

Quasy 2019

**First Workshop on Quantitative Syntax
(Quasy, SyntaxFest 2019)**

Proceedings

August 26, 2019
held within the **SyntaxFest 2019**, 26–30 August
Paris, France

©2019 The Association for Computational Linguistics

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 978-1-950737-65-9

Preface

The first edition of Quasy was part of the first SyntaxFest, a grouping of four events, which took place in Paris, France, during the last week of August:

- the Fifth International Conference on Dependency Linguistics (Depling 2019)
- the First Workshop on Quantitative Syntax (Quasy)
- the 18th International Workshop on Treebanks and Linguistic Theories (TLT 2019)
- the Third Workshop on Universal Dependencies (UDW 2019)

The use of corpora for NLP and linguistics has only increased in recent years. In NLP, machine learning systems are by nature data-intensive, and in linguistics there is a renewed interest in the empirical validation of linguistic theory, particularly through corpus evidence. While the first statistical parsers have long been trained on the Penn treebank phrase structures, dependency treebanks, whether natively annotated with dependencies, or converted from phrase structures, have become more and more popular, as evidenced by the success of the Universal Dependency project, currently uniting 120 treebanks in 80 languages, annotated in the same dependency-based scheme. The availability of these resources has boosted empirical quantitative studies in syntax. It has also led to a growing interest in theoretical questions around syntactic dependency, its history, its foundations, and the analyses of various constructions in dependency-based frameworks. Furthermore, the availability of large, multilingual annotated data sets, such as those provided by the Universal Dependencies project, has made cross-linguistic analysis possible to an extent that could only be dreamt of only a few years ago.

In this context it was natural to bring together TLT (Treebanks and Linguistic Theories), the historical conference on treebanks as linguistic resources, Depling (The international conference on Dependency Linguistics), the conference uniting research on models and theories around dependency representations, and UDW (Universal Dependency Workshop), the annual meeting of the UD project itself. Moreover, in order to create a point of contact with the large community working in quantitative linguistics it seemed expedient to create a workshop dedicated to quantitative syntactic measures on treebanks and raw corpora, which gave rise to Quasy, the first workshop on Quantitative Syntax. And this led us to the first SyntaxFest.

Because the potential audience and submissions to the four events were likely to have substantial overlap, we decided to have a single reviewing process for the whole SyntaxFest. Authors could choose to submit their paper to one or several of the four events, and in case of acceptance, the program co-chairs would decide which event to assign the accepted paper to.

This choice was found to be an appropriate one, as most submissions were submitted to several of the events. Indeed, there were 40 long paper submissions, with 14 papers submitted to Quasy, 31 to Depling, 13 to TLT and 16 to UDW. Among them, 28 were accepted (6 at Quasy, 10 at Depling, 6 at TLT, 6 at UDW). Note that due to multiple submissions, the acceptance rate is defined at the level of the whole SyntaxFest (around 70%). As far as short papers are concerned, 62 were submitted (24 to Quasy, 41 to Depling, 35 to TLT and 37 to UDW), and 41 were accepted (8 were presented at Quasy, 14 at Depling, 9 at TLT and 9 at UDW), leading to an acceptance rate for short papers of around 66%.

We are happy to announce that the first SyntaxFest has been a success, with over 110 registered participants, most of whom attended for the whole week.

SyntaxFest is the result of efforts from many people. Our sincere thanks go to the reviewers who thoroughly reviewed all the submissions to the conference and provided detailed comments and suggestions, thus ensuring the quality of the published papers.

We would also like to warmly extend our thanks to the five invited speakers,

- Ramon Ferrer i Cancho - Universitat Politècnica de Catalunya (UPC)
- Emmanuel Dupoux - ENS/CNRS/EHESS/INRIA/PSL Research University, Paris
- Barbara Plank - IT University of Copenhagen
- Paola Merlo - University of Geneva
- Adam Przepiórkowski - University of Warsaw / Polish Academy of Sciences / University of Oxford

We are grateful to the Université Sorbonne Nouvelle for generously making available the Amphithéâtre du Monde Anglophone, a very pleasant venue in the heart of Paris. We would like to thank the ACL SIGPARSE group for its endorsement and all the institutions who gave financial support for SyntaxFest:

- the "Laboratoire de Linguistique formelle" (Université Paris Diderot & CNRS)
- the "Laboratoire de Phonétique et Phonologie" (Université Sorbonne Nouvelle & CNRS)
- the Modyco laboratory (Université Paris Nanterre)
- the "École Doctorale Connaissance, Langage, Modélisation" (CLM) - ED 139
- the "Université Sorbonne Nouvelle"
- the "Université Paris Nanterre"
- the Empirical Foundations of Linguistics Labex (EFL)
- the ATALA association
- Google
- Inria and its Almanach team project.

Finally, we would like to express special thanks to the students who have been part of the local organizing committee. We warmly acknowledge the enthusiasm and community spirit of:

Danrun Cao, Université Paris Nanterre

Marine Courtin, Sorbonne Nouvelle

Chuanming Dong, Université Paris Nanterre

Yoann Dupont, Inria

Mohammed Galal, Sohag University

Gaël Guibon, Inria

Yixuan Li, Sorbonne Nouvelle

Lara Perinetti, Inria et Fortia Financial Solutions

Mathilde Regnault, Lattice and Inria

Pierre Rochet, Université Paris Nanterre

Chunxiao Yan, Université Paris Nanterre

Marie Candito, Kim Gerdes, Sylvain Kahane, Djamé Seddah (local organizers and co-chairs),
and Xinying Chen, Ramon Ferrer-i-Cancho, Alexandre Rademaker, Francis Tyers (co-chairs)

September 2019

Program co-chairs

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

- Quasy:
 - Xinying Chen (Xi’an Jiaotong University / University of Ostrava)
 - Ramon Ferrer i Cancho (Universitat Politècnica de Catalunya)
- Depling:
 - Kim Gerdes (LPP, Sorbonne Nouvelle & CNRS / Almanach, INRIA)
 - Sylvain Kahane (Modyco, Paris Nanterre & CNRS)
- TLT:
 - Marie Candito (LLF, Paris Diderot & CNRS)
 - Djamé Seddah (Paris Sorbonne / Almanach, INRIA)
 - with the help of Stephan Oepen (University of Oslo, previous co-chair of TLT) and Kilian Evang (University of Düsseldorf, next co-chair of TLT)
- UDW:
 - Alexandre Rademaker (IBM Research, Brazil)
 - Francis Tyers (Indiana University and Higher School of Economics)
 - with the help of Teresa Lynn (ADAPT Centre, Dublin City University) and Arne Köhn (Saarland University)

Local organizing committee of the SyntaxFest

Marie Candito, Université Paris-Diderot (co-chair)
Kim Gerdes, Sorbonne Nouvelle (co-chair)
Sylvain Kahane, Université Paris Nanterre (co-chair)
Djamé Seddah, University Paris-Sorbonne (co-chair)
Danrun Cao, Université Paris Nanterre
Marine Courtin, Sorbonne Nouvelle
Chuanming Dong, Université Paris Nanterre
Yoann Dupont, Inria
Mohammed Galal, Sohag University
Gaël Guibon, Inria
Yixuan Li, Sorbonne Nouvelle
Lara Perinetti, Inria et Fortia Financial Solutions
Mathilde Regnault, Lattice and Inria
Pierre Rochet, Université Paris Nanterre
Chunxiao Yan, Université Paris Nanterre

Program committee for the whole SyntaxFest

Patricia Amaral (Indiana University Bloomington)
Miguel Ballesteros (IBM)
David Beck (University of Alberta)
Emily M. Bender (University of Washington)
Ann Bies (Linguistic Data Consortium, University of Pennsylvania)
Igor Boguslavsky (Universidad Politécnica de Madrid)
Bernd Bohnet (Google)
Cristina Bosco (University of Turin)
Gosse Bouma (Rijksuniversiteit Groningen)
Miriam Butt (University of Konstanz)
Radek Čech (University of Ostrava)
Giuseppe Giovanni Antonio Celano (University of Pavia)
Çağrı Çöltekin (University of Tuebingen)
Benoit Crabbé (Paris Diderot University)
Éric De La Clergerie (INRIA)
Miryam de Lhoneux (Uppsala University)
Marie-Catherine de Marneffe (The Ohio State University)
Valeria de Paiva (Samsung Research America and University of Birmingham)
Felice Dell'Orletta (Istituto di Linguistica Computazionale "Antonio Zampolli" - ILC CNR)
Kaja Dobrovoljc (Jožef Stefan Institute)
Leonel Figueiredo de Alencar (Universidade federal do Ceará)
Jennifer Foster (Dublin City University, Dublin 9, Ireland)
Richard Futrell (University of California, Irvine)
Filip Ginter (University of Turku)
Koldo Gojenola (University of the Basque Country UPV/EHU)
Kristina Gulordava (Universitat Pompeu Fabra)
Carlos Gómez-Rodríguez (Universidade da Coruña)
Memduh Gökirmak (Charles University, Prague)
Jan Hajič (Charles University, Prague)
Eva Hajičová (Charles University, Prague)
Barbora Hladká (Charles University, Prague)
Richard Hudson (University College London)
Leonid Iomdin (Institute for Information Transmission Problems, Russian Academy of Sciences)
Jingyang Jiang (Zhejiang University)
Sandra Kübler (Indiana University Bloomington)
François Lareau (OLST, Université de Montréal)
John Lee (City University of Hong Kong)
Nicholas Lester (University of Zurich)
Lori Levin (Carnegie Mellon University)
Haitao Liu (Zhejiang University)
Ján Mačutek (Comenius University, Bratislava, Slovakia)
Nicolas Mazziotta (Université)
Ryan Mcdonald (Google)
Alexander Mehler (Goethe-University Frankfurt am Main, Text Technology Group)

Wolfgang Menzel (Department of Informatik, Hamburg University)
Paola Merlo (University of Geneva)
Jasmina Milićević (Dalhousie University)
Simon Mille (Universitat Pompeu Fabra)
Simonetta Montemagni (ILC-CNR)
Jiří Mírovský (Charles University, Prague)
Alexis Nasr (Aix-Marseille Université)
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 (independent)
Petya Osenova (Sofia University and IICT-BAS)
Jarmila Panevová (Charles University, Prague)
Agnieszka Patejuk (Polish Academy of Sciences / University of Oxford)
Alain Polguère (Université de Lorraine)
Prokopis Prokopidis (Institute for Language and Speech Processing/Athena RC)
Ines Rehbein (Leibniz Science Campus)
Rudolf Rosa (Charles University, Prague)
Haruko Sanada (Rissho University)
Sebastian Schuster (Stanford University)
Maria Simi (Università di Pisa)
Reut Tsarfaty (Open University of Israel)
Zdenka Uresova (Charles University, Prague)
Giulia Venturi (ILC-CNR)
Veronika Vincze (Hungarian Academy of Sciences, Research Group on Artificial Intelligence)
Relja Vulcanovic (Kent State University at Stark)
Leo Wanner (ICREA and University Pompeu Fabra)
Michael White (The Ohio State University)
Chunshan Xu (Anhui Jianzhu University)
Zhao Yiyi (Communication University of China)
Amir Zeldes (Georgetown University)
Daniel Zeman (Univerzita Karlova)
Hongxin Zhang (Zhejiang University)
Heike Zinsmeister (University of Hamburg)
Robert Östling (Department of Linguistics, Stockholm University)
Lilja Øvrelid (University of Oslo)

Additional reviewers

James Barry
Ivan Vladimir Meza Ruiz
Rebecca Morris
Olga Sozinova
He Zhou

Table of Contents

SyntaxFest 2019 Invited talk - Dependency distance minimization: facts, theory and predictions	1
<i>Ramon Ferrer-i-Cancho</i>	
Information-theoretic locality properties of natural language	2
<i>Richard Futrell</i>	
Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand?	16
<i>Jianwei Yan and Haitao Liu</i>	
A quantitative probe into the hierarchical structure of written Chinese	25
<i>Heng Chen and Haitao Liu</i>	
A Comparative Corpus Analysis of PP Ordering in English and Chinese	33
<i>Zoey Liu</i>	
Intervention effects in object relatives in English and Italian: a study in quantitative computational syntax	46
<i>Giuseppe Samo and Paola Merlo</i>	
An explanation of the decisive role of function words in driving syntactic development	57
<i>Anat Ninio</i>	
Extracting out of the subject in French: experimental evidence	68
<i>Anne Abeillé and Elodie Winckel</i>	
The relation between dependency distance and frequency	75
<i>Xinying Chen and Kim Gerdes</i>	
Full valency and the position of enclitics in the Old Czech	83
<i>Radek Cech, Pavel Kosek, Olga Navratilova and Jan Macutek</i>	
Dependency Length Minimization vs. Word Order Constraints: An Empirical Study On 55 Treebanks	89
<i>Xiang Yu, Agnieszka Falenska and Jonas Kuhn</i>	
Advantages of the flux-based interpretation of dependency length minimization	98
<i>Sylvain Kahane and Chunxiao Yan</i>	
Length of non-projective sentences: A pilot study using a Czech UD treebank	110
<i>Jan Macutek, Radek Cech and Jiri Milicka</i>	
Gradient constraints on the use of Estonian possessive reflexives	118
<i>Suzanne Lesage and Olivier Bonami</i>	
What can we learn from natural and artificial dependency trees	125
<i>Marine Courtin and Chunxiao Yan</i>	

Invited Talk

Monday 26th August 2019

Dependency distance minimization: facts, theory and predictions

Ramon Ferrer-i-Cancho

Universitat Politècnica de Catalunya

Abstract

Quantitative linguistics is a branch of linguistics concerned about the study of statistical facts about languages and their explanation aiming at constructing a general theory of language. The quantitative study of syntax has become central to this branch of linguistics. The fact that the distance between syntactically related words is smaller than expected by chance in many languages led to the formulation of a dependency distance minimization (DDm) principle.

From a theoretical standpoint, DDm is in conflict with another word order principle: surprisal minimization (Sm). In single head structures, DDm predicts that the head should be put at the center of the linear arrangement, while Sm predicts that it should be put at one of the ends. In spite of the massive evidence of the action of DDm and the trendy claim that languages are optimized, attempts to quantify the degree of optimization of languages according to DDm have been rather scarce. Here we present a new optimality measure indicating that languages are optimized to a 70%. We confirm two old theoretical predictions: that the action of DDm is stronger in longer sentences and that DDm is more likely to be beaten by Sm in short sequences (resulting in an anti-DDm effect), while shedding new light on the kind of tree structures where DDm is more likely to be shadowed. Finally, we review various theoretical predictions of DDm focusing on the scarcity of crossing dependencies. We challenge the belief that formal constraints on dependency trees (e.g., projectivity or relaxed versions) are real rather than epiphenomenal.

The talk is a summary of joint work with Carlos Gomez-Rodriguez, Juan Luis Esteban, Morten Christiansen, Lluís Alemany-Puig and Xinying Chen.

Short bio

Ramon Ferrer-i-Cancho is associate professor at Universitat Politècnica de Catalunya and the head of the Complexity and Quantitative Linguistics Lab. He is a language researcher in a broad sense. His research covers different levels of the organization of life: from human language to animal behavior and down farther to the molecular level. One of his main research objectives is the development of a parsimonious but general theory of language and communication integrating insights from probability theory, information theory and the theory of spatial networks. In the context of syntax, he pioneered the study of dependency lengths from a statistical standpoint putting forward the first baselines and the principle of dependency distance minimization. He also introduced the hypothesis that projectivity, the scarcity of crossings dependencies and consistent branching are epiphenomena of that principle.

Information-theoretic locality properties of natural language

Richard Futrell

Department of Language Science
University of California, Irvine
rfutrell@uci.edu

Abstract

I present theoretical arguments and new empirical evidence for an information-theoretic principle of word order: information locality, the idea that words that strongly predict each other should be close to each other in linear order. I show that information locality can be derived under the assumption that natural language is a code that enables efficient communication while minimizing information-processing costs involved in online language comprehension, using recent psycholinguistic theories to characterize those processing costs information-theoretically. I argue that information locality subsumes and extends the previously-proposed principle of dependency length minimization (DLM), which has shown great explanatory power for predicting word order in many languages. Finally, I show corpus evidence that information locality has improved explanatory power over DLM in two domains: in predicting which dependencies will have shorter and longer lengths across 50 languages, and in predicting the preferred order of adjectives in English.

1 Introduction

The field of functional linguistics has long argued that the distinctive properties of natural language are best explained in terms of what makes for an efficient communication system under the cognitive constraints particular to human beings. The idea is that the properties of language are determined by the pressure to enable efficient communication while minimizing the information-processing effort required for language production and comprehension by humans.

Within that field, a particularly promising concept is the principle of **dependency length minimization** (DLM): the idea that words linked in syntactic dependencies are under a pressure to be close in linear order. DLM provides a single unified explanation for many of the word order properties of natural language: Greenberg's harmonic word order universals (Greenberg, 1963; Dryer, 1992; Hawkins, 1994; Hawkins, 2004; Hawkins, 2014) and exceptions to them (Temperley, 2007); the rarity of crossing dependencies (Ferrer-i-Cancho, 2006; Ferrer-i-Cancho and Gómez-Rodríguez, 2016), which correspond to deviations from context-free grammar (Kuhlmann, 2013); ordering preferences based on constituent length such as Heavy NP Shift (Wasow, 2002; Gildea and Temperley, 2010); and the statistical distribution of orders in treebank corpora (Liu, 2008; Futrell et al., 2015). See Liu et al. (2017) and Temperley and Gildea (2018) for recent reviews. The theoretical motivation for DLM is based on efficiency of language processing: the idea is that long dependencies tax the working memory capacities of speakers and listeners (Gibson, 1998; Gibson, 2000); in line with this view, there is observable processing cost in terms of reading time for long dependencies (Grodner and Gibson, 2005; Bartek et al., 2011).

At the same time, there have been attempts to derive the properties of human language formally from information-theoretic models of efficiency (Ferrer-i-Cancho and Solé, 2003; Ferrer-i-Cancho and Díaz-Guilera, 2007). But it is not yet clear how a principle such as DLM, which appears to be necessary for explaining the syntactic properties of natural language, would fit into these theories, or more generally into the information-theoretic view of language as an efficient code. The motivation for DLM is based on heuristic arguments about memory usage and on empirical results from studies of online processing, and it is not clear how to translate this motivation into the mathematical language of information theory.

Here I bridge this gap by providing theoretical arguments and empirical evidence for a new, information-theoretic principle of word order, grounded in empirical findings from the psycholinguistic literature and in the theory of communication in a noisy channel. I assume that linguistic speakers and listeners are processing language incrementally using lossy memory representations of linguistic context. Under these circumstances, we can derive a principle of **information locality**, which states that an efficient language will minimize the linear distance between elements with high **mutual information**, an information-theoretic measure of how strongly two words predict each other. Furthermore, assuming a particular probabilistic interpretation of dependency grammar (Eisner, 1996; Klein and Manning, 2004), I show that DLM falls out as an approximation to information locality. Finally, I present two new pieces of empirical evidence that information locality provides improved explanatory power over DLM in predicting word orders in corpora.

The remainder of the paper is structured as follows. Section 2 reviews relevant psycholinguistic results and information-theoretic models of online processing difficulty, concluding that they are inadequate for predicting word order patterns. Section 3 shows how to derive the principle of information locality from a modified model of online processing difficulty, and how DLM can be seen as a special case of information locality. In Section 4, I give corpus evidence that information locality makes correct predictions in two cases where DLM makes no predictions: in predicting the distance between words in dependencies in general across 50 languages, and in predicting the relative order of adjectives in English.

2 Background: Efficient communication under information processing constraints

I am interested in the question: What would a maximally efficient communication system look like, subject to human information processing constraints? To answer this question, we need a model of those information processing constraints. Here I review a leading theory of information processing constraints operative during on-line language comprehension, called **surprisal theory** (Hale, 2001; Levy, 2008; Hale, 2016), which is mathematically grounded in information theory, and discuss the relevance of surprisal theory for word order patterns in languages. Perhaps surprisingly, it turns out surprisal theory has very little to say about word order, which will necessitate an update to the theory described in Section 3.

Surprisal theory holds that the incremental processing difficulty for a word w given preceding context c (comprising the previous words as well as extra-linguistic context) is proportional to the **surprisal** of the word given the context:

$$\text{Difficulty}(w|c) \propto -\log p(w|c), \quad (1)$$

where the surprisal is measured in bits when the logarithm is taken to base 2. This quantity is also interpretable as the **information content** of the word in context. It indicates the extent to which a word is unpredictable in context. Under surprisal theory, the average processing difficulty per word in language is proportional to the **entropy rate** of the language: the average surprisal of each word given an unbounded amount of context information.

There are multiple convergent theoretical motivations for surprisal theory (Levy, 2013), and it is in line with recent theories of information processing difficulty from robotics, artificial intelligence, and neuroscience in that it proposes a certain amount of cost per bit of information processed (Friston, 2010; Tishby and Polani, 2011; Genewein et al., 2015).

Surprisal theory also has excellent empirical coverage of psycholinguistic data: for example, taking word-by-word reading times as a measure of processing difficulty, Smith and Levy (2013) find that empirically observed reading times in naturalistic text are a robustly linear function of surprisal over 8 orders of magnitude. Levy (2008) shows that surprisal theory can explain many diverse phenomena studied in the previous psycholinguistic literature.

The fact that processing time is a linear function of surprisal will be important for deriving predictions about word order: it tightly constrains theories about the interactions of word order and processing difficulty. In fact, surprisal theory in the form of Eq. 1 leads to the prediction that the average processing difficulty per word is not at all a function of the word order rules of a language, provided that different word order rules do not affect the entropy rate of the language. To see this, consider a sentence of n

words w_1, \dots, w_n in some language L . The total information-processing difficulty for comprehending this sentence ends up being equal to the quantity of information content of the sentence in the language:

$$\begin{aligned}
\text{Difficulty}(w_1, \dots, w_n) &= \sum_{i=1}^n \text{Difficulty}(w_i | w_1, \dots, w_{i-1}) & (2) \\
&\propto \sum_{i=1}^n -\log p_L(w_i | w_1, \dots, w_{i-1}) \\
&= -\log \prod_{i=1}^n p_L(w_i | w_1, \dots, w_{i-1}) \\
&= -\log p_L(w_1, \dots, w_n). & (3)
\end{aligned}$$

Now let us consider how this sentence might look in another language L' with other rules for ordering words. As long as the total probability of the sentence in L' is the same as the equivalent sentence in L —regardless of the order of words—the predicted processing difficulty for the sentence is the same. For example, maybe L is English and L' is reversed-English: a language which is identical to English except that all sentences are reversed in order. Then the English sentence w_1, \dots, w_n would come out as the reversed-English sentence w_n, w_{n-1}, \dots, w_1 , with the same total probability and thus exactly the same predicted processing difficulty under surprisal theory.

The general expressed by Eq. 3 is that, under surprisal theory, the word order patterns of a language do not affect the overall processing difficulty of the language unless they increase or decrease the average total surprisal of sentences of the language, or in other words the entropy over sentences in a language. The predicted processing difficulty is not affected by word order rules except inasmuch as they decrease the entropy over sentences (by introducing ambiguities) or increase the entropy over sentences (by removing ambiguities) (Levy, 2005; Futrell, 2017). Essentially, all that surprisal theory has to say about word order is that less frequent orders within a language are more costly.

This invariance to order is problematic for theories that have attempted to explain word order patterns in terms of maximizing the predictability of words (Gildea and Jaeger, 2015; Ferrer-i-Cancho, 2017): such theories have derived predictions about word order by introducing auxiliary assumptions. For example, Gildea and Jaeger (2015) show that word order rules in languages minimize surprisal as calculated from a trigram model, rather than a full probability model; this ends up being a special case of the theory we advocate below in Section 3. Ferrer-i-Cancho (2017) implicitly assumes that the predictability of the verb is more impactful for processing difficulty than the predictability of other words, such that orders that minimize the surprisal of the verb are favorable.

There are at least two general ways to modify surprisal theory to break its order-invariance. The first would be to posit that processing difficulty is some non-linear function of surprisal. This route is not attractive, because the current state of empirical knowledge is that processing time is determined linearly by surprisal (Smith and Levy, 2013). The second way of modifying surprisal theory would be to posit that the relevant probability distribution of words given contexts does not take into account full information from the context, or is distorted in some way relative to the true distribution of words given contexts. As we will see below, this solution allows us to derive information locality.

3 Lossy-context surprisal and information locality

I propose to modify surprisal theory in the manner described in Futrell and Levy (2017). The contents of this section are a simplified exposition of the derivations presented in that paper.

In the modified surprisal theory, the predicted processing difficulty per word w is a function of the word’s expected log probability given a *lossy* or *noisy* **memory representation** m of the context c . That is:

$$\text{Difficulty}(w|c) \propto \mathbb{E}_{m|c} [-\log p(w|m)], \quad (4)$$

where $m|c$ indicates the conditional distribution of lossy memory representations given contexts, called the **memory encoding function**. I call this model **lossy-context surprisal**, because the predicted processing difficulty depends on a lossy memory m , rather than the objective context c . In general, due to the Data Processing Inequality (Cover and Thomas, 2006), m can be seen as a representation of c to which noise has been added. Taking c to be the sequence of word tokens leading up to a given token w_i , we can write Eq. 4 as:

$$\text{Difficulty}(w_i|w_{1:i-1}) \propto \mathbb{E}_{m|w_{1:i-1}} [-\log p(w_i|m)], \quad (5)$$

where $w_{1:i-1}$ denotes the sequence of words from index 1 to index $i - 1$ inclusive.

Unlike plain surprisal theory, lossy-context surprisal predicts that some systems of word order rules will result in more processing efficiency than others. In particular, it predicts locality effects (Gibson, 1998; Gibson, 2000) in the form of **information locality**: there will be difficulty when elements that have high mutual information are distant from each other in linear order. The basic intuition is that, when two elements that predict each other *in principle* are separated in time, they will not be able to predict each other *in practice* because by the time the processor gets to the second element, the first one has been partially forgotten. The result is that the second element is less predictable than it could have been, causing excess processing cost.

3.1 Derivation of information locality

Assume that the memory encoding function $m|c$ is structured such that some proportion of the information available in a word is lost depending on how long the word has been in memory. For a word which has been in memory for one timestep, the proportion of information which is lost is a constant e_1 ; for a word which has been in memory for two timesteps, the proportion of information lost is e_2 ; in general for a word which has been in memory for t timesteps, the proportion of information lost is e_t . Assume further that e_t is monotonically increasing in t : i.e. $t < \tau$ implies $e_t \leq e_\tau$. This process reflects the fact that information in a memory representation can only become degraded over time, in the spirit of the Data Processing Inequality (Cover and Thomas, 2006).

This memory model is equivalent to assuming that the context is subject to **erasure noise**, a commonly used noise model in information theory (Cover and Thomas, 2006). In erasure noise, a symbol x is stochastically *erased* (replaced with a special erasure symbol \mathbb{E}) with some probability e . The noise model here further assumes that the erasure rate increases with time: I call this noise model **progressive erasure noise**.

I will now show that subjective surprisal, under the assumption of progressive erasure noise, gives rise to information locality.

