and word order-independent metrics, e.g., mean hierarchical distance (Jing and Liu, 2015). LAL allows one to compute many scores of each of these two sorts.

The calculation of baselines is at the heart of Quantitative Dependency Syntax research as well as at the heart of LAL. For some random baselines, LAL offers exact algorithms or formulae to calculate the desired value under the null model, e.g., algorithms to calculate the expected sum of dependency distances (Ferrer-i-Cancho, 2004; Alemany-Puig and Ferrer-i-Cancho, 2021). The reality is, unfortunately, that algorithms and formulae to calculate exact expected values of certain scores might be difficult to derive.

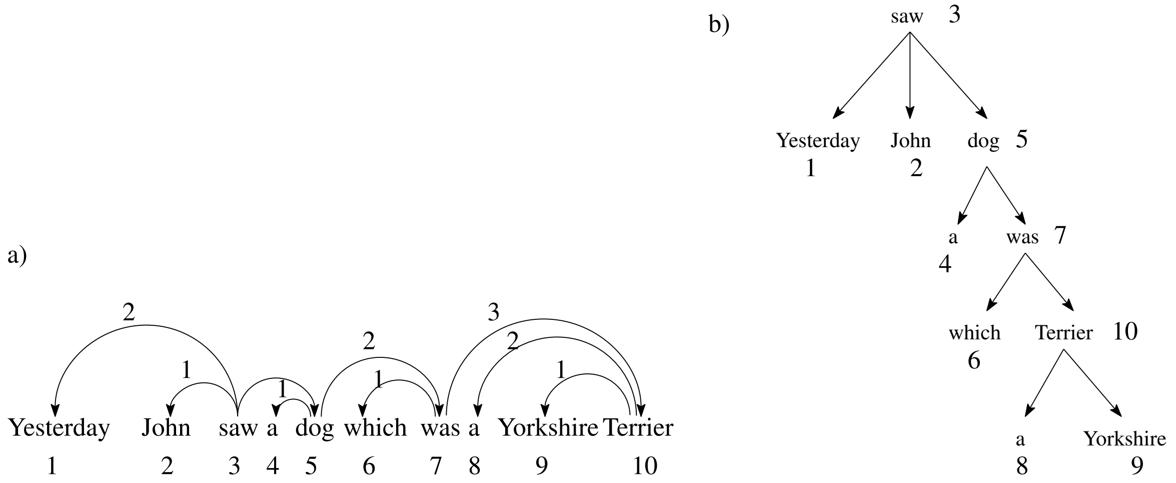


Figure 2: a) An example of syntactic dependency structure. Arc labels indicate edge lengths, each calculated as the absolute difference of the positions of the corresponding edge's endpoints. The numbers below the words indicate positions. b) The rooted tree underlying the sentence in a); the positions of the words are indicated below or to the right of each word. Adapted from (McDonald et al., 2005, Figure 2).

As an alternative, LAL allows researchers to resort to random sampling in order to calculate said values, which often involves tree and/or linear arrangement generation (Liu, 2008; Esteban and Ferrer-i-Cancho, 2017; Yadav et al., 2019). An example of an application of tree generation would be the calculation of the expected mean hierarchical distance (Jing and Liu, 2015) among n-vertex rooted trees[1]. Regarding linear arrangement generation, an example would be to calculate the expected flux weight (Kahane et al., 2017) over all arrangements of a tree.

Thanks to LAL the computation of random baselines can be restricted easily to two of the most frequently observed arrangements from a formal standpoint: planar orderings, where syntactic edges do not cross, and projective orders, namely planar orderings where the root is not covered (Sleator and Temperley, 1993; Kuhlmann and Nivre, 2006). For instance, the random baseline over the sum of dependency distances can be computed assuming unconstrained, projective and planar arrangements as shown in Table 1. In an unconstrained linear arrangement, edge crossings are allowed and the root may be covered.

The remainder of the article is organized as follows. Section 2 presents the design principles of LAL and its architecture. Section 3 gives further details about its functionalities and explains how to work with LAL following a standard research pipeline. We end with some suggestions for future development of LAL.

2 Design principles

2.1 Ease of use

Many measures/scores in Quantitative Dependency Syntax have been devised over the years (e.g., Jiang and Liu (2018)). Some of these are easy to calculate, e.g., the sum of dependency distances (or the sum of edge lengths) of an n-vertex syntactic dependency structure. This sum is easily computable in $O(n)$ time given the specification of the linear arrangement and that of the graph. However, the calculation of extremal values on linear arrangements (i.e. minimum or maximum values of a score), such as the solution to the Minimum Linear Arrangement (MLA) problem in unconstrained arrangements (Garey and Johnson, 1976; Shiloach, 1979; Chung, 1984) or of one of its constrained variants (Iordanskii, 1987; Hochberg and Stallmann, 2003; Gildea and Temperley, 2007; Bommasani, 2020; Alemany-Puig et al., 2022) are not straightforward because the algorithm is complex, it is hard to test or both. Likewise,

[1] This problem may be notoriously difficult to solve; take as a reference the work by Rényi and Szekeres (1967) where it is shown that the average labeled tree height H_n is such that $H_n \to \sqrt{2n\pi}$ as $n \to \infty$.

	D			C
	Unconstrained	Planar	Projective	Unconstrained
Minimum	Shiloach (1979), Chung (1984)	Hochberg and Stallmann (2003), Alemany-Puig et al. (2022)	Gildea and Temperley (2007), Alemany-Puig et al. (2022)	†
Complexity	$O(n^{2.2}), O(n^2)$	$O(n)$	$O(n)$	
Expected	Ferrer-i-Cancho (2004)	*	Alemany-Puig and Ferrer-i-Cancho (2021)	Verbitsky (2008)
Complexity	$O(1)$	$O(n)$	$O(n)$	$O(n)$
Maximum	Under study	In progress	In progress	Under study
Complexity		$O(n)$	$O(n)$	

Table 1: The baselines on D, the sum of dependency distances, and C, the number of syntactic dependency crossings, that can be calculated on a given input tree using LAL. Columns 'Unconstrained', 'Planar' and 'Projective' are the different constraints under which the 'Minimum', 'Expected' and 'Maximum' values can be calculated with LAL. *: available in LAL but article not published yet; †: the minimum value of C is trivially 0 for every tree.

performing statistical tests and calculating expected values (random baselines) require random sampling methods when exact algorithms/formulae are not known; such sampling is typically done uniformly at random over all possible trees (Ferrer-i-Cancho et al., 2018; Gómez-Rodríguez et al., 2020; Yadav et al., 2019) or random arrangements (Ferrer-i-Cancho et al., 2018; Ferrer-i-Cancho et al., 2021). LAL's main design principle is to make these algorithms (and random sampling methods) easily accessible and, therefore, LAL has been designed to be an easy-to-use tool for Quantitative Dependency Syntax researchers focused in the analysis of syntactic dependency trees, and Computer Scientists and Mathematicians specializing in Discrete Mathematics. Moreover, advanced programming knowledge is not required to use LAL. For example, a researcher in Quantitative Dependency Syntax can process a treebank as easily as shown in Code 1.

```python
import lal
err = lal.io.process_treebank("Cantonese.txt", "output_file.csv")
print(err)
```

Code 1: Python script for computing all scores on a single treebank with LAL. The input file is a series of head vectors, whose format is described in Section 3, and the output is a standard .csv file that can be loaded directly on a spreadsheet.

2.2 Connecting communities and traditions

LAL aims to unite research traditions and disciplines from all over the world as well as serving the distinct fields converging into Quantitative Dependency Syntax as much as possible. LAL integrates views, concepts and scores from the major research communities: Asia (China, Jing and Liu (2015); Japan, Komori et al. (2019)), North America (USA, Gildea and Temperley (2007) and Futrell et al. (2015)) and Europe (e.g., France, Kahane et al. (2017); Switzerland, Gulordava and Merlo (2016) and Gulordava and Merlo (2015)).

2.3 Openness and availability

After many years of research in Quantitative Dependency Syntax, the code used to calculate many of these metrics is not usually shared, thus forcing researchers in the Linguistics fields to repeat the same efforts as the original authors put into coding the algorithms. This repetition increases the probability of bugs in every researcher's code which lead to incorrect results used in their experiments to be published in scientific journals. We think that LAL is an answer to these challenges as it is an open-source project, licensed under the *GPL v3 Affero*[2]. The library's code is publicly available at `https://github.com/LAL-project/linear-arrangement-library.git`.

2.4 Robustness

Often, testing is not thorough or is not specified. Therefore, the increasing amount of research in Quantitative Dependency Syntax creates a need for a thoroughly-tested tool in which to find many if not most of the metrics devised so far. The continual testing that is applied to LAL solves this problem: tests are run periodically to ensure that all algorithms calculate correct values.

2.5 The architecture of LAL

The LAL project's architecture consists of a *core* and *extensions* (Figure 3). The core has two parts: the main branch and the testing branch (Figure 3). The latter is responsible for the robustness of the main branch. The test branch is not publicly available yet but it results from the transfer of knowledge and methods that enabled to test and correct classic algorithms with random and exhaustive methods (Esteban and Ferrer-i-Cancho, 2017; Alemany-Puig and Ferrer-i-Cancho, 2020; Alemany-Puig et al., 2022). The C++ language is used in LAL's core to implement its main functionalities and algorithms, some of which are parallelized internally to improve performance (a critical example is the function that computes (all) scores on a treebank collection). However, only the library branch is wrapped to Python (via SWIG (2020); Figure 3) given the fact that Python is usually the first choice of many scientists who are looking forward to automatizing their workflow as it is easy to use.

LAL extensions implement additional functionalities, e.g., dealing with the interface between existing treebanks and LAL. 'LAL extensions' is the place for contributing to the LAL ecosystem without knowledge on how LAL is implemented. A concrete LAL extension is introduced in the next section.

LAL's core is composed of 7 modules (Figure 3). Now follows a brief description of each of them.

- The *generate* module contains algorithms to generate trees uniformly at random or exhaustively, labeled or unlabeled, and free or rooted; algorithms to generate arrangements of trees under the projectivity or planarity constraint,
- In the *graphs* module users will find the implementations of different graph classes: undirected and directed graphs, free and rooted trees,
- The *io* (input/output) module contains algorithms to read data from a file in disk, but its current chief goal is to process input data, such as treebanks, to produce an output file with measures computable by the library,
- The *numeric* module contains wrappers of the GMP library (GMP, 2021) for arbitrary-precision integer and rational numbers to ensure exact precision in the calculations (mostly useful for the testing branch of the project or advanced research in mathematics),
- In the *linarr* module users will find many algorithms to compute measures of graphs that are defined on the linear ordering of the vertices, e.g., the sum of dependency distances (Gildea and Temperley, 2007),

[2] Many people think this license formally discourages commercial usage, but it certainly does not `https://www.gnu.org/licenses/agpl-3.0.en.html`.

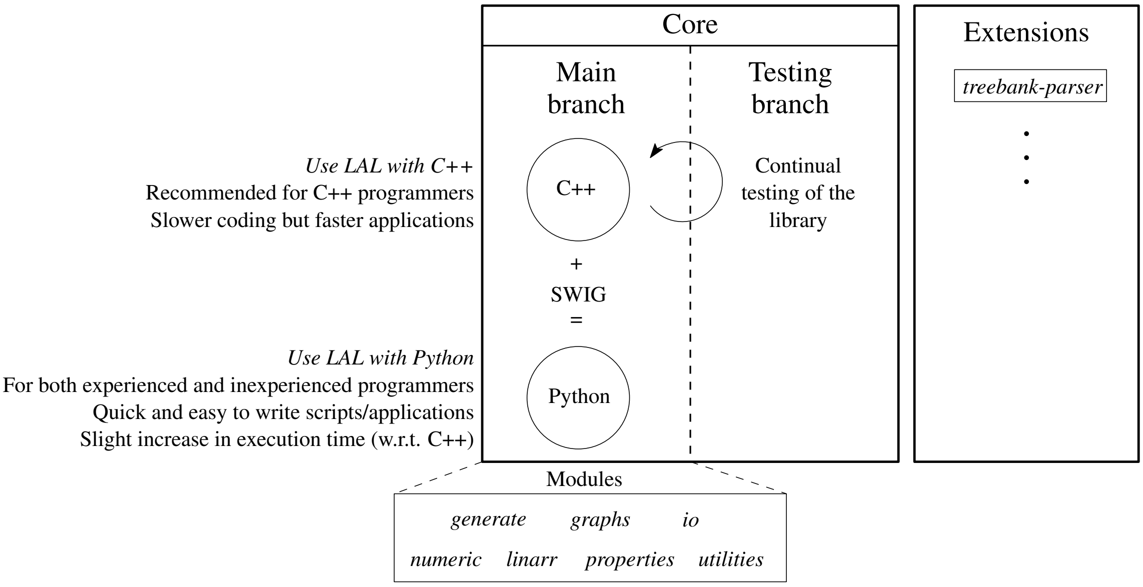

Core
Main branch
Testing branch
Extensions
treebank-parser
Use LAL with C++
Recommended for C++ programmers
Slower coding but faster applications
C++
Continual testing of the library
+
SWIG
=
Use LAL with Python
For both experienced and inexperienced programmers
Quick and easy to write scripts/applications
Slight increase in execution time (w.r.t. C++)
Python
Modules
generate
graphs
io
numeric
linarr
properties
utilities

Figure 3: The architecture of LAL project: *core* and *extensions*. The core consists of the library and testing branch. The library is developed in C++, and wrapped into Python via SWIG (2020). Testing the library is a crucial part of its development. The library is composed of 7 modules (both C++ and Python options): *generate*, *graphs*, *io*, *numeric*, *linarr*, *properties*, and *utilities*, each of which is briefly described in Section 2.

- The *properties* module implements algorithms to compute measures of graphs that do not depend on the linear ordering of their vertices, only on their structure, e.g., the Mean Hierarchical Distance (Jing and Liu, 2015),

- Finally, the *utilities* module contains algorithms that are interesting to the general public but cannot be classified into the categories above, such as the tree isomorphism test that tells whether or not two trees are the same. This test only takes into account the structure of the tree (nodes and edges) and not possible labels on the edges (e.g. grammatical relations between words in a sentence) or possible tokens in the nodes (e.g. words from a sentence). The algorithm implemented in LAL (Aho et al., 1974) is an $O(n)$-time algorithm (where n is the number of vertices of the trees) adapted to the case when the two trees are free or rooted.

3 Working with LAL

LAL's capabilities range over three kinds of operations depending on the entity to which they are applied: operations on *individual* syntactic dependency structures (modules *graphs*, *linarr*, *properties*), operations on individual *treebanks* or on treebank *collections* (*io* module).

LAL contains many functions with which one can calculate metrics/scores on a single tree. At present, the focus of LAL is on trees for simplicity and because of the high current interest on this kind of graphs. However, although fewer in number, LAL also contains functions that admit general graphs (e.g., graphs with cycles, and forests). Functions might depend on a given word ordering (*linarr* module) or not (*properties* module). Most functions in the *linarr* module typically require a linear arrangement as an input parameter. However, some functions return a linear arrangement rather than requiring one as an input parameter, as in the example given in Code 2. It is worth mentioning that LAL implements methods for a flexible construction of graphs (*graphs* module) by adding edges one by one, or in bulk, thus allowing complicated workflows, whenever those are needed. Nevertheless, LAL is also equipped with helper functions for simpler graph construction (e.g., construct a graph directly from its set of edges). Furthermore, it provides functions for reading graphs from a file in its *io* module.

```
import lal
t = lal.graphs.from_edge_list_to_free_tree([(0,1), ...])
Shiloachs_algorithm = lal.linarr.algorithms_Dmin.Shiloach
print(lal.linarr.min_sum_edge_lengths(t), Shiloachs_algorithm)
Chungs_algorithm = lal.linarr.algorithms_Dmin.Chung_2
print(lal.linarr.min_sum_edge_lengths(t), Chungs_algorithm)
```

Code 2: Python script to calculate the minimum baseline for the sum of dependency distances on a free tree with LAL applying two different algorithms: Shiloach's (Shiloach, 1979; Esteban and Ferrer-i-Cancho, 2017) and Chung's quadratic algorithm (Chung, 1984). LAL implements the correction of Shiloach's algorithm in (Esteban and Ferrer-i-Cancho, 2017). The free tree is specified as a list of pairs of edges. That baseline is the solution of the so-called minimum linear arrangement problem of computer science, hence the names of the algorithms mentioned above.

The library also implements treebank processing (*io* module). It provides its users with algorithms to generate data out of single treebank files and collections of treebanks. A collection of treebanks is simply a set of treebanks that are related to one another within some particular context. In the UD collection, texts from distinct languages annotated with the same formalism (Zeman et al., 2020), but there are many other possibilities: e.g., the novels of the same writer – where an individual treebank is a novel, all the articles in a given newspaper – where a single treebank is one of said articles.