Under progressive erasure noise, the context $w_{1:i-1}$ can be represented as a sequence of symbols $m_{1:i-1}$. Each symbol m_j , called a **memory symbol**, is equal either to the context word w_j or to the erasure symbol \mathbb{E} . The surprisal of a word w_i given the memory representation $m_{1:i-1}$ can be written in two terms:

$$-\log p(w_i|m_{1:i-1}) = -\log p(w_i) - \text{pmi}(w_i; m_{1:i-1}),$$

where $\text{pmi}(w_i; m_{1:i-1}) = \log \frac{p(w_i|m_{1:i-1})}{p(w_i)}$ is the **pointwise mutual information** (Fano, 1961; Church and Hanks, 1990) of the word and the memory representation, giving the extent to which the particular memory representation predicts the particular word. We can now use the chain rule to break the pointwise

mutual information into separate terms, one for each symbol in the memory representation:

$$\begin{aligned}
\text{pmi}(w_i; m_{1:i-1}) &= \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j | m_{1:j-1}) \\
&= \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) - \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j; m_{1:j-1}) \\
&= \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) - R,
\end{aligned} \tag{6}$$

where $\text{pmi}(x; y; z)$ is the three-way pointwise **interaction information** of three variables (Bell, 2003), indicating the extent to which the conditional $\text{pmi}(w_i; m_j | m_{1:j-1})$ differs from the unconditional $\text{pmi}(w_i; m_j)$. These higher-order interaction terms are then grouped together in a term called R .

Now substituting Eq. 6 into Eq. 5, we get an expression for processing difficulty in terms of the pmi of each memory symbol with the current word:

$$\begin{aligned}
\text{Difficulty}(w_i | w_{1:i-1}) &\propto \mathbb{E}_{m|w_{1:i-1}} [-\log p(w_i | m)] \\
&= \mathbb{E}_{m|w_{1:i-1}} \left[-\log p(w_i) - \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) + R \right] \\
&= -\log p(w_i) - \mathbb{E}_{m|w_{1:i-1}} \left[\sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) + R \right] \\
&= -\log p(w_i) - \sum_{j=1}^{i-1} \mathbb{E}_{m_j|w_j} [\text{pmi}(w_i; m_j)] + \mathbb{E}_{m|w_{1:i-1}} [R].
\end{aligned} \tag{7}$$

It remains to calculate the expected pmi of the current word and a memory symbol given the distribution of possible memory symbols. Recall that each m_j is either equal to the erasure symbol \mathbb{E} (with probability e_{i-j}) or to the word w_j (with probability $1 - e_{i-j}$). If $m_j = \mathbb{E}$, then $\text{pmi}(w_i; m_j) = 0$; otherwise $\text{pmi}(w_i; m_j) = \text{pmi}(w_i; w_j)$. Therefore the expected pmi between a word w_i and a memory symbol m_j is $(1 - e_{i-j})\text{pmi}(w_i; w_j)$. The effect of erasure noise on the higher-order terms collected in R is more complicated, but in general will have the effect of reducing their value, because a higher-order interaction information term will have a value of 0 if any single variable in it is erased. Therefore we can write the expected processing difficulty per word as:

$$\text{Difficulty}(w_i | w_{1:i-1}) \propto -\log p(w_i) - \sum_{j=1}^{i-1} (1 - e_{i-j})\text{pmi}(w_i; w_j) + o(R), \tag{8}$$

where $o(R)$ indicates a value that is upper-bounded by R . Assuming the higher-order terms $o(R)$ are negligible, then the expected processing difficulty as a function of word order is purely determined by the expression

$$-\sum_{j=1}^{i-1} (1 - e_{i-j})\text{pmi}(w_i; w_j). \tag{9}$$

As words w_i and w_j become more distant from each other, the value of the survival probability $(1 - e_{i-j})$ must decrease, so the value of (9) must increase, such that the theory predicts increased processing difficulty in proportion to the pairwise pmi between w_i and w_j .

In general, Eq. 8 holds that processing difficulty as a function of word order increases monotonically as elements with high pointwise mutual information are separated in linear order.¹ It will be minimized when elements with the highest pointwise mutual information are closest to each other. If word orders are shaped by a pressure for processing efficiency, then information locality comes out to a kind of attraction between words with high pmi.²

3.2 DLM as an approximation to information locality

The principle of information locality holds that groups of words with high mutual information will tend to be close to each other, in order to maximize online processing efficiency. I wish to argue that this result subsumes the principle of dependency length minimization (DLM), which holds that all words in syntactic dependencies will tend to be close to each other. This connection requires a linking hypothesis: that syntactic dependencies correspond to the word pairs with high mutual information within a sentence. I call this hypothesis the **Head-Dependent Mutual Information (HDMI) hypothesis**.

There are good theoretical and empirical reasons to believe the HDMI hypothesis is true. The empirically-measured mutual information of words pairs in head-dependent relationships has been found to be greater than various baselines in Futrell and Levy (2017) across languages. Theoretically, it makes sense for word pairs in dependency relationships to have the highest mutual information because mutual information is a measure of the strength of covariance between two variables, and words in dependencies are by definition those word pairs whose covariance is directly constrained by grammatical rules. More formally, in distributions over dependency trees generated by head-outward generative models (Eisner, 1996; Klein and Manning, 2004), heads and dependents will have the highest mutual information of any word pairs (Futrell et al., 2019). The basic idea that dependencies correspond to high-mutual-information word pairs has a long history in computational linguistics (Resnik, 1996; de Paiva Alves, 1996; Yuret, 1998).

If we assume the strongest form of the HDMI hypothesis—that mutual information between words *not* linked in a dependency is completely negligible—then Eq. 8 implies that the expected processing cost for a sentence as a function of word order is a monotonically increasing function of dependency length, which is exactly the claim underlying DLM. This strong form of the HDMI hypothesis is surely false, but it shows how DLM can serve as an approximation to the predictions of information locality.

The notion of information locality is also linked to more general notions of complexity, such as the theory of statistical complexity (Crutchfield and Young, 1989), which apply to any stochastic process. The **statistical complexity** of a process is the entropy of the maximally compressed representation of the past of a process required to predict the future of the process with optimal accuracy. Among processes with the same entropy rate, processes with poor information locality properties (where elements with high mutual information are far from each other) will have higher statistical complexity, because each bit of predictive information will need to be retained in memory over more timesteps. If we view DLM as a special case of information locality, then that means that minimizing dependency length has the effect of lowering statistical complexity. Thus it may be the case that the word order properties of human language are a very general consequence of minimization of statistical complexity.

¹If we include the effects of the higher-order terms collected in R , then Eq. 7 also implies that processing difficulty will increase when groups of elements with high interaction information are separated from each other in time. Here "high interaction information" refers to a large positive value in the case of even-cardinality groups of elements, and a large negative value in the case of odd-cardinality groups of elements. See Bell (2003) for the relevant technical details on interaction information.

²If words are under a pressure to be close as a function of their pmi, then this raises the question of what is to be expected for nonce and novel words, for which no corpus co-occurrence statistics are available. This issue was raised as an objection to information locality by Dyer (2017). While the probabilities that go into practically calculating pmi come from corpora, the probabilities that are truly important from the perspective of processing difficulty are the listener's subjective probabilities, which are only approximated by corpus-derived probabilities (Smith and Levy, 2011). A listener encountering a nonce word will have some hypothesis about its syntax and meaning, which means that the listener will have expectations about what words the nonce word will co-occur with, and thus the nonce word will have a nonzero (subjective) pmi value with other words for the listener. In an experimental study, the pmi values for nonce words could be measured using techniques such as the Cloze task (Taylor, 1953), which measures these subjective probabilities.

4 Information locality beyond DLM

Here I give new empirical evidence that natural languages exhibit information locality in a way that goes beyond the predictions of DLM.

4.1 Strength of locality effect for different dependencies

DLM predicts that all words in dependencies will be under a pressure to be close to each other, but it does not make any predictions about *which* dependencies will be under especially strong pressure. However, empirically, DLM effects in word order preferences and also in online processing difficulty show asymmetries based on the details about particular dependencies (Stallings et al., 1998; Demberg and Keller, 2008).

Here I propose that the strength of attraction between two words linked in a dependency is modulated by the pointwise mutual information of the two words, as predicted by information locality.

Language	β_{pmi}	p value	Language	β_{pmi}	p value
Ancient Greek	-0.18	< .001	Japanese	-0.32	<.001
Arabic	-0.26	<.001	<i>Kazakh</i>	<i>-1.18</i>	<i>0.01</i>
Basque	-0.22	<.001	Korean	-0.14	<.001
Belarusian	-0.20	<.001	Latin	-0.18	<.001
Bulgarian	-0.29	<.001	Latvian	-0.32	<.001
Catalan	-0.29	<.001	Lithuanian	-0.41	<.001
Church Slavonic	-0.23	<.001	Mandarin	-0.19	<.001
Coptic	-0.35	<.001	Modern Greek	-0.25	<.001
Croatian	-0.32	<.001	Norwegian	-0.37	<.001
Czech	-0.27	<.001	Persian	-0.19	<.001
Danish	-0.38	<.001	Polish	-0.23	<.001
Dutch	-0.10	<.001	Portuguese	-0.23	<.001
English	-0.38	<.001	Romanian	-0.36	<.001
Estonian	-0.32	<.001	Russian	-0.18	<.001
Finnish	-0.29	<.001	<i>Sanskrit</i>	<i>0.10</i>	<i>0.28</i>
French	-0.33	<.001	Slovak	-0.30	<.001
Galician	-0.35	<.001	Slovenian	-0.38	<.001
German	-0.25	<.001	Spanish	-0.37	<.001
Gothic	-0.19	<.001	Swedish	-0.35	<.001
Hebrew	-0.21	<.001	Tamil	-0.18	<.001
Hindi	-0.26	<.001	Turkish	-0.22	<.001
Hungarian	-0.11	<.001	Ukrainian	-0.29	<.001
Indonesian	-0.22	<.001	Urdu	-0.22	<.001
Irish	-0.37	<.001	<i>Uyghur</i>	<i>-0.04</i>	<i>0.79</i>
Italian	-0.35	<.001	Vietnamese	-0.27	<.001

Table 1: Regression coefficients predicting dependency length as a function of pmi between head and dependent. A negative sign indicates that words with higher pmi are closer to each other. Languages where the effect is not significant at $p < .001$ are in italics.

I tested this hypothesis in 50 languages of the Universal Dependencies 2.1 treebanks (Nivre et al., 2017). I excluded all punctuation and root dependencies, and collapsed all strings of words linked by “flat”, “fixed”, and “compound” dependencies (which indicate multiword expressions) into single tokens. For each word pair $r = (h, d)$ in a head–dependent relationship, I fit a linear regression model to predict the distance between the two words y_r from their pmi:

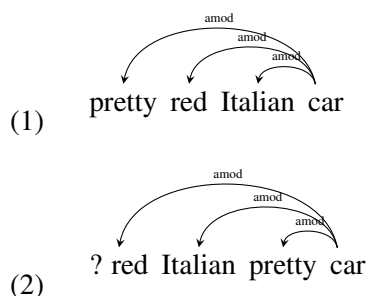
$$y_r = \beta_0 + \beta_{\text{pmi}}\text{pmi}(h; d) + S_i + S_{i,\text{pmi}}\text{pmi}(h; d) + \epsilon_r, \quad (10)$$

where S_i and $S_{i,\text{pmi}}$ are by-sentence random intercepts and slopes subject to L_2 regularization, making this a mixed-effects regression model (Gelman and Hill, 2007; Baayen et al., 2008; Barr et al., 2013). These extra terms account for any per-sentence idiosyncratic behavior of dependencies (for example, effects of sentence length). The key coefficient is β_{pmi} which, if significantly negative, indicates that words with high pmi are especially attracted to each other. The pmi values were calculated between part-of-speech tags rather than wordforms in order to avoid data sparsity issues in the estimation of mutual information. For computational tractability, I include only at most 10,000 sentences per language and exclude sentences of length greater than 20 words.³

Table 1 shows the values of β_{pmi} and their significance. In all except 3 languages, I find the significant negative effect at $p < .001$, indicating information locality effects beyond DLM. The effect size is relatively stable across languages. In particular, the average value of β_{pmi} across languages is around -0.3 , with standard error 0.02, indicating that for every bit of pmi between parts-of-speech, words in dependencies are about 0.3 words closer together on average, robustly across languages.

4.2 Adjective order

Speakers of many languages show robust, stable patterns of preferences in terms of how they order attributive adjectives that simultaneously modify a noun (Dixon, 1982; Scontras et al., 2017; Scontras et al., 2019). For example, English speakers generally prefer the order in Example 1 over 2, or they perceive Example 2 as expressing a different meaning which is marked relative to the first. As the dependency structures show, the classical theory of DLM would not make any predictions for the relative ordering of these adjectives, as all are in equivalent dependency relationships with the head noun. Classical syntactic theories of adjective order have assumed that adjectives can be sorted into semantic classes (e.g., value, color, nationality) and that there is a universal order of semantic classes in terms of which adjectives go closer to the noun (e.g., adjectives are placed close to the noun in the order nationality > color > value) (Cinque and Rizzi, 2008).



Here I suggest that the preferred order of adjectives is determined by information locality: that is, adjectives with higher mutual information with a noun go closer to that noun.

Previous work has shown that the best empirical predictor of adjective order is the rating of **subjectivity** given to adjectives by experimental participants, with more subjective adjectives going farther out from the head noun (Scontras et al., 2017; Scontras et al., 2019), but this work did not compare predictions with mutual information. Simultaneously, Kirby et al. (2018) compared size adjectives and color adjectives—where color adjectives are preferred to be farther out from the noun in English—and found that color adjectives have lower pmi with the noun than size adjectives.

Here I compare subjectivity and mutual information as predictors of adjective order in large corpora of English. For subjectivity ratings, I use the data from Scontras et al. (2017). For co-occurrence and order data, I use the Google Syntactic n -grams corpus (Goldberg and Orwant, 2013). From this corpus, I collect all cases of two adjectives modifying a single following noun with relation type *amod*, where an adjective counts for inclusion if its part of speech is JJ, JJR, or JJRS and it is listed as an adjective in the CELEX database (Baayen et al., 1995), and a noun counts for inclusion if its part of speech is NN or NNS and it is listed as a noun in CELEX. The result is a dataset of adjective–adjective–noun (AAN) triples, containing 1,604,285 tokens and 16,681 types.

³The code for this analysis is available online at <http://github.com/langprocgroup/cliqs>.

Subjectivity	PMI	Both
68.4%	66.9%	72.9%

Table 2: Accuracy of subjectivity and pmi as predictors of adjective order in logistic regressions, for held-out types of adjective–adjective–noun triples.

For the calculation of pointwise mutual information, we need the conditional probability of adjectives given nouns. I find this probability by maximum likelihood estimation, collecting all instances of single adjectives modifying a following noun, by the same criteria as above.

I tested the relationship between pmi, subjectivity, and order statistically using logistic regression. Given two adjectives preceding a noun, the question is which of the two is closer to the noun, as a function of the *difference* in pmi or subjectivity between the two. Given an observed pair of adjectives (A_1, A_2), ordered alphabetically, I fit a logistic regression to predict whether the alphabetically-second adjective A_2 is ordered closer to the noun N in the corpus, as a function of the pmi and subjectivity difference between A_1 and A_2 . The regression equation is:

$$\log \frac{p(A_2 \text{ closer to } N)}{p(A_1 \text{ closer to } N)} = \beta_0 + \beta_S(S(A_1) - S(A_2)) + \beta_{\text{pmi}}(\text{pmi}(A_1; N) - \text{pmi}(A_2; N)) + \epsilon,$$

where $S(A)$ is the subjectivity rating of adjective A . This regression setup was used to predict order data in Morgan and Levy (2016). A positive coefficient β_S indicates that the adjective with greater subjectivity is likely to be farther from the noun. A negative coefficient β_{pmi} indicates that an adjective with greater pmi with the noun is likely to be closer to the noun.⁴

To evaluate the accuracy of subjectivity and information locality as theories of adjective order, I separated out 10% of the AAN types as a test set (1,668 types), with the remaining types forming a training set (15,013 types). I fit the logistic regression to the token-level data for the AAN types in the training set, a total of 1,473,269 tokens. I find $\beta_0 = -0.7$, $\beta_S = 14.1$, and $\beta_{\text{pmi}} = -0.6$, with all coefficients significant at $p < .001$. The results mean that for each bit of pmi between an adjective A and a noun, beyond the pmi of the other adjective with the noun and controlling for subjectivity, the log-odds that A is closer to the noun increase by .6.⁵

I used the held-out AAN triples to test how well subjectivity and pmi would generalize when predicting adjective order in unseen data (a total of 131,016 tokens). Table 2 shows test-set accuracy of logistic regressions predicting adjective order using subjectivity, pmi, or both as predictors. Subjectivity and pmi have roughly equal accuracy on the held-out types, with pmi slightly lower. The highest accuracy is achieved when both subjectivity and pmi are included as predictors. This result shows that mutual information has predictive value for adjective order beyond what is accounted for by subjectivity.

The regression shows that both subjectivity and mutual information are good predictors of adjective order, so the question arises of whether the two predictors make overlapping or divergent predictions. For 53% of the test-set tokens, subjectivity and pmi make the same (correct) prediction. I qualitatively examined cases where pmi got the order right and subjectivity did not, and found that these usually consist of cases where two adjectives are in the same semantic class, and yet strong ordering preferences exist in the corpus, such as *big long beard* (preferred) vs. *long big beard* (dispreferred).

The results here do not adjudicate between subjectivity and mutual information as better predictors of adjective order. The two may be independent factors predicting adjective order—a hypothesis explored by Hahn et al. (2018)—or they may be related. I posit that subjectivity and mutual information are conceptually related. The reasoning is: if an adjective is more subjective, then its applicability to any given noun is determined by some external factor outside than the noun itself—the speaker’s subjective state. In contrast, the applicability of a less subjective adjective is more strongly determined by the noun itself due to the inherent properties of the noun. Mutual information is calculated from co-occurrence

⁴The code for this analysis is available online at <http://github.com/langprocgroup/adjorder>.

⁵The values of subjectivity tend to be smaller than the values of pmi (average subjectivity of adjectives in the corpus is 0.5; average pmi is 2.3), so the larger coefficient β_S should not necessarily be interpreted as meaning that the effect of subjectivity is larger.

statistics, where the speaker’s subjective state is unknown and therefore appears as a noise variable affecting the distribution of adjectives. So from the perspective of co-occurrence statistics, the distribution of more subjective adjectives is noisier, and therefore has less mutual information with the head noun. The relationship between subjectivity, mutual information, and adjective order may be the following: subjectivity determines the joint distribution of adjectives and nouns, which in turn dictates the mutual information, which then determines the preferred order via the principle of information locality.

In support of this idea, I found that the subjectivity score for an adjective is moderately anticorrelated with its average pmi with nouns at $r = -.32$, Spearman’s $\rho = -.35$; the relationship between the two is shown in Figure 1. Note that the estimates of mutual information obtained from corpora are noisy: it is notoriously difficult to estimate quantities such as mutual information from count data (Paninski, 2003). Better estimates of mutual information, obtained through more data or more sophisticated estimation techniques, may show stronger correlations with subjectivity and with adjective order.

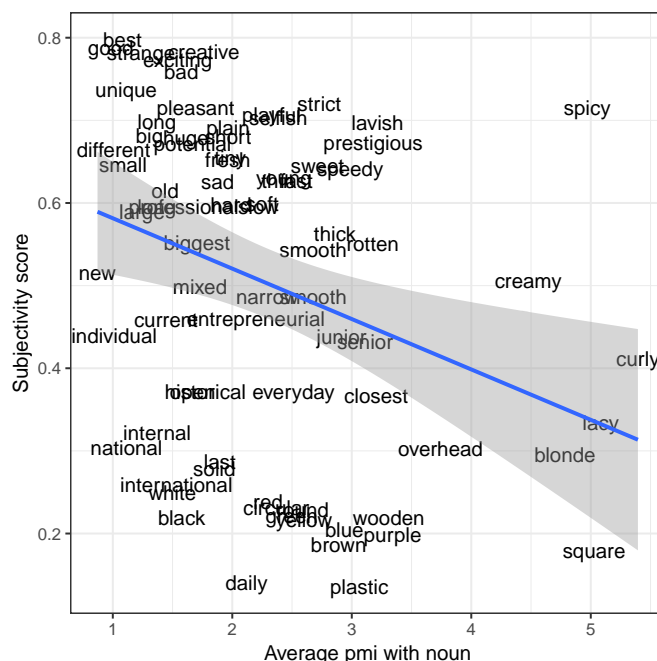


Figure 1: Relationship between adjective subjectivity score and average pmi with nouns in Google Syntactic n -Grams corpus.

5 Conclusion

I presented a theoretical argument that, if languages are organized for efficient communication subject to human information processing constraints, then they will have the property of information locality: words that predict each other will appear close to each other in time. I presented two pieces of novel evidence in favor of information locality over previous theories of word order. I believe the principle of information locality will enrich the growing link between theories of syntax and notions of processing efficiency. By deriving the principle of dependency length minimization in an information-theoretic setting and demonstrating improved predictive power over simple dependency length minimization, it opens the way for unified information-theoretic models of human language.

Acknowledgments

I thank Roger Levy, Greg Scontras, and Ted Gibson for many conversations on these topics and the anonymous reviewers for helpful comments on the paper.

References

- R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania.
- R. Harald Baayen, D.J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.
- Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.
- B. Bartek, Richard L. Lewis, Shravan Vasishth, and M. R. Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178–1198.
- Anthony J. Bell. 2003. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 921–926.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Guglielmo Cinque and Luigi Rizzi. 2008. The cartography of syntactic structures. *Studies in linguistics*, 2:42–58.
- Thomas M. Cover and J. A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.
- James P Crutchfield and Karl Young. 1989. Inferring statistical complexity. *Physical Review Letters*, 63(2):105.
- Eduardo de Paiva Alves. 1996. The selection of the most probable dependency structure in Japanese using mutual information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 372–374.
- Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.
- Robert M. W. Dixon. 1982. *Where have all the adjectives gone? And other essays in semantics and syntax*. Mouton, Berlin, Germany.
- Matthew S Dryer. 1992. The Greenbergian word order correlations. *Language*, 68(1):81–138.
- William E. Dyer. 2017. *Minimizing integration cost: A general theory of constituent order*. Ph.D. thesis, University of California, Davis, Davis, CA.
- Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 340–345.
- Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communication*. MIT Press, Cambridge, MA.
- Ramon Ferrer-i-Cancho and Albert Díaz-Guilera. 2007. The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06009.
- Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.
- Ramon Ferrer-i-Cancho and R.V. Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788.
- Ramon Ferrer-i-Cancho. 2006. Why do syntactic links not cross? *Europhysics Letters*, 76(6):1228.
- Ramon Ferrer-i-Cancho. 2017. The placement of the head that maximizes predictability: An information theoretic approach. *Glottometrics*, 39:38–71.
- Karl Friston. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127.
- Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

- 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, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (DepLing 2019)*.
- Richard Futrell. 2017. *Memory and locality in natural language*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
- A. Gelman and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.
- Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun. 2015. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- E. Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O’Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126.
- Daniel Gildea and T. Florian Jaeger. 2015. Human languages order information efficiently. *arXiv*, 1510.02823.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.
- Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 241–247.
- Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.
- Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2):261–290.
- Michael Hahn, Judith Degen, Noah Goodman, Dan Jurafsky, and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*.
- John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*, pages 1–8.
- John T. Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.
- John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press, Cambridge.
- John A. Hawkins. 2004. *Efficiency and complexity in grammars*. Oxford University Press, Oxford.
- John A. Hawkins. 2014. *Cross-linguistic variation and efficiency*. Oxford University Press, Oxford.
- Simon Kirby, Jennifer Culbertson, and Marieke Schouwstra. 2018. The origins of word order universals: Evidence from corpus statistics and silent gesture. In *The Evolution of Language: Proceedings of the 12th International Conference (Evolangxii)*. NCU Press.
- Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 478.
- Marco Kuhlmann. 2013. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387.
- Roger Levy. 2005. *Probabilistic Models of Word Order and Syntactic Discontinuity*. Ph.D. thesis, Stanford University, Stanford, CA.

- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In Roger P. G. van Gompel, editor, *Sentence Processing*, page 78–114. Hove: Psychology Press.
- Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- Emily Morgan and Roger Levy. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:382–402.
- Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phê Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Mackentanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Măranduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lê Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Östling, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamel Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uribe, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Tak-sum Wong, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.
- Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.
- Gregory Scontras, Judith Degen, and Noah D. Goodman. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, 1(1):53–65.
- Gregory Scontras, Judith Degen, and Noah D. Goodman. 2019. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*.
- Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

- Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.
- Lynne M. Stallings, Maryellen C. MacDonald, and Padraig G. O’Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417.
- Wilson L. Taylor. 1953. “Cloze procedure”: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.
- David Temperley and Dan Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15.
- David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.
- Naftali Tishby and Daniel Polani. 2011. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer.
- Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications, Stanford, CA.
- Deniz Yuret. 1998. *Discovery of linguistic relations using lexical attraction*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand?

Jianwei Yan
Department of Linguistics
Zhejiang University, China
jwyan@zju.edu.cn

Haitao Liu
Department of Linguistics
Zhejiang University, China
lhtzju@gmail.com

Abstract

This paper investigates which annotation scheme of dependency treebank is more congruent for the measurement of syntactic complexity and cognitive constraint of language materials. Two representatives of semantic- and syntactic-oriented annotation schemes, the Universal Dependencies (UD) and the Surface-Syntactic Universal Dependencies (SUD), are under discussion. The results show that, on the one hand, natural languages based on both annotation schemes follow the universal linguistic law of Dependency Distance Minimization (DDM); on the other hand, according to the metric of Mean Dependency Distances (MDDs), the SUD annotation scheme that accords with traditional dependency syntaxes are more expedient to measure syntactic difficulty and cognitive demand.

1 Background and Motivation

Dependency grammar deals with the syntactically related words, i.e. the governor and the dependent, within sentence structure (Hinger, 1993; Hudson, 1995; Liu, 2009). It can be dated back to the seminal work of *Eléments de Syntaxe Structurale* by Tesnière (1959), and developed through different theories, including Word Grammar, Meaning-Text-Theory, Lexicase, etc. (e.g. Hudson, 1984; Mel'čuk, 1988; Starosta, 1988; Eroms, 2000). Thus far, there are many representations of dependency grammar. **Figure 1** displays two typical dependency representations of one sample sentence *We walked along the lake*.

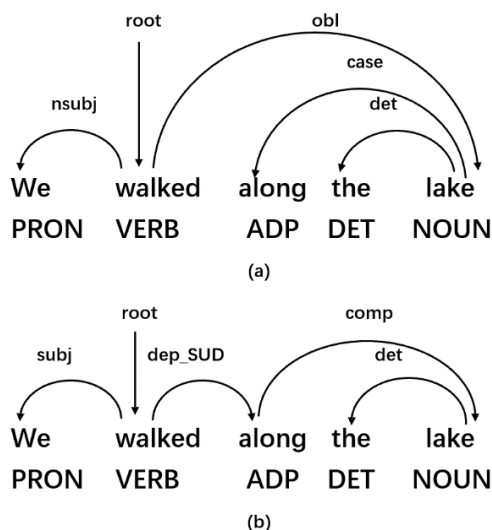


Figure 1. Dependency Representations of One English Sentence *We walked along the lake* Based on UD and SUD Annotation Schemes.

The dependency representation based on the Universal Dependencies (UD¹), as shown in **Figure 1 (a)**, is one of the most eminent models by now under the framework of dependency grammar. It attempts at establishing a multilingual morphosyntactic scheme to annotate various languages in a consistent manner (Nivre, 2015; Osborne and Gerdes, 2019). Thus, the UD annotation scheme holds a semantic over

¹See also <http://universaldependencies.org/>.

syntactic criteria to put priorities to content words to maximize “crosslinguistic parallelism” (Nivre, 2015; de Marneffe and Nivre, 2019). On the contrary, the Surface-Syntactic Universal Dependencies (SUD²) annotation scheme, as shown in **Figure 1 (b)**, follows the syntactic criteria to define not only the dependency labels but also the dependency links. It aims to make the annotation scheme close to the dependency traditions, like Meaning-Text-Theory (MTT) (Mel’čuk, 1988), Word Grammar (Hudson, 1984), etc. Hence, the SUD annotation scheme is a syntactic-oriented dependency representation that seeks to promote the syntactic motivations (Gerdes et al., 2018; Osborne and Gerdes, 2019). Therefore, the UD and SUD annotation schemes signify two typical preferences of dependency grammar, one is semantic-oriented, and the other is syntactic-oriented.

As shown in **Figure 1**, the linear sentence in both representations can be divided into several words; and the labelled arcs, directed from the governors to the dependents, represent different dependency types indicating the syntactic relations between elements within the sentence. Hence, the dependency representations indicate both the functional role of each word as well as the syntactic relations between different elements. More importantly, based on dependency representations, linguists have proposed several measurements for linguistic analysis. For one thing, **dependency distance** is defined as the linear distance of the governor and the dependent (Hudson, 1995). For another, the linear order of the governor and the dependent of each dependency type is referred to as **dependency direction** (Liu, 2010). When a governor appears before a dependent, the dependency direction is governor-initial or negative. Otherwise, it is governor-final or positive. For instance, in **Figure 1 (a)**, the arc above the dependent *we* and the governor *walked* forms a governor-final relation; and the dependency distance between these two elements is $2 - 1 = 1$ (the number 2 and 1 in the subtraction represent the linear order of the governor and dependent, respectively). Detailed calculating method will be shown in **Section 2**. Therefore, the dependency representations and the measures of dependency relations are both explicit and clear-cut. This explains the reason why dependency treebanks, i.e. corpora with annotations (Abeillé, 2003), are widespread among linguists in big-data era. As a result, the variations and universals of human languages are explored and unveiled through statistical and mathematical tools (Hudson, 1995; Liu et al., 2017). What is noteworthy is that previous studies have shown that **dependency distance** is an important indicator in demonstrating the notion of syntactic complexity and cognitive demand (Hudson, 1995; Gibson, 2000; Liu, 2008).

Under the framework of dependency grammar, Hudson (1995) characterized the definition of dependency distance based on the theories of memory decaying and short-term memory (e.g. Brown, 1958; Levy et al., 2013). The notion of syntactic difficulty and cognitive demand have been subsequently related to the linear distance between the governors and the dependents in cognitive science (Gibson, 1998; Hawkins, 2004). Based on a Romanian dependency treebank, Ferrer-i-Cancho (2004) hypothesized and proved that the mean distance of a sentence is minimized and constrained. These paved the way for Liu’s (2008) empirical study on dependency distance which provides a viable treebank-based approach towards the metric of syntactic complexity and cognitive constraint. Afterwards, series of studies exploring the relationship between dependency distance and syntactic and cognitive benchmarks have been conducted (e.g. Jiang and Liu, 2015; Wang and Liu, 2017; Liu et al., 2017). These studies share some similarities. First, it is well-corroborated that the frequency of dependency distance decreases with the increase of the dependency distance, viz., the distribution of dependency distance follows the linguistic law of the Least Effort Principle (LEP) or Dependency Distance Minimization (DDM) (Zipf, 1965; Liu et al., 2017). Second, it is believed that the greater the dependency distance is, the more difficult the sentence structure (Gibson, 1998; Hiranuma, 1999; Liu et al., 2017). Thus, the arithmetic average of all dependency distances of one sentence or a treebank or the **mean dependency distances** (MDDs) (Liu, 2008) has been an important index of memory burden, demonstrating the syntactic complexity and cognitive demand of the language concerned (Hudson, 1995; Liu et al., 2017).

Previous studies have shown that there are several factors that have effects on the measurement of dependency distance of a sentence, including sentence length, genre, chunking, language type, grammar, annotation scheme and so forth (e.g. Jiang and Liu, 2015; Wang and Liu, 2017; Lu et al., 2016; Hiranuma, 1999; Liu and Xu, 2012; Gildea and Temperley, 2010). Most of these factors have been well-investigated, however, the factor of annotation scheme has rarely been studied. Liu et al. (2009), for instance, investigated Chinese syntactic and typological properties based on five different Chinese

²See also <https://gitlab.inria.fr/grew/SUD>.

treebanks with different genres and annotation schemes, yet the treebanks adopted with different annotation schemes were used to avoid the corpus influences to ensure a reliable conclusion. Hence, the question as to the effects of annotation scheme on the distribution of dependency distance and MDD remains open.

Moreover, investigations into the benchmark of syntactic complexity and cognitive demand introduced above were primarily based on traditionally syntactic-oriented dependency models, for instance, the Stanford Typed Dependencies annotation scheme (de Marneffe and Manning, 2008) or other annotation schemes that specifically designed for each individual language. Thus, there is no consistency among different treebanks. In addition, although there are some qualitative investigations on the distinctions between the UD annotation scheme and various traditional syntactic-oriented annotation schemes (e.g. Osborne and Maxwell, 2015), and the existing studies also include some empirical studies focusing primarily on the consistently annotated UD scheme (e.g. Chen and Gerdes, 2017; 2018), it is still of our interest that, compared with those based on consistently annotated traditionally syntactic-oriented schemes, whether linguistic analysis based on the UD annotation scheme can still function as a metric of syntactic difficulty and cognitive demand, and if it can, what are the reasons for these distinctions?

Therefore, the deficiency of investigations into annotation scheme of treebanks leads to the inquiry of current study. We attempt at making comparisons of dependency distances based on two different annotation schemes, UD and SUD. Aimed to address the issues mentioned above, the following questions are under discussion based on UD and SUD treebanks:

- (1) Will the probability distribution of dependency distances of natural texts change when they are based on different annotation schemes? Do they still follow the linguistic law of DDM?
- (2) Based on MDDs, which annotation scheme is more congruent for the measurement of syntactic complexity and cognitive demand?
- (3) Which dependency types account most for the distinctions between UD and SUD annotation schemes?

2 Materials and Methods

Taking English language as an example, we adopt the Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017) in UD 2.2 and SUD 2.2 projects. Both versions of the treebank are consisted of seven genres, viz. *academic writing*, *biographies*, *fiction*, *interviews*, *news stories*, *travel guides* and *how-to guides*. Since the treebanks are balanced in term of genres, it would better demonstrate the general features of the probability distribution of dependency distance when we adopt different annotation schemes.

To measure the effectiveness of MDDs as a metric of syntactic difficulty and cognitive demand in a broad sense, the *testing* sets of 20 languages with two versions of annotations were drawn from the UD 2.2 and SUD 2.2 to form 20 corresponding treebanks. There 20 languages are *Arabic (ara)*, *Bulgarian (bul)*, *Catalan (cat)*, *Chinese (chi)*, *Czech (cze)*, *Danish (dan)*, *Dutch (dut)*, *Greek (ell)*, *English (eng)*, *Basque (eus)*, *German (ger)*, *Hungarian (hun)*, *Italian (ita)*, *Japanese (jpn)*, *Portuguese (por)*, *Romanian (rum)*, *Slovenian (slv)*, *Spanish(sp)*, *Swedish (swe)* and *Turkish (tur)*. These 20 treebank-pairs would help to demonstrate the features and distinctions of syntactic- and semantic-oriented annotation schemes in measuring syntactic complexity and cognitive constraint.

As for the calculation of dependency distance, we adopted Jiang and Liu’ (2015) approach. Formally, let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_x and W_y ($x \geq 1, y \leq n$), if W_x is a head and W_y is its dependent, then the dependency distance between them is defined as the difference $x - y$; by this measure, the dependency distance of adjacent words is 1.

The MDD of the entire sentence can be defined as:

$$\text{MDD (sentence)} = \frac{1}{n-1} \sum_{i=1}^{n-1} |DD_i| \quad (1)$$

In this formula, n is the number of words in the sentence and DD_i is the dependency distance of the i -th syntactic relation of the sentence. Usually in a sentence there is one word (the *root* verb) without a head, whose DD is defined as zero.

The MDD of a treebank can be defined as:

$$\text{MDD}(\text{treebank}) = \frac{1}{n-s} \sum_{i=1}^{n-s} |\text{DD}_i| \quad (2)$$

Here, n is the total number of words in the sample, s is the total number of sentences in the sample and DD_i is the dependency distance of the i -th syntactic link of the sample.

When it comes to the MDD for a specific type of dependency relation in a sample, the formula can be shown as follows:

$$\text{MDD}(\text{dependency type}) = \frac{1}{n} \sum_{i=1}^n \text{DD}_i \quad (3)$$

In this case, n is the number of examples of that relation in the sample. DD_i is the dependency distance of the i -th dependency type.

For both UD and SUD annotations, the formats of their representations are CoNLL-X (de Marneffe & Manning, 2008). **Table 1** is a simplified CoNLL-X version of the sample sentence with UD annotation scheme.

Order	Word	Dependent			Head Order	Relation Dependency Type
		Lemma	POS	Feature		
1	<i>We</i>	we	PRON	PRP	2	nsubj
2	<i>walked</i>	walk	VERB	VBP	0	root
3	<i>along</i>	along	ADP	IN	5	case
4	<i>the</i>	the	DET	DT	5	det
5	<i>lake</i>	lake	NOUN	NN	2	obl

Table 1. Simplified Annotation of *We walked along the lake* in UD Treebank.

Take the first line in **Table 1** for example. It shows that the second word *walked* in the sentence has a dependent *we*, which is the first word of the sentence. The type of this dependency is *nsubj*, or *nominal subject*. As for the second line, it indicates that the *root* of the sentence is *walked*, signifying the head of the whole sentence rather than demonstrating a dependency relation; hence it is removed during computation. Regarding the sample sentence above, the DD of *nsubj* (line one) is $2 - 1 = 1$; *case* (line three) is $5 - 3 = 2$; *det* (line four) is $5 - 4 = 1$; *obl* (line five) is $2 - 5 = -3$. Hence, following formula (1), the MDD of the sentence can be obtained as follows: $(|1|+|2|+|1|+|-3|)/4=1.75$. Similarly, the MDD of the sample sentence based on SUD annotation scheme in **Figure 1 (b)** is $(|1|+|-1|+1+|-2|)/4=1.25$.

3 Results and Discussion

Taking English language as an example, we would first focus on the probability distribution of dependency distance to investigate whether it follows the linguistic law of DDM when we adopt two distinctive annotation schemes. Following what Liu (2008) did, we would then calculate MDDs of 20 languages based on two annotation schemes to demonstrate which annotation is more effective to measure syntactic difficulty and cognitive demand. Finally, specific dependency types in the treebank of GUM would be under investigation to display the possible underlying explanation beneath the distinctions between these two annotation schemes.

3.1 Annotation Scheme and Probability Distribution of Dependency Distance

It is believed that dependency distance is cognitively restrained by human working memory (Liu et al., 2017). Therefore, human beings tend to minimize the dependency distances while interpreting or producing languages. Hence, based on different syntactic-oriented annotation schemes, it has been found that the probability distribution of dependency distances of natural languages follows similar distributional patterns, including right truncated zeta (Jiang and Liu, 2015; Wang and Liu, 2017; Liu et al., 2017) and right truncated waring (Jiang and Liu, 2015; Lu and Liu, 2016; Wang and Liu, 2017).

Following these researches, we fitted dependency distances of all 95 texts of GUM to these two probability distributions by the fitting program of probability distributions, Altmann-Fitter³. Since the determination coefficient R^2 can indicate the goodness-of-fit (Wang and Liu, 2017; Wang and Yan, 2018), the mean values of the determination coefficient R^2 in all seven genres were calculated. Conventionally, the excellent, good, acceptable and not acceptable goodness-of-fit for determination coefficient R^2 are 0.90, 0.80, 0.75 and less than 0.75, respectively.

It was found that the mean determination coefficient R^2 in the model fitting of right truncated waring and right truncated zeta based on both UD and SUD are larger than 0.80, indicating that the fitting results are good. In other words, the frequencies of dependency distances based on both UD and SUD treebanks can well capture the models of right truncated waring and right truncated zeta with a good coefficients of determination R^2 .

To conclude, the probability distributions of dependency distances of natural texts based on both UD and SUD annotation schemes share similar power law distribution, viz. the frequency of dependency distance decreases with the increase of the dependency distance. The results reveal that dependency distance distributions of all texts based on both UD and SUD follow the same regularity, supporting the Least Effort Principle (LEP) (Zipf, 1965) or the linguistic law of DDM (Liu, 2008; Futrell et al., 2015; Liu et al., 2017).

3.2 Annotation Scheme and Mean Dependency Distance

Except the probability distribution of dependency distance, the syntactic and cognitive parameter of MDDs is also of our interest. Hence, the MDDs of all 20 corresponding treebanks based on UD and SUD were calculated in accordance with formula (2). Our results show that although these two annotation schemes are divergent from each other, what is in common is that the MDDs of all 20 languages based on both annotation schemes are within 4 (Cowan, 2001), showing that the syntactic complexity of human languages is constrained by human cognitive limitation or LEP rather than annotation scheme itself. This is consistent with what we've discovered in **Section 3.1**. Moreover, what is noteworthy is that MDDs based on 20 UD treebanks are always larger than those based on SUD for each individual language. This means that language materials based on UD annotation scheme lead to the interpretation of larger MDDs. Theoretically, it is believed that annotation schemes that lead to shorter MDDs is more linguistically applicable (Osborne and Gerdes, 2019). Hence, the SUD annotation scheme seems to be more suitable for reflecting the human cognitive demand and the syntactic complexity of the language under processing.

This was followed by a dependent-samples t test. The result shows that MDDs based on UD ($M = 2.86$, $SD = .32$) are significantly longer than MDDs that based on SUD ($M = 2.52$, $SD = .39$), $t(19) = 11.10$, $p < .05$, $d^f = 2.48$. The p -value of .000 is less than .05, the null hypothesis that the means of MDD based on different annotation schemes are equal is rejected. Moreover, according to Cohen's conventions, the effect size of 2.48 corresponds to a large effect in practice and indicates that the MDDs based on UD was rated 2.48 standard deviations longer in distance than was SUD. Hence, with the distinction of annotation schemes, natural language texts based on UD annotation scheme tend to have longer MDD than that based on SUD.

In addition, the results of MDDs based on SUD are closer to those of Liu (2008: 174). This might be due to the fact that both the SUD annotation scheme and what Liu (2008) based on belong to the category of syntactic-oriented annotation schemes. Although some languages have larger MDDs (e.g. *Hungarian (hun)* and *Chinese (chi)*) and the other have smaller ones (e.g. *Turkish (tur)* and *Japanese (jpn)*) in Liu (2008) than those based on SUD, this might be attributed to that the annotation schemes of Liu (2008) are not consistently annotated across languages. Hence, when it comes to the relationship between annotation scheme and MDD, although still within a threshold of 4, MDD of language materials based on UD annotation scheme tends to be longer than that based on SUD, and the difference is significant. Moreover, the MDDs based on SUD share great similarities with those based on Liu (2008). Thus, to some extent, it can be summarized that the syntactic-oriented SUD is comparatively more expedient annotation scheme to researches concerning syntactic complexity and cognitive demand.

³See also <https://www.ram-verlag.eu/software-neu/software/>.

⁴A commonly used effect size statistic for the dependent-samples t test is d . In accordance with Cohen's (1988) conventions, small, medium, and large effect sizes for the dependent samples t test are .20, .50, and .80, respectively.

3.3 Annotation Scheme and Annotating Preference

In **Section 3.1** and **Section 3.2**, we investigated the universal inclination of DDM for natural languages and MDD as an indicator of syntactic complexity as well as cognitive demand based on two different annotation schemes. In **Section 3.3**, the reasons for the similarities as well as distinctions are of our interest. Since it is impossible to make detailed analysis based on all 20 corresponding treebanks, the English treebank of GUM is under investigation as a representative.

Since the SUD annotation scheme is near-isomorphic to the UD initiative (Gerdes et al., 2018), treebanks based on UD and SUD are very similar to a large extent. The greatest difference between UD and SUD treebanks is the direction of the dependency types used to indicate the relations between function words and content words. In this case, UD’s *aux*, *cop*, *mark* and *case* dependencies indicate dependency relations pointed from content words to function words (e.g. the *case* relation between *lake* and *along* as shown in **Figure 1 (a)**), while their directions are inverted in SUD and renamed as *comp* as shown in **Table 2** (e.g. the *comp* relation between *along* and *lake* in **Figure 1 (b)**) (Gerdes et al., 2018: 71). Meanwhile, other subordinate dependency relations remain intact.

UD Dependency	Corresponding SUD Dependency
<i>aux</i> , <i>cop</i> , <i>mark</i> , <i>case</i> , <i>xcomp</i> , <i>ccomp</i> , <i>obj</i> , <i>iobj</i> , <i>obl:arg</i>	<i>comp</i>

Table 2. General Corresponding Dependency Relations in UD and SUD Annotation Schemes.

As **Table 2** shows, the *comp* relation in SUD is consisted of more than four UD types (i.e. *aux*, *cop*, *mark* and *case*). Hence, according to Gerdes et al. (2018: 72), we nailed down the actual 4 corresponding pairs that differentiate the UD and SUD annotation schemes in **Table 3**. They are *aux* and *comp:aux*, *aux:pass* and *comp:pass*, *cop* and *comp:cop*, and finally, *mark & case* and *comp*.

UD		SUD	
Type	Relation	Type	Relation
<i>aux</i>	auxiliary	<i>comp:aux</i>	complement: auxiliary
<i>aux:pass</i>	passive auxiliary	<i>comp:pass</i>	complement: passive auxiliary
<i>cop</i>	copula	<i>comp:cop</i>	complement: copula
<i>mark</i>	marker	<i>comp</i>	complement: subordinating conjunction
<i>case</i>	case marking		complement: adposition

Table 3. Detailed Corresponding Dependency Relations in UD and SUD Annotation Schemes.

In **Table 3**, all the dependency relations in UD are head-final, and in SUD head-initial. In other words, nearly all *comp:aux*, *comp:pass*, *comp:cop* and *comp* in SUD designate the function words as heads over content words, hence the dependency directions are negative. Correspondingly, nearly all the *aux*, *aux:pass*, *cop* and *mark & case* in UD choose content words as head; hence the dependency directions are also altered. This shows that the underlying reason for the distinction between UD and SUD annotation scheme is that the UD annotation scheme favours taking the content words as the head of function words while the SUD annotation scheme chooses the function words as heads over content words in dependency relations (Nivre, 2015; Gerdes et al., 2018; Osborne and Gerdes, 2019). To be specific, the UD treebanks first connect the content words and then the function words to emphasize the semantic similarities of all languages, while the SUD treebanks connect content words mediated by function words to complete the functional roles of function words. Consequently, the more the number of modifiers before noun (between the apposition and the noun) is, the longer the dependency distance between the verb (*root*) and the noun. For instance, the distance between *walked* and *lake* in **Figure 1 (a)** would

enlarge if there are more modifier before the noun *lake*. Hence, the longer MDDs in UD treebanks can be attributed to the emphasis on semantic relations within sentence structure.

In fact, designating the head of linguistic structure has always been a focus of modern grammar, not only for dependency grammar but also for constituency-based frameworks (Jackendoff, 1977; Zwicky, 1985; Pollard and Sag, 1994), especially when it comes to the function words within sentence structure (de Marneffe and Nivre, 2019). The design of the SUD representation that prioritizes function words as heads over content words is in line with most traditional syntaxes (Hudson, 1984; Mel'čuk, 1988; Starosta, 1988; Eroms, 2000). Moreover, under the framework of dependency grammar, the status of function words has also been discussed by many theoretical studies (e.g. Groß and Osborne, 2015; Osborne and Maxwell, 2015). However, most studies focus on one aspect of function words or emphasize the qualitative features of dependency relations that related to function words. The current section provides some empirical evidence of the status of function words in semantic-oriented UD and syntactic-oriented SUD treebanks.

4 Conclusions and Implications

Through observation and calculation, it can be found that based on UD and SUD annotation schemes, natural English texts exhibit similar probability distribution. No matter what the genre of the text is, both share a power law distribution with a trend of minimizing dependency distance. This is also consistent with the well-exemplified DDM theory corroborated by Liu et al. (2017), showing the limitation of human working memory capacity.

When 20 corresponding treebanks are under investigation, the MDDs of Liu (2008)'s study and our study based on SUD annotation scheme are similar with each other, and they are significantly shorter than those based on UD, showing the consistency of syntactic-oriented annotation schemes and the possibility of applying SUD to language materials to measure syntactic complexity and cognitive demand.

Moreover, the reason underlying for the distinctions between UD and SUD annotation schemes is the dependency types indicating the relations between apposition and noun. The UD annotation scheme prefers a semantic orientation, while the SUD annotation scheme favours a syntactic orientation which holds a function-word priority. To be specific, the corresponding pairs, *aux* and *comp:aux*, *aux:pass* and *comp:pass*, *cop* and *comp:cop*, and *mark & case* and *comp* in UD and SUD annotation schemes lead to longer MDDs of UD treebanks.

Therefore, the current study suggests that, to some extent, the consistently syntactic-oriented annotation scheme (SUD) is better than the consistently semantic-oriented one (UD) in linguistic analysis of syntactic complexity and human cognitive demand. However, it is still worthwhile to spare more efforts to assess the effectiveness of consistently annotated syntactic-oriented representation to capture both the variations and universals of natural human languages.

References

- Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581-612. DOI: 10.1007/s10579-016-9343-x
- Anne Abeillé. 2003. *Treebank: Building and Using Parsed Corpora*. Kluwer Academic Publisher, Dordrecht.
- Arnold M. Zwicky. 1985. Heads. *Journal of Linguistics*, 21:1-29.
- Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2): 286-310.
- Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1-76. DOI: 10.1016/s0010-0277(98)00034-1
- George K. Zipf. 1965. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner Publishing Company, New York.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159-191.

- Haitao Liu. 2009. *Dependency Grammar: From Theory to Practice*. Science Press, Beijing.
- Haitao Liu. 2010. Dependency direction as a means of word-order typology a method based on dependency treebanks. *Lingua*, 120(6):1567-1578.
- Haitao Liu and Chunshan Xu. 2012. Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics*, 48(4):597-625.
- 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, Yiyi Zhao, and Wenwen Li. 2009. Chinese syntactic and typological properties based on dependency syntactic Treebanks. *Poznań Studies in Contemporary Linguistics*, 45(4):509-523.
- Hans Jürgen Heringer. 1993. Dependency syntax-basic ideas and the classical model. *Syntax-An International Handbook of Contemporary Research*, volume 1, 298-316.
- Hans-Werner Eroms. 2000. *Syntax der deutschen Sprache*. Walter de Gruyter, Berlin.
- Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.
- 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.
- Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. *Computational Linguistics and Intelligent Text Processing*, 3-16. DOI: 10.1007/978-3-319-18111-0_1
- Joel E. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Johanna Nichols. 1986. Head-marking and dependent-marking grammar. *Language*, 62:56-119.
- John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford.
- John Brown. 1958. Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10:173-189.
- 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 Universal Dependencies Workshop 2018*, 66-74. Brussels.
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.
- Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Typed Dependencies Manual. *Technical Report*, 338-345.
- Marie-Catherine de Marneffe and Joakim Nivre. 2019. Dependency Grammar. *Annual Review of Linguistics*, 5:197-218.
- Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87-185.
- Peter H. Matthews. 1981. *Syntax*. Cambridge University Press, Cambridge.
- Qian Lu, Chunshan Xu, and Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages? *Complexity*, 21(S2):33-41.
- Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135.
- Ray S. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS*, 112:10336-10341.
- Richard Hudson. 1984. *Word Grammar*. Basil Blackwell, New York.
- Richard Hudson. 1995. Measuring syntactic difficulty. <http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>
- Roger Levy, Evelina Fedorenko, and Edward Gibson. 2013. The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, 69 (4):461-495.

- So Hiranuma. 1999. Syntactic difficulty in English and Japanese: A textual study. *UCL Working Papers in Linguistics*, 11:309-322.
- Stanley Starosta. 1988. *The Case for Lexicase: An Outline of Lexicase Grammatical Theory*. Pinter Publishers, New York.
- Thomas Groß and Timothy Osborne. 2015. The dependency status of function words: auxiliaries. In Eva Hajičová and Joakim Nivre (eds.), *Proceedings of the 3rd International Conference on Dependency Linguistics*, pp. 111-120. Stroudsburg, PA: Assoc. Comput. Linguist.
- Timothy Osborne and Daniel Maxwell. 2015. A historical overview of the status of function words in dependency grammar. In Eva Hajičová and Joakim Nivre (eds.), *Proceedings of the 3rd International Conference on Dependency Linguistics*, pp. 241-250. Stroudsburg, PA: Assoc. Comput. Linguist.
- Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: A Journal of General Linguistics*, 4(1):17.1-28. DOI: 10.5334/gjgl.537
- Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135-147.
- Yaqin Wang and Jianwei Yan. 2018. A quantitative analysis on a literary genre *Essay*'s syntactic features. In Jingyang Jiang & Haitao Liu (eds.), *Quantitative Analysis of Dependency Structures*, pp. 295-314. Berlin/Boston: De Gruyter Mouton.

A quantitative probe into the hierarchical structure of written Chinese

Heng Chen

Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China
chenheng@gdufs.edu.cn

Haitao Liu

Department of Linguistics, Zhejiang University, Hangzhou, China
lhtzju@yeah.net

Abstract

Language unit is a fundamental conception in modern linguistics, but the boundaries are not clear between language levels both in the past and present. As language is a multi-level system, quantification rather than microscopic grammatical analysis should be used to investigate into this question. In this paper, Menzerath-Altmann law is used to make out the basic language units in written Chinese. The results show that “stroke > component > word > clause > sentence” is the hierarchical structure of written Chinese.

1 Introduction

Language levels and language units are critical conceptions in a language system, and they are highly related with the entities in a language, as well as the methods in linguistics. The conception of language unit is definitely put forward by Saussure in the first half of the 20th century. In his seminal book representing the birth of modern linguistics, Saussure puts forward the conceptions of language entities or language units and analyzes the methods as well as difficulties of dividing the spoken chain into language units. Moreover, Saussure distinguished the concept of language units from speech units. Language unit becomes the fundamental problem in modern linguistics. The conception of language level is introduced by American descriptive linguistics. Gleason (1956) distinguishes three types of language levels: language levels of structure, analysis and speech. Later, a number of other linguistic theories (Halliday, 1985; Hudson, 2010; Miyagawa et al., 2013; Nordström, 2014) treat language as a multi-level system.

Generally, five language units are commonly recognized by grammarians: morpheme, word, phrase, clause and sentence (Lyons, 1968). However, different linguistic schools have different opinions upon the systematicness of language, therefore, the methods and standards they use to divide language levels and units are different. Mackey (1967) lists seven sets of language levels from different linguistic schools, and the maximum is Bulundaer’s 14 levels, and the minimum is Harris’s 2 levels. These language units include sound, word, phrase, sentence, phone, phoneme, morph, morpheme, syllable, affix, word-group, and so on. The boundaries between language levels are not clear in the past for the lack of a common standard, however, it is the same after the introduction of the conception of language level.