In LAL a collection is represented by a plain text file listing all the treebank files in the collection. Treebanks are typically written in CoNLL-U format (Buchholz and Marsi, 2006), but LAL requires a simpler format with the essential information. In particular, LAL requires treebanks to be provided as a series of *head vectors* (or *head sequences*); a head vector of an n-word sentence is a sequence of n non-negative integer numbers in the positions from 1 to n where each number indicates the position of its parent word, and the number 0 indicates the root word of the sentence. The most important reason to use the intermediate format is the existence of many previous formats, and the possibility of many new formats coming to life; we believe it will help in reducing the library's usage complexity. This approach is similar to that taken by compiler developers and designers who use the so-called 'Three address code': it is an intermediate language into which many programming languages can be transformed, and is then compiled to the target machine. Nevertheless, we have also developed a LAL extension (Figure 3), the *treebank-parser* that applies core LAL functions to convert CoNLL-U-formatted files into the head vector format after applying optional preprocessing: (a) removal of punctuation marks, e.g., for crosslinguistic generality, (b) removal of functions words, e.g., to compare languages with many such words against languages where these are scarce and (c) removal of sentences shorter or longer than a given length (Ferrer-i-Cancho et al., 2021). This tool can be found online at `https://github.com/LAL-project/treebank-parser.git`.

The processing of treebanks (and collection of treebanks) can be done automatically by LAL, but it can also be customized by its users. In other words, the automatic processing of a treebank (or a collection of treebanks) is performed by the library applying the metrics/scores selected by the user, with optional internal parallelization, and other customizable options on the format of the output. One obvious drawback of automatic processing is that it is limited to the metrics/scores implemented in LAL. Nevertheless, should users look forward to calculating a new score on a treebank (or in a collection), they can still use LAL to carry out such a task since LAL implements algorithms to easily iterate over a file containing only head vectors, thus removing the aforementioned limitation. Therefore, LAL facilitates working on a single treebank or on a collection of treebanks.

Figure 4 illustrates one possible pipeline when working with a single treebank. That pipeline can easily be adapted to any treebank collection. A description now follows.

Phase 1 Firstly, one chooses a source of the input data. Such data may come from a handwritten text, or a typeset text, which is to be analyzed by either a computer or by a human; the data may come from already-analyzed data such as the UD collection (Zeman et al., 2020; Gerdes et al., 2018), or treebanks annotated with the Stanford (de Marneffe et al., 2014) or Prague (Hajič et al., 2006) conventions.

Phase 2 Secondly, some preprocessing of the data might be required, e.g., removal of punctuation marks, removal of function words, or the exclusion of sentences that are too short or too long. See Ferrer-i-Cancho et al. (2021) for a complete example of such preprocessing. We call the resulting treebank ‘Treebank (2)’.

Phase 3 Then, one transforms ‘Treebank (2)’ into ‘Treebank (3)’, a file containing only the so-called head vectors so that LAL can understand the data.

Phase 4 Here, LAL is used to generate a `.csv` file containing all the measures that are interesting for one’s own research. Staying true to its efficiency design principle, LAL can evaluate all metrics it implements on every tree of the UD 2.6 treebank (Zeman et al., 2020) in ~ 23 seconds, while producing only the primary data (on only the UD 2.6 treebank) of a recent article (Ferrer-i-Cancho et al., 2021) takes ~ 6 seconds[3].

Phase 5 The last phase consists of using the data to perform statistical analyses on the treebank. These analyses comprise hypothesis and theoretical prediction testing (Ferrer-i-Cancho and Gómez-Rodríguez, 2021) as well as evaluation of the quality of a treebank (Alzetta et al., 2017; Heinecke, 2019). When the pipeline is adapted to a treebank collection, LAL supports research in typology (Croft et al., 2017; Alzetta et al., 2018) or on comparisons of annotation schemes (Osborne and Gerdes, 2019; Passarotti, 2016).

LAL extensions can be used for tasks in Phases 2 and 3. For the time being, the aforementioned extension *treebank-parser* (Figure 3) allows one to perform Phase 2 and Phase 3 in a row over treebanks in CoNLL-U format. In the future, we hope that LAL extensions grow in number with contributions that researchers wish to make to the LAL ecosystem for Phase 2 or Phase 3.

3.1 Capabilities

To begin with, LAL implements several algorithms to calculate relevant word order-dependent metrics (module *linarr*) in Quantitative Dependency Syntax. These include the sum of dependency distances (Gildea and Temperley, 2007); the number of edge crossings (Ferrer-i-Cancho et al., 2018), with the option to choose among several algorithms; the prediction of the number of crossings based on the length of the edges (Ferrer-i-Cancho, 2014); the class of syntactic dependency structure (such as WG_1 (Gómez-Rodríguez et al., 2011), 1-Endpoint Crossing (Satta et al., 2013), projective and planar (Sleator and Temperley, 1993; Kuhlmann and Nivre, 2006)); the solution to the MLA problem under several constraints (Table 1); the computation of dependency fluxes (Kahane et al., 2017); the proportion of head initial dependencies (Liu, 2010).

Researchers will find in LAL many word order-independent metrics (metrics that do not depend on the ordering of the vertices) in the *properties* module, including the Mean Hierarchical Distance (Jing and Liu, 2015); the variance and expected number of crossings in unconstrained arrangements of trees, useful for variable standardization (Alemany-Puig and Ferrer-i-Cancho, 2020); the expected sum of edge lengths under three different formal constraints (Table 1) to cover different views about the nature of formal constraints (Yadav et al., 2021), and the variance of said sum in unconstrained linear arrangements (Ferrer-i-Cancho, 2019); the calculation of the m-th moment of degrees about zero used to calculate the well-known hubiness coefficient (Ferrer-i-Cancho et al., 2018), extended to the out- and in-degrees[4]; the number of pairs of independent edges (two edges are independent when they share no vertices); the

[3] Running times are estimated on a brand-new PC with a 3.30 GHz, 6-core (2 threads/core) *i5-10600* CPU (16 GB); the program uses 6 threads and runs on Ubuntu 20.04.

[4] Trivially, the m-th moment of in-degree about zero of any n-vertex tree is $\langle k_{in}^m \rangle = (n-1)/n$. Nevertheless, a function to calculate it is provided.

Sampling method	Type of tree		References
Exhaustive	Labeled	Free	Prüfer (1918)
		Rooted	Based on E-L-F*
	Unlabeled	Free	Wright et al. (1986)
		Rooted	Beyer and Hedetniemi (1980)
Random	Labeled	Free	Prüfer (1918)
		Rooted	Based on Rn-L-F**
	Unlabeled	Free	Wilf (1981)***
		Rooted	Nijenhuis and Wilf (1978)

Table 2: The eight different possibilities of tree generation in LAL. By 'random' we mean 'uniformly random over the complete set of the respective kind of trees'. * Based on Exhaustive-Labeled-Free tree generation. ** Based on Random-Labeled-Free tree generation. *** The algorithm includes the correction pointed out in (Marohnić, 2018).

central and centroidal vertices of trees (Harary, 1969). Furthermore, one can classify trees into classes according to their structure. These classes are: *linear*, *star*, *quasistar* and *bistar* (San Diego and Gella, 2014), *caterpillar* (Harary and Schwenk, 1973), and *spider* (Bennett et al., 2019) trees.

In order to overcome the research limitations arising from the lack of research on every possible expected value that can be elicited, LAL is equipped with several algorithms for the generation of trees (Table 2) and of linear arrangements of trees under several formal constraints (projectivity and planarity[5] (Kuhlmann and Nivre, 2006)), and, for the sake of ease of use, it also provides exhaustive and random generation of unconstrained arrangements, i.e., exhaustive and random generation of permutations. Random generation allows one to estimate, via random sampling of the appropriate structure, the expected value of a certain existing or new metric. For example, one could easily approximate the expected Mean Hierarchical Distance (MHD) over the set of uniformly random unlabeled rooted trees by sampling said trees and averaging their MHD. The same can be said about those metrics dependent on word order. It goes without saying that said approximation is not limited to what LAL can calculate: researchers with sufficient programming skills can implement their own metrics either in C++ or in Python and approximate that metric's expectation and other aspects of its distribution under null models; the online documentation at `https://cqllab.upc.edu/lal/guides` explains how these estimates can be calculated and provide examples that inexperienced programmers can easily adapt to the metric of their choice. In previous research, the methods of generation of random trees do not warrant uniform sampling of the space of possible trees (Liu, 2007; Courtin and Yan, 2019). LAL simplifies research where uniformity is required.

3.2 Applications of LAL

LAL's domain of application revolves primarily around studies on Quantitative Dependency Syntax. We have dealt with various applications of LAL when describing Phase 5 of the pipeline (Section 3, Figure 4). LAL is also a convenient tool for reproducing results from previous research. For instance, LAL allows one to reproduce the results of the analyses to predict the actual number of dependency crossings (Gómez-Rodríguez and Ferrer-i-Cancho, 2017), the results on the scaling of sum of dependency distances in minimum linear arrangements (Esteban et al., 2016), or recent findings on the degree of optimality of languages (Ferrer-i-Cancho et al., 2021). This research eventually converged into the formal development of LAL.

LAL offers many possibilities for research on the similarity among trees. For instance, LAL can be used to find how many isomorphic tree structures are present in two treebanks. Also, given any treebank, one can also find the amount of unique trees (up to graph isomorphism). Furthermore, rooted trees are

[5] Random and exhaustive generation of planar arrangements is available in LAL but not published at the time of submission.

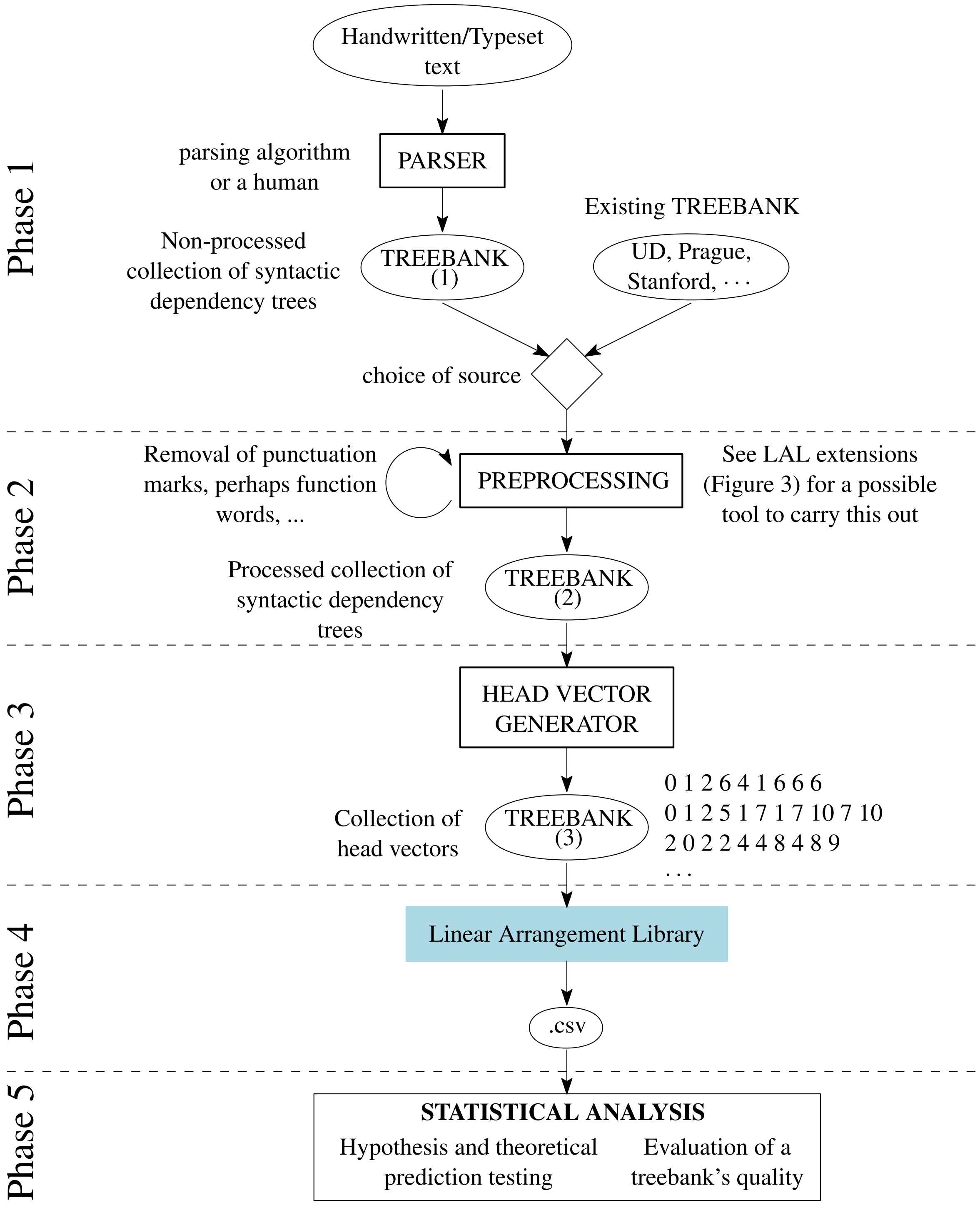

Phase 1
Handwritten/Typeset text
parsing algorithm or a human
PARSER
Existing TREEBANK
Non-processed collection of syntactic dependency trees
TREEBANK (1)
UD, Prague, Stanford, · · ·
choice of source
Phase 2
Removal of punctuation marks, perhaps function words, ...
PREPROCESSING
See LAL extensions (Figure 3) for a possible tool to carry this out
Processed collection of syntactic dependency trees
TREEBANK (2)
Phase 3
HEAD VECTOR GENERATOR
Collection of head vectors
TREEBANK (3)
0 1 2 6 4 1 6 6 6
0 1 2 5 1 7 1 7 10 7 10
2 0 2 2 4 4 8 4 8 9
. . .
Phase 4
Linear Arrangement Library
.csv
Phase 5
STATISTICAL ANALYSIS
Hypothesis and theoretical prediction testing
Evaluation of a treebank's quality

Figure 4: A standard pipeline for research on a single treebank. The pipeline comprises five phases that start with the choice of an existing treebank or producing it from raw text (Phase 1) and ends with analyses of the output produced by LAL (Phase 5). In Phase 4, LAL receives a treebank that has been preprocessed and transformed into a format that LAL can digest.

implemented so that users can extract rooted subtrees and perform similarity tests based on subgraph isomorphism. These have already been tackled for general graphs (Cordella et al., 2004; Jüttner and Madarasi, 2018).

Beyond the realm of measurements on treebanks, LAL contains tree-generation (or linear-arrangement generation) methods that can help one to test new algorithms. With this one can easily implement the testing protocol to assert the correctness of the implementation of the algorithm to compute the minimum sum of dependency distances (Esteban et al., 2016; Esteban and Ferrer-i-Cancho, 2017).

4 Future work

LAL is a growing project to support current and future research. We plan to extend the library with algorithms to calculate extreme or random baselines for scores for which a fast exact algorithm is not available yet. For instance, we plan to add efficient algorithms for the calculation of the maximum sum of edge lengths in unconstrained arrangements and also on said maximum under the projectivity and planarity constraint (Table 1). We will also work on the classification of linear arrangements into different classes of formal constraints and also extend the library with functionalities that ease common tasks in the analysis of syntactic dependency structures or that reflect consolidated results from the distinct disciplines involved. We are also open to update the library based on demands of engaged users.

Acknowledgements

We thank Aleksandra Petrova for helpful comments. LAP is supported by Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the Social European Fund. RFC and LAP are supported by the grant TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad). RFC is also supported by the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya). JLE is funded by the grant PID2019-109137GB-C22 from MINECO.

References

Alfred V. Aho, Jeffrey E. Hopcroft, and John D. Ullman. 1974. *The Design and Analysis of Computer Algorithms*. Addison-Wesley series in computer science and information processing. Addison-Wesley Publishing Company, Michigan University, 1st edition.

Lluís Alemany-Puig and Ramon Ferrer-i-Cancho. 2020. Edge crossings in random linear arrangements. *Journal of Statistichal Mechanics*, 2020:023403.

Lluís Alemany-Puig and Ramon Ferrer-i-Cancho. 2021. Linear-time calculation of the expected sum of edge lengths in projective linearizations of trees. *arXiv*.

Lluís Alemany-Puig, Juan Luis Esteban, and Ramon Ferrer-i-Cancho. 2022. Minimum projective linearizations of trees in linear time. *Information Processing Letters*, 174:106204.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 201–210, Prague, Czech Republic.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal Dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Patrick Bennett, Sean English, and Maria Talanda-Fisher. 2019. Weighted Turán problems with applications. *Discrete Mathematics*, 342:2165–2172, 8.

Karl-Heinz Best and Otto Rottmann. 2017. *Quantitative Linguistics, an Invitation*. RAM-Verlag, Lüdenscheid, Germany, 1 edition.

Terry Beyer and Sandra Mitchell Hedetniemi. 1980. Constant time generation of rooted trees. *SIAM Journal on Computing*, 9(4):706–712.

Rishi Bommasani. 2020. Generalized optimal linear orders. Master's thesis, Cornell University.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, 06. Association for Computational Linguistics.

Fan R. K. Chung. 1984. On optimal linear arrangements of trees. *Computers & Mathematics with Applications*, 10(1):43–60.

Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. 2004. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1367–1372.

Marine Courtin and Chunxiao Yan. 2019. What can we learn from natural and artificial dependency trees. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 125–135, Paris, France, 08. Association for Computational Linguistics.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets universal dependencies. In *TLT*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*.

Matt DeVos and Kathryn Nurse. 2018. A maximum linear arrangement problem on directed graphs. *arXiv*.

Juan Luis Esteban and Ramon Ferrer-i-Cancho. 2017. A correction on shiloach's algorithm for minimum linear arrangement of trees. *SIAM Journal on Computing*, 46(3):1146–1151.