The most characteristic feature of modern linguistics is structuralism. Briefly, this means that language is not a haphazard conglomerate of words and sounds but a tightly knit and coherent whole. However, linguistics is traditionally preoccupied with the fine detail of language structure (Hudson, 2010:104), or in other words, the language phenomena at the microscopic scale rather than at the system level (Liu and Cong, 2014). Therefore, it is not ordinarily feasible to analyze each language level separately, and the work must be carried on simultaneously on all levels. Moreover, the results should be stated in terms of an orderly hierarchy of levels (Lyons, 1968).

Quantification is necessary in the inquiry into the structure of the language system (Altmann, 1987, 1996). Without quantification, it would be extremely difficult to investigate language as a multi-level system empirically. Unfortunately, systems thinking in linguistics is generally unaffected by quantitative methods. Liu & Cong (2014) characterize modern Chinese as a multi-level system from the complex

network. However, their emphasis is on the levels of grammatical analysis, for example, syntax and semantics, but not language levels. In this paper, we try to analyze the language levels of written Chinese as a multi-level system using Menzerath-Altmann law.

Menzerath-Altmann law is a general statement about the natural language constructions which says: the longer is a construction, the shorter are its constituents. Language is a whole complex system, and it is a set of relations. The language units correlate with each other in different levels and in complex ways through the relations. The whole is composed by its parts, and they restrains mutually. Language units of the same levels are relatively homogeneous. Therefore, the relation between two adjacent language levels is “whole-part”.

Actually, in quantitative linguistics, the relationship between “whole-part” has been extensively investigated (Menzerath, 1954; Krott, 1996; Uhlířová, 1997; Mikros and Milička, 2014; Milička, 2014). The relation was investigated and tested on many linguistic levels and in many languages and even on some non-linguistic data (Baixeries et al., 2013). Köhler (1984) conducted the first empirical test of the Menzerath-Altmann law on “sentence > clause > word”, analyzing German and English short stories and philosophical texts. The tests on the data confirmed the validity of the law with high significance. Heups (1983) evaluates 10,668 sentences from 13 texts separated with respect to text genre and her results also confirm the Menzerath-Altmann law with high significance. The law has also been used to study phenomena on the supra-sentential level (Hřebíček, 1990, 1992) and fractal structures of text (Hřebíček, 1994; Andres, 2010). This is why this law is considered one of the most frequently corroborated laws in linguistics. The law is a good example of the importance of the quantitative linguistic methodology since it clearly shows that the “independent language subsystems” are in fact interconnected by relationships which are hard to detect by a qualitative research.

In this paper, we will test the construction units in written Chinese, which includes stroke, component, character, word, clause and sentence. We do not include phrase in our language units list because it is hard to divide a sentence into one or several independent phrases in written Chinese. Despite of this, it can be inferred from the Menzerathian results of “sentence-clause-word”. That is to say, if “sentence-clause-word” fits well with Menzerath’s law, then the unit phrase in written Chinese can be left out, or we should reconsider phrase as an indispensable language unit in written Chinese.

The remainder of this paper is organized as follows. Section 2 introduces the materials and methods of the present study. Section 3 presents the results of the tests for different hierarchical language units. Section 4 concludes the study and makes suggestions for further research.

2 Materials and Methods

We use the Lancaster Chinese corpus (LCMC) as the testing material. The corpus is segmented and part of speech (POS) tagged, and its basic information is in table 1.

Language units	scale
Character (tokens)	1,314,058
Character (types)	4,705
Clauses (types)	126,455
Sentence (types)	45,969
Word (types)	847,521

Table 1. Basic information of LCMC

The language units we will test in this paper are stroke, component, character, word, clause and sentence. The reason why we do not include phrase here is that a complete sentence or clause cannot be divided into several sequential phrases, both theoretically and practically.

All the language units are easy to get in LCMC by using some tools except clause. Therefore, in the following, we will first define the other language units, and then give our methods of defining phrase.

The stroke is a segment written with one uninterrupted movement. The component is the constructing units of characters which have more than one strokes. The character are logograms used in the writing of Chinese, which is called hanzi in Chinese. For example, the word “语言”(“yǔ yán”, which means “language”) consists of two characters “语, 言”(“yǔ, yán”, which means “language, parole”), and the two characters have nine strokes “丶, 丿, 一, |, 冫, 一, |, 冫, 一” and seven strokes “丶, 一, 一, 一,

丨, 丿, 一” respectively, eleven in total. “语” “言” have five components “讠” “五” “口” and one component “言” (means “parole”), respectively. To measure the number of strokes and components of a word, we used a list consisting of 20,902 characters (CJK Unified Ideographs) with numbers of strokes and components of each character.

In written Chinese, sentences are separated from one another by using special marks of punctuation (full-stop, question-mark, exclamation-mark). As for our case, the sentences are tagged in LCMC, so here there is no difficulties distinguishing sentence.

Clause is not tagged in LCMC, nor in any other corpus available. Xing (1997:13) states that clause is the smallest independent grammatical unit of expression. But this definition can hardly be used to obtain the clauses in LCMC. Lu (2006) analyzes a long sentence from a literary book and claims that the constituents just between two punctuations (comma and period) can be defined as clauses roughly. We believe that although this method is not so exact in grammatical analyses, it can in large-scale-corpus studies. But we need to state that, since in LCMC sentences are tagged, we choose comma and semicolon as our marks of clause boundaries.

After obtaining all the statics with respect to language units in LCMC, we use the Menzerath-Altmann law to fit the hierarchical data.

Menzerath-Altmann law (short for Menzerathian function) describes the mathematical relation between two adjacent language units, and its model function is

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

In this function, y represents the length of the upper language unit, and x represent the mean length of the lower language unit; a, b, c are parameters which seem to depend mainly on the level of the language units under investigation: much more than on language, the kind of text, or author as previously expected, and e is natural constant, which equals 2.71828 approximately. The goodness of fit can be seen from determination coefficient R^2 . We say the result is accepted for $R^2 > 0.75$, good for $R^2 > 0.80$, and very good for $R^2 > 0.90$.

3 Results

The language units we will examine in this paper are stroke > component > character > word > clause > sentence (here we use “>” to direct to a higher-rank unit in written Chinese). Since the Menzerath-Altmann law is only used to fit the data of two adjacent language units, we corroborate that the fitting results of these five groups, namely “component> character > word”, “stroke > character > word”, “stroke > component > word”, “component > word > clause”, “word > clause > sentence”, can answer the question of the hierarchical structure in written Chinese. We will give the result of each group in the following. We begin with the word level since it is regarded as the most basic language unit in all languages.

3.1 Component > Character > Word

The Menzerathian data of “component > character > word” can be seen in Table 2. In this group, word length is measured in character, and character length is measured in component. Mean character length can be calculated with this function:

$$M_i = \frac{F_i}{F_i' * i} \quad (2)$$

In this function, i refers to word length class, i.e. the first column in Table 2; M_i represents mean character length of word length class i (if a word’s length is 1, then it belongs to word length class 1, and the like), i.e. the second column in Table 2; F_i represents the sum length of all the characters (measured in component) in the words (based on tokens) of word length class i ; F_i' represents the number of words (based on tokens) of word length class i .

Word length (in character)	Mean character length (in component)	Word length (in character)	Mean character length (in component)
1	2.4592	6	2.2054
2	2.5899	7	2.1860
3	2.5435	8	2.1354
4	2.5372	9	2.4222
5	2.1536	10	2.7000

Table 2. Hierarchical data of “component > character > word”

In table 2, we can see that there are ten word length classes and their corresponding mean character lengths. We fit the Menzerathian function introduced in section 2 to the two groups of variables. The goodness of fit indicator $R^2 = 0.1625$ means that the fitting result is unaccepted, which indicate that the hierarchical group “component > character > word” does not line with Menzerath-Altmann law. Therefore, next, we need to test two other possible groups, “stroke > character > word” and “stroke > component > word” to find out the hierarchy in word level.

3.2 Stroke > Character > Word

We replace component in “component > character > word” with stroke, and the Menzerathian data can be seen in Table 3. In this group, word length is measured in character, but character length is measured in stroke instead.

Word length (in character)	Mean character length (in stroke)	Word length (in character)	Mean character length (in stroke)
1	6.9359	6	6.1622
2	7.4136	7	6.2326
3	7.2189	8	6.2708
4	7.1969	9	6.5778
5	6.2356	10	6.4000

Table 3. Hierarchical data of “stroke > character > word”

We fit the Menzerathian function to the two groups of variables in Table 3, and the goodness of fit indicator $R^2 = 0.5009$ means that the fitting result is also unaccepted. This indicates that the group “stroke > character > word” does not line with Menzerath-Altmann law. Then the only possible group in the word level is “stroke > component > word”.

3.3 Stroke > Component > Word

The Menzerathian data of “stroke > component > word” is displayed in Table 4. In this group, word length is measured in component, and component length is measured in stroke.

Word length (in component)	Mean component length (in stroke)	Word length (in component)	Mean component length (in stroke)
1	3.45959	13	1.72858
2	2.80834	14	1.62894
3	2.44086	15	1.71641
4	2.21272	16	1.62715
5	2.00806	17	1.55203
6	1.86860	18	1.66435
7	1.81350	19	1.90789
8	1.80166	20	1.350
9	1.80735	21	1.71428
10	1.78970	22	1.98484
11	1.80674	23	1.34782
12	1.74935	25	1.960

Table 4. Hierarchical data of “stroke > component > word”

Then we fit the Menzerathian function to the data in Table 4, and we have a good result this time. The fitting is displayed in Figure 1, and the fitting results, i.e. parameters (with 95% confidence bounds) and determination coefficient are shown in the bottom of Table 4. The value of the goodness of fit indicator R^2 is 0.8982, which means that the result is good, and the group “stroke > component > word” lines with Menzerath-Altman law.

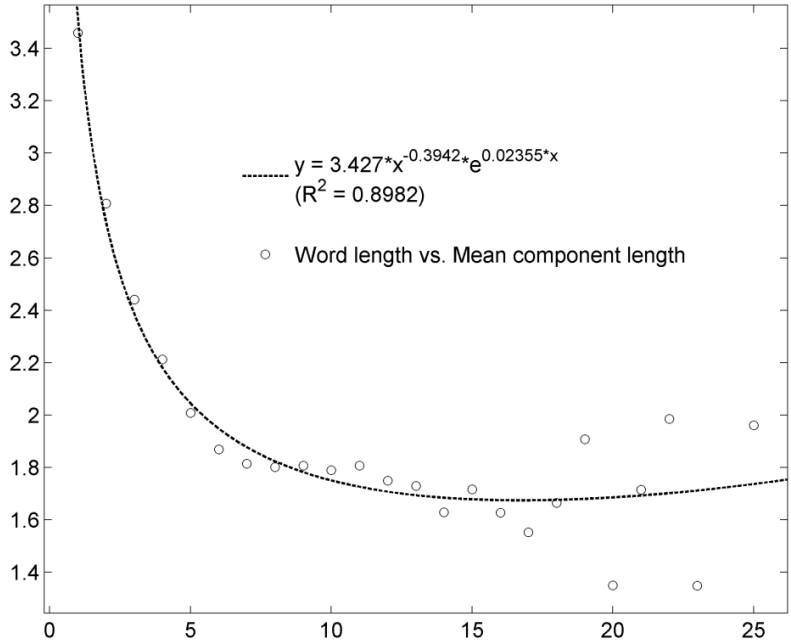


Figure 1. Fitting Menzerath-Altman law to the hierarchical data of “stroke > component > word”

In sum, in word-level, component is its immediate lower basic language unit. Since above word level, there are two other language units, we first need to examine the group “component > word > clause” to determine if we need go into “component > word > sentence”.

3.4 Component > Word > Clause

Table 5 shows the Menzerathian data of “component > word > clause”. In this group, clause length is measured in word, and word length is measured in component.

Clause length (in word)	Mean word length (in component)	Clause length (in word)	Mean word length (in component)	Clause length (in word)	Mean word length (in component)
1	5.5445	12	3.9150	23	4.0552
2	4.5248	13	3.9402	24	4.1348
3	4.1405	14	3.9494	25	4.1948
4	3.9387	15	3.9944	26	4.1137
5	3.8897	16	3.9733	27	4.2187
6	3.8444	17	4.0052	28	3.8613
7	3.8383	18	4.0247	29	4.1614
8	3.8458	19	4.0453	30	4.2573
9	3.8657	20	4.0729	31	4.1608
10	3.8738	21	4.0674	32	4.0275
11	3.8966	22	4.1309	33	4.3384

Table 5. Hierarchical data of “component > word > clause”

We then fit the Menzerathian function to the two groups of variables in Table 5. The fitting is displayed in Figure 2, and the results is shown in the bottom of Table 5. As can be seen from the value of R^2 (0.7657) in Table 5, the fitting result is accepted.

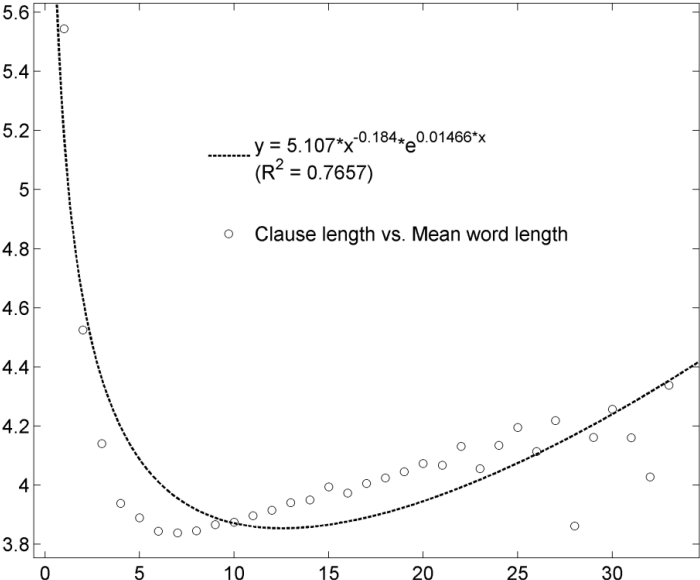


Figure 2. Fitting Menzerath-Altman law to the hierarchical data of “component > word > clause”

Although the fitting result ($R^2 = 0.7657$) in this group is not as good as in “stroke > component > word” ($R^2 = 0.8982$), the group “component > word > clause” lines with Menzerath-Altman law. This means that clause is the immediate higher language unit of word, thus we need not go into the group “component > word > sentence”. Ultimately, we only have “word > clause > sentence” to be tested.

3.5 Word > Clause > Sentence

Table 6 shows the Menzerathian data of this group. As can be seen in Table 6, the sentence length is measured in clause, and the clause length is measured in word.

Sentence length (in clause)	Mean clause length (in word)	Sentence length (in clause)	Mean clause length (in word)
1	7.7407	9	6.2194
2	7.0465	10	6.3932
3	6.7162	11	5.8068
4	6.4866	12	5.7661
5	6.3357	13	6.1723
6	6.2485	14	6.5510
7	6.1646	15	6.4500
8	6.2296		

Table 6. Hierarchical data of “word > clause > sentence”

The Menzerathian function is again used, and the fitting is displayed in Figure 3. As can be seen from the fitting results in Table 6, the goodness of fit indicator R^2 (0.8498) indicates the result is good. This means that the group “word > clause > sentence” lines with Menzerath-Altman law.

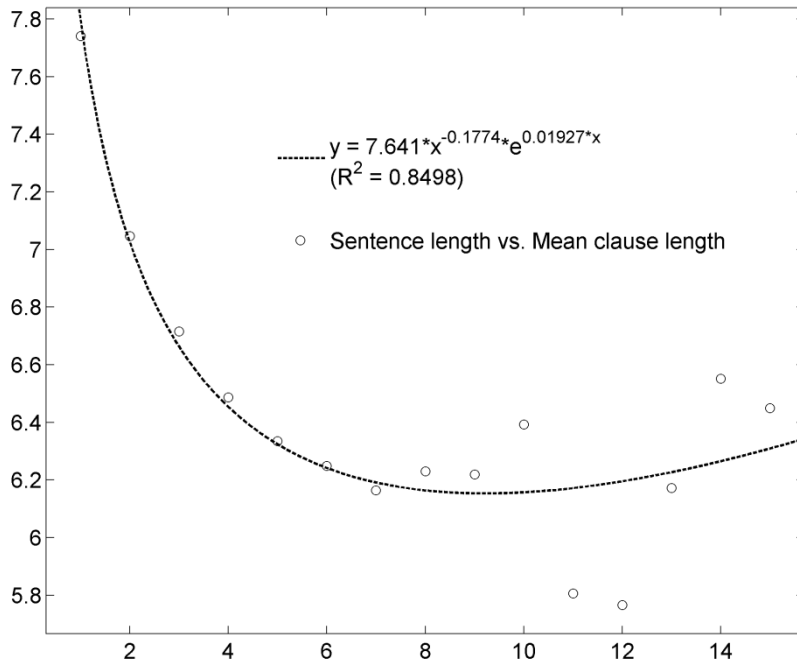


Figure 3. Fitting Menzrath-Altman law to the hierarchical data of “word > clause > sentence”

4 Discussions and Conclusions

In section 3 we tested five Menzrathian groups, namely “component > character > word”, “stroke > character > word”, “stroke > component > word”, “component > word > clause”, “word > clause > sentence”. The results shows that only “stroke > component > word”, “component > word > clause” and “word > clause > sentence” line with Menzrath-Altman law. However, the fitting results of “component > word > clause” and “word > clause > sentence” are not as good as that of “stroke > component > word”. We think that there are two possible reasons. One reason is the data sparseness problem: clause length distribution is sparser than word length distribution because the length range of word is more fixed than that of clause. The other reason may be the rough way of segmenting clauses by means of punctuations: the clauses obtained in this way may be a little bigger or smaller than the practical situation. Generally, the results indicate that “stroke > component > word > clause > sentence” is a Menzrathian hierarchy in written Chinese.

Character is an easy-to-distinguish language unit in written Chinese; clause is commonly regarded as one level of language unit by grammarians. However, they are not included in the Menzrathian hierarchy, i.e. they are not basic language units. For character, the reason may be that although there are thousands of single-character words, they are not enough for communication. The combinations of characters into multi-character words makes ends meet. In classic Chinese, Character may be a basic language unit, however, it is replaced by word in modern Chinese, because the classic Chinese habitually uses mono-syllable words while the modern Chinese prefers to choose multi-syllable words to express the same meaning. As for phrase, first, it is difficult to segment a sentence into several phrase sequences; secondly, from a quantitative perspective, the main reason may be that clause can directly be composed of words, but not via one level of phrase.

That language is a system has been put forward for about 100 years, however, it has never been realized until quantification is introduced into linguistics. In this paper, we shows that Menzrath-Altman law can be an efficient way of finding the basic language units in a language. In the future, we will investigate into this question from a diachronic perspective to see if the basic language units have changed with time.

Acknowledgements

This work was supported by the National Social Science Fund of China (Grant No. 18CYY031) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

Reference

- Andrea Krott. 1996. Some Remarks on the Relation between Word Length and Morpheme Length. *Journal of Quantitative Linguistics*, 3 (1): 29-37.
- G. Heups (1983). Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In R. Kohler & J. Boy (Eds.), *Glottometrika 5* (pp. 113 – 133). Bochum: Brockmeyer.
- Gabriel Altmann. 1987. The Levels of Linguistic Investigation. *Theoretical Linguistics*, 14(2-3): 227-240.
- Gabriel Altmann. 1996. The Nature of Linguistic Units. *Journal of Quantitative Linguistics*, 3(1): 1-7.
- Georgios Mikros and Jiří Milička. 2014. Distribution of the Menzerath's law on the syllable level in Greek texts. In Altmann, G., Čech, R., Mačutek, J., & Uhlířová, L. (Eds.), *Empirical approaches to text and language analysis*, pp. 180-189, RAM-Verlag.
- Haitao Liu and Jin Cong. 2014. Empirical Characterization of Modern Chinese As A Multi-Level System From The Complex Network Approach. *Journal of Chinese Linguistics*, 42(1): 1-38.
- Jackie Nordström. 2014. Language as a Discrete Combinatorial System, rather than a Recursive-Embedding One. *The Linguistic Review*, 31(1): 151-191.
- Jan Andres. 2010. On a Conjecture about the Fractal Structure of Language. *Journal of Quantitative Linguistics*. 17: 101–122.
- Jaume Baixeries , Antoni Hernández-Fernández , Núria Fornas & Ramon Ferrer-i-Cancho. 2013. The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*, 20: 94-104.
- Jiří Milička. 2014. Menzerath's Law: The Whole is Greater than the Sum of its Parts. *Journal of Quantitative Linguistics*, 21(2): 85-99.
- John Lyons. 1968. *Introduction to Theoretical Linguistics*. London: Cambridge university press.
- Jr Henry A. Gleason. 1955. *An Introduction to descriptive linguistics*. New York : Holt, Rinehart and Winston.
- Ludk Hřebíček. 1990. The constants of Menzerath-Altmann's Law. In R. Hammerl (ed.), *Glottometrika 12*. Bochum: Brockmeyer.
- Ludk Hřebíček. 1992. *Text In Communication: Supra-Sentence Structures*. Universitätsverlag Dr. N. Brockmeyer.
- Ludk Hřebíček. 1994. Fractals in Language. *Journal of Quantitative Linguistics*.1: 82-86.
- Ludmila Uhlířová. 1997. Length vs. Order: Word Length and Clause Length from the Perspective of Word Order. *Journal of quantitative linguistics*, 4(1-3): 266-275.
- Michael A. K. Halliday. 1985. *An Introduction to Functional Grammar*, Edward Arnold Ltd.
- Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.
- Reinhard Köhler. 1984. Zur Interpretation des Menzerathschen Gesetzes. In R. Köhler, J. Boy (Eds.), *Glottometrika 6*, pp. 177-183. Bochum: Brockmeyer.
- Richard, Hudson. 2010. *An introduction to word grammar*. Cambridge University Press.
- Shigeru Miyagawa, Robert C. Berwick, and Kazuo Okanoya. 2013. The emergence of hierarchical structure in human language. *Frontiers in Psychology*, 4:71.
- William F. Mackey. 1967. *Language Teaching Analysis*. Longman.

A comparative corpus analysis of PP ordering in English and Chinese

Zoey Liu

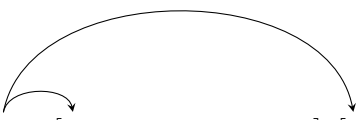
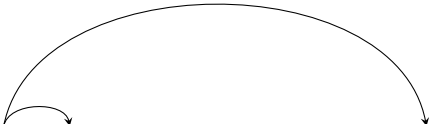
Department of Linguistics
University of California, Davis
yiliu@ucdavis.edu

Abstract

We present a comparative analysis of PP ordering in English and (Mandarin) Chinese, two languages with distinct typological word order characteristics. Previous work on PP orderings have mainly focused on English using data of relatively small size. Here we leverage corpora of much larger scale with straightforward annotations. We use the Penn Treebank for English, which includes three corpora that cover both written and spoken domains, and the Chinese Penn Treebank for Chinese. We explore the individual effect of dependency length, the argument status of the PP (argument or adjunct) and the traditional adverbial ordering rule, Manner before Place before Time. In addition, we evaluate the predictive power of dependency length and argument status with weights estimated from logistic regression models. We show that while dependency length plays a strong role across genre for English, it only exerts a mild effect in Chinese. On the other hand, the argument status of the PP has a pronounced role in both languages, that is, there exists a strong tendency for the argument-like PP to appear closer to the head verb than the adjunct-like PP. Our work contributes empirically to the long-standing proposal in linguistic typology that crosslinguistic word ordering preference is driven by cooperating and competing principles.

1 Introduction

Recent research has presented typological evidence that the overall or average dependency lengths between syntactic heads and their dependents tend to be minimized by their grammars as a whole (Futrell et al., 2015). Other experiments looking at specific syntactic constructions of individual languages that have alternative constituent orderings have also shown that speakers opt for constituents of shorter length to appear closer to their syntactic heads and thus shorten overall dependency distance in the sentence (Jaeger and Norcliffe, 2009). It has been argued as well as demonstrated in psycholinguistic and corpus studies that the preference for shorter dependencies is driven by processing efficiency (Gibson, 1998; Levy, 2013) and ease of communication (Hawkins, 2014; Gibson et al., In Press). As an illustration of how dependency length minimization (DLM) applies to constituent orderings, consider the following sentences in English:

- (1) a. *Dylan presented [on something linguistic] [to her colleagues and friends].*
- 
- b. *Dylan presented [to her colleagues and friends] [on something linguistic].*
- 

Both (1a) and (1b) have two PPs, shown within square brackets: *on something linguistic* and *to her colleagues and friends*. Switching the order of the two PPs does not change the grammaticality nor the semantic meaning of the sentence. As indicated by the syntactic dependency arcs, we consider the prepositions in both PPs to be the heads of their respective constituents, and to be dependents of the verb *presented*, which is the head of the VP in each sentence. The length of the dependency that attaches each PP to its corresponding VP is then the linear distance between the head of the dependency relation (the verb *presented*) and the preposition, which serves as the dependent. In both (1a) and (1b), the dependency length between *presented* and its closest PP is the same; however, the distance between *presented* and the farther PP is shorter in (1a), where the PP of shorter length is placed closer to the verb. From this example, we can see that in cases where the VP has two PP dependents occurring on the same side of the head verb, DLM predicts that there is a preference for placing the shorter PP closer to its head.

The effect of dependency length on syntactic preferences has been examined in various ways (Gibson, 2000; Gildea and Temperley, 2007; Gildea and Temperley, 2010; Temperley, 2007). Although strong evidence for DLM has been found, it is clear that it is not the only motivation in determining preferred word orders. Other competing and/or cooperating factors must also be at play that govern ordering preferences. The interaction between DLM and other principles and constraints in different contexts is currently under investigation (Gulordava et al., 2015; Wiechmann and Lohmann, 2013).

This study makes a contribution to the aforementioned research direction. We present a comparative analysis of PP orderings in English and Chinese, two languages with distinct typological properties. We focus in particular on VP instances with exactly two PP dependents appearing on the same side of the head verb, the ordering of which permits flexibility. Previous work on PP orderings has mainly focused on English with relatively small amounts of data (Hawkins, 1999; Wiechmann and Lohmann, 2013). Here we resort to corpora of much larger scale with straightforward annotations. For English, we use the Penn Treebank (PTB) (Marcus et al., 1993), which includes syntactic structures for approximately one million words of text from each of: the Wall Street Journal (WSJ), the Brown corpus (Kučera and Francis, 1967) and transcriptions of spontaneous spoken conversations from the Switchboard corpus (Godfrey et al., 1992). For Chinese, we exploit the Penn Chinese Treebank (CTB) (Xue et al., 2005), which has a total of 500K words. We probe to what extent dependency length, the argument status of the PP (argument or adjunct), and the traditional adverbial ordering rule, Manner before Place before Time, explain the observed PP ordering patterns. We explore how the effects of the three factors and their interactions differ across genres for English, and between the two languages.