Juan Luis Esteban, Ramon Ferrer-i-Cancho, and Carlos Gómez-Rodríguez. 2016. The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(6):063401, jun.

Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez. 2021. Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1):50–76.

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, and Juan Luis Esteban. 2018. Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, 493:311–329.

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, Juan Luis Esteban, and Lluís Alemany-Puig. 2021. The optimality of syntactic dependency distances. *Physical Review E*, page in press.

Ramon Ferrer-i-Cancho. 2003. *Language Universals: Principles and origins*. Ph.D. thesis, Universitat Politécnica de Catalunya - BarcelonaTech. (unpublished).

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5):5.

Ramon Ferrer-i-Cancho. 2014. A stronger null hypothesis for crossing dependencies. *EPL (Europhysics Letters)*, 108(5):58003, 12.

Ramon Ferrer-i-Cancho. 2019. The sum of edge lengths in random linear arrangements. *Journal of Statistical Mechanics: Theory and Experiment*, 2019:053401, 05.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Roger Park Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Michael R. Garey and David Stifler Johnson. 1976. Some simplified NP-complete graph problems. *Theoretical Computer Science*, pages 237–267.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium, 11. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic, 06. Association for Computational Linguistics.

David Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310.

GMP. 2021. The GNU multiple precision arithmetic library. `https://gmplib.org/`. Accessed: 2021-09-16.

Mark K. Goldberg and Israel A. Klipker. 1976. A Minimal Placement of a Tree on the Line. Technical report, Physico-Technical Institute of Low Temperatures. Academy of Sciences of Ukranian SSR, USSR. in Russian.

Carlos Gómez-Rodríguez and Ramon Ferrer-i-Cancho. 2017. Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96:062304.

Carlos Gómez-Rodríguez, John Carroll, and David Weir. 2011. Dependency Parsing Schemata and Mildly Non-Projective Dependency Parsing. *Computational Linguistics*, 37(3):541–586.

Carlos Gómez-Rodríguez, Morten H. Christiansen, and Ramon Ferrer-i-Cancho. 2020. Memory limitations are hidden in grammar. *Arxiv*, page under review.

Kristina Gulordava and 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)*, pages 121–130, Uppsala, Sweden, 08. Uppsala University.

Kristina Gulordava and Paola Merlo. 2016. Multi-lingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0. CDROM CAT: LDC2006T01, ISBN 1-58563-370-4. Linguistic Data Consortium.

Frank Harary and Allen J. Schwenk. 1973. The number of caterpillars. *Discrete Mathematics*, 6:359–365.

Frank Harary. 1969. *Graph Theory*. Addison-Wesley, Reading, MA.

Refael Hassin and Shlomi Rubinstein. 2000. Approximation algorithms for maximum linear arrangement. In *Scandinavian Workshop on Algorithm Theory - Algorithm Theory - SWAT 2000*, volume 1851, pages 231–236.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93, Paris, France, 08. Association for Computational Linguistics.

Robert A. Hochberg and Matthias F. Stallmann. 2003. Optimal one-page tree embeddings in linear time. *Information Processing Letters*, 87(2):59–66.

Richard Hudson. 1995. Measuring syntactic difficulty. *Unpublished paper*.

Mikhail Anatolievich Iordanskii. 1987. Minimal numberings of the vertices of trees — approximate approach. In Lothar Budach, Rais Gatič Bukharajev, and Oleg Borisovič Lupanov, editors, *Fundamentals of Computation Theory*, pages 214–217, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jingyang Jiang and Haitao Liu, editors. 2018. *Quantitative Analysis of Dependency Structures*. De Gruyter Mouton, Berlin, 1 edition.

Yingqi Jing and Haitao Liu. 2015. Mean hierarchical distance. Augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics*, pages 161–170.

Alpár Jüttner and Péter Madarasi. 2018. VF2++ – An improved subgraph isomorphism algorithm. *Discrete Applied Mathematics*, 242:69–81. Computational Advances in Combinatorial Optimization.

Sylvain Kahane, Chunxiao Yan, and Marie-Amélie Botalla. 2017. What are the limitations on the flux of syntactic dependencies? evidence from ud treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 73–82, 9.

Reinhard Köhler and Gabriel Altmann. 2012. *Quantitative Syntax Analysis*. Quantitative linguistics. De Gruyter Mouton.

Saeko Komori, Masatoshi Sugiura, and Wenping Li. 2019. Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 130–135, Paris, France, 08. Association for Computational Linguistics.

Alex Kramer. 2021. Dependency lengths in speech and writing: A cross-linguistic comparison via YouDePP, a pipeline for scraping and parsing YouTube captions. In *Proceedings of the Society for Computation in Linguistics*, volume 4, pages 359–365.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, COLING-ACL '06, pages 507–514, 07.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Haitao Liu. 2007. Probability distribution of dependency distance. *Glottometrics*, 15:1–12, 06.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: a method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Luka Marohnić. 2018. Graph theory package for giac/xcas - reference manual. `https://usermanual.wiki/Document/graphtheoryusermanual.346702481/view`. Accessed: 2020-01-13.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany, NY, USA.

Albert Nijenhuis and Herbert S. Wilf. 1978. *Combinatorial Algorithms: For Computers and Hard Calculators*. Academic Press, Inc., Orlando, FL, USA, 2nd edition.

Joakim Nivre. 2006. Constraints on non-projective dependency parsing. In *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 73–80.

T. Osborne and K. Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: A Journal of General Linguistics*, 4(1):17.

Y. Albert Park and Roger Park Levy. 2009. Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the 10th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) conference*, pages 335–343, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Carlo Passarotti. 2016. How far is stanford from prague (and vice versa)? comparing two dependency-based annotation schemes by network analysis. *L'ANALISI LINGUISTICA E LETTERARIA*, pages 21–46.

Daniele Pighin. 2012. The Ti*k*Z-`dependency` package. `https://osl.ugr.es/CTAN/graphics/pgf/contrib/tikz-dependency/tikz-dependency-doc.pdf`. Accessed: 2021-06-17.

Heinz Prüfer. 1918. Neuer Beweis eines Satzes über Permutationen. *Arch. Math. Phys*, 27:742–744.

Alfréd Rényi and George Szekeres. 1967. On the height of trees. *Journal of the Australian Mathematical Society*, 7(4):497–507.

Immanuel T. San Diego and Frederick S. Gella. 2014. The b-chromatic number of bistar graph. *Applied Mathematical Sciences*, 8(116):5795–5800.

Giorgio Satta, Emily Pitler, Sampath Kannan, and Mitchell Marcus. 2013. Finding optimal 1-Endpoint-Crossing trees. In *Transactions of the Association for Computational Linguistics*, pages 13–24.

Yossi Shiloach. 1979. A minimum linear arrangement algorithm for undirected trees. *SIAM Journal on Computing*, 8(1):15–32.

Daniel Sleator and Davy Temperley. 1993. Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies (IWPT'93)*, pages 277–292. ACL/SIGPARSE.

SWIG. 2020. Swig 4.0.2. `http://www.swig.org/`. Accessed: 2021-06-09.

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4(1):67–80.

Oleg Verbitsky. 2008. On the obfuscation complexity of planar graphs. *Theoretical Computer Science*, 396(1):294–300.

Herbert S. Wilf. 1981. The uniform selection of free trees. *Journal of Algorithms*, 2:204–207.

Robert Alan Wright, Bruce Richmond, Andrew Odlyzko, and Brendan D. McKay. 1986. Constant time generation of free trees. *SIAM Journal on Computing*, 15:540–548, 05.

Himanshu Yadav, Samar Husain, and Richard Futrell. 2019. Are formal restrictions on crossing dependencies epiphenominal? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 2–12, Paris, France, 08. Association for Computational Linguistics.

Himanshu Yadav, Samar Husain, and Richard Futrell. 2021. Do dependency lengths explain constraints on crossing dependencies? *Linguistics Vanguard*, 7(s3):20190070.

Xiang Yu, Agnieszka Falenska, and Jonas Kuhn. 2019. Dependency length minimization vs. word order constraints: An empirical study on 55 treebanks. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 89–97, Paris, France, 08. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni

Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Ethan Chi, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Oxford, England.

Attributivity and Subjectivity in Contemporary Written Czech

Miroslav Kubát[1], Radek Čech[1], Xinying Chen[2]
[1]Department of Czech Language, University of Ostrava
[2]School of Foreign Studies, Xi'an Jiaotong University
miroslav.kubat@gmail.com cechradek@gmail.com cici13306@gmail.com

Abstract

The study focuses on two syntactic indices (attributivity, subjectivity) in various text types and genres in the contemporary written Czech. Index of attributivity is defined as the ratio of the frequency of attributes to the sum of frequencies of nouns, pronouns, and attributes. Index of subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects. The goal is (a) to find out the utility of the proposed indices in stylometry and (b) to enrich stylistics with new quantitative findings. The research is based on the corpus SYN2020 belonging to the Czech National Corpus. The results show that both indices can distinguish different styles and genres. In general, non-fiction texts tend to have higher values of both indices compared to fiction literature.

1 Introduction

The Czech stylistics is mainly focused on the lexical features of text styles. Phonetic, morphological, and syntactic features are usually rather out of the main interest of scholars (cf. Čechová et al. 2008; Hoffmannová et al. 2016). The exception is Bečka (1992) who paid extraordinary attention to syntax. Czech stylistics is also rather based on qualitative than quantitative approach. Only a few quantitative studies deal with syntactic functions or parts of speech in Czech from a stylistic point of view (e.g. Kubát 2016, Těšitelová 1985, Uhlířová 1974). Since these studies are usually limited to (a) small samples and (b) few analyzed styles or genres, we aim to tackle these issues differently. First, our research is based on a large corpus. Second, we analyze not only the main style groups such as fiction and non-fiction, but we focus also on particular genres such as novels, short stories, etc.

In this study, we focus on two stylometric indices. Index of attributivity is defined as the ratio of the frequency of attributes to the sum of frequencies of nouns, pronouns, and attributes. Index of subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects. Proposing these indices is inspired by similar indices successfully applied in stylometry such as nominality, activity, descriptivity (cf. Zörnig 2015).[1] In contrast to these indices based on morphological level (part-of-speech), we focus on writing style in terms of syntactic functions. The goal is to investigate how the resulting values of attributivity and subjectivity vary among different styles and genres in a large corpus. We use data from the Czech National Corpus, namely the corpus SYN2020 consisting of 100 million tokens. We aim to test the utility of these new indices in stylometric research and to enrich Czech stylistics with new findings.

We expect that (a) sentences with higher complexity are generally longer and prefer using more attributes, thus higher attributivity might appear in more formal texts with longer sentences; (b) since Czech has rich morphological features, subjects can be omitted in expressions, therefore, higher subjectivity would also appear in more formal texts which are more syntactically well-formed.

[1] Further research should be done to investigate possible correlations between nominality and descriptivity on the one side and subjectivity and attributivity on the other side. Research of this kind is beyond the scope of this paper.

2 Material

Since we need a big syntactically annotated corpus containing diverse text types and genres of Czech texts, the corpus SYN2020 is used as a dataset for this research. SYN2020 is a synchronous representative and reference corpus of contemporary written Czech containing 100 million tokens (Křen et al. 2020). It is the latest corpus of the representative corpora SYN series (SYN2000, SYN2005, SYN2010, SYN2015, SYN2020), released every five years. Each of the SYN series corpora primarily covers the language of the last five years; thus, SYN2020 consists of the texts published in the 2015–2019 period. Corpus SYN2020 is lemmatized, morphologically tagged, and syntactically annotated.

The syntactic annotation is based on the principles of the annotation used in the Prague Dependency Treebank (cf. Hajič et al. 2020). It marks dependency relations between two words in a sentence and the analytical functions of individual words. The annotation procedure is described in detail on the corpus website[2]. The accuracy rates of SYN2020: UAS = 92.39%, LAS = 88.73%.[3] The error rate is higher for less common syntactic functions and constructions, whereas the most frequent functions in expected contexts have an error rate lower than 5% (cf. corpus website[4]). We have to therefore take into account possible errors when dealing with this dataset. Although the accuracy of syntactic annotation is not perfect, we consider the error rate acceptable for our research.

Since we deal with stylometry in this research, the text style diversity of the corpus is important. SYN2020 consists of texts of various text types and genres. There are three main text type groups (fiction, non-fiction, newspapers and magazines) that consist of several subcategories/genres (see Table 1). The main three groups are equally covered in the corpus.

Fiction (FIC)	
Novels (NOV)	Novels and novellas.
Short stories (COL)	Collections of short stories and other shorter prose texts.
Poetry (VER)	Collections of poetry, marginally song lyrics.
Drama, screenplays (SCR)	Theatre plays, marginally also screenplays for film.
Non-fiction (NFC)	
Scientific (SCI)	Scientific texts, academic publications, university textbooks.
Professional (PRO)	Texts intended for professionals in a given field.
Popular (POP)	Texts intended for a lay audience with an interest in the field.
Memoirs, autobiographies (MEM)	Memoirs, (auto)biographies.
Administrative texts (ADM)	Rules and regulations, meeting minutes, annual reports, etc.
Newspapers and magazines (NMG)	
Newspapers (NEW)	Daily newspapers (current news from home and abroad).
Magazines (LEI)	Special interest magazines focused on thematic groups such as home, garden, hobbies, lifestyle, sports…

Table 1: Text type groups and subcategories/genres in SYN2020.

It is important to mention that SYN2020 is already one of the best quality corpora of such size and style diversity not only for Czech but for all languages. Releasing a large corpus with such a high quality of syntactic annotation and text variety was one of the main motivations for this research.

3 Methodology

We propose two syntactic stylometric indices in this study: attributivity and subjectivity.

[2] https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace

[3] UAS (unlabeled attachment score) is the rate of successful parent identification. LAS (labeled attachment score) is the rate of successful identification of both parent and syntactic function.

[4] https://wiki.korpus.cz/doku.php/cnk:syn2020:automaticka_anotace

3.1 Attributivity

Attributivity expresses a magnitude of depicting/describing things in the text. The more detailed description of things, the higher the index of attributivity. The meaning of nouns and pronouns can be modified by attribute (modifier) that provides an extra detail. Attribute is typically realized by adjective (e.g. *nové* auto) [a *new* car] but can be also expressed by pronoun (e.g. *naše* auto) [*our* car], numeral (e.g. *druhé* auto) [a *second* car], noun (úpravy *textu*) [*text* correction]), nonfinite verb (přání *zdokonalit*) [a desire *to improve*]), or adverb (cesta *domů*) [the way *home*]. Attribute can be also realized by a dependent clause. We expect that more formal texts tend to detailed descriptions of things. Thus, an author needs to use more attributes for the description than in less formal texts.

The index of attributivity is defined as the ratio of the frequency of attributes to the sum of frequencies of nouns, pronouns, and attributes. The formula is as follows.

$$\text{attributivity} = \frac{\text{attributes}}{\text{nouns} + \text{pronouns} + \text{attributes}}$$

3.2 Subjectivity

Subjectivity indicates a level of expressing subjects. Subject is typically realized by noun or pronoun. In Czech grammar, subject can be omitted whereas a predicate has to be explicitly expressed in every clause. This is caused by the rich morphology of Czech language where the person can be easily identified by the ending morpheme of the predicate, especially for the first and second person. In case of a clause with the predicate in the third person, the subject can be also omitted if it is known from the context. That is why we expect higher subjectivity in more formal texts that tend to explicitly express subjects. On the other hand, less formal texts, especially those close to spoken language, should prefer omitting subjects.

Subjectivity is defined as the ratio of the frequency of subjects to the sum of frequencies of predicates and subjects. The formula is as follows.

$$\text{subjectivity} = \frac{\text{subjects}}{\text{predicates} + \text{subjects}}$$

Both indices are simple ratios expressing style features that can be interpreted straightforwardly. We expect that these features considerably differ in various text groups and genres. It should be noted that the nature of the data (big data, results are not based on average values) prevents us from applying a statistical test.

CQL (corpus query language) queries for searching predicates, attributes, subjects, nouns, and pronouns the corpus SYN2020 used in this research can be found in the Appendix of this paper.

4 Results

4.1 Attributivity

The resulting values show that fiction tends to have much lower attributivity compared to non-fiction and journalism (see Figure 1). This can be explained by the fact that more formal texts need a precise and detailed description of nouns and pronouns. This is also visible in differences between genres inside each text type group. In fiction, drama reaches the lowest attributivity (see Figure 2). Drama is close to spoken language which is generally less formal and has a simpler structure. We can see the same pattern also in the case of non-fiction literature where memoirs and autobiographies are less attributive because of their style close to fiction (see Figure 3). Interestingly, there are no big differences in journalism (see Figure 4).