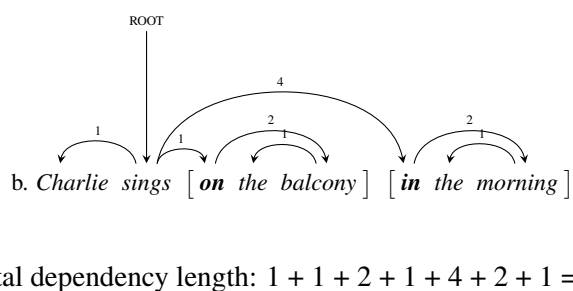
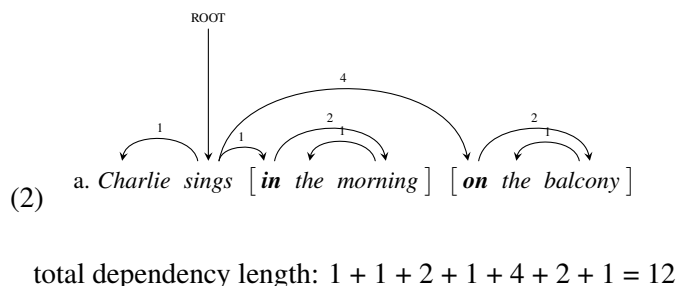
2 Related Work

2.1 Dependency length

Preceding DLM, the preferences for shorter syntactic dependencies have been formulated in various principles, including Early Immediate Constituent (Hawkins, 1994), Minimize Domains (Hawkins, 2004) and Dependency Locality Theory (Gibson, 2000). These principles all suggest the same idea that if grammatical alternatives exist for the syntactic constructions, there is a tendency to put shorter constituents closer to the syntactic heads and to avoid longer dependencies. Empirical support for the significant effects of dependency length in constituent ordering preferences has been found in various studies. Most work has focused on one specific or few syntactic structures in English, ranging from heavy NP shift (Wasow, 1997a; Arnold et al., 2000), dative alternation (Wasow and Arnold, 2003; Bresnan et al., 2007), verb particle constructions (Lohse et al., 2004), to postverbal PP orderings (Hawkins, 1999; Wiechmann and Lohmann, 2013) and so on. Some studies have extended their investigations to constructions in a small number of languages other than English, including Japanese (Yamashita and Chang, 2001; Yamashita, 2002), Korean (Choi, 2007), Russian (Kizach, 2012), Persian (Rasekh-Mahand et al., 2016) and certain Romance languages (Gulordava and Merlo, 2015).

As powerful as its effects are, dependency length itself will not suffice for predicting syntactic orderings across languages. First of all, dependency length is not able to indicate which ordering structure might be preferred when switching the order of constituents does not change the overall dependency length. As seen in the following examples, both (2a) and (2b) have two PPs, which are of equal length,

occurring after the head verb *sings*. Here we calculate dependency length as the distance from the head to its dependents, including the head verb. Changing the order of the two PPs does not appear to affect the total dependency lengths in (2a) and (2b) ¹.



Additionally, previous studies have shown that longer dependencies are preferred in certain syntactic constructions in corpora (e.g. preverbal adjuncts ordering in English (Rajkumar et al., 2016; Temperley, 2007)). Other psycholinguistic experiments presented that subordinate clausal structures with longer dependencies are easier to process in both rigid OV (e.g. Hindi (Vasishth and Lewis, 2006)) or non-rigid OV languages (e.g. German (Konieczny, 2000; Konieczny and Döring, 2003)).

What's more, the efficacy of DLM appears to vary crosslinguistically. Comparing German and English, Gildea and Temperley (2007) showed that German tends to have longer dependencies and minimizes dependency lengths to a lesser extent. They argued that the prevalent OV structures in German, where the verbs are in the final position, enlarge the dependency distance between the verb and its preverbal dependents. One other possible explanation that they discussed was that German has relatively free word order, which means that the constituent orderings in German may be driven more by considerations other than DLM. Looking at 37 languages, Futrell et al. (2015) suggested that head-final languages such as Japanese have longer dependencies compared to head-initial languages like English and Arabic. They conjectured that rich case marking systems in head-final languages allow more word order freedom, which lead to longer dependencies. Regardless of the proposed explanations, the fact that dependency length is minimized to different extents, and that it is not always minimized in certain cases indicate there are other cooperating and competing biases, cognitive or structural, that are effective and interact with DLM (Hawkins, 2014; MacWhinney et al., 2014)

2.2 Argument status

The role of argument status in constituent orderings is hardly new. Arguments prefer to be adjacent to their syntactic heads compared to adjuncts, which has been shown extensively in English (Culicover et al., 2005; Jackendoff, 1977; Pollard and Sag, 1994) as well as in other languages (Tomlin, 1986; Dyer, 2017).

Previous literature has distinct ways of deciding whether a constituent is an argument or an adjunct when investigating its effects on syntactic ordering preferences. For instance, in an examination of heavy NP shift, Wasow (1997b) found different shifting patterns when the verb and the PP are collocations than

¹Following the measure from Hudson (1995), which approximates dependency length as the number of intervening tokens between the head and its dependents, the total dependency length is also the same for both sentences, except that the value will be $0 + 0 + 1 + 0 + 3 + 1 + 0 = 5$.

when they are not. When the verb and the PP are collocational, in other words, when the PP is considered to be an argument of the verb, (e.g. *take into account*), there is a greater tendency to shift the NP and place the PP immediately after the verb. On the other hand, when the verb and the PP are not collocational (e.g., *take to the store*), the proportion of examples where the NP is shifted is much smaller. Using 394 relevant sentences in English, Hawkins (1999) noted the significant roles of syntactic dependency and the argument status of the PP, namely that the PP which is a complement of the head verb tends to appear closer to the verb. Wiechmann and Lohmann (2013) found similar results with 1,256 sentences from both the written and spoken sections of the International Corpus of English. Both Hawkins (1999) and Wiechmann and Lohmann (2013) used entailment tests to define the argument status of the PP in relation to the verb. For instance, the PP *on his family* in the sentence *He counts on his family* is an argument of the verb *counts*, since the sentence does not entail *He counts*. By contrast, the PP *in the park* in the sentence *He played in the park* is an adjunct of the verb *played* because the sentence does entail *He played*. With the same entailment tests, Lohse et al. (2004) showed that the length of the object NP as well as the argument status of the particle in relation to the verb influence the orders of verb particle constructions in English.

2.3 Manner Place Time (MPT)

Proposed in Quirk et al. (1985), the traditional ordering rule for PPs and adverbials in postverbal position in English appears to follow Manner before Place before Time (MPT), as in *Zoey danced [manner elegantly] [place on the dance floor] [time at night]*. In contrast, this rule applies in the opposite direction when the PPs and adverbials occur in preverbal positions. That is, the ordering of preverbal PPs and adverbials follows Time before Place before Manner (TPM) (Hawkins, 1999). While Hawkins (1999) found that MPT plays no significant role in PP ordering in English, Wiechmann and Lohmann (2013) showed that it has a statistically significant yet weak effect.

3 Experiments²

3.1 Data

We searched for sentences in PTB and CTB with verb phrases containing exactly two PPs attached to the same side of the same head verb, where the ordering of the PPs allows certain flexibility.

Corpus	Total
WSJ	3596
Brown	3033
Switchboard	1187
Ctb	250

Table 1: Total VP instances for each corpus

3.2 Measures of each factor

3.2.1 Dependency length

Though different metrics have been applied to approximate dependency length in the literature, these measures have been demonstrated to be closely correlated (Gildea and Jaeger, 2015). To estimate the effect of dependency length on PP ordering, we followed the simple procedure as Hawkins (1999). We measured the lengths of the PP closer to the verb and of the PP farther from the verb as the number of tokens in each PP. We approximated phrase length using the number of tokens according to the treebank tokenization. We then calculated the proportion of cases where the shorter PP occurs closer to the head verb, the longer PP appears closer and when the two PPs are of equal length, for each corpus separately.

²Codes available at <https://zoeyliu18@bitbucket.org/zoeyliu18/pp-order-in-english-vs-chinese.git>

3.2.2 Argument status

To decide the argument status of a PP constituent, we borrowed the coding scheme from Merlo and Ferrer (2006), which carefully distinguishes PP arguments and adjuncts given their annotated grammatical function and semantic tag from the treebanks, shown in Table 2. As described in their paper, the motivation to include untagged PPs as arguments is due to that in the corpora, NPs (direct object & indirect object) and sentential constituents that are clearly arguments of the verb are left untagged (Marcus et al., 1994; Bies et al., 1995). The difference between argument and adjunct is gradient and not a binary distinction. Rather than looking at each PP as strictly an argument or an adjunct, we interpret the notion as an approximation for how argument-like and adjunct-like each PP is relative to the head verb. To analyze the effect of argument status, we only examined VP instances that have one argument-like PP and one adjunct-like PP (WSJ: $n = 1371$, Brown: $n = 1048$, Switchboard: $n = 470$, CTB: $n = 68$). We then computed the proportion of cases when the argument-like PP occurs closer. Statistical significance of the effects for both dependency length and argument status in each language were evaluated with Monte Carlo permutation test for 1,000,000 iterations.

Argument-like PPs	
-CLR	dative object if dative shift not possible (e.g., donate); phrasal verbs; predication adjuncts
-EXT	marks adverbial phrases that describe the spatial extent of an activity
-PUT	locative complement of <i>put</i>
-DTV	dative object if dative shift possible (e.g., give
-BNF	benefactive (dative object of <i>for</i>
-PRD	non VP predicates
untagged PPs	
Adjunct-like PP	
-DIR	direction and trajectory
-LOC	location
-MNR	manner
-PRP	purpose and reason
-TMP	temporal phrases

Table 2: Grammatical functions and semantic tags of PP constituents in the treebanks

3.2.3 Manner Place Time

In the treebanks, certain PPs have function tags that denote manner (PP-MNR), place (PP-LOC) or time (PP-TMP). We restricted our analysis to sentences that have both PPs annotated with these function tags. For English, we calculated whether the ordering of the two PPs follow MPT. For Chinese, we computed whether the ordering of the two PPs follow TPM.

3.3 Logistic Regression Models

We further compare the predictive power of dependency length and argument status in PP ordering with logistic regression modeling, which has been widely applied to model structural preferences (Bresnan and Ford, 2010; Gulordava et al., 2015; Levy and Jaeger, 2007; Morgan and Levy, 2015; Wasow et al., 2011). We did not include the rule of MPT in the model as the number of cases where it applies is quite small (see Section 4.5). Following similar methods in Rajkumar et al. (2016), we trained the logistic

regression models to predict the original observations in the corpora. For each model, we evaluated its prediction accuracy with Monte Carlo permutation test for 10,000 iterations.

We randomly selected half of the original instances extracted from the corpora and left them the way they were. For the other half, we constructed their structural variants simply by switching the order of the two PPs. Hence for the dataset of each corpus, half of the sentences are the originals while the other half are the constructed variants. The outcome binary variable is the ordering of the two PPs, represented as *Order*. We code *Order* as 1 for all original sentences, and 0 for all variants. Dependency length and argument status are included as the predictors in the model. For dependency length, we code it as 1 when the shorter PP is closer to the head verb, -1 when the longer PP is closer, and 0 when the two PPs have the same length. For argument status, we code it as 1 when the argument-like PP appears closer, -1 when the adjunct-like PP occurs closer, and 0 when the argument status of the two PPs is the same. A summary of our coding for the predictors is presented in Table 3.

Factor	1	-1	0
dependency length	short PP closer	long PP closer	equal length
argument status	argument-like PP closer	adjunct-like PP closer	same argument status

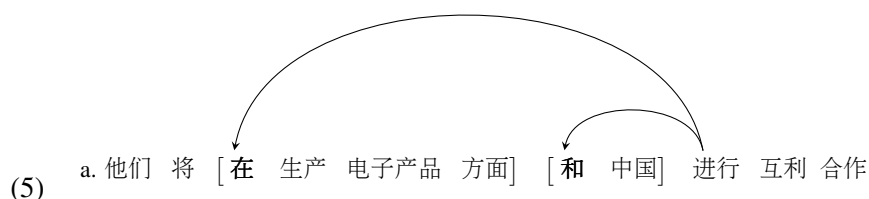
Table 3: Coding for Predictors in Logistic Regression Models

4 Results & Analysis

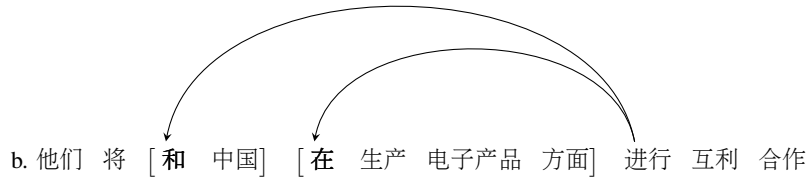
4.1 PP Ordering in Chinese

Previous work has demonstrated empirically that Chinese has a dominant SVO order (Sun and Givón, 1985; Liu et al., 2009; Mei, 1980). Nevertheless, compared to English, which has more consistent head-dependent orderings, the headedness of different structures in Chinese is profoundly inconsistent. The adposition system in Chinese has been argued and shown to have both prepositions and postpositions (Hawkins, 1994).

The VP instances that fit our search criteria in CTB (i.e. cases with exactly two PP dependents attached to the same side of the head verb) appear as (5), where two head-initial PPs are placed before the head verb. Different from the PP orderings in English (see Section 2.1), where both the VP and the PP are head-initial, here we observed inconsistent headedness between the VP and the PP. Though based on predictions by DLM, the structure of (5a) will be more preferred to that of (5b), as the shorter PP is closer to the head verb in (5a). Nevertheless, when the head verb has head-initial PP dependents, to derive optimal overall dependency lengths, the PPs should occur after, rather than before the head verb like Chinese. In the cases below, the longest dependency length between the first PP and the head verb is already incurred regardless of the orderings of the two PPs, so it may not matter as much whether the shorter PPs are closer to the head verb or not. Accordingly, we expect there to be much weaker or even no effect for dependency length in PP orderings in Chinese.



They will [in the aspects of electronic device production] [with China] conduct mutually beneficial collaboration.



They will [with China] [in the aspects of electronic device production] conduct mutually beneficial collaboration.

They will collaborate in a mutually beneficial fashion with China in the production of electronic devices.

4.2 Effect of dependency length

As shown in Figure 1³, the order predicted by DLM is strongly preferred in English. The number of sentences that have the shorter PP closer to the verb is 1.8 to 3.5 times larger than the number of sentences that have the longer PP closer to the verb. However, in roughly 20% of all sentences, DLM makes no prediction, since the two PPs have the same number of tokens. Although these numbers suggest that the preference for DLM is not as strong in spoken data as it is in written text, the preference for shorter dependencies is substantial across all three domains.

On the contrary, Chinese shows only a mild tendency for DLM. The number of cases when the shorter PP appears closer is not significantly much larger than that of instances when the longer PP is closer. This aligns with what we expected originally, that when inconsistent headedness exists between the VP and the PP, as in Chinese, dependency length does not seem to play a strong role.

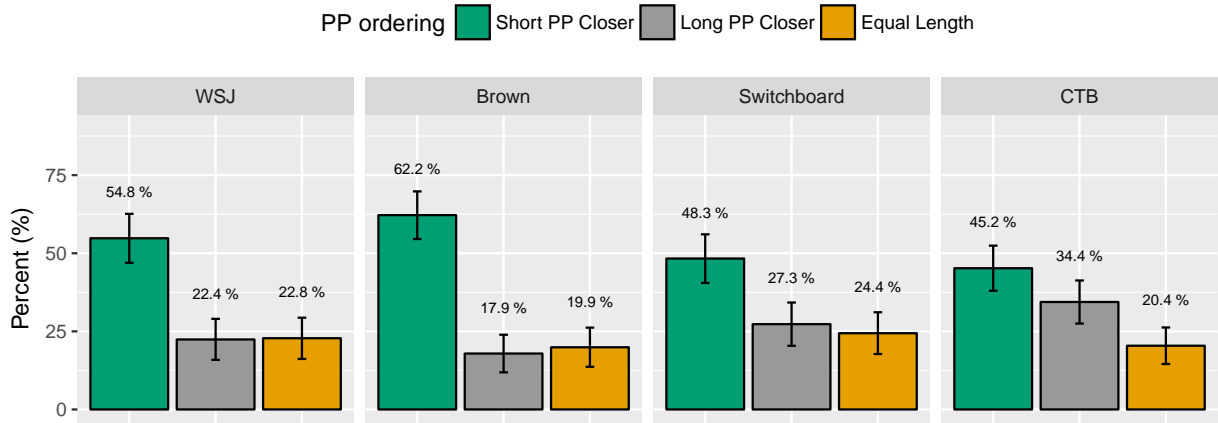


Figure 1: Effect of dependency length

To acquire a better understanding of why the efficacy of DLM is weaker in spoken genre than in written texts for English, we took a closer examination at the PP lengths of the extracted instances from the three corpora in PTB. We conjectured two possible reasons. First, compared to written texts, the average PP length for spoken genre might be much shorter. Second, spoken data might have more cases where the length difference between the two PPs is relatively small. Both indicate it might be less necessary to put the shorter PP closer to the head verb in Switchboard, leading to overall weaker preference for DLM. To test our conjectures, we computed the average PP lengths as well as the number of cases where the lengths of the two PPs differ by only 1-2 words. Nevertheless, as shown in Table 4, the average PP length in Switchboard is comparable to that in Brown, and only mildly shorter than that of WSJ (by 0.3 word). The proportion of cases where the two PPs have small length difference in Switchboard is similar to Brown, while slightly higher than WSJ (by 1.2%). This suggests that there are other potential constraints

³We estimated effects of dependency length after removing punctuation, as well as repetition in Switchboard for comparison, which did not appear to affect the results much. Thus we included punctuation for our calculation.

possibly competing with dependency length and working in different directions. They play stronger roles in the spoken than written domains in English and have overruled the impact of dependency length.

Corpus	Average PP length	% with small PP length difference
WSJ	5.4 ± 0.6	34.7 ± 6.7
Brown	4.7 ± 0.6	42.8 ± 6.9
Switchboard	4.0 ± 0.5	49.5 ± 6.9

Table 4: Comparisons of PP lengths

4.3 Effect of argument status

Consistent across domains for English and for Chinese, there appears to be a strong preference for argument-like PP to be close to the head verb. The number of instances where an argument-like PP is more adjacent is 1.5 to 2.7 times larger than when the adjunct-like PP occurs closer. By comparison, argument status has stronger effect in Switchboard and CTB than WSJ and Brown.

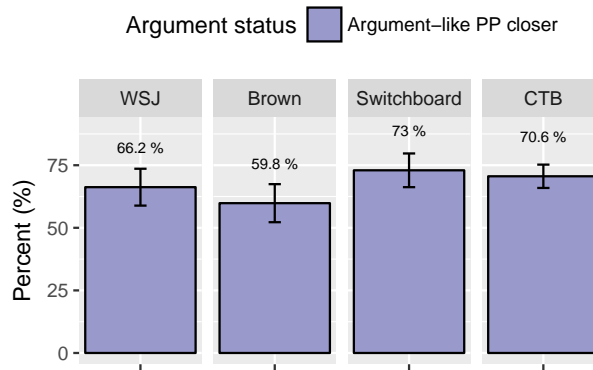


Figure 2: Effect of argument status of the PP

Now it is natural to ask how argument status interacts with dependency length pertaining the order of the two PPs. We estimated and compared the effects of argument status in sentences when the shorter PP appears closer versus when the longer PP is closer. In particular, in cases where shorter PPs are closer, it might matter less whether these shorter PPs are argument-like or not, since dependency length is already exerting a positive effect. Comparatively, in instances where longer PPs are closer, it is possible that most of the longer PPs are arguments of the verb, and tend to be more adjacent. Though results from Figure 3 do not align exactly with our initial thoughts, we observe some interesting patterns. In WSJ, when the longer PPs are close, the number that those longer PPs are arguments of the head verbs is significantly much larger. The preference for argument-like PP to be adjacent even when it is the longer PP suggests that when dependency length and argument status have the opposite effect, there will be strong competition between the two factors. This indicates that in WSJ, dependency length and argument status might have comparable predictive power in deciding what the PP ordering will be. In Switchboard, on the other hand, the argument-like PP is more adjacent to the head verb regardless of whether it is the shorter or the longer PP. The consistently pronounced effect for argument status here suggests that it might bear a stronger role than dependency length. When the two factors are pulling in different directions, the PP ordering might abide more by predictions of argument status than of DLM. However, in Brown, the number of the argument-like PPs being close is not much higher than by chance in spite of its length. This suggests there might be more cooperation rather than competition between the two constraints.

4.4 Cooperation and competition between dependency length and argument status

To further compare the cooperation and competition between the dependency length and argument status, we turn to evaluate and quantify the predictive power of the two factors with logistic regression mod-

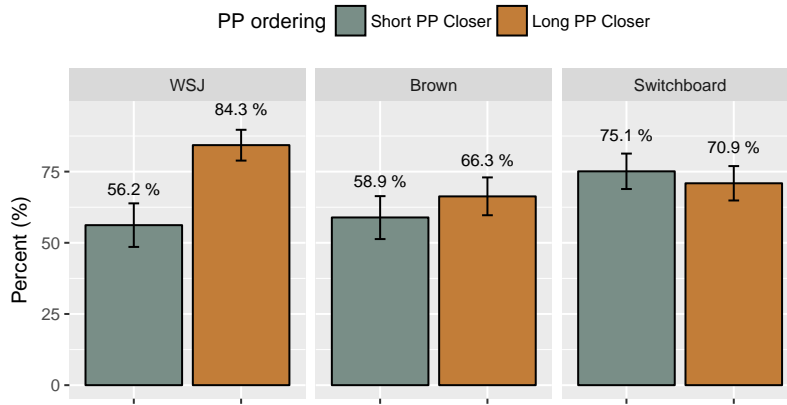


Figure 3: Effect of argument status when short vs. long PP is closer

els. We examined cases where at least one of the two constraints has an effect. Results from Figure 4 demonstrate that dependency length and argument status cooperate as well as compete with each other to different extents in the treebanks. The relative strengths of the two factors vary across domains in English and across the two languages. The most strongly preferred order is when the PP that is both shorter and argument-like to be adjacent to the head verb. On the other hand, competition between the two factors arise when they pull in the opposite directions (i.e. when the shorter PP is an adjunct or when the longer PP is an argument). The comparable predictive power for the two constraints in WSJ speaks to what we suggested earlier (see Section 4.3), that there is strong competition when dependency length and argument status are working against each other. In Brown, dependency length appears to be more predictive than argument status, indicating that the shorter PP is still more likely to be closer even when it is not an argument. In other words, the PP orderings in Brown will align more with predictions by DLM. In both Switchboard and CTB, argument status has a more pronounced role. This contrast suggests that in these two corpora, the orders tend to put the argument-like PP adjacent to the head verb, even if it is the longer PP between the two PPs.

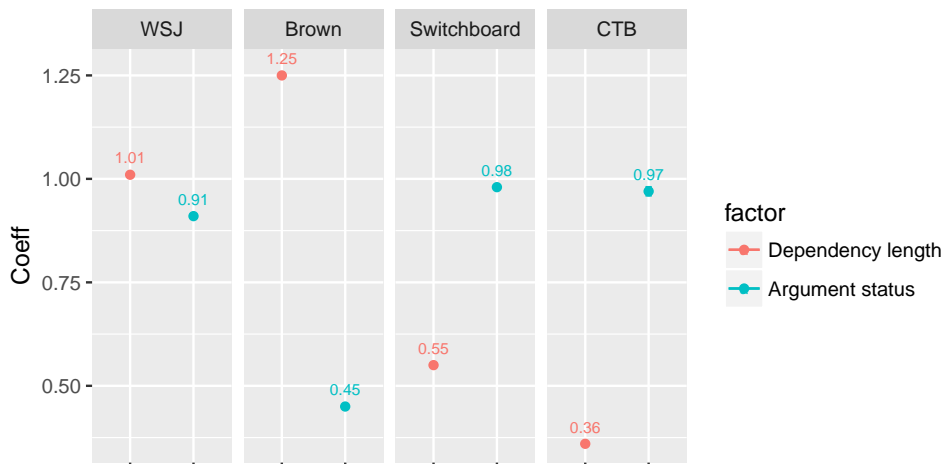


Figure 4: Coefficient estimates for dependency length and argument status in each corpus. Dots show point estimates with bars indicating standard errors.

4.5 Effect of Manner Place Time

Though there were not enough cases that TPM applies in Chinese, we found a significant role for MPT in English. This differs from Hawkins (1999) and Wiechmann and Lohmann (2013), which have shown no or weak effect for MPT, respectively. It is possible that their results are due to the use of smaller language

samples. In our dataset, MPT applies to about 6% of all instances. Within this set, it correctly accounts for the order of 89.3% of sentences in WSJ, 100% in Brown, and 100% in Switchboard. However, because it applies so infrequently, its overall impact is much smaller than that of dependency length and argument status.

5 Discussion & Conclusion

We analyzed the effects of dependency length, argument status and MPT in PP orderings for both English and Chinese. Consistent with previous studies, dependency length serves as a strong predictor for PP ordering across domains in English. Nevertheless, it only exerts a mild effect in Chinese. This relates to previous studies, which have shown that whether the preference for DLM exists and its efficacy are dependent on the headedness of the specific structures for languages with different typological characteristics (Lohmann and Takada, 2014; Faghiri and Samvelian, 2014; Faghiri et al., 2014). The argument status of the PP also has a pronounced effect on the orderings. It appears to play a comparable or even stronger role when compared to dependency length using logistic regression modeling. Overall, our results provide direct and quantitative evidence that dependency length and argument status are competing and cooperating motivations in PP ordering preferences across English and Chinese.

As effective as dependency length and argument status are, it is clear that around 30% of the data in English and around 40% of the data in Chinese remain unexplained based on model prediction accuracy presented in Table 5. Other constraints and their interactions with dependency length and argument status await to be discovered. One other factor that has been addressed previously on PP orderings is pragmatic information status (Hawkins, 1999; Wiechmann and Lohmann, 2013). Though Hawkins (1999) found no significant role for pragmatic information, it seems to have a mild effect based on results from Wiechmann and Lohmann (2013).

Corpus	Accuracy (%)
WSJ	67.3 ± 0.03
Brown	73.4 ± 0.03
Switchboard	66.4 ± 0.04
CTB	63.3 ± 0.10

Table 5: Model prediction accuracy with dependency length and argument status

Finally, previous experiments have presented contrary evidence regarding whether shorter dependencies will facilitate processing in Chinese relative clauses. Different from the head-initial relative clause structure in English, the head noun of relative clauses in Chinese comes in the final position. This results in longer dependencies in subject-extracted (SR) than object-extracted relative clauses (OR), whether the relative clause is modifying the subject or the object of the sentence. Certain studies have found that ORs are easier to process than SRs (Gibson and Wu, 2013; Hsiao and Gibson, 2003), providing support for predictions by DLM. On the other hand, findings from others have shown significantly shorter reading times for SRs for both adults (Hsiao and MacDonald, 2013; Vasishth et al., 2013; Chen et al., 2012; Chen et al., 2010) and children of different ages (Hu et al., 2016). As argued in Jäger (2015), expectation-based accounts are able to offer more thorough explanations. It is possible that SRs are processed faster due to its higher conditional probabilities given the preceding context in the sentence. Following this line of thought, it is likely that one’s probabilistic knowledge of the grammars for a language as well as the overall structural distributions of the language affect constituent ordering preferences. Extensions of these predictions to word order variations across languages will lead to a more fruitful research direction.

Acknowledgements

We would like to thank Kenji Sagae for his extensive feedback on the paper, as well as Annie Barbarika and the three anonymous reviewers for their helpful comments.

References

- Jennifer E Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.
- Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. *University of Pennsylvania*, 97:100.
- Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.
- Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94.
- Zhong Chen, Qiang Li, Kuei-Lan Kuo, and Shraavan Vasishth. 2010. Processing chinese relative clauses: Evidence for the universal subject preference. *Unpublished manuscript*.
- Zhong Chen, Kyle Grove, and John Hale. 2012. Structural expectations in chinese relative clause comprehension. In *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pages 29–37. Cascadilla Proceedings Project Somerville, MA.
- Hye-Won Choi. 2007. Length and order: A corpus study of korean dative-accusative construction. *Discourse and Cognition*, 14(3):207–227.
- Peter W Culicover, Ray S Jackendoff, Ray Jackendoff, et al. 2005. *Simpler syntax*. Oxford linguistics. Oxford University Press.
- William Edward Dyer. 2017. *Minimizing Integration Cost: A General Theory of Constituent Order*. Ph.D. thesis, University of California, Davis.
- Pegah Faghiri and Pollet Samvelian. 2014. Constituent ordering in persian and the weight factor.
- Pegah Faghiri, Pollet Samvelian, and Barbara Hemforth. 2014. Accessibility and word order: The case of ditransitive constructions in persian. In *The 21th International Conference on Head-Driven Phrase Structure Grammar*, pages pages–217.
- 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.
- Edward Gibson and H-H Iris Wu. 2013. Processing chinese relative clauses in context. *Language and Cognitive Processes*, 28(1-2):125–155.
- Edward Gibson, Richard Futrell, Steven Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. In Press. How efficiency shapes human language. *Trends in Cognitive Sciences*.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.
- Daniel Gildea and T Florian Jaeger. 2015. Human languages order information efficiently. *arXiv preprint arXiv:1510.02823*.
- 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. Association for Computational Linguistics.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.
- John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages 517–520. IEEE.
- Kristina Gulordava and Paola Merlo. 2015. Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257.

- Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 477–482.
- John A Hawkins. 1994. *A performance theory of order and constituency*, volume 73. Cambridge University Press.
- John A. Hawkins. 1999. The relative order of prepositional phrases in english: Going beyond manner–place–time. *Language variation and change*, 11(3):231–266.
- John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.
- John A. Hawkins. 2014. *Cross-linguistic variation and efficiency*. OUP Oxford.
- Franny Hsiao and Edward Gibson. 2003. Processing relative clauses in chinese. *Cognition*, 90(1):3–27.
- Yaling Hsiao and Maryellen C MacDonald. 2013. Experience and generalization in a connectionist model of mandarin chinese relative clause processing. *Frontiers in psychology*, 4:767.
- Shenai Hu, Anna Gavarró, Mirta Vernice, and Maria Teresa Guasti. 2016. The acquisition of chinese relative clauses: contrasting two theoretical approaches. *Journal of Child Language*, 43(1):1–21.
- Richard Hudson. 1995. Measuring syntactic difficulty. *Manuscript, University College, London*.
- Ray S Jackendoff. 1977. *X-bar syntax: A study of phrase structure*. Linguistic inquiry monographs 2. Cambridge, Massachusetts: MIT Press.
- T. Florian Jaeger and Elisabeth J. Norcliffe. 2009. The cross-linguistic study of sentence production. *Language and Linguistics Compass*, 3(4):866–887.
- Lena Jäger, Zhong Chen, Qiang Li, Chien-Jer Charles Lin, and Shravan Vasishth. 2015. The subject-relative advantage in chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, 79:97–120.
- Johannes Kizach. 2012. Evidence for weight effects in russian. *Russian linguistics*, 36(3):251–270.
- Lars Konieczny and Philipp Döring. 2003. Anticipation of clause-final heads: Evidence from eye-tracking and srns. In *Proceedings of the ICCS/ASCS Joint International Conference on Cognitive Science*, pages 13–17.
- Lars Konieczny. 2000. Locality and parsing complexity. *Journal of psycholinguistic research*, 29(6):627–645.
- Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Dartmouth Publishing Group.
- Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS)*.
- Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In Roger P. G. van Gompel, editor, *Sentence Processing*, page 78–114. Hove: Psychology Press.
- Haitao Liu, Yiyi Zhao, and Wenwen Li. 2009. Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics*, 45(4):509–523.
- Arne Lohmann and Tayo Takada. 2014. Order in np conjuncts in spoken english and japanese. *Lingua*, 152:48–64.
- Barbara Lohse, John A Hawkins, and Thomas Wasow. 2004. Domain minimization in english verb-particle constructions. *Language*, 80(2):238–261.
- Brian MacWhinney, Andrej Malchukov, and Edith Moravcsik. 2014. *Competing motivations in grammar and usage*. OUP Oxford.
- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.
- Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics.
- Kuang Mei. 1980. Is modern chinese really a sov language? *Cahiers de Linguistique-Asie Orientale*, 7(1):23–45.

- Paola Merlo and Eva Esteve Ferrer. 2006. The notion of argument in prepositional phrase attachment. *Computational Linguistics*, 32(3):341–378.
- Emily Morgan and Roger Levy. 2015. Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, page 1649–1654.
- Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.
- Randolph Quirk, Sidney Greenbaum, and Geoffrey Leech. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.
- Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written english syntactic choice phenomena. *Cognition*, 155:204–232.
- Mohammad Rasekh-Mahand, Mojtaba Alizadeh-Sahraie, and Raheleh Izadifar. 2016. A corpus-based analysis of relative clause extraposition in persian. *Ampersand*, 3:21–31.
- Chao-Fen Sun and Talmy Givón. 1985. On the so-called sov word order in mandarin chinese: A quantified text study and its implications. *Language*, pages 329–351.
- David Temperley. 2007. Minimization of dependency length in written english. *Cognition*, 105(2):300–333.
- Russell S Tomlin. 1986. *Basic word order: Functional principles*. London: Croom Helm.
- Shravan Vasishth and Richard L. Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.
- Shravan Vasishth, Zhong Chen, Qiang Li, and Gueilan Guo. 2013. Processing chinese relative clauses: Evidence for the subject-relative advantage. *PloS one*, 8(10):e77006.
- Thomas Wasow and Jennifer Arnold. 2003. Post-verbal constituent ordering in english. *Topics in English Linguistics*, 43:119–154.
- Thomas Wasow, T. Florian Jaeger, and David Orr. 2011. Lexical variation in relativizer frequency. *Expecting the unexpected: Exceptions in grammar*, pages 175–96.
- Thomas Wasow. 1997a. End-weight from the speaker’s perspective. *Journal of Psycholinguistic research*, (3):347–361.
- Thomas Wasow. 1997b. Remarks on grammatical weight. *Language variation and change*, (1):81–105.
- Daniel Wiechmann and Arne Lohmann. 2013. Domain minimization and beyond: Modeling prepositional phrase ordering. *Language Variation and Change*, 25(1):65–88.
- Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.
- Hiroko Yamashita and Franklin Chang. 2001. “long before short” preference in the production of a head-final language. *Cognition*, 81(2):B45–B55.
- Hiroko Yamashita. 2002. Scrambled sentences in japanese: Linguistic properties and motivations for production. *Text – Interdisciplinary Journal for the Study of Discourse*, 22(4):597–634.

Intervention effects in object relatives in English and Italian: a study in quantitative computational syntax

Giuseppe Samo

Department of Linguistics
Beijing Language and Culture University
samo@blcu.edu.cn

Paola Merlo

Department of Linguistics
University of Geneva
Paola.Merlo@unige.ch

Abstract

Discontinuous dependencies are one of the hallmarks of human languages. The investigation of the locality constraints imposed on such long-distance dependencies is a core aspect of syntactic explanations. The aim of this work is to investigate locality constraints in object relative clauses adopting a theory-driven and quantitative point of view. Based on a comparison of the theoretically expected and the observed counts of features of object relative clauses, we study which set of features plays a role in the syntactic computation of locality (type, number, animacy). We find both effects predicted by a narrow and a broad view of intervention locality. For example, in Italian the feature number triggers a numerically stronger effect than in English, a prediction of the narrow, grammar-driven view of locality. We also find that the feature animacy plays a role in the frequency of object relative clauses, an effect predicted by a broader view of locality.

1 Introduction

The aim of this work is to investigate locality issues adopting a quantitative computational syntax point of view (Merlo, 2016). Quantitative computational syntax uses large-scale resources and simple computational models in order to answer quantitative linguistic questions. In this paper, we concentrate on the quantitative aspect of long-distance dependencies according to a theory of intervention. Based on a comparison of the theoretically expected and the observed counts of features in grammatical structures, we study which set of features plays a role in the syntactic computation of long-distance dependencies.

A core distinguishing property of human languages is the ability to interpret discontinuous elements as if they were a single element. Sometimes these elements are distant in the string. These are called long-distance dependencies. For example, sentence (1a) is an object-oriented restrictive relative clause, where the object of the verb *show* is also the semantic object of the verb *wash*, connecting two distant elements. The sentence (1b) is a subject-oriented restrictive relative clause, where the semantic object of the verb *show* is also the subject of the verb *wash*.¹

(1a) Show me the elephant that **the lion** is washing <the elephant>.

(1b) Show me the elephant that <the elephant> is washing the lion.

Long-distance dependencies are not all equally acceptable (Rizzi, 2004). The facts involving them are complex, and a precise description encompassing all phenomena is one of the major topics of research in current linguistic theory (Rizzi, 1990; Gibson, 1998). We study here the predictions of an intervention theory of locality (Rizzi, 1990). In a nutshell, a long-distance dependency between two elements in a sentence is difficult, and often impossible in child grammar (Friedmann et al., 2009), if a *similar* element intervenes. For example, sentence (1a) causes trouble for children while (1b) does not, because in (1a) the *lion* intervenes between the two discontinuous occurrences of *elephant* (one pronounced one silent), while in (1b) there is no intervener.

Core to the explanation of these facts is the notion of *intervener*. An intervener is an element that is *similar* to the two elements that are in a long-distance relation, and structurally intervenes between the two, blocking the relation. In our examples, the intervener is *the lion*, shown in bold.

¹The unpronounced element(s) in the long-distance relation are indicated by < >.

Notice that here and in all the following, intervention is defined structurally and not linearly. Linear intervention that does not structurally hierarchically dominate (technically *c*-command) does not matter, as shown by the contrast **When do you wonder who won?/You wonder who won at five* compared to *When did the uncertainty about who won dissolve?/The uncertainty about who won dissolved at five* (Rizzi, 2013). Notice also that the non *c*-commanding, more acceptable alternative creates a linearly longer dependency than the *c*-commanding more difficult one, therefore also showing that length of the dependency does not directly affect acceptability.

Defining and justifying which properties come into play in computing whether two elements are similar or not is therefore a crucial element in this explanation. In this paper, we briefly review some results from the theoretical and experimental literature that have attempted to characterise precisely this notion of intervener in the case of object relative clauses. Based on their findings, we develop hypotheses of the expected corpus distributions.

2 Object relatives and intervention locality

A robust set of experimental studies and results on both production and comprehension of relatives clauses, both subject relatives and object relatives on acquisition (Friedmann and Novogrodsky, 2004), on adult processing (Frauenfelder et al., 1980), and on pathology (Grillo, 2008) confirms that object relatives are harder than subject relatives, in various respects both in children and adult grammar, as shown in example (1) above. The intervention locality explanation ascribes this difficulty to the fact that the subject acts as intervener between the head of the relative clause and the object position in object relatives, while in subject relative clauses no intervention occurs (Friedmann et al., 2009).

According to this theory, the crucial property in intervention is not the amount of material that can be considered as intervener, but rather its quality. If the head of the relative clause and the intervener share some computationally relevant features, this leads to slower processing for adults.

One important aspect in verifying the intervention-based explanation of the difficulty of object relatives, then, is determining which features trigger the intervention. Object relatives are grammatical structures and thus the type of intervention could be qualitatively different from other long-distance constructions creating ungrammatical sentences, such as long distance complex questions (*wh*-islands). According to recent studies (Belletti et al., 2012), the relevant features in intervention in object relative clauses are those features which could be considered syntactically relevant in the language. In particular, those features able to trigger the movement of syntactic elements such as the subject and the verb. In English and Italian, features such as *number* and *person* (Bentea, 2016) have been investigated and the various forms of noun phrases, such as pronouns (head) vs. maximal projection, indicated as XP, (we will call it *type*) (Friedmann et al., 2009).² Finally, the status of an *animacy* feature remains controversial; some results argue in favour of an ameliorative effect (Brandt et al., 2009), some suggest animacy has no effect (Adani, 2012). Some recent studies show a clear effect of animacy as an intervention feature in *wh*-islands, another kind of long-distance dependency (Villata and Franck, 2016). Further evidence for the need for a finer theory of locality comes from studies in language pathology and language acquisition, where, within the same language, the grammar of different populations (e.g. the grammar of adults vs. the grammar of children) exhibits different locality effects (Grillo, 2008; Friedmann et al., 2009; Belletti et al., 2012).

3 Quantifying the hypotheses

We choose to investigate the features of *type*, *number* and *animacy*. We show in Table 1 some examples of relatives clauses with these features. We select these features to explore several dimensions of variation. First of all, the notion of *type* (head or maximal projection) goes back to the core formulation of

²Several pieces of work in language acquisition, adult processing and language pathology have investigated a set of morphosyntactic features such as *number*, *animacy*, *gender*, *case* and *lexical restriction*. A non-exhaustive list of reference is to be found at the ERC Syncart website <https://www.unige.ch/lettres/linguistique/syncart/cartographylocality/references/thematic-order/th/> edited by Karen Martini. According to some proposals these sets of features may be organised in a structural typology expressed as morphosyntactic features (Rizzi, 2004). Argumental: *person*, *number*, *gender*, *case*; Quantificational: *Wh*, *Neg*, *measure*, *focus*; Modifier: *evaluative*, *epistemic*, *Neg*, *frequentative*, *celerative*, *measure*, *manner*; *Topic*.

head of relative		subject	
the debate	which	we	held
XP, singular, inanimate		head, plural, animate	
these lovely little chocolates	that	we	get
XP, plural, inanimate		head, plural, animate	
Il terreno	che	l' acqua	copre
<i>the ground</i>	<i>that</i>	<i>the water</i>	<i>covers</i>
XP, singular, inanimate		XP, singular, inanimate	

Table 1: Examples of object relative clauses in several featural configurations. The examples are naturally occurring clauses extracted from the Universal Dependency corpora. XP=maximal projection.

intervention locality theory and has been shown to be active in the acquisition of object relative clauses (Rizzi, 1990; Friedmann et al., 2009). The morphosyntactic feature of number (singular or plural) has been studied because it is related to the richness of the verbal morphological system (Bentea, 2016). It has been argued that a rich morphological system triggers greater verb movement (Pollock, 1989). If this is the case, then morphosyntactic features in Italian may show a different strength of intervention than in English, since Italian has richer morphology and a greater movement of the verb. Animacy is a lexical semantic feature, whose influence on intervention is still controversial, as indicated above, and for which there is no reason to expect a cross-linguistic difference. Based on the findings in the theoretical and experimental literature, we can formulate the following questions.

1. Do the features *type*, *number* and *animacy* play a role in intervention effects?
2. If the features play a role in intervention effects, are these effects stronger in one of the two languages?

To answer these questions quantitatively based on corpus counts, we need to define the concept of similarity, central to the notion of intervention, and a linking hypothesis.

- *Similarity* The head of the relative clause and the intervener are *similar* if their features match.
- *Feature match* A *feature match*, $match_f(C, I)$, is true iff, for a given feature f , the head of the relative C and the intervener I have the same value.
- *Linking hypothesis* If a feature is a stronger intervener, we expect it to create greater inacceptability and hence surface less often in a corpus in a match configuration.

In this work, we make use of observational data provided by corpora, and operate on counts. We will refer to the notion of *observed counts*, as usual, as the counts in the corpus, and to *expected counts* as the counts of the features that we would expect based on their distribution in a setting where intervention is not at play and, therefore, they do not interact with each other. That is, the expected counts are the counts we would expect given the probability of the two features to cooccur independently of intervention, proportionally to the size of the corpus. Specifically, an object-oriented relative clause brings into play the object of the verb and its features, the noun phrase that is being relativized, and the subject of the sentence and its features, the intervener. Precisely, let C_s^f be the counts of a subjects feature and C_o^f be the counts of an object feature in a sample of size S . Let T be the total number of observations. Then, the expected counts of subject and object features occurring in a sentence with intervention are calculated as $C_s^f/S \times C_o^f/S \times T$, namely the product of the relative frequencies of these two elements, counted independently, in a sample, scaled by the total size of the corpus.

As we said at the beginning, we use corpus counts and frequencies in the spirit of the computational quantitative syntax framework: *differentials in counts are the expression of underlying grammatical properties*. In this respect, our quantitative hypotheses below are to be contrasted to an H_0 hypothesis that would predict that grammatical properties are uncorrelated to observed counts in a corpus, because corpus counts are effects of usage, while grammar makes no predictions about them, and as such there is

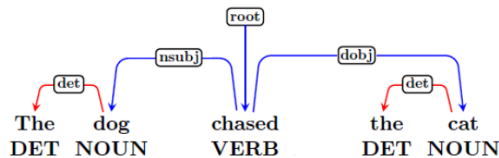


Figure 1: Canonical order

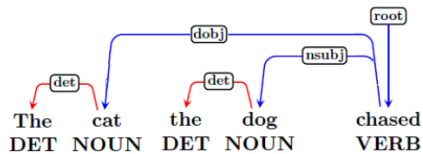


Figure 2: Non-canonical order

no expectation of distribution of counts beyond the observed ones. Based on these notions, we formulate then the following quantitative hypotheses.

H_1 Both in Italian and English, if the features *type*, *number* or *animacy* trigger intervention effects, we expect match configurations to be less frequent than expected. (Possibly, non-match configurations are more frequent than expected.)

H'_1 If the features *number* triggers intervention effects, the effect (the difference between expected and observed matches) should be larger in Italian than in English.

4 Materials and methods

Our hypotheses above follow a common schema that requires calculating the observed counts of a feature in the corpus and compare it to the counts we would expect if intervention was not at play. The annotated corpora we use are the universal dependency treebanks for Italian and English.

4.1 The corpus

We extract our counts from the Italian and English Universal Dependencies (UD) treebanks (Nivre, 2015) version 2.0 (<http://universaldependencies.org/>).³ The data comes from five different treebanks: English ParTut (Bosco and Sanguinetti, 2014), English LinEs (Ahrenberg, 2015), English UD (Bies et al., 2012), Italian ParTut (Bosco and Sanguinetti, 2014), Italian UD (Bosco et al., 2013). They comprise a variety of text genres. For English: blogs, social media, reviews, fiction, nonfiction, spoken legal, news, wiki For Italian : legal, news, wiki. In what follows, the analysis of the phenomenon will not be presented according to the different corpora, but all the different treebanks for each languages will be merged and thus we will refer to Italian and English data without the specification of the treebanks. The reason for this merge is the observation that different corpora show very large fluctuations in distributions of grammatical constructions, in general, and specifically for relative clauses (Roland et al., 2007; Belletti and Chesi, 2014, 3-5). By merging the different corpora, we limit the corpus-related biases in our counts and increase the generality of our results.

We choose UD because a comparative investigation is made possible by the uniformity of UD annotation across languages, although comparative syntax investigations are not the purported goal for which Universal Dependencies was developed.⁴ Our theory is formulated in terms of discontinuous dependents, which are not explicitly encoded in UD, so we need to adapt our searches to a dependency annotation. UD dependencies can be defined by the syntactic relation label and its direction. Syntactic relations, in

³To extract the construction of interest here, object relative clauses, we used the tool SETS, a query UD treebank search tool developed and maintained by the University of Turku (Haverinen et al., 2013).

⁴Universal Dependencies (UD) aim to provide a cross-linguistically uniform syntactic representation to advance multilingual applications of natural languages processing (Nivre, 2015).

Treebank	objects	left objects	OR	%OR
English ParTut (Bosco and Sanguinetti, 2014)	3186	51	44	86
English LinEs (Ahrenberg, 2015)	5985	139	16	11
English UD (Bies et al., 2012)	15259	403	191	47
Italian ParTut (Bosco and Sanguinetti, 2014)	3142	56	49	71
Italian UD (Bosco et al., 2013)	14639	549	216	39

Table 2: Object dependencies to the left (left objects) and object relatives (OR) in English and Italian.

	English		Adjusted English		Italian		Adjusted Italian	
	Subject	Object	Subject	Object	Subject	Object	Subject	Object
XP	.49	.91	.49	1.0	.62	.86	.62	1.0
head	.48	.09	.48	.00	.05	.14	.05	.00
null	.03	-	.03	.00	.33	-	.33	.00
singular	.70	.73	.70	.73	.74	.67	.74	.67
plural	.30	.27	.30	.27	.26	.33	.26	.33
animate	.93	.22	.93	.22	.78	.20	.78	.20
inanimate	.07	.78	.07	.78	.22	.80	.22	.80

Table 3: The proportions of expected counts and adjusted expected counts of the different features, in English and Italian, for subjects and objects.

a given language, have a canonical direction, for example subjects are left directed arcs and objects right directed arcs in an SVO language as shown in Figure 1. Non-canonical directions, left for object, for example, can be indicators of discontinuous dependency, as illustrated in Figure 2. The query search ‘VERB > obj@L_’ provides all the occurrences of the objects on the left of the verb. Statistics on the collected constructions and the number of object relatives after manual inspection and validation are shown in Table 2. This table shows that the occurrences of objects whose dependency is on the left represent a very small proportion in corpora; a sizable proportion of these left dislocated objects are object relatives (ORs).⁵

4.2 The counts

Like many experimental settings, our predictions are formulated in terms of differences in numerical observations. In this case, we study the difference between the expected counts and the observed counts.

Expected counts To validate our hypotheses, we establish the expected counts of morpho-syntactic and animacy features in all syntactic configurations. We sampled randomly one-hundred sentences in the English UD treebanks and one-hundred sentences in the Italian UD treebanks, selected with the SETS treebank search tool (the search query ‘_ <obj _’ provided all the results of sentences having an overt object). We then coded the features of subjects and objects. The subjects in a normal construction become the interveners in an object relative clause and the objects are the element that undergoes the relative clause long-distance dependency. Post-verbal subjects in Italian were considered as having the same intervening effects as preverbal subjects and were included in the subject counts.

The three features we code are *type*, *number* and *animacy*. Number is indicated in the morphology of the noun phrase and agreement in the sentence and leads to unambiguous classification. Type and animacy require judgement. Type can have the values head or maximal projection: pronouns were coded as heads, other noun phrases were coded as maximal projections. For animacy, elements such as *the parliament*, *the commission* (animate collectives) were labelled as animate elements. Those relative heads and subjects involving human beings, animals or groups of human beings and/or animals (e.g. the

⁵Other than object relatives, left dislocated objects have been analysed as Topics, anaphoric clitics, resumptive clitics, Wh-elements, Embedded Wh-elements, Wh-DP and rare cases of imperfect annotations with heterogenous distributions in the different treebanks.

Match			Relative head			Intervener			Sentence
type	num	an	type	num	an	type	num	an	
0	0	0	XP	sg	in	head	pl	an	<i>the foreign investment</i> that they need to help their economies grow
0	1	0	XP	pl	in	head	pl	an	<i>the fees</i> that they charge
1	0	0	XP	sg	in	XP	pl	an	<i>a luxury</i> that only rich countries can afford
1	0	1	XP	sg	an	XP	pl	an	<i>a better person</i> that people are wanting to hire
1	1	0	XP	sg	in	XP	sg	an	<i>a realist technique</i> which French novelist Marcel Proust later named <i>retrospective illumination</i>
1	1	1	XP	sg	in	XP	sg	in	a format that Access recognizes

Table 4: Examples of OR clauses in several featural configurations in English. The examples show the values of the features and if they match (1) or not (0) between head of the relative clause (in italics) and the intervener (in bold). The examples are naturally occurring clauses extracted from the UD corpora.

Match			Relative head			Intervener			Sentence
type	num	an	type	num	an	type	num	an	
0	0	0	XP	pl	in	null	sg	an	<i>i luoghi</i> che [0] aveva visitato spesso (<i>the places that (s/he) had visited often</i>)
0	0	1	XP	pl	in	head	sg	in	<i>i seri problemi</i> che ciò genera (<i>the serious problems that this engenders</i>)
0	1	0	XP	pl	in	null	pl	an	<i>la prima cosa</i> che [0] vide (<i>the first thing that (s/he) saw</i>)
0	1	1	XP	sg	an	null	sg	an	<i>l'associazione per l'abolizione della pena di morte</i> che [0] aveva fondato (<i>the association for the abolition of the death penalty that (s/he) had founded</i>)
1	0	0	XP	pl	in	XP	sg	an	<i>i sonetti</i> che Shakespeare intendeva pubblicare (<i>the sonets that Shakespeare meant to publish</i>)
1	0	1	XP	pl	in	XP	sg	in	<i>le limitazioni</i> che la legge stabilisce (<i>the limitations that the law dictates</i>)
1	1	0	XP	sg	in	XP	sg	an	<i>il tipo di effetto</i> che Balzac tentava di ottenere nelle sue opere (<i>the type of effect that Balzac attempted to obtain in his works</i>)
1	1	1	XP	sg	in	XP	sg	in	Il terreno che l' acqua copre (<i>the ground that the water covers</i>)

Table 5: Examples of OR clauses in several featural configurations in Italian. The examples show the values of the features and if they match (1) or not (0) between head of the relative clause (in italics) and the intervener (in bold). Empty subjects are indicated as [0]. The examples are naturally occurring clauses extracted from the UD corpora. A literal translation of the Italian clauses is given in parentheses.

government, the European Union, Russia), were labelled as animate. The nouns denoting non-human beings and non-animals were considered inanimate.

The expected counts of the different features, in English and Italian, for subjects and objects are shown in the contingency Table 3. We include a column indicating adjusted counts. This is based on the observation that relatives clauses with a pronoun head or a null head are extremely rare or impossible. This is because neither English nor Italian are null-object languages and, thus, a null relative head will result in an ungrammatical sentence. So, in fact, the counts of these features in a relative clause are different from their distribution in a simple transitive sentence. We will use the adjusted expected counts for our comparisons.

Observed counts We also need to collect the observed counts of cooccurrence of features. The statistics of observed relative clauses are given in Table 2, which indicates that we have 251 object relative clauses for English and 265 for Italian. A manual analysis by the first author coded the features and the match vs. mismatch conditions. We show some examples of relatives clauses with the feature coding in Tables 4 and 5.⁶ The Boolean values indicate the feature coding in a summarised way, by indicating if the features of the head of the relative clause and of the intervener match (1) or not (0). This encoding is just a shorthand for illustratory purposes. It was not added to the coding of features.

For example, the sentences *the foreign investment that they need to help their economies grow* is coded $\langle 0,0,0 \rangle$, as none of the features type, number, and animacy match, as *the foreign investment* is of type maximal projection, number singular, and inanimate, while *they* is a plural, animate pronoun head. The example *the fees that they charge* is coded as $\langle 0,1,0 \rangle$, as *the fees* is a plural, inanimate maximal projection, while *they* is a plural, animate pronoun head. Finally, the example *Il terreno che l'acqua copre (the ground that the water covers)* is coded as $\langle 1,1,1 \rangle$, as both *il terreno (the ground)* and *l'acqua (the water)* are singular, inanimate maximal projections.