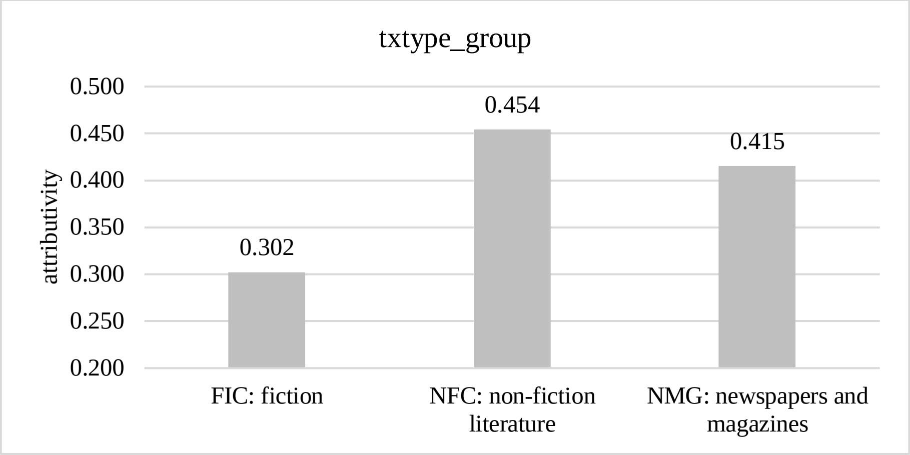


Figure 1: Attributivity in text type groups.

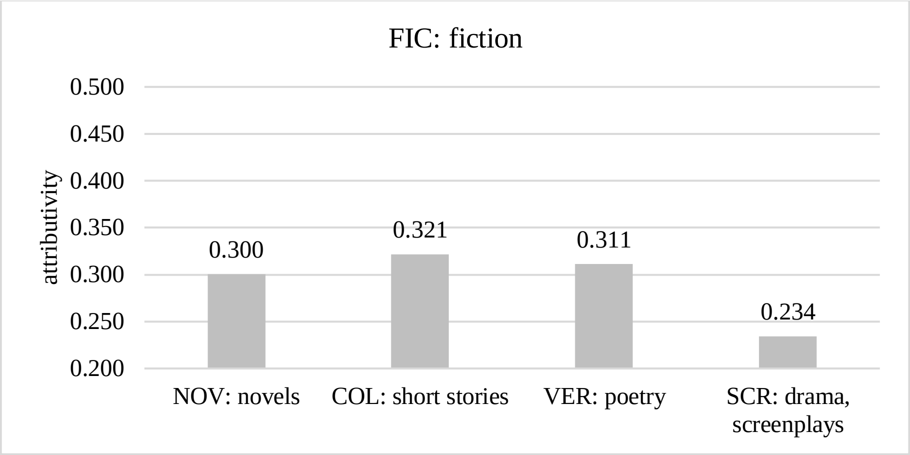


Figure 2: Attributivity in fiction.

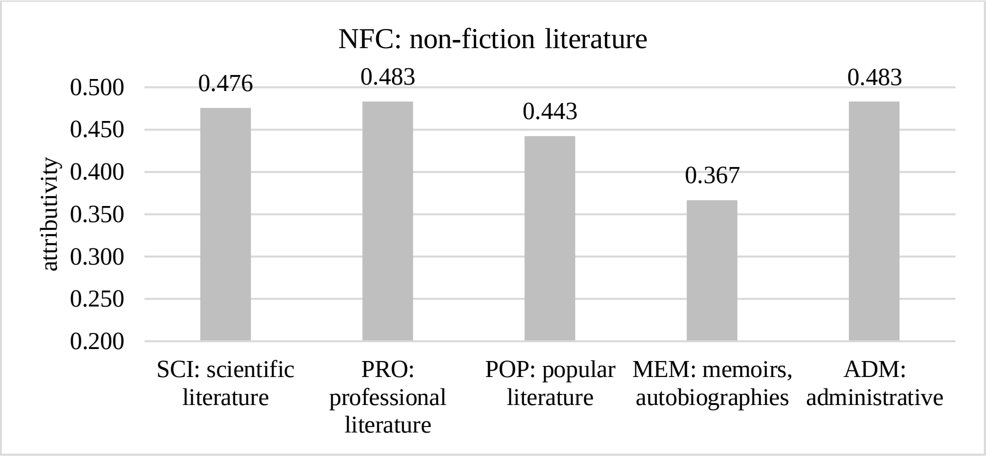


Figure 3: Attributivity in non-fiction literature.

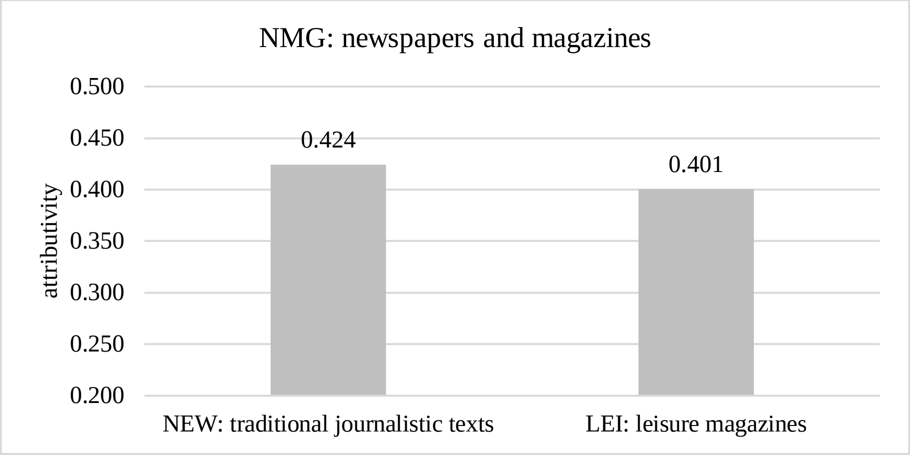


Figure 4: Attributivity in newspapers and magazines.

4.2 Subjectivity

The resulting values of subjectivity in Figure 5 show quite a clear difference between more formal texts (non-fiction, journalism) on the one side and less formal texts (fiction) on the other side. This could be explained by the fact that more formal texts generally tend to have explicit expressions and redundancy, whereas less formal texts prefer simpler forms. Features typical for spontaneous informal spoken language are very common in Czech contemporary fiction literature. The language is simple, sentences are rather short and there are lots of ellipses as well (cf. Hoffmannová et al. 2016). We can also see this tendency in Figure 7 in non-fiction texts where the lowest subjectivity has the genre of memoirs and autobiographies which are close to fiction literature. All the analyzed genres in fiction literature have very similar values of subjectivity (see Figure 6). In journalism (see Figure 8), we can see that magazines have lower subjectivity than daily newspapers. This is in line with our expectations because newspapers have rather formal texts compared to leisure magazines.

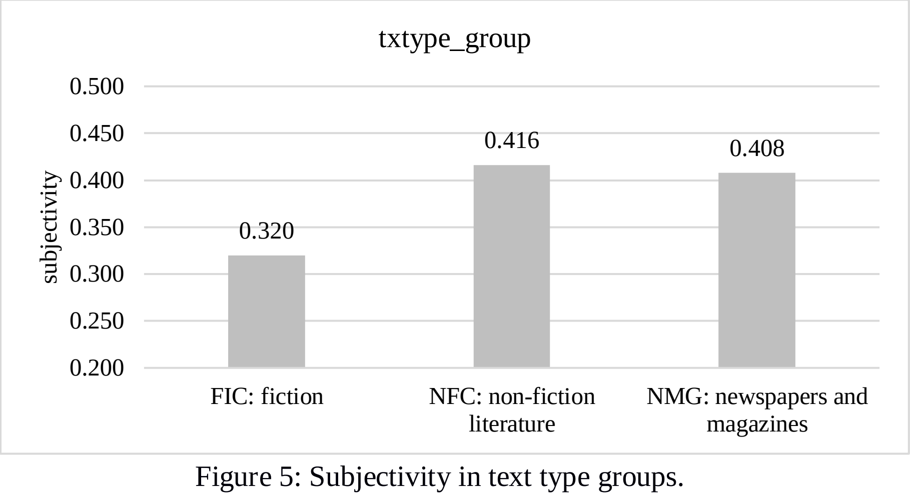


Figure 5: Subjectivity in text type groups.

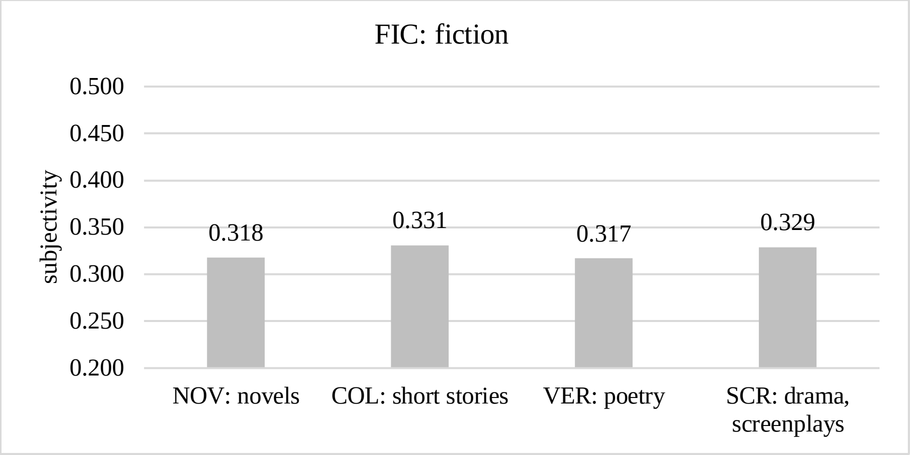


Figure 6: Subjectivity in fiction.

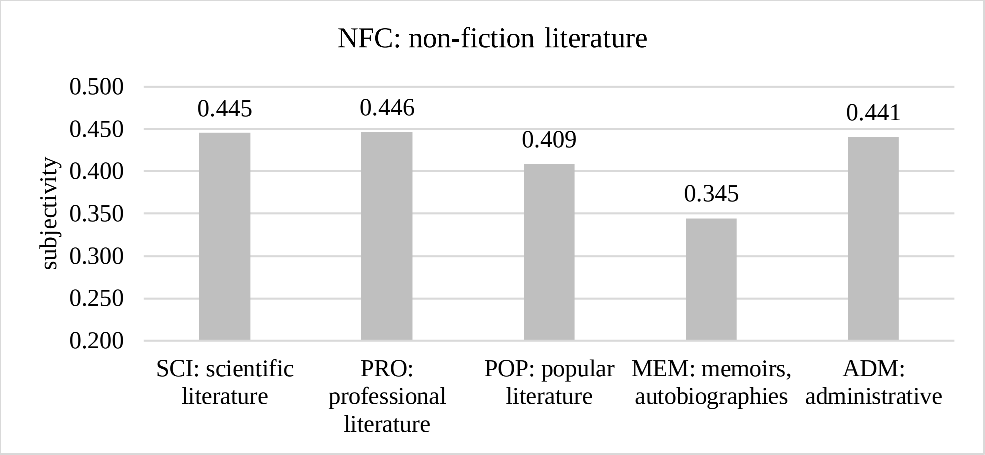


Figure 7: Subjectivity in non-fiction.

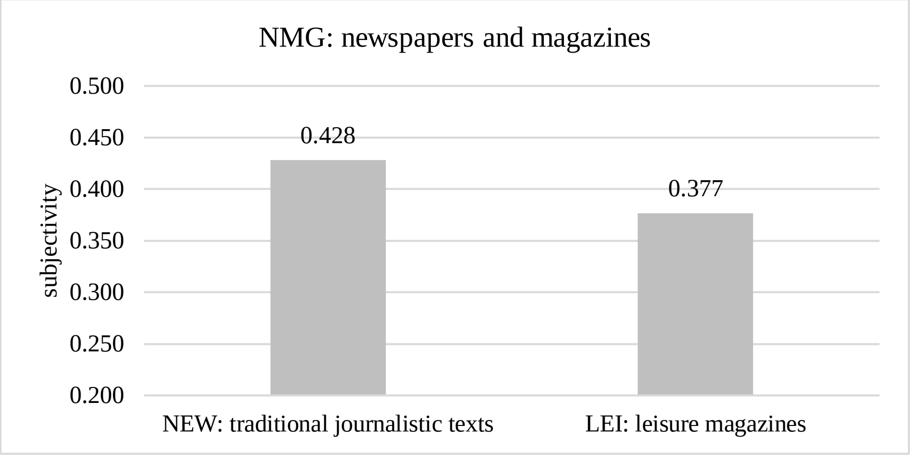


Figure 8: Subjectivity in newspapers and magazines.

5 Conclusion

The proposed syntactic indices (attributivity, subjectivity) seem to be sensitive to different styles and genres. The results show that both indices express stylistic characteristics of texts that can distinguish various text types. We can therefore preliminary conclude that they can be applied in stylometric research as other indices such as nominality, activity, descriptivity, lexical richness.

Fiction literature reached considerably lower values of both indices whereas non-fiction writing reached higher values. Journalism is between them (closer to non-fiction). The more formal text, the higher attributivity and subjectivity. Fiction literature and less formal texts tend to lower subjectivity because of the preference for omitting subjects. Non-fiction literature and more formal texts tend to be more attributive due to the need of precise and detailed description of the nouns and pronouns.

Although the study comes with promising results, we must emphasize that this is just a first attempt to apply indices of attributivity and subjectivity in stylometry. Further research must be done to confirm our preliminary findings. These methods can be also applied in the authorship attribution domain to discover whether attributivity and subjectivity are sensitive to the writing style of different authors. Since our study is limited only to Czech, it is also important to investigate other languages.

Acknowledgements

This work is partly supported by the National Social Science Fund of China (18CYY031).

References

Josef V. Bečka. 1992. *Česká stylistika*. Academia, Praha, Czechia.

Marie Čechová, Marie Krčmová, and Eva Minářová. 2008. *Současná stylistika*. NLN, Praha, Czechia.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank – Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020),* pages 5208–5218. Marseille, France

Jana Hoffmannová, Jiří Homoláč, Eliška Chvalovská, Lucie Jílková, Petr Kaderka, Petr Mareš, and Kamila Mrázková. 2016. *Stylistika mluvené a psané češtiny*. Academia, Praha, Czechia.

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. 2020. *SYN2020: reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague, Czechia. Available at http://www.korpus.cz.

Miroslav Kubát. 2016. *Kvantitativní analýza žánrů.* University of Ostrava, Ostrava, Czechia.

Marie Těšitelová. 1985. *Kvantitativní charakteristiky současné češtiny*. Academia, Praha, Czechia.

Ludmila Uhlířová. 1974. O frekvenci větných členů v souvislém textu. *Slovenská reč*: 39(3), 141–146.

Peter Zörnig. 2015. *Descriptiveness, activity and nominality in formalized text sequences*. RAM-Verlag, Lüdenscheid, Germany.

Appendix

We used the following CQL (corpus query language) queries for searching predicates, attributes, subjects, nouns, and pronouns in the corpus SYN2020.
Predicates: [tag="V[B,i,p,q,s,t].*"&afun!="AuxV|AuxT|AuxR"]
Attributes: [afun="Atr" | afun="Atr_Co" | afun="Atr_Ap" | afun="Atr_Pa" | afun="AtrAdv" | afun="AtrAdv_Co" | afun="AtrAdv_Ap" | afun="AtrAdv_Pa" | afun="AtrAtr" | afun="AtrAtr_Co" | afun="AtrAtr_Ap" | afun="AtrAtr_Pa" | afun="AtrObj" | afun="AtrObj_Co" | afun="AtrObj_Ap" | afun="AtrObj_Pa" | afun="AtrAdv" | afun="AtrAdv_Co" | afun="AtrAdv_Ap" | afun="AtrAdv_Pa"]
Subjects: [afun="Sb" | afun="Sb_Co" | afun="Sb_Ap" | afun="Sb_Pa"]
Nouns: [tag="N.*"]
Pronouns: [tag="P.*"]

Dependency distance minimization predicts compression

Ramon Ferrer-i-Cancho
Complexity and Quantitative Linguistics Lab,
LARCA Research Group
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya
Campus Nord, Edifici Omega
Jordi Girona Salgado 1-3
08034 Barcelona, Catalonia, Spain
rferrericancho@cs.upc.edu

Carlos Gómez-Rodríguez
Universidade da Coruña, CITIC
FASTPARSE Lab, LyS Research Group
Departamento de Ciencias
de la Computación y Tecnologías
de la Información
Facultade de Informática, Elviña,
15071, A Coruña, Spain
cgomezr@udc.es

Abstract

Dependency distance minimization (DDm) is a well-established principle of word order. It has been predicted theoretically that DDm implies compression, namely the minimization of word lengths. This is a second order prediction because it links a principle with another principle, rather than a principle and a manifestation as in a first order prediction. Here we test that second order prediction with a parallel collection of treebanks controlling for annotation style with Universal Dependencies and Surface-Syntactic Universal Dependencies. To test it, we use a recently introduced score that has many mathematical and statistical advantages with respect to the widely used sum of dependency distances. We find that the prediction is confirmed by the new score when word lengths are measured in phonemes, independently of the annotation style, but not when word lengths are measured in syllables. In contrast, one of the most widely used scores, i.e. the sum of dependency distances, fails to confirm that prediction, showing the weakness of raw dependency distances for research on word order. Finally, our findings expand the theory of natural communication by linking two distinct levels of organization, namely syntax (word order) and word internal structure.

1 Introduction

According to the dependency distance minimization (DDm) principle, the distance between heads and their dependent words in a sentence has to be reduced (Ferrer-i-Cancho, 2004; Gildea and Temperley, 2007). The principle has been supported widely by studies whose coverage of languages and families is increasing over time (e.g., Liu (2008), Futrell et al. (2015), Futrell et al. (2020), Ferrer-i-Cancho et al. (2021)). For simplicity, such distance is usually measured in words (see Ferrer-i-Cancho (2015b) for an exception) but it could be measured with more precision in syllables or phonemes (Ferrer-i-Cancho, 2015b). In that way, the distance of a dependency would be a function of the length of the words defining the dependency and that of the words in-between. For this reason, it was predicted that word lengths should be minimized to minimize dependency distances (Ferrer-i-Cancho, 2017a; Ferrer-i-Cancho, 2017b), namely the DDm principle predicts compression, i.e., another principle whereby L, the mean word length, has to be minimized. However, to our knowledge, such a prediction has never been tested in spite of its great theoretical importance. First, it is crucial for the construction of a theory of language and other natural communication systems (Semple et al., 2021). A critical component of a theory are the predictions that it can make. For instance, DDm predicts the scarcity of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017) and compression of word lengths predicts Zipf's law of abbreviation (Ferrer-i-Cancho et al., 2019). These are examples of first order predictions, namely manifestations that are predicted by a certain principle. The focus of this article are second order predictions, namely, principles that are predicted by other principles, such as (a) the prediction of DDm from Dm, a general principle of distance minimization (Ferrer-i-Cancho et al., 2021; Ferrer-i-Cancho, 2003) in the spirit of Behaghel's pioneering views (Behaghel, 1932), or (b) a principle of surprisal or entropy minimization (Hm) from compression (Ferrer-i-Cancho, 2018). Table 1 summarizes these and other first order and second order predictions from previous research stemming from a general principle of energy minimization (Em) in the spirit of Zipf's least effort principle (Zipf, 1949) but extended to other biological systems (Semple et al., 2021).

In the quantitative linguistics tradition, syllables are considered to be one of the best units (if not the best one) for measuring word length, mainly to warrant cross-linguistic validity (Popescu et al., 2013; Grzybek, 2013). In addition, phonemes (and graphemes) are considered to not be appropriate because they are not immediate constituents of words (words are made of syllables that are made in turn of phonemes). Here we will revise this view arising from research on word lengths and check if it still applies for research on the interaction between dependency distance (in words) and word length.

This article is aimed at testing the hypothesis that DDm implies compression, which we refer to as repertoire unit weight minimization (RUWm). We follow the convention that a lower case m at the end of the abbreviation of a principle indicates "minimization" (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). RUWm is the minimization of the weight (or cost) of units in a repertoire. In the context of word types and their length as their weight, RUWm corresponds to the minimization of word lengths. RUWm, predicts that more likely units in a repertoire (e.g., more frequent words in a vocabulary) should be lighter (e.g., be shorter). In particular, RUWm has been argued to lead to the emergence of the law of abbreviation across linguistic levels (Ferrer-i-Cancho et al., 2019): e.g., the lexical level where frequent word types or meanings tend to have shorter forms (Zipf, 1949; Kanwal et al., 2017; Brochhagen, 2021), and the sublexical level where more frequent cases or case marker types tend to have shorter forms (Liu, 2021). RUWm is a specialization of a more general principle of unit weight minimization (UWm) arising from research on unifying compression with the origins of both the law of abbreviation and Menzerath's law (Gustison et al., 2016). The law of abbreviation is the tendency of more frequent words to be shorter (Zipf, 1949) while Menzerath's law is the tendency of linguistic constructs with more parts to be made of smaller parts (Altmann, 1980). A specialization of UWm on sequences, sequence unit weight minimization (SUWm) sheds light on the possible origin of Menzerath's law (Gustison et al. (2016); Table 1). In the context of words as sequences of syllables and the length of as syllable as its weight (or cost), SUWm corresponds to the minimization of the lengths of syllables in the words they appear. SUWm predicts that units in longer sequences (e.g., syllables in words with more syllables) should have smaller weight (e.g., be shorter).

Formally, we aim to test

$$DDm \longrightarrow RUWm. \tag{1}$$

As an alternative to that prediction one could consider a compensation hypothesis, where less optimized languages at the level of dependency distances would have more pressure for shorter words (or more optimized languages at the level of dependency distances would tolerate longer words). The idea of compensation has been applied in word order research in different ways (Ferrer-i-Cancho, 2018; Ferrer-i-Cancho, 2015b). For instance, the suboptimal placement of the verb in SOV orders with respect to DDm (DDm predicts SVO or OVS) has been argued to be compensated by the short length of clitics (Ferrer-i-Cancho, 2015b).

2 Methods

To measure dependency distance in a sentence, we considered two scores, both computed using dependency distances measured in words. The first score is Ω, a recently introduced normalized score that takes the value 1 when the dependency distances are fully optimized and is expected to take a value of 0 if there is no bias on dependency distances, for or against DDm (Ferrer-i-Cancho et al., 2021). Ω can be seen as score of the intensity of DDm, which is maximum when $\Omega = 1$, and missing when $\Omega \approx 0$. $\Omega < 0$ indicates that DDm is surpassed by other word order principles. The score has the virtue of satisfying a series of mathematical and statistical properties: dual normalization (i.e. normalization with respect to both the minimum and the random baseline), constancy under minimum linear arrangement, stability under random linear arrangement, invariance under linear transformation and boundedness under maximum linear arrangement (Ferrer-i-Cancho et al., 2021). These properties are particularly useful when calculating an average score over the sentences of a treebank. For comparison, we consider also D, the

Em	⟶	***Dm***	⟶	***SDm***	⟶	pairs of primarily alternating word orders (Ferrer-i-Cancho, 2016)
			⟶	***DDm***	⟶	acceptability (Morrill, 2000)
					⟶	word order preferences (Morrill, 2000)
					⟶	scarcity of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017)
					⟶	tendency to uncover the root (Ferrer-i-Cancho, 2008)
					⟶	projectivity & planarity with high probability
				DDm (+ projectivity)	⟶	medial placement of the root (Gildea and Temperley, 2007; Alemany-Puig et al., 2022)
				DDm (+ planarity)	⟶	medial placement of the central vertex (Iordanskii, 1987; Hochberg and Stallmann, 2003; Alemany-Puig et al., 2022)
				DDm + extreme V placement in SVO triples	⟶	placement of adjectives with respect to nominal heads (Ferrer-i-Cancho, 2008; Ferrer-i-Cancho, 2015a)
					⟶	placement of auxiliary V with respect to main V (Ferrer-i-Cancho, 2008)
					⟶	consistent branching for dependents of nominal heads (Ferrer-i-Cancho, 2015b)
					⟶	”unnecessity” of headedness parameter (Ferrer-i-Cancho, 2015b)
					⟶	***RUWm*** (to be tested in this article)
	⟶	***UWm***	⟶	***RUWm***	⟶	Zipf's law of abbreviation (Shannon, 1948; Ferrer-i-Cancho et al., 2019)
						reduction (Ferrer-i-Cancho, 2017a)
				RUWm + unique segmentation	⟶	***Hm*** (Ferrer-i-Cancho, 2018)
				SUWm	⟶	Menzerath's law (Gustison et al., 2016; Ferrer-i-Cancho et al., 2019)

Table 1: Optimization principles and their predictions. Arrows link principles with their predictions. Predictions can take the form of principles or manifestations. Principles are marked in boldface. The two principles whose relationship is the target of this article are marked in blue. *Em*: energy minimization. *Dm*: distance minimization. *UWm*: unit weight minimization, popularly known as *compression*. *RUWm*: repertoire unit weight minimization. *SUWm*: sequence unit weight minimization. *SDm*: swap distance minimization. *DDm*: dependency distance minimization. *Hm*: entropy (or surprisal) minimization. We use parentheses for assumptions that are likely to be predictions of DDm and thus likely to be unnecessary assumptions to a large extent (see Gómez-Rodríguez and Ferrer-i-Cancho (2017) for further details about the argument). Awareness of such unnecessary assumptions is vital for the construction of a parsimonious theory.

sum of dependencies of a sentence. While Ω is a measure of closeness, D is a measure of distance. D is the most widely used score (Gildea and Temperley, 2007; Gildea and Temperley, 2010; Futrell et al., 2015; Futrell et al., 2020) but does not satisfy any of the remarkable mathematical properties enumerated above. The intensity of DDm is difficult to assess just from the value of D.

For each treebank considered in this study, we calculated an average Ω and average D over all the sentences of the treebank. To measure word length, we considered two different units: phonemes and syllables. For each language, we calculated L_s, the mean word length in syllables (mean syllables per word token) and L_p, the mean word length in phonemes (mean phonemes per word token).

To control for the content or the source text of the treebanks, we used a parallel collection of treebanks, in particular, the Parallel Universal Dependencies (PUD) collection version 2.6 (Zeman et al., 2017). PUD contains 20 languages from 9 distinct families. PUD follows the UD annotation style (Zeman et al., 2020). To control for annotation style, we also use SUD, i.e. Surface-Syntactic Universal Dependencies (Gerdes et al., 2018). We use PSUD to refer to the PUD collection following the SUD annotation (Ferrer-i-Cancho et al., 2021). The PUD and PSUD treebank collections are borrowed from a recent study (Ferrer-i-Cancho et al., 2021). The preprocessing of these treebanks involves the removal of punctuation marks and reparalellization to warrant there is no loss of parallelism after punctuation mark removal (Ferrer-i-Cancho et al., 2021). The data are available from `https://github.com/lluisalemanypuig/optimality-syntactic-dependency-distances` in two levels of preprocessing: (a) the preprocessed treebanks as head vectors and (b) the raw text tables that were extracted from them and used to feed the statistical analyses. The transformation of the raw head vectors into the raw text tables can be replicated easily with the Linear Arrangement Library (Alemany-Puig and Ferrer-i-Cancho, 2022).

The PUD data does not include syllable or phoneme annotations that would allow us to measure word lengths in these units, and obtaining them would be highly costly. Thus, we instead borrowed mean word lengths from the dataset of another recent study (Fenk-Oczlon and Pilz, 2021). In that study, mean word lengths were estimated from 22 simple declarative sentences encoding one proposition and using basic vocabulary. Three languages from PUD/PSUD are missing in that dataset: Arabic (Afro-Asiatic), Indonesian (Austronesian) and Swedish (Indo-European). As a result, the final collection has 17 languages from 7 distinct families that are displayed in Table 2.

We consider two approaches to investigate the relationship between dependency distance and word length. First, a Kendall τ correlation test between the mean dependency distance score (D or Ω) and mean word length (L_s or L_p). With respect to plain Pearson correlation, τ is more robust to extreme observations and to non-linearity (Newson, 2002). A significant negative correlation between mean Ω and L_s or L_p would confirm the prediction in Eq. 1. Note that Ω and D have opposite interpretations (larger values of Ω imply shorter dependencies whereas larger values of D imply longer dependencies) hence a positive correlation with respect to D would be analogous to a negative correlation with respect to Ω. Therefore, a positive correlation between D and L_s or L_p could also be interpreted as confirming the prediction in Eq. 1 but D lacks the mathematical and statistical properties that are required for robust assessment (Ferrer-i-Cancho et al., 2021).

Second, the fact that the Indo-European family is over-represented and the only one that is represented by more than one language (Table 2) motivates the need to control for the effect of family size. Accordingly, we also consider a couple of kinds of generalized linear models. First, a null model with mean word length as response (L_s or L_p) and language family as random factor. Second, a mixed effects model with mean word length as response (L_s or L_p), language family as random effect and a dependency distance score (mean D or mean Ω) as fixed effect. To test the prediction with these models, we use information theoretic model selection (Burnham and Anderson, 2002; Winter, 2019). AIC of the mixed effects model being lower than that of the null model would confirm the prediction provided that the weight of the association between the predictor and the response has a sign that matches that of the prediction. Again, some caution is needed when D is involved because of its technical limitations (Ferrer-i-Cancho et al., 2021).

The linear models were fitted with the `lme4` R package. Confidence intervals for the weight of the

Family	Languages
Turkic (1)	Turkish
Indo-European (11)	Czech, English, French, German, Hindi, Icelandic, Italian, Polish, Portuguese, Russian, Spanish
Japonic (1)	Japanese
Koreanic (1)	Korean
Sino-Tibetan (1)	Chinese
Tai-Kadai (1)	Thai
Uralic (1)	Finnish

Table 2: The 17 languages from 7 families used in the present study. The counts attached to each family name indicate the number of different languages included in the present study.

Collection	distance	length	n	τ	p
PUD	Ω	L_s	17	-0.052	0.773
		L_p	17	-0.37	0.039
PSUD	Ω	L_s	17	-0.111	0.536
		L_p	17	-0.37	0.039
PUD	D	L_s	17	-0.258	0.149
		L_p	17	-0.459	0.011
PSUD	D	L_s	17	-0.185	0.303
		L_p	17	-0.385	0.032

Table 3: The correlation between the mean dependency distance score and mean word length. We show the annotation style, the distance score, the word length score, the value of the Kendall τ correlation statistic and p, the p-value of the corresponding two-sided test. We assume a significance level of 0.05.

fixed effect were computed using the parametric bootstrapping method of function `confint` of the `lme4` package.

3 Results

Fig. 1 shows the relationship between word length and Ω. Table 3 indicates that the correlation between Ω and L_p is negative and statistically significant whereas the correlation between Ω and L_s is also negative but not significant. That is, the shorter the syntactic dependencies of a language upon dual normalization (Ω), the shorter the words when their length is measured in phonemes. Table 4 indicates that the result is confirmed by a linear mixed effects model that predicts L_s or L_p based on Ω with family as random effect. The null model (only family as random factor) always yields an AIC value that is larger

Collection	distance	length	AIC mixed effects	AIC null
PUD	Ω	L_s	23.15	23.9
		L_p	42.16	47.64
PSUD	Ω	L_s	23.54	23.9
		L_p	42.96	47.64
PUD	D	L_s	32.28	23.9
		L_p	45.23	47.64
PSUD	D	L_s	32.22	23.9
		L_p	48.03	47.64

Table 4: Information theoretic selection of models to predict the mean word length. We show the annotation style, the distance score, the word length score, the Akaike Information Criterion (AIC) of the mixed effects model and the AIC of the null model.

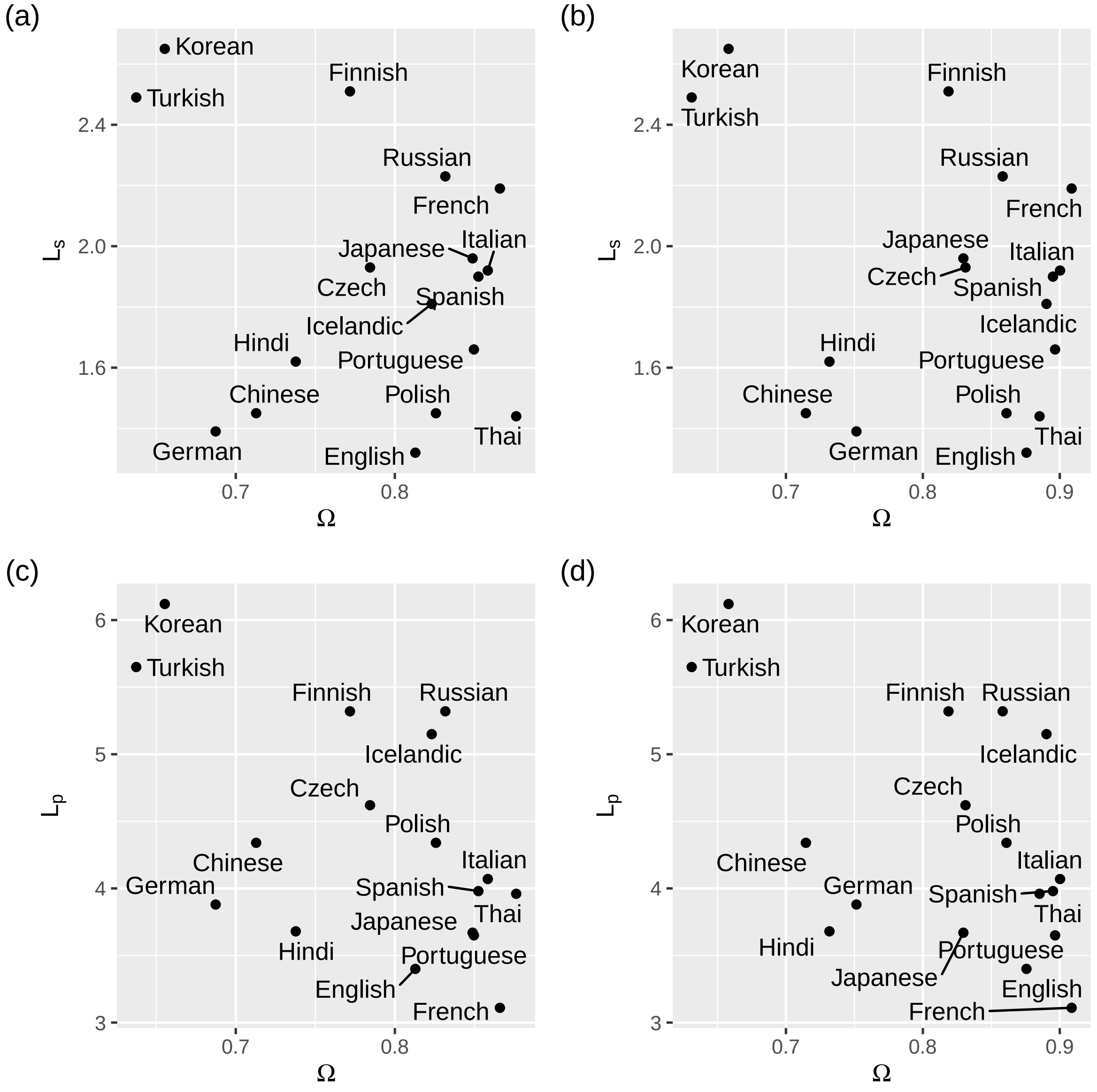


Figure 1: Mean word length (L) as a function of the mean degree of optimality of syntactic dependency distances (mean Ω). (a) L_s as a function of mean Ω in PUD. (b) L_s as a function of mean Ω in PSUD. (c) L_p as a function of mean Ω in PUD. (d) L_p as a function of mean Ω in PSUD.