5 Results and discussion

The calculations of expected counts and actual observed counts, the probabilities of these observations under a binomial distribution and their statistical significance are shown in Table 6. The binomial test gives us the probability of k successes in n independent trials, given a base probability p of an event. The event in our case is the cooccurrence of two features. So, for example, the binomial distribution tells us the probability of the (anim, anim) pair of features in English. Specifically, it tells us the probability of 20 successes in 251 trials given a base probability of the event of $.93 \times .22 = 0.2046$. The base probability of the event is, in our case, the product of the probabilities of the subject and object features, that is the probability of cooccurrence of these two features if they were independent and not in an intervention configuration. If certain conditions are met, the binomial distribution can be approximated by the normal distribution and a significance test can be performed. We calculate the cumulative probability distribution: the probability that the observed counts are exactly as observed, or greater, if the observed counts are larger than the expected counts, or the probability that the observed counts are exactly as observed, or smaller, if the observed counts are smaller than the expected counts. The z -score gives us the (one-tailed) probability of exactly, or greater/smaller counts than the expected counts.

The results that confirm the hypotheses, because they are in the right direction numerically and statistically significant, are shown in bold. These results are mixed, but have some interesting sub-regularities. In the match configurations, hypothesis H_1 is confirmed for the features *type* and *animacy* in most cases, for both English and Italian. Only the (inanimate, inanimate) pair in English is numerically smaller than expected, and as such confirming the hypothesis, but not significantly so statistically. For these features, we also observe an increase of observed non-match configurations, where statistically valid conclusions can be drawn. Mismatches are robustly more frequent than expected, especially in Italian. This is possibly compatible with an intervention effect, if we take these preferences for non-matching configurations as preference for alternative forms to avoid matches. We also observe that for these features, both in the match and mismatch configuration, the hypothesis is not confirmed only in the smaller or zero observed

⁶The supplementary materials with all the coded data are also available from the first author.

Match condition English						
HRel	Interv	Expected	Observed	p	Binomial p	$z-p$
XP	XP	123.0	108	0.490	0.033	0.033
sing	sing	128.7	132	0.511	0.341	0.341
plur	plur	20.3	22	0.081	0.382	0.393
anim	anim	51.4	20	0.205	0.000	< .000001
inan	inan	13.7	12	0.055	0.399	0.384

Match condition Italian						
HRel	Interv	Expected	Observed	p	Binomial p	$z-p$
XP	XP	164.3	149	0.62	0.0313	0.03053
sing	sing	131.4	138	0.496	0.218	0.218543
plur	plur	22.7	34	0.86	0.011	0.007814
anim	anim	41.3	23	0.156	0.0006	0.001263
inan	inan	46.6	27	0.176	0.0006	0.001009

Mismatch condition English						
HRel	Interv	Expected	Observed	p	Binomial p	$z-p$
XP	head	120.5	135	0.480	0.383	0.038
XP	null	7.5	0	0.030	0.0005	<i>n.v.</i>
sing	plur	47.4	49	0.219	0.203	0.202
plur	sing	53.2	40	0.189	0.131	0.132
anim	inan	3.9	0	0.015	0.022	<i>n.v.</i>
inan	anim	182.1	211	0.725	0.00001	0.00003

Mismatch condition Italian						
HRel	Interv	Expected	Observed	p	Binomial p	$z-p$
XP	head	13.3	29	0.050	0.000075	0.000009
XP	null	87.5	101	0.330	0.0453	0.044109
sing	plur	46.2	59	0.174	0.0249	0.022341
plur	sing	64.7	48	0.244	0.0088	0.010407
anim	inan	11.7	0	0.044	0.000007	0.000415
inan	anim	165.4	229	0.624	0.00000001	0.000001

Table 6: Expected counts and observed counts. Expected counts are based on adjusted proportions. English $N = 251$, Italian $N = 265$. p is the prior probability of the event. Binomial p indicates the probability of the observed counts under a binomial distribution (the binomial test). $z-p$ is the statistical significance of the binomial probability. *n.v.* indicates that conditions are not met for a valid calculation of statistical significance. The $z-p$ gives us the (one-tailed) probability of exactly the observed, or greater/smaller counts than the expected counts, for $\alpha = 0.5$. Results confirming the hypotheses are in bold.

counts. We reserve to investigate further if this result is due to a too small sample size. Notice that the feature *animacy* clearly triggers intervention effects, both in Italian and English, with a big preference for the mismatch configuration and dispreference for the match. This is quite interesting, as the results concerning this features are still not entirely converging. Experimental work on *wh*-islands indicate that the feature is relevant (Villata and Franck, 2016), showing a similar preference for mismatches over controls as what we found here, but results from acquisition seem to indicate it is not (Adani, 2012). Our results show that, at least in the adult grammar in written text, animacy makes a difference to the preference for choice of relative head and intervener in an object relative.

Instead, neither H_1 nor H'_1 are convincingly confirmed for the feature *number*. For H_1 , none of the predictions in the match configurations are confirmed and only half of the mismatch configurations are. With respect to the cross-linguistic hypothesis H'_1 about the feature *number*, the numerical differences do show a greater differential in Italian than in English, but not always in the right direction. All aspects of the hypotheses that concern the feature *number*, then, need further investigation.

The corpus investigation reported here provides a new contribution to the debate about what features count in intervention and what do not. As discussed in work by Franck and Villata, one approach defines the relevant notions of similarity as *narrow similarity*, where only morphosyntactic features count (Rizzi, 2004; Belletti et al., 2012). Another approach defines a notion of *broad similarity*, where any syntactic or semantic features can count, as long as they can be related to verb argument relations. (See, for example, Villata and Franck, which also show an effect of animacy in *wh*-islands). Our results seem to indicate that a more articulate characterisation of intervention locality is needed, as we find results compatible, but only partially, with both approaches. The distinction in strength of effect of the *number* feature between Italian and English and the effect of *type* feature is predicted by a narrow theory of similarity, that ties the effects and its strength to the morpho-syntactic make up of the language. The effect of *animacy*, though, extends the set of features relevant to intervention to lexical semantic aspects of the actants in grammatical long-distance dependencies.

These corpus results also join the rich current debate on the exact nature of structural dependencies and locality in computational method, and like other approaches, show for the moment, mixed conclusions. While some experiments have shown that Recursive Neural Networks can learn the main descriptive properties of long-distance dependencies in English, for example the fact that they obey a uniqueness constraint (only one gap per filler) and also that they obey island constraints (Wilcox et al., 2018), work attempting to replicate finer-grained human judgments for French have failed to show a correlation with human behaviour (Merlo and Ackermann, 2018), while other work on English has found mixed results (Chowdhury and Zamparelli, 2018). Lack of correlation with human grammaticality judgments has also been found in *wh*-islands and object relative clauses for both French and English (Merlo, 2019). More work will be needed to establish the exact boundaries of quantitative properties in long-distance dependencies across several languages.

6 Conclusions and future work

The contributions of this treebank study are many-fold. First, we formulate quantitative predictions about object-oriented relative clauses based on intervention theory. These predictions aim to identify which features come into play in defining the notion of intervener, and with what strength. Our results corroborate some previous findings concerning morphosyntactic features and animacy, but not all, opening the door to further investigation.

Future work will have to extend the investigation to other features and to other constructions that have been proposed and discussed in the theory and develop more complex models of intervention similarity. Current work is investigating the morpho-syntactic feature *person* and models of similarity related to word embeddings (Merlo, 2019).

Finally, thanks to the resources such as UD, we can also envisage to extend the investigation to the many languages for which theoretical predictions already exists and help formulate new ones in new languages.

Acknowledgments

The work was partially done while the first author was participating in the ERC Advanced Grant n. 340297 SynCart, which we gratefully acknowledge.

References

- Flavia Adani. 2012. Some notes on the acquisition of relative clauses: new data and open questions. In Valentina Bianchi and Cristiano Chesi, editors, *ENJOY LINGUISTICS! Papers offered to Luigi Rizzi on the occasion of his 60th birthday*, pages 6–13. CISCLPress.
- Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to universal dependencies. In *Third International Conference on Dependency Linguistics (DepLing 2015)*, pages 10–19, Uppsala, Sweden, August. Association for Computational Linguistics.
- Adriana Belletti and Cristiano Chesi. 2014. A syntactic approach toward the interpretation of some distributional frequencies: Comparing relative clauses in Italian corpora and in elicited production. *Rivista di Grammatica Generativa*, 36:1–28.
- Adriana Belletti, Naama Friedmann, Dominique Brunato, and Luigi Rizzi. 2012. Does gender make a difference? Comparing the effect of gender on children’s comprehension of relative clauses in Hebrew and Italian. *Lingua*, 122(10):1053–1069.
- Anamaria Bentea. 2016. *Intervention effects in language acquisition: the comprehension of A-bar dependencies in French and Romanian*. Ph.D. thesis, University of Geneva.
- Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.
- Cristina Bosco and Manuela Sanguinetti. 2014. Towards a Universal Stanford Dependencies parallel treebank. *CLARIN-D*, page 14.
- Cristina Bosco, Montemagni Simonetta, and Simi Maria. 2013. Converting Italian treebanks: Towards an Italian Stanford Dependency treebank. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.
- Silke Brandt, Evan Kidd, Elena Lieven, and Michael Tomasello. 2009. The discourse bases of relativization: An investigation of young German and English-speaking children’s comprehension of relative clauses. *Cognitive Linguistics*, 20(3):539–570.
- Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING’18)*, pages 133–144. Association for Computational Linguistics.
- Ulrich Hans Frauenfelder, Juan Segui, and Jacques Mehler. 1980. Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, 19(3):328–337.
- Naama Friedmann and Rama Novogrodsky. 2004. The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child language*, 31(3):661–681.
- Naama Friedmann, Adriana Belletti, and Luigi Rizzi. 2009. Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, 119(1):67 – 88.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.
- Nino Grillo. 2008. *Generalized minimality: Syntactic underspecification in Broca’s aphasia*. Ph.D. thesis, University of Utrecht.
- Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2013. Building the essential resources for Finnish: the Turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.
- Paola Merlo and Francesco Ackermann. 2018. Vectorial semantic spaces do not encode human judgments of intervention similarity. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018*, pages 392–401, Brussels, Belgium, October.

- Paola Merlo. 2016. Quantitative computational syntax: some initial results. *Italian Journal of Computational Linguistics*, 2(1):11–30.
- Paola Merlo. 2019. Probing word and sentence embeddings for long-distance dependencies effects in French and English. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy, July. Association for Computational Linguistics.
- Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16. Springer.
- Jean-Yves Pollock. 1989. Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry*, 20:365–424.
- Luigi Rizzi. 1990. *Relativized Minimality*. MIT Press, Cambridge, MA.
- Luigi Rizzi. 2004. Locality and left periphery. In Adriana Belletti, editor, *The cartography of syntactic structures*, number 3 in Structures and beyond, pages 223–251. Oxford University Press, New York.
- Luigi Rizzi. 2013. Locality. *Lingua*, 130(1):69 – 86.
- Douglas Roland, Frederic Dick, and Jeffrey L Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of memory and language*, 57(3):348–379.
- Sandra Villata and Julie Franck. 2016. Semantic similarity effects on weak islands acceptability. In *41st Incontro di Grammatica Generativa Conference*, Perugia, Italy. <https://archive-ouverte.unige.ch/unige:82418>.
- Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221. Association for Computational Linguistics.

An explanation of the decisive role of function words in driving syntactic development

Anat Ninio

Department of Psychology, The Hebrew University of Jerusalem, Jerusalem 91905, Israel
Anat.Ninio@huji.ac.il

Abstract

The early mastery of function words (FWs) better predicts children's concurrent and subsequent syntactic development than their acquisition of content words (CWs). Wishing to understand why the advantage of the early mastering of a FW vocabulary, we tested the hypothesis that the learning of FWs involves learning their syntax to a higher degree than is the case for CWs. English-language parental (N=506) and young children's speech samples (N=350) were taken from the CHILDES archive. We mapped the use of words of different form-classes in parental speech, comparing the words' occurrence as single-word utterances and as the heads of two-word long syntactically structured sentences. The distributions showed a dramatic effect of form-class: the four FW categories subordinators, determiners, prepositions and auxiliary verbs are used by parents almost exclusively in multiword utterances. By contrast, words in the four CW categories verbs, nouns, adjectives and adverbs appear both as single-word utterances and as roots of two-word sentences. Analysis of children's talk had similar results, the proportions correlating very highly with parents'. Acquisition of FWs predicts syntactic development because they must be learned as combining words, whereas CWs can be learned as stand-alone lexemes, without mastering their syntax.

1. The research question

1.1 FWs predict syntactic development better than CWs

Grammatical words such as determiners, auxiliary verbs and prepositions had long been considered marginal for the early stages of syntactic development. Many authorities such as Radford (1990) believed that such 'function words' (FWs) are acquired late by typically-developing children, and that, at the early stages of acquisition, syntactic development relies on 'content words' (CWs) or 'lexical words' (nouns, verbs, adjectives and adverbs), that carry semantic relations which can be expressed as patterned speech (Brown, 1973).

In the last few years the trend has turned, as recently developmental studies have been offering some new evidence for the importance of the early mastery of FWs for syntactic development. In several studies it was found that in children acquiring various languages, the early mastery of FWs such as subordinators, auxiliary verbs, prepositions and determiners strongly predicts children's concurrent and especially subsequent syntactic development. Kedar et al. (2006) found that 18- and 24-month-old infants acquiring English oriented faster and more accurately to a visual target following sentences in which the referential expression included determiners. They concluded that by 18 months of age, infants use their knowledge of determiners when they process sentences and establish reference. Le Normand et al. (2013) examined the speech of French-speaking children aged 2-4 years and correlated the diversity of word types in various form-classes with the children's mean length of utterance in words (MLU). They found that the diversity of word types in FWs was the best predictor of MLU, much surpassing CW categories. Szagun and Schramm (2016) studied young deaf children acquiring German who received cochlear implants at an early age. They found that the early type and token frequencies of determiners predict MLU two years later more strongly than the early frequency of lexical words. In addition, Ninio (2019) established that FWs are indeed participants in the sentence's syntactic structuring. Taking the determiner-noun relation as an exemplar of the FW-CW relationship, Ninio demonstrated that the syntactic relation of FWs to the CWs they are

associated with is probably identical to the syntactic relation of Head-Dependent complementation (or Merge) that constitutes the major building process of syntax. The determiner-noun combination appears to be a type of complementation, the same syntactic operation that underlies, for example, the combination of a verb with its direct object. Possibly, some generalized knowledge of how words relate to one another is learned when producing early multiword utterances headed by FWs, and, through transfer and facilitation, it drives the construction of grammar. It seems that children learn syntactic principles through specific constructions; in the case of determiner-noun combinations, they apparently learn the principle of head-complement relation, which then can be transferred to other syntactic constructions employing the same basic combinatory operation.

Although the correlational results are a good foundation for theorizing, it is actually still unclear what is the explanation for the higher predictive power of learning FWs for syntactic development. At the early stages of syntax, children learn both a vocabulary of FWs and of CWs; and children do learn to produce verb-object combinations early, not only determiner-noun ones. Nevertheless, it is the acquisition of determiners as vocabulary items, and not of verbs, that best predicts a child's syntactic development. In the present study, our goal is to provide an explanation for FWs' advantage over CWs in providing a learning environment for principles of syntax while being learned as vocabulary items.

Our starting point is the most significant difference between CWs and FWs which is that CWs have rich semantic content whereas FWs are grammatical words that have little or none. This implies that CWs require a learning format where words are matched to their meaning in some non-linguistic context (Macnamara, 1972), but may not necessarily need a syntactic context to acquire their meaning (Ninio, 2016). By contrast, FWs cannot be learned from the non-linguistic context; they are words whose whole purpose is to participate in some combinatory process with another word. It follows that learning them requires multiword input that will make possible to observe their syntactic behaviour.

Thus, our hypothesis is that the acquisition of FWs is better connected to syntactic development than the learning of CWs because their learning necessarily involves multiword input, hence it necessitates the mastery of syntactic principles. CWs, as words possessing semantic content, can be learned from single-word utterances, as their meaning is learned from the nonlinguistic context. We therefore predict that CWs are more likely to be learned from single-word input than FWs; the latter are more likely to be learned from multiword input sentences.

To make the comparison between single-word and multiword as equitable as possible, we took as multiword input sentences the shortest possible multi-word combinations, which are two-words long sentences possessing syntactic connectivity.

The hypotheses were tested by two different comparisons: first, we compared the parental input of content words (CWs) and of function words (FWs) as either single-word utterances or as heads of two-word syntactic combinations. Second, we looked at words occurring in children's single-word sentences and compared them with the heads of their two-word syntactic combinations. The hypothesis was that FWs are presented, and learned, as strongly syntactic elements, while CWs, less so. We expect a significant difference between FWs and CWs in proportion of tokens in single-word and multiword uses, both in the parental input and in children's own spontaneous productions.

2. Method

2.1 Participants

For English-language parental and child samples we systematically sampled the English transcripts in the CHILDES (Child Language Data Exchange System) archive (MacWhinney, 2000). The CHILDES is a public domain database for corpora on first and second language acquisition. The publicly available, shared archive contains documentation of the speech of more than 500 English-speaking parents addressed to their young children. The CHILDES archive stores the transcribed observations collected in various different research projects. In building our corpora, we followed closely the principles established in linguistics for constructing systematically assembled large corpora (Francis and Kučera, 1979).

We selected projects among the ones available using the criteria that the observations were of normally developing young children with no diagnosed hearing or speech problems, and of their parents, native speakers of English, their speech produced in the context of naturalistic, dyadic parent-child interaction. We restricted the child's age during the observed period to three years and six months. This process resulted in the selection of parents and children from 33 research projects in the CHILDES archive: the British projects Belfast, Howe, Korman, Manchester, and Wells, and the American projects Bates, Bernstein-Ratner, Bliss, Bloom 1970 and 1973, Brent, Brown, Clark, Cornell, Demetras, Feldman, Gleason, Harvard Home-School, Higginson, Kuczaj, MacWhinney, McMillan, Morisset, New England, Peters-Wilson, Post, Rollins, Sachs, Suppes, Tardif, Valian, Van Houten, and Warren-Leubecker (MacWhinney, 2000). From these projects, we selected 471 observational studies involving a target child of the desired young age range, namely, below three years and six months.

2.2 Parents' corpus

We built a corpus of parental utterances containing single-word and two-word sentences. Each parent was selected individually, so that from the same research project involving the same target child, we included in the study either the mother, or the father, or both parents as separate speakers, as long as either or both passed the criteria for inclusion. In 35 of the 471 studies there were two active parents interacting with the target child, resulting in a parental sample of 506 different parents.

In order to avoid severely unequal contributions to the pooled corpus, the number of utterances included from each parent was restricted to a maximum of 3,000, counting from the beginning of observations. We have excluded the speech of parents addressed to other adults present in the observational session or on the telephone, as this speech may be ignored by young children because of unfamiliar subjects. Contextual comments were checked in order to ascertain that we included only spontaneous utterances from target parent to target child. The resultant parental corpus contains almost 1.5 million (1,470,811) running words of transcribed speech based on naturalistic observations of interaction between parents and their young children, representing several hundred hours of transcribed speech. Most of the children addressed were under three years of age, and 93% of the parents in the sample talked to a child between one year and two and a half years of age in all or the majority of the observations we included in the corpus. The mean age of the children addressed was 2.25 years.

The corpora of English-language parental child-directed speech represent the linguistic input that young children receive when acquiring syntax. Although each separate study is by necessity limited in its coverage of the phenomenon, the different studies pooled together can provide the requisite solid database for generalization. The use of pooled corpora of unrelated parents as a representation of the linguistic input is a relatively conventional move in child language research (e.g., Goodman et al., 2008). Multiple speakers of child-directed speech may provide a good estimate of the total linguistic input to which children are exposed, which includes, besides the speech of the individual mother or father, also the speech of grandparents, aunts and uncles, older siblings and other family members, neighbours, care professionals, and so forth, represented in our corpus by the speech of mothers and fathers unrelated to the individual child. The pooled database represents the language behaviour exhibited by the community as a whole when addressing young children.

The analyses of a study using corpus data do not attempt to demonstrate that particular children learned particular patterns of use from their own parents. When working with a corpus pooling the speech data of a large number of, respectively, parents and children, the aim is, rather, to create a data set of typical child directed speech and use this to make predictions about children's contrastive mastery of different patterns, thus finding out which of a possible set of factors are most predictive of development. Thus, the variability exploited for statistical testing is not individual differences but, rather, contrasts between the effects of different potential sources of input.

As our analytic plan was to find what kind of input, single-word or multiword sentences, was more likely to serve as a model for learning words of the various form-classes by young children, we manually checked the transcribed dialogue and the action and other contextual comments in the CHILDES archive in order to ascertain that we include only spontaneous utterances from target parent to target child. This means we

excluded parental utterances if they imitated the child's previous utterance, if in order to ask a clarification question or to provide feedback. We wanted parental utterances that can serve as models for the child's learning -- if the parent imitates the child, this cannot be considered a model for new learning. This was in particular important for single-word sentences by parents that were likely to be verbatim repetitions of children's single-word sentences. The exclusion of such utterances alongside children's imitation ensured that we did not arrive at a positive correlation between parental and child frequencies for particular types of words because of mutual imitation of participant speakers.

We focused on parents' single-word utterances and their two-word long sentences possessing syntactic structure, excluding utterances consisting of vocatives or an interjections, or where one of the two words of an utterance was a vocative or an interjection. Unfinished or cut off utterances, or containing words not transcribed in the original were also excluded. Besides these exclusions, we used the original transcripts' separation into sentences as our criterion for identifying single-word and two-word utterances. This corpus represents the shortest linguistic input that young children receive when acquiring syntax. Parents produced 25,694 single-word utterances and 23,141 two-words long sentences. The total number of sentences by parents processed in this study was 48,834.

2.3 Children's corpus

Samples of one-word and two-word utterances from 471 children were taken from the same English transcripts in the CHILDES archive from which we took the parents' speech. We restricted the contribution of each individual child to 300 multiword sentences, starting from the first observation in which they produced multiword utterances. Children's utterances were included only if they were spontaneous, namely, not immediate imitations of preceding adult utterances. For each utterance marked in the original transcriptions as one uttered by the child, we checked the context to make certain that the line was indeed child speech (and not, for example, an action description or parental sentence erroneously marked as child speech). The size of the resulting pooled child corpus is 194,359 running words. It contains 101,064 utterances; this makes the mean length of utterance (MLU) in words 1.92. Similar to the group of parents, we are treating young children acquiring English as their first language as a homogeneous group, as far as the important characteristics of their syntax is concerned. In this, we follow the tradition of researchers who examine pooled corpora of child speech for various characteristics thought to reflect on the relevant class of child speakers (Radford, 1990 ; Serratrice *et al.*, 2003).

For this study, we selected the children who have well-mastered two-word speech. Our criterion was that they not only produced some two-word sentences but that they have already started to combine three words in syntactic combinations. Of the total sample of 471 children, 350 children produced at least three sentences of three words in syntactic combination, that is, excluding vocatives, interjections, or syntactically unrelated words, during the period of observation sampled in the study. The mean age of the children was 2 years and 18 days ($SD = 4$ months 8 days); range 14-42 months). They produced 24,429 single-word utterances and 11,642 two-word long syntactically structured sentences. The total number of sentences by children processed in this study was 36,070.

3. Data analysis

3.1 Lemmatizing verbs and nouns

We lemmatized all verbs in the corpus into their respective stem-groups. Lemmatization is the grouping of related verb forms that share the same stem and differ only in inflection or spelling. For example, *eat*, *eats*, *ate*, *eaten*, and *eating* all belong to the stem-group or lemma of *eat*. In case of irregular verbs changing their shape when inflected such as *had* and *has* of the verb *have*, these forms were also included in the lemma of the relevant stem. This process neutralizes differences in morphological shape irrelevant for the syntactic behaviour of verbs, such as differences of tense, aspect, and person. This analysis assumes that young children ignore the differences in morphological form between verbs belonging to the same lemma, so that

they treat an inflected form such as *eats* as equivalent to an uninflected form such as *eat*. Similarly, we collapsed singular and plural nouns into a single noun-stem category.

3.2 Syntactic annotation for grammatical relations.

Sentences were parsed manually for syntactic structure. We based our dependency analyses on the detailed descriptions of Hudson’s English Word Grammar (Hudson, 1990) with its online update (Hudson, 2014). We also consulted descriptive grammars of English, and in particular Quirk, Greenbaum, Leech, and Svartvik (1985).

For each sentence, the root (namely, the highest element syntactically) of the dependency structure was identified and subsequently tagged for form class membership (see below).

Syntactic annotation of the sentence was done by graduate students at the Hebrew University with training in linguistics. It relied on extensive coding instructions and a very large collection of annotated exemplars. We checked for reliability by having three pairs of coders blindly recode 1,900 utterances produced by four different parents and two children. A checking of all reliability codes showed that the agreement of each coder with the others was above 95%, based on codes actually given by the relevant pairs of coders. Throughout coding, all problem cases were discussed and resolved. Ultimately, each coded utterance was double-checked by another coder.

Classifying roots for form-class: The root-words were classified according to categories of form-classes. Table 1 presents the form-classes employed, with CWs separated from FWs.

Symbol	Definition of form-class	Examples of words in category
Content words (CWs)		
VB	Lexical verbs	come, go, want, get, see, eat
NN	Common nouns and proper nouns	baby, ball, bottle, bear, bird , Anna, David, Mommy, Nina
AJ	Adjectives	big, little, red, wet, good, bad
AV	Adverbs	very, here, there, now, again, beautifully, slowly, gently
Function words (FWs)		
AuxV	Auxiliary verbs, including copulas and the dummy verb ‘do’	be, was, can, may, have (auxiliary), seem, do
PR	Pronouns, e.g. demonstratives, indefinites, interrogatives	I, he, they, this, that, somebody, something, one, who, which
DT	Determiners, e.g. articles, numerals, possessive pronouns, possessors	a, an, the, this, that, some, two, my, your, John’s, no
PP	Prepositions	to, from, in, on, like, for, by
PT	Particles	up, off, on, in
SU	Subordinators	that, which, where, when, because

Table 1. Form-Class Categories Used in the Study to Classify Roots of Sentences.

Table 2 presents the distribution of parents' single-word utterances and roots of two-word sentences by part of speech (POS)

POS	Tokens of syntactic roots	
	Single-word utterances	Two-word sentences
Noun	11,034	2,708
Verb	4,235	12,351
Pronoun	3,398	391
Adjective	3,306	1,359
Adverb	3,642	676
Particle	76	0
Determiner	0	4,541
Preposition	0	825
Auxiliary Verb	0	289
Subordinator	0	1
Total	25,694	23,141

Table 2. Distribution of Parents' Single-Word Utterances and Roots of Two-Word Sentences by Part of Speech (POS)

POS	Tokens of syntactic roots	
	Single word utterances	Two-word sentences
Noun	14,559	1,500
Verb	2,482	3,481
Pronoun	2,280	118
Adjective	1,794	385
Adverb	3,109	298
Particle	187	0
Determiner	14	4,001
Preposition	0	1,365
Auxiliary Verb	4	493
Subordinator	0	1
Total	24,429	11,642

Table 3. Distribution of Children's Single-Word Utterances and Roots of Two-Word Sentences by Part of Speech (POS)

4. Results and discussion

First, we compared parents' use of words belonging to each form-class as single-word utterances or as the roots of two-word long syntactically connected utterances. Figure 1 presents the proportion of single-word and of two-word utterances with root-words belonging to each form-class.