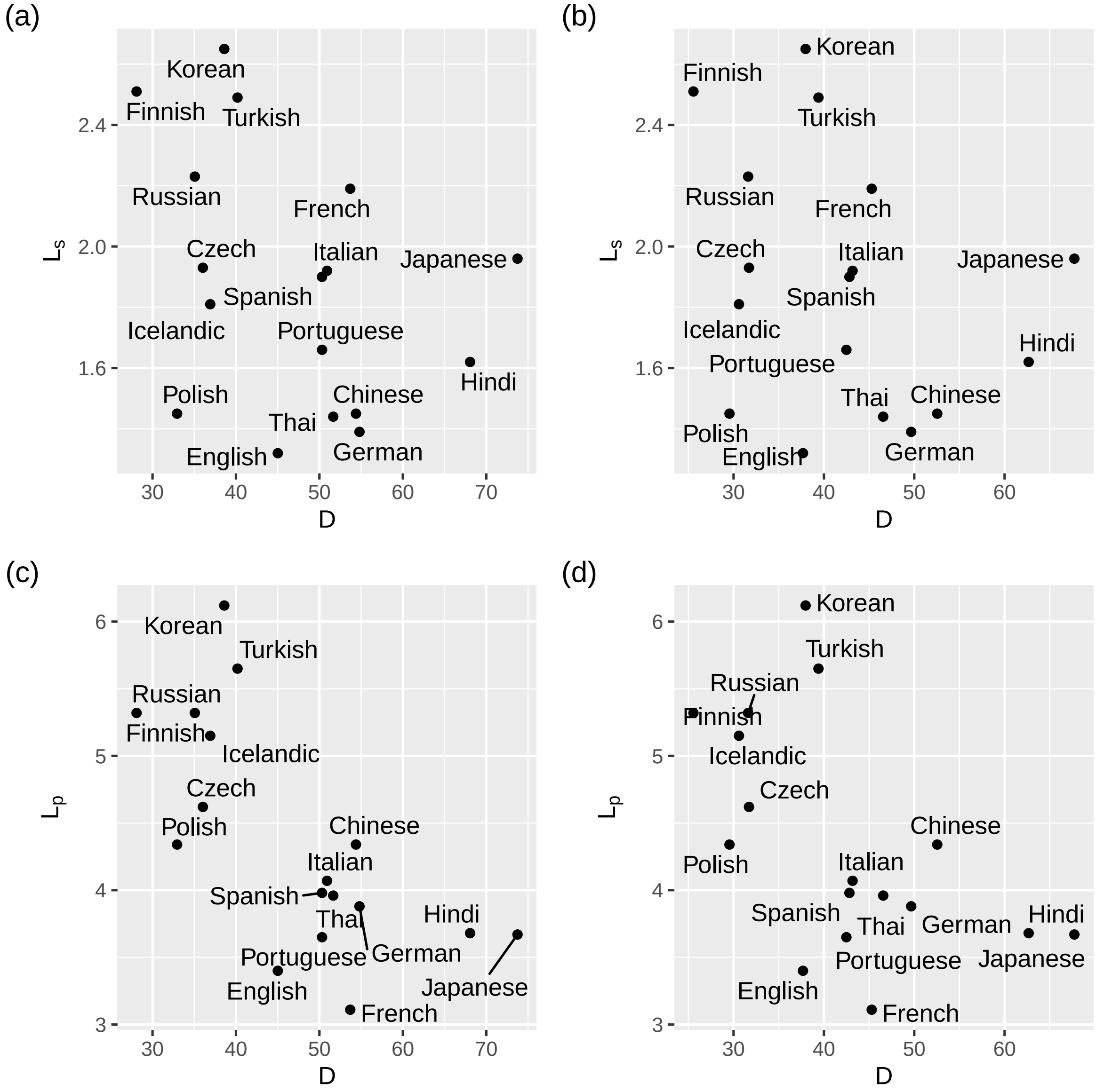


Figure 2: Mean word length (L) as a function of the mean sum of dependency distances (mean D). (a) L_s as a function of D in PUD. (b) L_s as a function of D in PSUD. (c) L_p as a function of mean D in PUD. (d) L_p as a function of mean D in PSUD.

Collection	Fixed effect	Response	Lower	Upper
PUD	Ω	L_p	-10.85	-1.19
PSUD	Ω	L_p	-9.21	-0.69
PUD	D	L_s	-0.03	0
		L_p	-0.07	-0.02
PSUD	D	L_s	-0.03	0.01
		L_p	-0.07	-0.02

Table 5: Lower and upper bounds of the 95% confidence interval for the weight of the fixed effect (Ω or D) in the mixed effects models. The confidence intervals for Ω with L_p as response could not be computed due to numerical problems.

than that of the mixed effects model (Ω as fixed effect and family as random effect). However, the AIC of the null model is only sufficiently large (i.e., differing by various units) when the predictor is L_s, in agreement with the plain correlation analysis (Table 3). The analysis of the confidence intervals supports the conclusions: the confidence interval for the weight of Ω with L_p as response comprises exclusively negative values, consistently with a negative correlation between Ω and L_p (Table 5).

If Ω is replaced by D in the preceding analyses, opposite conclusions are drawn concerning the direction of the correlation. First, Figure 2 suggests that mean word length tends to decrease as D increases when word length is measured in phonemes in both PUD and PSUD. Accordingly, Table 3 confirms that the correlation between D and L_p is negative and statistically significant while the correlation between D and L_s turns out to not be significant. That is, the longer the syntactic dependencies of a language, the shorter the words when their length is measured in phonemes, the opposite conclusion that is reached with Ω. Table 4 indicates that the results are confirmed by linear mixed effects linear models that predict L_s or L_p based on D with family as random effect. When measuring word lengths in syllables, the AIC of the null model is much smaller than that of the mixed effects model with D as predictor, supporting that D and L_s are uncorrelated. In contrast, the AIC of the null model is larger (by more than unit) than that of the mixed effects model with D as predictor, supporting that D and L_p are correlated. The analysis of the confidence intervals (Table 5) confirms a negative correlation between D and L_p because the confidence intervals for the weight of D with L_p as response comprise exclusively negative values; that does not happen when L_s is the response.

4 Discussion

We have confirmed the prediction that DDm leads to compression (Eq. 1) by showing that Ω and L_p are negatively correlated: the closer the syntactically related words, the shorter the words. Although syllables are the preferred unit of measurement of word length over phonemes in quantitative linguistics (Popescu et al., 2013; Grzybek, 2013), we have failed to confirm the prediction with that unit. We cannot exclude the possibility that a negative correlation between Ω and L_s exists but has not surfaced because it is weaker and the sample of languages is not large enough. However, the failure with syllables may be a confirmation of the higher capacity of phonemes over syllables to capture the so-called 'phonological complexity' (Pimentel et al., 2020). Furthermore, we suspect that word lengths in phonemes lead eventually to a more accurate estimation of the true distance between words, currently measured in words, than syllables. To see the relationship between word length and dependency distance and to build an explanation, suppose that δ_p is the phonemic distance between two words, δ_s is their syllabic distance and d is their distance in words ($d = 1$ if two words are consecutive). Then assuming that a word and its dependent word are connected by the middle of their phonemic or syllabic sequence, a mean field approximation yields $\delta_p \approx dL_p$ (the head and their dependent contribute with $L_p/2$ phonemes; the words in-between the head and the governor contribute with $(d-1)L_p$ phonemes). Analogously, $\delta_s \approx dL_s$. Then δ_p may be more strongly correlated with the time that is needed to keep unresolved dependencies in memory (Morrill, 2000) than δ_s because syllables vary in phonemic length. That time was considered to be the key to understand what determines the acceptability of sentences and other word order phenomena

(Morrill, 2000). Then δ_p looks like a better proxy for the combination of memory decay and interference that is believed to cause DDm (Liu et al., 2017; Temperley and Gildea, 2018). To recap, the argument that syllables are more appropriate units than phonemes (or graphemes) because syllables are immediate word constituents (Popescu et al., 2013; Grzybek, 2013) may not be appropriate for research on dependency distances because distances (or time) could be measured more accurately with constituents located farther down in the hierarchy of constituents.

We have also shown that D, a widely used dependency score (Gildea and Temperley, 2007; Gildea and Temperley, 2010; Futrell et al., 2015; Futrell et al., 2020), fails to confirm the prediction that DDm implies compression. It is not the first time that Ω shows a superior performance when testing theoretical predictions. When testing that DDm should be surpassed by other word order principles in short sentences (Ferrer-i-Cancho, 2014; Ferrer-i-Cancho, 2017a), Ω was able to find many more languages with anti-DDm effects than D (Ferrer-i-Cancho et al., 2021). Our findings reinforce the view that raw dependency distances are a poor reflect of DDm and that advanced scores such as Ω are crucial for progress in research on that optimization principle and related memory constraints (Ferrer-i-Cancho et al., 2021), and eventually, for the construction of a general theory of natural systems with human language as a particular case and energy minimization at the center (Semple et al., 2021). An early sketch of that theory is shown in Table 1.

For each language, we have measured dependency distances (Ω and D) on the collection of sentences included in PUD/PSUD whereas mean word lengths (L_p and L_s) come from an independent collection of sentences (Fenk-Oczlon and Pilz, 2021). It could be argued that mean word lengths should have been estimated on PUD/PSUD, too. This would make the whole study fully parallel, and word length data potentially more accurate, as it would come from a larger sample and, more specifically, from the same sentences where dependency lengths have been measured. It is conceivable, for example, that a given kind of syntactic construction could produce shorter dependency distances in language A than in language B, while being prevalent in PUD/PSUD but not in the word length dataset (e.g. due to genre, style or topic differences). This could cause us to observe increased dependency distance minimization in language A with respect to B without being able to observe the associated compression, as our word length data would not include sentences with that specific phenomenon. In turn, this could cause us to underestimate the correlation between Ω and L_s or L_p. Unfortunately, the technical complexity of replicating the present study with actual phonemic or syllabic lengths on PUD/PSUD goes beyond the scope of the present article. These considerations notwithstanding, notice that the stress of this article is on testing a prediction rather than theoretically agnostic data description. In this context, it is rather astonishing that the theoretical prediction (Eq. 1) is confirmed even though the collections of sentences for dependency distances and the collections for word lengths are independent. That confirmation offers two major interpretations: (a) the results are due to biases in the sentences, either in PUD or in Fenk-Oczlon's dataset (Fenk-Oczlon and Pilz, 2021) or, crucially, (b) there is actually a deep reason for the prediction to hold. Further research on a fully parallel scenario or alternative sources is required.

By having shown that DDm implies compression of word lengths, we do not mean that compression is produced exclusively by DDm. Our work does not rule out other sources for compression. Following previous research testing successfully the prediction that DDm weakens in small sequences (Ferrer-i-Cancho and Gómez-Rodríguez, 2021; Ferrer-i-Cancho et al., 2021), we formulate a new prediction, namely that compression *per se* (independently from DDm) may surface in short sequences. Future research should clarify the weight of the contribution to compression from DDm, compression itself and other principles.

Once one integrates our findings into the piece of a mathematical theory of communication in Table 1, it turns out that we have actually uncovered the following chain,

$$Dm \longrightarrow DDm \longrightarrow RUWm \longrightarrow \text{ Zipf's law of abbreviation}$$

Namely, a general principle of minimization of the distance between elements leads to the prediction of the law of abbreviation.

Our findings have implications for competing views and frameworks. First, the compensation hypothesis outlined in the introduction, i.e. less optimized languages at the level of dependency distances would have more pressure for shorter words (or more optimized languages at the level of dependency distances would tolerate longer words), would predict that the correlation between Ω and word length should not be negative. Instead, a zero or positive correlation would be expected depending on the strength or the nature of the compensation effect. Our findings of a negative correlation rule out the compensation hypothesis as a primary explanation for the global trend. However, we cannot exclude that compensation has some secondary role in general, an important role in specific languages or an important role with certain domains of a language, as suggested for the suboptimal placement of clitics with respect to DDm in Romance languages (Ferrer-i-Cancho, 2015b). Second, DDm predicts compression and compression in turn predicts reduction (see Section 3.4 of Ferrer-i-Cancho (2017a)). By reduction, here we mean the shortening or omission of predictable utterances, a phenomenon that has been used to justify the uniform information density and related hypotheses (see Lemke et al. (2021) and references therein). Therefore, the need for uniform information density and related hypotheses as standalone hypotheses needs to be revised. Third, our finding of an association between word length and dependency distance also challenges the recent conclusion that *(compositional) morphology and graphotactics can sufficiently account for most of the complexity of natural codes – as measured by code length* (Pimentel et al., 2021) as if no strong independent pressure for compression (beyond morphology and graphotactics) really existed. Our findings suggest that the two components of the problem of compression as defined in standard information theory, i.e. the minimization of word length and the conditions of the coding scheme (Ferrer-i-Cancho et al., 2019), may not be as easy to dissociate from (compositional) morphology and graphotactics as expected from the traditional reductionistic division into linguistic levels. Indeed, our article demonstrates how constraints on the “syntactic level” (DDm) may be shaping the “(sub)lexical level” (compression on word lengths).

Our work has implications beyond the current state of development of the theory of the communication efficiency reviewed in Table 1. We have confirmed that DDm implies compression but the outcome of our correlation analysis does not exclude that it could also the other way around, namely that compression is actually leading to DDm. A possible track could be that pressure for shorter words may lead to a loss of information that would require more words to convey the same message, which in turn would imply longer sentences and then higher pressure for DDm. Another track could be that a general principle of compression operates on top both at the level of words and at higher levels (phrases, clauses, sentences) and then close packaging (DDm) at all these levels is simply a consequence of maximizing compression, in line with the now-or-never bottleneck (Christiansen and Chater, 2016). We hope that our work stimulates further research on general principles and their predictions.

Acknowledgements

We are grateful to S. Semple and M. Gustison for extremely valuable and inspiring discussions on theory construction (Semple et al., 2021). We are also grateful to two anonymous reviewers, T. Pimentel and G. Fenk-Oczlon for helpful comments. RFC is supported by the grant TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad) and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya). CGR is supported by the European Research Council (ERC), under the European Union’s Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), the ANSWER-ASAP (TIN2017-85160-C2-1-R) and SCANNER-UDC (PID2020-113230RB-C21) projects from ERDF/MICINN/AEI, Xunta de Galicia (ED431C 2020/11 and an Oportunius program grant to complement ERC grants); the CITIC research center is funded by ERDF and Xunta de Galicia (ERDF - Galicia 2014-2020 program, grant ED431G 2019/01).

References

Lluís Alemany-Puig and Ramon Ferrer-i-Cancho. 2022. The Linear Arrangement Library. A new tool for research on syntactic dependency structures. In *Proceedings of the SyntaxFest 2022*, Sofia, Bulgaria. Association for Computational Linguistics.

Lluís Alemany-Puig, Juan Luis Esteban, and Ramon Ferrer-i-Cancho. 2022. Minimum projective linearizations of trees in linear time. *Information Processing Letters*, 174:106204.

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10.

Otto Behaghel, 1932. *Deutsche Syntax: eine geschichtliche Darstellung*, chapter Bd 4: Wortstellung. Periodenbau. Carl Winters Univeritätsbuchhandlung, Heidelberg, Germany.

Thomas Brochhagen. 2021. Brief at the risk of being misunderstood: Consolidating population-and individual-level tendencies. *Computational Brain & Behavior*, Feb.

Kenneth P. Burnham and David R. Anderson. 2002. *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, New York, 2nd edition.

Morten H. Christiansen and Nick Chater. 2016. The now-or-never bottleneck: a fundamental constraint on language. *Behavioral and Brain Sciences*, 39:1 – 72.

Gertraud Fenk-Oczlon and Jürgen Pilz. 2021. Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication*, 6:66.

Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez. 2021. Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1):50–76.

Ramon Ferrer-i-Cancho, Christian Bentz, and Caio Seguin. 2019. Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, page in press.

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, Juan Luis Esteban, and Lluís Alemany-Puig. 2021. The optimality of syntactic dependency distances. *Physical Review E*, page in press.

Ramon Ferrer-i-Cancho. 2003. *Language: universals, principles and origins*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135.

Ramon Ferrer-i-Cancho. 2008. Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, 11(3):393–414.

Ramon Ferrer-i-Cancho. 2014. Why might SOV be initially preferred and then lost or recovered? A theoretical framework. In E. A. Cartmill, S. Roberts, H. Lyn, and H. Cornish, editors, *The Evolution of Language - Proceedings of the 10th International Conference (EVOLANG10)*, pages 66–73, Vienna, Austria. Wiley. Evolution of Language Conference (Evolang 2014), April 14-17.

Ramon Ferrer-i-Cancho. 2015a. The placement of the head that minimizes online memory. A complex systems approach. *Language Dynamics and Change*, 5(1):114–137.

Ramon Ferrer-i-Cancho. 2015b. Reply to the commentary "Be careful when assuming the obvious", by P. Alday. *Language Dynamics and Change*, 5(1):147–155.

Ramon Ferrer-i-Cancho. 2016. Kauffman's adjacent possible in word order evolution. In *The evolution of language: proceedings of the 11th International Conference (EVOLANG11)*.

Ramon Ferrer-i-Cancho. 2017a. The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics*, 39:38–71.

Ramon Ferrer-i-Cancho. 2017b. Towards a theory of word order. comment on "dependency distance: A new perspective on syntactic patterns in natural language" by Haitao Liu et al. *Physics of Life Reviews*, 21:218 – 220.

Ramon Ferrer-i-Cancho. 2018. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3):207–237.

Richar Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium, November. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic, June. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Carlos Gómez-Rodríguez and Ramon Ferrer-i-Cancho. 2017. Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96:062304.

Peter Grzybek. 2013. Word length. In John R. Taylor, editor, *The Oxford Handbook of the Word*. Oxford University Press.

Morgan L. Gustison, Stuart Semple, Rarmon Ferrer-i-Cancho, and Thore Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences USA*, 13:E2750–E2758.

Robert A. Hochberg and Matthias F. Stallmann. 2003. Optimal one-page tree embeddings in linear time. *Information Processing Letters*, 87:59–66.

Mikhail A. Iordanskii. 1987. Minimal numberings of the vertices of trees — Approximate approach. In Lothar Budach, Rais Gatič Bukharajev, and Oleg Borisovič Lupanov, editors, *Fundamentals of Computation Theory*, pages 214–217, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.

Robin Lemke, Lisa Schäfer, and Ingo Reich. 2021. Modeling the predictive potential of extralinguistic context with script knowledge: The case of fragments. *PLoS One*, 16:e0246255.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:159–191.

Zoey Liu. 2021. A multifactorial approach to crosslinguistic constituent orderings. *Linguistics Vanguard*, page in press.

Glyn Morrill. 2000. Incremental processing and acceptability. *Computational Linguistics*, 25(3):319–338.

Roger Newson. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal*, 2(1):45–64.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18, 01.

Tiago Pimentel, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021. How (non-)optimal is the lexicon? In *North American Chapter of the Association for Computational Linguistics*.

Ioan-Iovitz Popescu, Sven Naumann, Emmerich Kelih, Andrij Rovenchak, Haruko Sanada, Anja Overbeck, Reginald Smith, Radek Cech, Panchanan Mohanty, Andrew Wilson, and Gabriel Altmann. 2013. Word length: Aspects and languages. In *Studies in Quantitative Linguistics*, volume 3, pages 224–281. RAM-Verlag, Lüdenscheid.

Stuart Semple, Ramon Ferrer-i-Cancho, and Morgan Gustison. 2021. Linguistic laws in biology. *Trends in Ecology and Evolution*, page in press.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 623–656.

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/Cognitive universal? *Annual Review of Linguistics*, 4(1):67–80.

Bodo Winter. 2019. *Statistics for linguists: an introdution using R*. Routledge, New York & London.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, et al. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

George Kingsley Zipf. 1949. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge (MA), USA.

The properties of rare and complex syntactic constructions in English A corpus-based comparative study

Ruochen Niu
Zhejiang University
China
niuruochen@126.com

Yaqin Wang
Guangdong University of
Foreign Studies
China
wyq322@126.com

Haitao Liu
Zhejiang University
China
htliu@163.com

Abstract

The study adopts a corpus-based approach to investigate rare and complex constructions in English, such as it-clefts and topicalization. Two dependency-based metrics, namely dependency distance (DD) and hierarchical distance (HD) were used to measure and compare the syntactic complexities at the linear and hierarchical levels of three treebanks, i.e., one specially-designed corpus containing many difficult and infrequent constructions and two reference corpora containing normal constructions. It was found that compared to normal constructions, syntactically infrequent and complex constructions may enjoy higher complexity at the linear level, i.e., longer dependencies, but they are not necessarily more complex at the hierarchical level. In fact, the results suggest that syntactically marked constructions are less tolerant of structural or hierarchical complexity, which may be motivated by a mechanism to avoid self-embedding or recursion driven by limited cognitive resources of human beings.

1 Introduction

Complex and rare syntactic constructions, such as it-clefs, subject-extracted relative clauses and topicalization, constitute an important part of English languages. Generally, such a syntactically marked construction has one or a combination of the following characteristics: (1) a special word order that is different from the canonical word order of the language; (2) non-local dependencies; (3) crossing dependencies that violate the principle of projectivity (also referred to as "discontinuities"). Due to these unique features, these complex constructions have been a central topic of study in many fields of linguistics: in psycholinguistics, they are known for being difficult to process, corresponding to longer reading times in language processing experiments (e.g., Grodner and Gibson, 2005); in syntax, they pose great challenges to grammarians attempting to describe and theorize their structures (e.g., Hudson, 2010; Osborne, 2019). While the above studies are conducive to our understanding of the complex and rare constructions, they are limited in two aspects: (1) the materials used to draw conclusions are often limited in number and range; (2) different constructions are studied independently as individual phenomena. This has hindered our understanding of complex constructions as a whole.

To address the above issue, the current study adopts a corpus-based approach to analyzing the properties of syntactically complex and rare constructions in English. Three corpora were adopted, one specially designed to contain as many of these hard-to-process constructions as possible (Futrell et al., 2021) and two self-built reference corpora sampled from the British National Corpus (Burnard, 2000). Comparisons were drawn on the syntactic complexities of the treebanks at the linear and hierarchical dimensions, which were measured by *dependency distance* (DD) (e.g., Hudson, 1995; Ferrer-i-Cancho, 2004; Liu, 2008; Liu et al., 2017) and *hierarchical distance* (HD) (e.g., Jing and Liu, 2015; Liu and Jing, 2016; Komori et al., 2019), respectively. By comparing these metrics and their distributions in the three treebanks, we are able to gain insights into the structural properties of complex and rare constructions in English and natural languages at large. The following of the manuscript is organized as follows: Section 2 and Section 3 introduces the methods and materials, Section 4 reports and discusses the results, and Section 5 draws a conclusion.

2 Dependency Distance and Hierarchical Distance

This section defines and illustrates the syntactic complexity measures we used to make comparisons among the corpora. They are dependency distance (DD) and hierarchical distance (HD) from the theoretical framework of dependency grammar.

2.1 Two dimensions of a syntactic tree

The study adopts *dependency grammar* as opposed to constituency grammar to analyzing syntactic structure. Dependency grammar views sentence structure as composed of direct links between words, i.e., dependencies (e.g., Tesnière, 1959/2015; Heringer et al., 1980; Mel'čuk, 1988; Hudson, 2010; Nivre, 2015; Osborne, 2019). Between the two words building a dependency relation, the word that expresses the core meaning or licenses the appearance of the other word is called the *head* (or governor), and the word that complements or modifies is the *dependent* (or subordinator).

A dependency structure of a sentence can be shown by a two-dimensional tree that clearly illustrates the two ordering principles. Between them, the horizontal or *x* axis represents the linear order of words that is based on the left-to-right occurrence and the vertical or *y* axis the hierarchical order of words according to the head-dependent relation (c.f., Tesnière, 1959/2015; Osborne, 2019). To illustrate, the dependency tree of an example sentence is given as Figure 1:

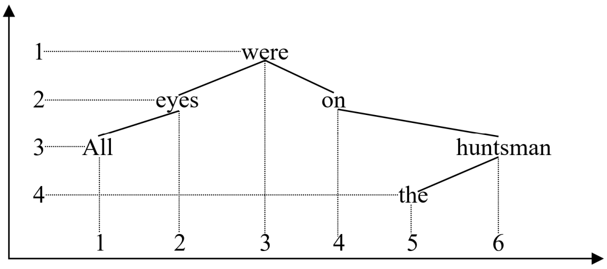


Figure 1. A Two-Dimensional Dependency Tree

In Figure 1, words are connected by concrete lines representing dependencies. Each word projects upon two axes, and the numbers of a word at the *x* and *y* axes stand for its linear and hierarchical orders within the sentence, respectively. Note that we define the hierarchical order of the sentence root, i.e., "were", as 1. As a result, the hierarchy of its dependents, i.e., "eyes" and "one", is 2.

2.2 DD and HD

Based on the above background, DD and HD are two metrics proposed and used for the measurement of the syntactic complexity at the linear and hierarchical level, respectively (e.g., Hudson, 1995; Ferrer-i-Cancho, 2004; Liu, 2008; Jing and Liu, 2015; Komori et al., 2019). Their definitions are given as follows:

DD
The absolute value of the linear order difference between a head and its dependent. [1]
HD
The hierarchical order difference between a word and the sentence root.
MDD
The mean of DD of a sample, e.g., a treebank.
MHD
The mean of HD of a sample, e.g., a sentence.

To give an example, there are six words and five dependencies in the sentence shown in Figure 1.

[1] In previous studies (e.g., Jiang and Liu, 2015; Wang and Liu, 2017), DD is a directed measure by which a positive and negative value denotes a head-final and head-initial relation, respectively; but it is always the absolute value of DD that is used when measuring the syntactic complexity. As the current study is not concerned with head-dependent ordering, we simply define DD as its absolute value to avoid confusion.

The DDs for these dependencies are: 1 (between "All" and "eyes"), 1 (between "eyes" and "were"), 1 between "were" and "on"), 2 (between "on" and "huntsman") and 1 (between "the" and "huntsman"). Thus, the MDD of the sentence is their mean, i.e., 1.2. In the meantime, the HDs for "All", "eyes", "on", "the" and "huntsman" are 2, 1, 1, 3, and 2, respectively. Thus, the MHD of the sentence is their mean, i.e., 1.8.

The motivation of using DD and HD as syntactic complexity metrics is related to the general cognitive constraints underlying language processing. While DD has been found to be related to working memory capacity limits (see Liu et al., 2017 for a review), HD has been associated with the decay of spreading activation (e.g., Hudson, 2010; Jing and Liu, 2015). Similar metrics have also been proposed and acknowledged within other syntactic frameworks (e.g., Yngve, 1960; Köhler and Altmann, 2000; Hawkins, 2004; Grodner and Gibson, 2005).

3 Dependency Treebanks

This section introduces the dependency-annotated corpora, i.e., treebanks, used in our study. They are the Natural Stories Corpus (hereafter abbreviated as the NS Corpus) that contain many rare and complex syntactic constructions (Futrell et al., 2021) and the two reference treebanks that we built based on the British National Corpus (Burnard, 2000).

3.1 Natural Stories Corpus (NS)

The NS Corpus is a "constructed-natural" corpus of English (Shain et al., 2016), i.e., it is designed to contain many infrequent and hard-to-process syntactic constructions while still sounding fluent to native speakers. Including a high proportion of syntactically marked constructions—non-local VP conjunction, it-cleft, topicalization, etc., the corpus is suitable for exploring the features of complex and rare syntactic constructions.

The NS Corpus is composed of ten edited children's stories, e.g., the Bradford's Boar. It provides three types of hand-corrected syntactic parses by Stanford Parser, from which we adopted the Stanford Dependencies parses and the Penn style PoS tags. Before data analysis, a thorough consistency check of all the parses was conducted. The trimmed treebank contains 464 sentences and 10,257 word tokens in total. Figure 2 illustrates the format of the treebank:

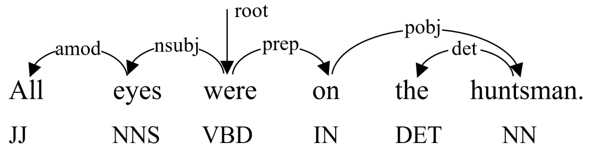


Figure 2. Format of the Treebank

In which the dependency relation is represented by the directed arc pointing from the head to the dependent. Upon each arc marks the type of the dependency, with *amod* denoting adjective modifier, *nsubj* noun subject, *prep* preposition, *pobj* object of a preposition, *det* determiner and *root* sentence root. Below each word is the word's category, with *JJ* standing for adjective, *NNS* plural noun, *VBD* verb in past tense, *IN* preposition, *DET* determiner and *NN* singular noun. This information provides a basis for further analysis.

3.2 British National Corpus (IMA and INFO)

The comparable corpora were built on the British National Corpus (also called the BNC Corpus), which is a corpus of contemporary English containing a variety of domains. BNC's written component (as opposed to spoken) can be largely divided into two genres, namely imaginative (hereafter abbreviated as the IMA Corpus), which is composed of fiction or other literary works, and informative (hereafter abbreviated as the INFO Corpus), which includes a variety of domains, e.g., science, word affairs and leisure (Burnard, 2000).

To build the reference treebanks, i.e., IMA and INFO Corpus, we randomly selected approximately 100,000 word tokens for each of the two genres from the BNC Corpus. Then, these two corpora were automatically annotated using Stanford Parser (version 3.9.2, de Marneffe et al., 2006) and hand-corrected by experienced annotators. The standardized IMA and INFO treebanks enjoy the same annotation scheme as the NS Corpus shown in Figure 2, but they are much larger in size—the IMA Corpus has 103,171 word tokens in 7,669 sentences and the INFO Corpus has 120,974 word tokens in 6,471 sentences. We think it is appropriate to have two comparable corpora given the potential effects of genre on syntactic complexity measures (e.g., Wang and Liu, 2017). Between the two treebanks we built, the IMA Corpus has the similar genre to the NS Corpus, and the INFO Corpus is a more general corpus of a wider coverage of the language which makes it a more suitable reference corpus (Leech, 2002).

3.3 Quantitative Properties of the Treebanks

A preliminary analysis was conducted to obtain a quick overview of the treebanks. Properties such as *mean sentence length* (MSL), *mean dependency distance* (MDD) and *mean hierarchical distance* (MHD) were focused. The results are presented in Table 1:

	Word Tokens	Sentences	MSL	MDD	MHD
NS	10257	464	22.1034	2.5719	3.0789
IMA	103171	7669	13.4530	2.2614	3.1030
INFO	120974	6471	18.6948	2.3466	3.2841

Table 1. Quantitative Properties of the Treebanks

In which *sentence length* (SL) is measured by the number of words in a sentence, and MSL the mean of all SL of a corpus. From Table 1, it is clear that (1) the MDD and MHD of the three treebanks are all below 4. This is supportive of previous findings proposing a threshold of MDD and MHD equal to working memory capacity limits (e.g., Liu, 2008; Liu and Jing, 2016); (2) the NS Corpus has a greater MSL than the INFO and IMA Corpus. This means that the three treebanks are not suitable for direct comparison because longer sentences usually lead to larger MDD and MHD (Jiang and Liu, 2015; Jing and Liu, 2015); (3) despite a larger MSL, NS has a smaller MHD than the other two treebanks, which is contradictory to our expectation and deserves more investigation.

4 Results and Discussion

In Section 3.3, we found that the mean sentence length (MSL) of the NS Corpus is greater than that of the IMA and INFO Corpus. To control for the effects of SL on the results, four SL groups were selected based on the distribution of SLs in the treebanks.[2] This section reports and discusses the findings in comparing the DD and HD related properties of the three treebanks at different sentence lengths.

4.1 DD Distribution

In previous studies, the DD distributions of natural languages have been revealed to exhibit a long tail, which indicates a preference for short dependencies driven by limited capacity of working memory (e.g., Liu, 2008; Jiang and Liu, 2015; Wang and Liu, 2017). To investigate whether the NS Corpus demonstrates any universalities or peculiarities in this regard, the frequencies of the dependencies at different

[2] Because the NS Corpus is small in size, we have to make sure that after controlling for SL it still has enough data for analysis. The selected SL groups are therefore the top four groups of SL that have the most number of sentences in the NS Corpus, i.e., (sentences made of)16-20 (words), 21-25, 11-16 and 26-30.

DD for different SL groups were extracted from the three treebanks; these frequencies were then transformed to corresponding probabilities (or proportions) for comparison among the treebanks.[3] The results are shown in Figure 3:

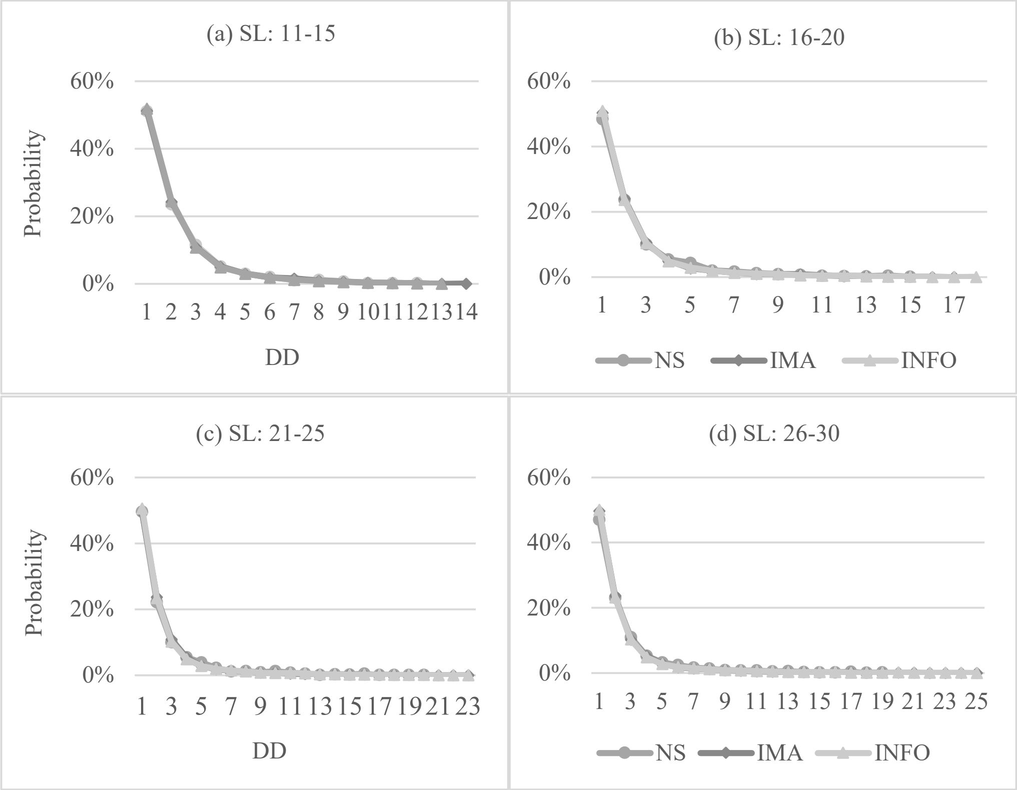


Figure 3. DD Distribution of the Treebanks at Varying SLs

Figure 3 shows that for all four sentence length (SL) groups, the dependency distance (DD) distributions of the three corpora exhibit a similar, almost identical long tail, i.e., the proportion of dependencies is the highest when DD = 1, and drops significantly when DD is increased. This indicates that (1) the preference for short dependencies is not affected by sentence lengths, which corroborates previous findings (Jiang and Liu, 2015); (2) as a corpus that includes many complex syntactic constructions, the NS Corpus is not distinctly different from the reference corpora containing normal constructions in terms of the DD distributions. These findings provide further support to the account that the preference to minimize DD is a language universal driven by general cognitive constraints of human beings rather than intra-linguistic factors (c.f., Liu et al., 2017).

4.2 HD Distribution

In Section 4.1, we found that the NS Corpus is not different from the two reference corpora in terms of the tendency to minimize syntactic difficulty at the linear level, i.e., DD. In this section, we explore their potential similarities and differences at the hierarchical level, i.e., in terms of the HD distributions. Unlike the DD distribution, the HD distribution of natural languages has attracted little attention from the academia. The only exception is Liu (2017) who studied the distribution of hierarchies in three languages and attributed the universalities found to the valency patterns in natural languages proposed by Tesnière (1959/2015). Since HD has been used as a syntactic complexity metric at the hierarchical level in both

[3] Because the treebanks vary greatly in terms of size, it is preferable to use probabilities (or proportions) rather than frequencies for comparison. The probability of the dependencies at a given DD is calculated by dividing the real frequency of dependencies at that DD in a treebank by the total frequency of all dependencies under such circumstance. For example, when SL is between 11 and 15, the frequency of the dependencies with a DD of 1 is 480 in the NS Corpus, and the total frequency of the dependencies with all possible DDs is 938. Thus, the probability of the dependencies at DD = 1 in this case is 480/938 = 51.17%, as shown in Figure 3(a).

general and applied linguistics (e.g., Komori et al., 2019), direct investigation into its distribution may yield new insights into the hierarchical complexities of human languages.

To obtain the HD distributions of the three treebanks, frequency data at different HD for the four sentence length (SL) groups were extracted from the three corpora and transformed into probabilities for cross-corpus comparison. The results are illustrated using Figure 4:

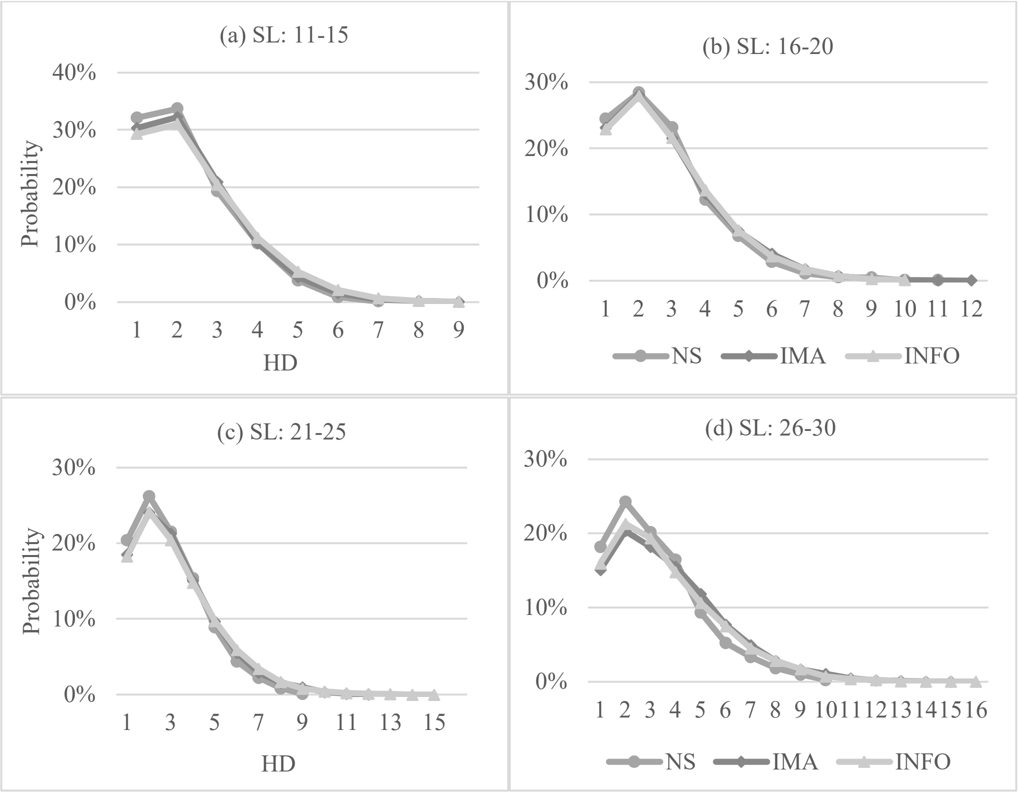


Figure 4. HD Distribution of the Treebanks of Varying SLs

It was found that (1) for all SLs and treebanks, the probability (of words at a given HD) increases when HD is increased from 1 to 2, reaches the peak when HD = 2, and then decreases sharply when HD keeps increasing. These results accord well with Liu (2017)'s findings using another metric called hierarchy (which equals to HD + 1), and suggests that the hierarchical syntactic complexity of human languages is also constrained; (2) for all three treebanks, the proportions of shorter HD (HD ≤ 3) exhibit a decreasing trend when sentences become longer. From Figure (4a) to (4d), the peaks of the curves at HD = 2 decreases from around 30% when SL is between 11 and 15 to around 20% when SL is between 26 and 30. This indicates that unlike the DD distribution that is hardly affected by SL, the HD distribution is influenced by SL to a greater extent; (3) when SL is increased, the NS Corpus seems to exhibit a stronger tendency to avoid longer HDs compared to the two reference corpora. As shown in Figure (4a) and (4b), the HD distributions of the three treebanks are similar when SL is rather small, i.e., between 11-15 and 16-20. However, the slope of the curve for the NS Corpus seems to be steeper than the other corpora when sentences become longer; this is particularly evident in Figure (4d), in which NS has higher proportions of words with shorter HDs (HD ≤ 3) but lower proportions of words with longer HDs (HD ≥ 4).

In general, the results in this section suggest that although the three treebanks share some commonalities in terms of their HD distributions, the NS Corpus that enjoys a higher proportion of complex and infrequent syntactic constructions such as relative clause and it-clefts, seems to demonstrate a greater tendency for short HDs when SL is increased. In addition, our results indicate that unlike the DD distribution, the HD distribution is more likely to be affected by SL. To further validate the phenomenon observed above, we investigated the MDD and MHD of the three treebanks at different sentence lengths, which is reported and discussed in Section 4.3.

4.3 Relation between MDD and MHD

In 4.2, we found that while the three treebanks share some similarities in their HD distributions, the NS Corpus seems to have a stronger tendency to avoid longer HDs when SL is increased. This indicates that complex and rare syntactic constructions may have a stronger tendency to avoid complexity at the hierarchical level compared to normal syntactic constructions. This subsection further explores the phenomenon by examining the relation between MDD and MHD of the three treebanks.

To begin with, we calculated the MSL, MHD and MDD of the three treebanks for the above-mentioned sentence lengths groups.[4] The results are presented in Table 2:

	SL group	NS	IMA	INFO
MSL	11-15	13.2078	12.8437	13.0315
MHD	11-15	2.2281	2.3317	2.4201
MDD	11-15	2.1098	2.0903	2.0509
MSL	16-20	18.0968	17.795	17.8541
MHD	16-20	2.6728	2.7805	2.7859
MDD	16-20	2.3808	2.2601	2.2491
MSL	21-25	22.7386	22.8515	22.9785
MHD	21-25	2.9314	3.1844	3.2437
MDD	21-25	2.4819	2.4023	2.3691
MSL	26-30	27.8553	27.8283	27.8601
MHD	26-30	3.2032	3.7066	3.5953
MDD	26-30	2.646	2.4582	2.4842

Table 2. MHD and MDD of the Treebanks at Different SLs

In Table 2, the MHD and MDD of all three treebanks both increase with MSL, and the increase of MSL brings more gain on MHD than on MDD. This is consistent with previous findings of English (Liu and Jing, 2016). In addition, the NS Corpus has the highest MDD and lowest MHD among the three treebanks within each SL group. This corroborates the reciprocal relationship between MDD and MHD found in previous studies (e.g., Jing and Liu, 2015), and suggests that syntactically complex structures are not necessarily more complex in both linear and hierarchical dimensions. More importantly, the difference in MHD among the NS Corpus and the two reference corpora becomes larger when SL is increased. In other words, the other two corpora seem to be less constrained in MHD than the NS Corpus when sentence length is increased. This coincides with our finding in Section 4.2, and suggests that syntactically marked structures may be more complex at the linear level, but they tend to avoid complexity at the hierarchical level compared to normal constructions.

To visualize the trade-off relation between MDD and MHD and the differences of the treebanks in this regard, we created a bubble chart based on the MDD and MHD of all sentences in the three treebanks.[5] In Figure 5, the direction of the *x* axis denotes larger MHD and the direction of the *y* axis lager MDD. It is clear that the NS Corpus has greater overall MDD (the upper part of the chart is occupied mostly by dark grey bubbles), whereas the IMA and INFO Corpus tend to have greater overall MHD (the right side of the chart is taken up mostly by grey and light gray bubbles). This is supportive of our findings above and indicates that syntactically rare and complex constructions in English may have slightly

[4] Sentence length is controlled because it has an impact on MDD and MHD (e.g., Jing and Liu, 2015; Liu and Jing, 2016). See footnote 2 for how these four SL groups were selected.

[5] A bubble chart is a variation of a scatter chart in which the data points are replaced with bubbles. In addition to the two axes, a third piece of information about the data is shown by the size of the bubbles (here the number of sentences at a given MHD or MDD). What's more, bubble charts are helpful for analyzing the trend of the data.

longer dependencies, but they are not more complex hierarchically. In fact, our results suggest the reverse situation, i.e., syntactically complex structures, e.g., it-clefts and subject-extracted relative clauses, are less lenient on the complexity at the hierarchical level compared to normal structures.

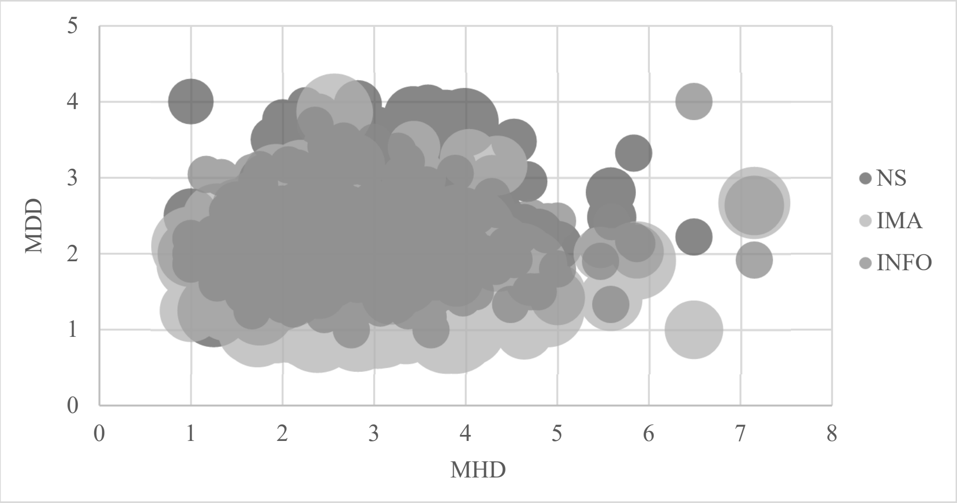


Figure 5. Reciprocal Relation between MDD and MHD in the Treebanks

Previous studies have related a word's difficulty at the hierarchical level (or HD) to the decay of spreading activation from the sentence root to that word (Jing and Liu, 2015; c.f., Hudson, 2010). This, however, does not explain why syntactically marked constructions as a whole are less lenient on the hierarchical complexity than normal constructions. After examining the distributions of grammatical relations at different HDs of the three treebanks, we found that this property of the NS Corpus may be motivated by a mechanism to avoid self-embedding, i.e., embedding of structures of the similar kind. Self-embedding (or recursion) is known as a fundamental property of human languages (e.g., Hauser et al., 2002), but researchers from different fields have found constraints on its use in natural languages, driven perhaps by limited cognitive resources of human beings (e.g., Karlsson, 2007; Christiansen and MacDonald, 2009). In other words, self-embedded structures are cognitively challenging. In our study, the syntactically complex and rare constructions in the NS Corpus have been found to be more difficult to process at the linear level. Presumably, the embedding of these constructions will induce more cognitive effort than that of normal constructions, which can at times lead to processing breakdown. Thus, the stricter constraint on the hierarchical complexity of the NS Corpus may be an attempt to avoid self-embedding to counteract the cognitive burden imposed by the presence of rare and complex constructions. In general, the above findings and discussions support the idea of language as a human-driven complex adaptive system (e.g., Christiansen and MacDonald, 2009; Liu, 2018).

5 Conclusion

Our study investigates the properties of syntactically marked constructions based on a specially designed corpus containing many infrequent and complex constructions and two reference corpora in the same annotation scheme. By examining and comparing their syntactic complexities at the linear and hierarchical levels (measured by DD and HD, respectively), we found that syntactically complex constructions may be more difficult to process at the linear level, but their hierarchical structures are not necessarily more complex than normal constructions. We attribute the stricter constraint on the hierarchical complexity of the complex constructions to an attempt to avoid self-embedding or recursion in natural language processing, which is ultimately motivated by limited cognitive resources. Altogether these findings indicate properties of natural languages as a human-driven self-adaptive complex system, which calls for more interdisciplinary research.

Acknowledgements

We are grateful to the three anonymous reviewers, whose insightful comments helped improve the content of the manuscript.

References

Lou Burnard (ed.). 2000. *Users Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Morten H. Christiansen and Maryellen C. MacDonald. 2009. A usage-based approach to recursion in sentence processing. *Language Learning*, 59: 126–161.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454.

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70: 056135.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55: 63–77.

Daniel. J. Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29: 261–290.

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298: 1569–1579.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammar*. Oxford: Oxford University Press.

Hans J. Heringer, Bruno Strecker, and Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. München: Wilhelm Fink Verlag.

Richard Hudson. 1995. Measuring syntactic difficulty. Unpublished paper. Available at http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf (Accessed 5 June 2019).

Richard Hudson. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel English-Chinese dependency treebank. *Language Sciences*, 50: 93–104.

Yingqi Jing and Haitao Liu. 2015. Mean hierarchical distance: Augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170.

Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43: 365–392.

Reinhard Köhler and Gabriel Altmann. 2000. Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics*, 7(3): 189–200.

Saeko Komori, Masatoshi Sugiura, and Wenping Li. 2019. Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 130–135.

Geoffrey Leech. 2002. The importance of reference corpora. In *Hizkuntza-corpusak. Oraina eta geroa*. Available at https://uzei.eus/online/dokumentazioa/biltzarrak/corpus-jardunaldia-2002/(Accessed 26 September).

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9: 159–191.

Haitao Liu. 2017. *Juzi jiegou cengji de fenbu guilv* "The hierarchical distribution of sentence structure". *Foreign Language Teaching and Research*, 49(3): 345–352.

Haitao Liu. 2018. Language as a human-driven complex adaptive system. *Physics of Life Reviews*, 26-27: 149–151.

Haitao Liu and Yingqi Jing. 2016. *Yingyu juzi cengji jiegou jiliang fenxi* "A quantitative analysis of English hierarchical structure". *Journal of Foreign Languages*, 6: 2–11.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21: 171–193.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2015, Part I, LNCS 9041*), pages 3–16.

Timothy Osborne. 2019. *A Dependency Grammar of English. An Introduction and Beyond*. Amsterdam: John Benjamins.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC 2016)*, pages 49–58.

Lucien Tesnière. 2015. *Elements of Structural Syntax* (original work published in 1959, translated by Timothy Osborne and Sylvain Kahane). Amsterdam: John Benjamins.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59(866): 135–147.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society (Vol. 104, No. 5)*, pages 444–466.