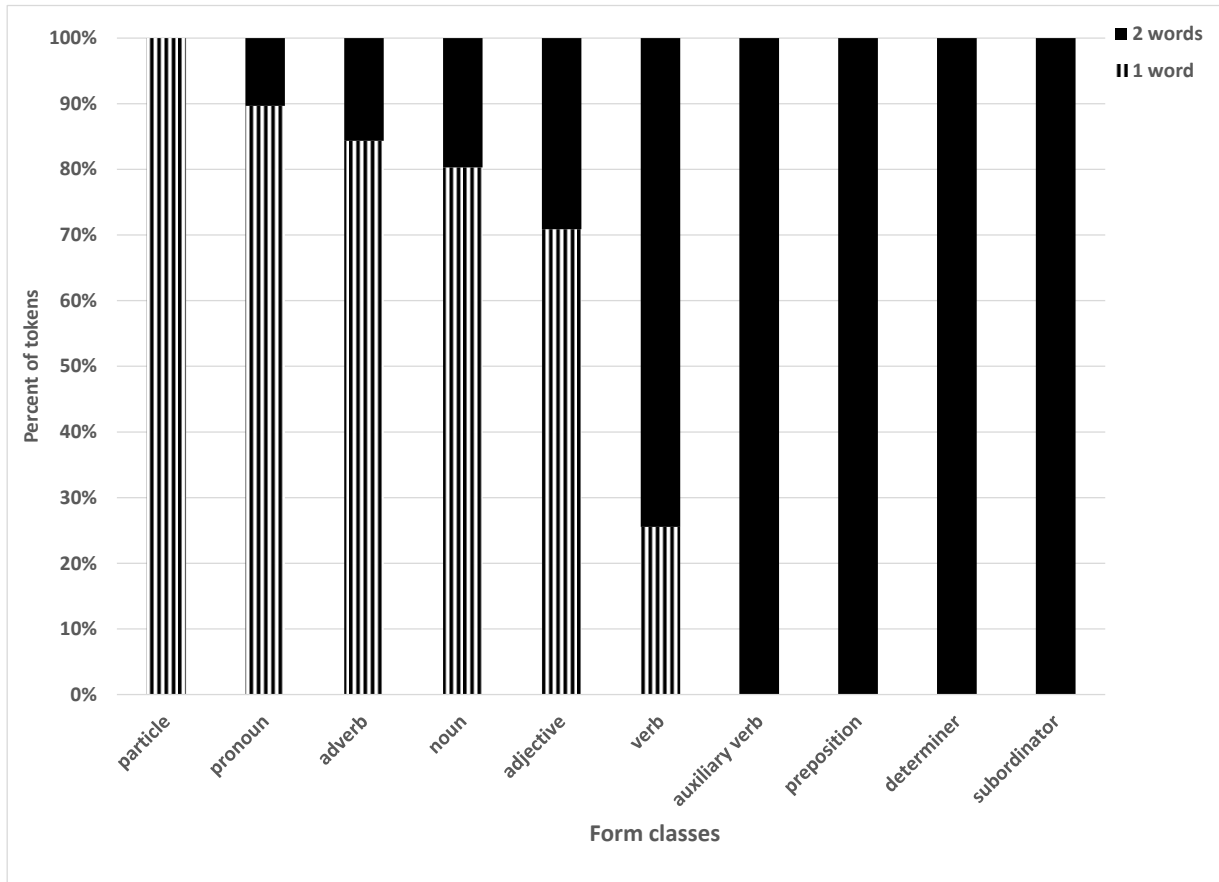


Figure 1. Percent of Tokens of Parents' Use of Words in a Given Form-Class as Single-Word Utterances or as Roots of Two-Word Sentences.

The comparison revealed that in four categories of function words, namely subordinator, determiner, preposition and auxiliary verb, there were almost no single-word utterances produced. The two remaining closed-class categories -- particle and pronoun -- were also extreme in their distribution but to the other direction, as almost all tokens were single-word utterances. The four open classes occurred in a mixture of single-word and multiword sentences, at least 20% in the minority category.

Next, we did the same analysis of children's talk. The results were very similar. Figure 2 presents the results.

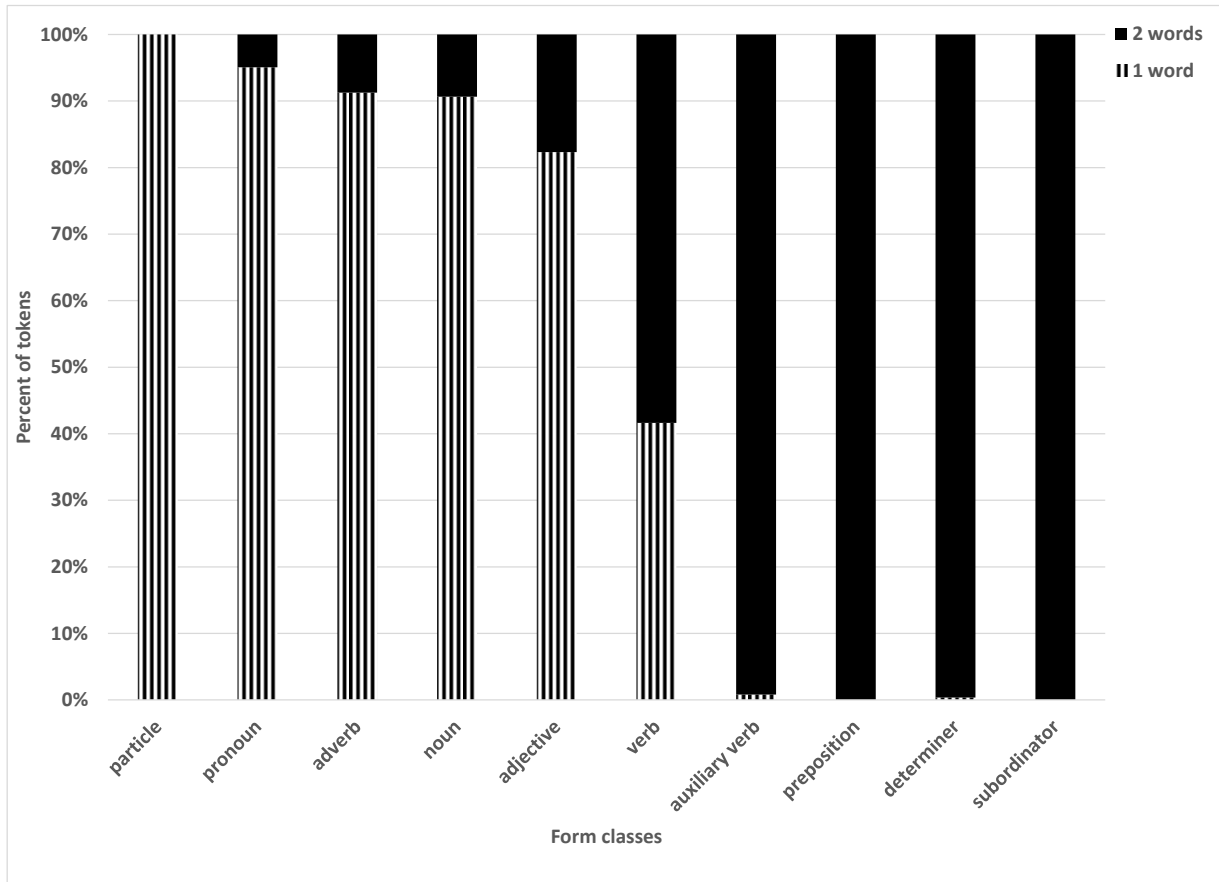


Figure 2. Percent of Tokens of Children’s Use of Words in a Given Form-Class as Single-Word Utterances or as Heads of Two-Word Sentences.

Table 4 compares the tendency to express words belonging to various form-classes as single-word sentences, relative to using the words as heads of two-word long sentences. The table presents the proportion of single-word utterances out of the total tokens of single-word and two-word sentences for head-words belonging to various form-classes.

POS	Percent of tokens of single word uses of vocabulary items out of single-word plus two-word tokens as heads	
	Children's speech	Parents' speech
Particle	100.00%	100.00%
Pronoun	89.68%	95.08%
Adverb	84.34%	91.25%
Noun	80.29%	90.66%
Adjective	70.87%	82.33%
Verb	25.55%	41.62%
Auxiliary Verb	0.00%	0.80%
Determiner	0.00%	0.35%
Preposition	0.00%	0.00%
Subordinator	0.00%	0.00%

Table 4. Proportion of Single-Word Tokens in Children's and Parents' Use of Words as Head-words out of Total One and Two-word Long Sentences, by Part of Speech (POS).

To estimate the similarity of the two distributions, we computed Pearson correlation coefficients between the proportion of single-word tokens of children's and parents' use of words as head-words, out of the total tokens of single-word and two-word sentences in various form-classes. We found that the correlation is very high, with a correlation coefficient of 0.99. That means children closely follow parental models in their use-patterns of words of various type as single words or as roots of two-word long sentences.

Conclusions

Our results offer a simple explanation for the high correlation found in many studies between the learning of such FWs as subordinators, determiners, prepositions and auxiliary verbs, and syntactic development in languages in which such FWs are used in many contexts. These languages are mostly analytic languages such as English, but include also relatively more synthetic languages such as French and German that nevertheless do use FWs quite extensively. We have shown that there is a strong relation between the acquisition of FWs and the development of syntax because such words need multiword input to learn them, hence mastering syntax is a condition for their acquisition. FWs must be learned from multiword sentences because their all of their content is concerned with connection between words; namely, they must be learned as combining words. By contrast, CWs are often learned from single-word utterances as they have semantic content and that can be learned from the nonlinguistic context by semantic matching of the word to the world.

The studies finding a stronger correlation of FWs with syntactic development than CWs (e.g. Le Normand et al., 2013) used as predictors words that children have learned into their active vocabulary. For this measure, the precise form of employment of the words is irrelevant and it is not given in the publications. Our study provides the missing information which is that children not only learn FW from multiword and not single-word input sentences, but that they also use them exclusively in syntactically connected multiword sentences. Thus, we have shown that when a child has learned a FW, he or she has also learned its syntax, while the same is not necessarily true for CWs. Apparently such learning facilitates syntactic development in general in the relevant languages.

These results also explain a paradoxical finding in the field according to which the learning of CWs such as verbs does not correlate significantly with the mastery of syntax. The lack of correlation is unexplainable on a well-accepted theory called the syntactic bootstrapping hypothesis according to which the meaning of verbs cannot be gleaned from the interactive context but need the syntactic context to be learned (Gleitman,

1990). If this theory were correct, we would expect that the learning of lexical verbs and other content-words would be closely correlated to syntactic development, which does not happen according to the relevant studies. Our findings account for this lack of correlation by showing that CWs occur as single-word utterances in large numbers both in parental input and in children's productions and hence the syntactic context is not crucial for learning their use (see also Ninio, 2016). The opposite is the case for those FWs that do not function as single-word utterances, namely determiners, prepositions, subordinators, and auxiliary verbs; such words and not CWs are the ones that require syntactic bootstrapping for their acquisition.

Although the preceding explanation is sufficient for accounting for the findings that FWs are more strongly correlated with syntactic development than CWs, this cannot be the whole story. FWs (and not CWs) are also crucially connected with language loss, not only with language development. In particular, it was repeatedly found that in patients suffering from Broca's aphasia there a strong correlation between the loss of syntax and the loss of the FW vocabulary (e.g., Bock, 1989; Garrett, 1982). The connection with both development of syntax and its loss suggests that FWs must have a central role in the syntactic structuring of sentences.

Our findings provide an opening for such a model. We have shown that FWs are distinctive in lacking semantic meaning and being solely defined by their kind of connection to other words. The promising possibility is that they work as **interface elements**, serving communication with other units of the total system, while the specific semantic content of their complement content-word is "encapsulated". This is the role elements possessing behavioral content but lacking semantic content fill, in, for example, Object Oriented Programming languages (The Java Tutorials, 2017). In our ongoing research, we are currently testing this intriguing possibility.

Acknowledgements

Portions of the research reported here were presented at the Departmental Colloquium, Department of Language and Linguistic Science, University of York, UK, September, 2018. Thanks are due to the parents and children who participated in the observational studies, and to the researchers who contributed the corpora to the CHILDES archive. Various aspects of the research were supported under Grant GR2007 043 by the Center for Complexity Science (CCS) and Grant 200900206 by the Spencer Foundation.

References

- Kathryn Bock. 1989. Closed-class immanence in sentence production. *Cognition*, 31:163-186.
- Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press, Cambridge, M.A.
- Winthrop Nelson Francis and Henry Kučera. 1979. *Brown Corpus Manual of Information to Accompany a Standard Corpus of Present-Day American English, Revised and Amplified*. Brown University, Department of Linguistics, Providence, R.I.
- Merrill F. Garrett. 1982. Production of speech: Observations from normal and pathological language use. In Andrew W. Ellis, editor, *Normality and Pathology in Cognitive Functions*, pp. 19-76. Academic Press, London, UK.
- Leila Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1:3-55.
- Judith C. Goodman, Philip S. Dale, and Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, 35:515-531.
- Richard Hudson. 1990. *English Word Grammar*. Basil Blackwell, Oxford, UK.
- Richard Hudson. 2014. *An Encyclopedia of Word Grammar and English Grammar*. Online book. Available at <http://www.phon.ucl.ac.uk/home/dick/enc2010/frames/frameset.htm/>

- Yarden Kedar, Marianella Casasola, and Barbara Lust. 2006. Getting there faster: 18-and 24-month-old infants' use of function words to determine reference. *Child Development*, 77:325-338.
- Marie-Thérèse Le Normand, Ignacio Moreno-Torres, Christophe Parisse, and Georges Dellatolas. 2013. How do children acquire early grammar and build multiword utterances? A corpus study of French children aged 2 to 4. *Child Development*, 84:647-661.
- John Macnamara. 1972. Cognitive basis of language learning in infants. *Psychological Review*, 79:1-14.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk, 3rd edition*. Lawrence Erlbaum, Mahwah, N.J.
- Anat Ninio. 2016. Learning transitive verbs from single-word verbs in the input by young children acquiring English. *Journal of Child Language*, 43:1103-1130.
- Anat Ninio. 2019. Complement or adjunct? The syntactic principle English-speaking children learn when producing determiner-noun combinations in their early speech. *First Language*, 39:33-44.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English language*. Longman, London, UK and New York, N.Y.
- Andrew Radford. 1990. *Syntactic Theory and the Acquisition of English Syntax*. Basil Blackwell, Oxford, UK.
- Ludovica Serratrice, Kate L. Joseph, and Gina Conti-Ramsden. 2003. The acquisition of past tense in preschool children with specific language impairment and unaffected controls: regular and irregular forms. *Linguistics*, 41-42:321-349.
- Gisela Szagun and Satyam Antonio Schramm. 2016. Sources of variability in language development of children with cochlear implants: age at implantation, parental language, and early features of children's language construction. *Journal of Child Language*, 43:505-536.
- The Java Tutorials. 2017. *What Is an Interface?* Oracle. Available at <https://docs.oracle.com/javase/tutorial/java/concepts/interface.html>

Extraction out of subject phrases in French: experimental evidence

Elodie Winckel

Humboldt-Universität zu Berlin
Inst. für dt Sprache und Linguistik
10117 Berlin, Germany

&

Université de Paris, LLF, CNRS
F-75013 Paris, France
elodie.winckel@hu-berlin.de

Anne Abeillé

Université de Paris, LLF, CNRS
F-75013 Paris, France
anne.abeille@
linguist.univ-paris-diderot.fr

Abstract

Extracting the complement outside a subject has been claimed to be more difficult than out of the object (Chomsky, 1973, a.o). In a series of controlled experiments, we compare extraction out of subject and out of object in French, and we also compare relative clauses and wh-questions. As it turns out, relativizing out of the subject is easier than out of the object, while the opposite pattern holds for wh-questions. We claim that our results cast doubt on a universal syntactic constraint on this type of operation, and call for a discourse based explanation, since the discourse function of relative clauses is different from that of wh-questions.

1 Introduction

Processing (Kluender and Kutas, 1993) and discourse-based (Goldberg, 2013) approaches to islands pose a challenge to syntax-based approaches to the problem. We focus here on the “subject island constraint” (Ross, 1967), which is supposed to ban extraction out of the subject. In Ross (1967)’s original account, only extraction out of sentential subject in English is concerned by the constraint, but followers (Chomsky, 1973; Chomsky, 1986; Chomsky, 2008; Huang, 1982; Boeckx, 2012) extended the constraint to nominal subjects as well. The constraint is also claimed to be universal, despite Rizzi (1982), Godard (1988) and Stepanov (2007). Experimental data is still rare (Kravtchenko et al., 2009; Polinsky et al., 2013; Sprouse et al., 2016; Abeillé et al., 2018).

In a series of experiments, we compare relative clauses and interrogatives in French. Our results show that extraction out of subjects in relative clauses is felicitous (and not restricted to *dont* relative clause, as claimed by Tellier, 1991). They also show that wh-questions differ from relative clauses in this respect, and we provide an explanation which takes into account the different discourse function of the constructions.

All our experiments are controlled online acceptability judgement studies (Gibson and Fedorenko, 2013), in which participants have to rate sentences on a scale from 1 (bad) to 10 (good), with experimental items mixed with other items serving as distractors. The items are randomized and each participant sees each experimental item in only one condition. Participants also have to answer a comprehension question after each item. Results of participant with a low score in comprehension questions are excluded for the statistical analysis. In the following, we report only the significant results ($p > .05$) of linear mixed-effect models (Bates, 2010). We use the R.I.S.C. website (<http://experiences.risc.cnrs.fr/>) and social media to gather volunteers for these experiments.

2 Experiment 1: *dont* Relative Clauses

We investigate relative clauses introduced by *dont*, which have been reported as felicitous for relativizing the complement of a subject (Godard, 1988).

- (1) C’est un philosophe dont le portrait se trouve au Louvre.
‘It’s a philosopher of who the portrait stands in the Louvre.’

We compare relativizing the complement of a subject noun with that of an object noun, using closely related verbs (enchanter ‘delight’/ aimer ‘like’). We compare relative clauses (extraction) with clausal

coordination (no extraction) and ungrammatical control conditions (*que* instead of *dont*), both for subject and object (2x3 design). We have 24 target items and 24 distractors. It is an acceptability judgement study conducted on internet (Ibex) with 48 participants.

Material for Experiment 1

- (2) a. subj, PP-ext: Le concessionnaire a une décapotable dont [la couleur –] enchante le footballeur à cause de sa luminosité.
 ‘The dealer has a sportscar of which [the color –] delights the football player because of its luminance.’
- b. obj, PP-ext: Le concessionnaire a une décapotable dont le footballeur adore [la couleur –] à cause de sa luminosité.
 ‘The dealer has a sportscar of which the football player loves [the color –] because of its luminance.’
- c. subj, no-ext: Le concessionnaire a une décapotable et sa couleur enchante le footballeur à cause de sa luminosité.
 ‘The dealer has a sportscar and its color delights the football player because of its luminance.’
- d. obj, no-ext: Le concessionnaire a une décapotable et le footballeur adore sa couleur à cause de sa luminosité.
 ‘The dealer has a sportscar and the football player loves its color because of its luminance.’
- e. subj, ungram: Le concessionnaire a une décapotable que la couleur enchante le footballeur à cause de sa luminosité.
 ‘The dealer has a sportscar which [the color -] delights the football player because of its luminance.’
- f. obj, ungram: Le concessionnaire a une décapotable que le footballeur adore la couleur à cause de sa luminosité.
 ‘The dealer has a sportscar which the football player loves the color because of its luminance.’

On average, relativization from subject position (2a) is rated higher than relativization from object position (2b). There is however no interaction effect, because clausal coordination (no-extraction) is rated higher in the subject condition (2c) than in the object condition (2d). Ungrammatical controls (2e,f) are rated much lower than the other ones (see Figure 1 with z-scores).

These results confirm the corpus findings of Abeillé et al. (2016) in written and spoken French: *dont* relative clauses are more frequent for relativizing the complement of the subject than of the object. This confirms the theoretical proposals of Godard (1988), Sportiche (1981) : there is no difficulty to extract the complement of a subject in *dont* relative clauses in French.

3 Experiment 2: *de qui* Relative Clauses

Tellier (1991) claimed that *dont* relative clauses are an exception in French, and that extraction out of the subject with a relative pronoun like *qui* (‘who’) is ruled out.

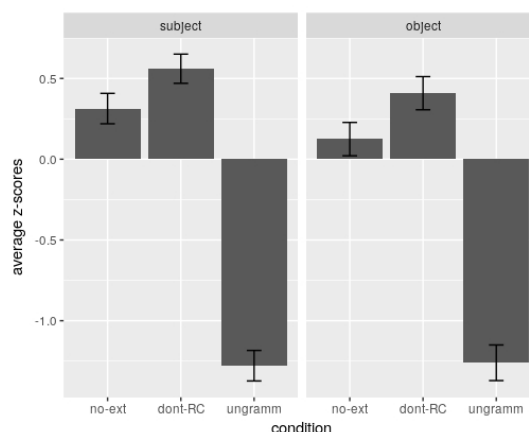


Figure 1: Results of E1

- (3) a. C'est un linguiste dont/*de qui [les parents -] ont déménagé à Chartres. (Tellier, 1991)
 'It is a linguist of whom the parents moved to Chartres'
 b. C'est un linguiste dont/ de qui vous avez rencontré [les parents -]. (Tellier, 1991)
 'It is a linguist of whom you have met the parents'

In a generative perspective, Tellier (1991) proposed that *dont* as a complementizer does not involve syntactic movement, contrary to true relative pronouns like *duquel* or *de qui*. Heck (2009), in the same theoretical school, proposed that *dont* is a specifier which may belong to the subject Noun Phrase, but it is not the case for *de qui*.

We therefore ran a second experiment to investigate relative clauses introduced by *de qui*. We compare extracted variants with clausal coordination (no extraction) and ungrammatical control (missing preposition) conditions, both for subject and object (2x3 design). We have 24 target items and 24 distractors. It is an acceptability judgement study conducted on internet (Ibex) with 28 participants. We choose animate human nouns for both subjects and objects, to avoid animacy mismatch, and reversible transitive verbs (aimer 'like', connaître 'know').

Material for Experiment 2

- (4) a. subj, PP-ext: J'ai un voisin de qui [la compagne -] connaît ma cousine.
 'I have a neighbour of whom [the partner -] knows my cousin'
 b. obj, PP-ext: J'ai un voisin de qui ma cousine connaît [la compagne -].
 'I have a neighbour of whom my cousin knows [the partner -].
 c. subj, no-ext: J'ai un voisin, et la compagne de ce voisin connaît ma cousine.
 'I have a neighbour, and the partner of this neighbour knows my cousin.'
 d. obj, no-ext: J'ai un voisin, et ma cousine connaît la compagne de ce voisin.
 'I have a neighbour, and my cousin knows the partner of this neighbour.'
 e. subj, ungram: J'ai un voisin qui [la compagne -] connaît ma cousine.
 'I have a neighbour who the partner knows my cousin'
 f. obj, ungram: J'ai un voisin qui ma cousine connaît [la compagne -].
 'I have a neighbour who my cousin knows the partner'

We obtained no significant difference between extraction out of subject (4a) and out of object (4b), and both were rated higher than ungrammatical controls (4c,d). We conclude that Tellier (1991)'s contrast in grammaticality should be revised as a preference difference: extraction out of subject is possible in both *dont* and *de qui* relative clauses, and actually preferred with *dont* relative clause.

4 Experiment 3: interrogatives

In a similar kind of experiments, Sprouse et al. (2016) show that extraction out of NP subject in Italian is felicitous in relative clauses and degraded in wh-questions but they do not explain this difference. Abeillé et al. (2018) find a similar contrast in English between relative clauses (with pied-piping) and wh-questions. This is why we test wh-questions as well.

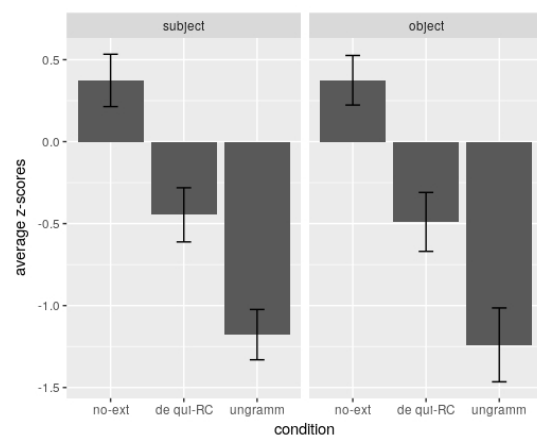


Figure 2: Results of E2

In Experiment 3, we compare questioning the complement of a subject and of the object. We avoid subject-verb inversion and use [*de quel(le) + N*] (litt. ‘of which N’) as extracted element. We compare wh-questions (extraction condition) with polar questions (no extraction) and ungrammatical control (missing preposition) conditions, both for subject and object (2x3 design). We have 24 target items and 32 distractors. It is an acceptability judgement study conducted on internet (Ibex) with 47 participants.

Material for Experiment 3

- (5) a. subj, PP-ext: De quelle décapotable est-ce que [la couleur _] enchante le footballeur à cause de sa luminosité ?
 ‘Of which sportscar does [the color _] delight the football player because of its luminance?’
- b. obj, PP-ext: De quelle décapotable est-ce que le footballeur adore [la couleur _] à cause de sa luminosité ?
 ‘Of which sportscar does the football player love [the color _] because of its luminance?’
- c. subj, no-ext: Est-ce que la couleur de la décapotable enchante le footballeur à cause de sa luminosité ?
 ‘Does the color of the sportscar delight the football player because of its luminance?’
- d. obj, no-ext: Est-ce que le footballeur adore la couleur de la décapotable à cause de sa luminosité ?
 ‘Does the football player love the color of the sportscar because of its luminance?’
- e. subj, ungram: Quelle décapotable est-ce que la couleur enchante le footballeur à cause de sa luminosité ?
 ‘Which sportscar does the color delight the football player because of its luminance?’
- f. obj, ungram: Quelle décapotable est-ce que le footballeur adore la couleur à cause de sa luminosité ?
 ‘Which sportscar does the football player love the color because of its surprising luminance?’

We find that, on average, extraction out of subjects (5a) was rated lower than out of objects (5b), and also lower than polar (non-extraction) questions (5c,d), but better than ungrammatical controls (5e,f). On the other hand, polar questions are rated higher in subject than in object condition: there is an interaction between extraction and subject/object condition (see Figure 3 with z-score).

Discussion Contrary to relative clauses, which show no subject penalty both with *dont* and *de qui*, we find a subject penalty with wh-questions. This is in line with corpus data: Abeillé and Winckel (2018) find no examples of questioning the complement of the subject in Frantext (2000-2010).

These results are difficult to explain under syntactic theories which analyse relativization and wh-question via the same movement operation. Why would movement be easy in relative clause and difficult in wh-question if the syntactic structures are similar? They are easier to explain under discourse based theories which view locality constraints as coming from a discourse status clash (Erteschick-Shir, 2007; Goldberg, 2013): under discourse based theories, extraction makes an element more salient and is more felicitous for elements belonging to the “focal domain” of the sentence, i.e. not backgrounded elements (like sentential subjects) nor topical elements (like nominal subjects). Following Abeillé et al. (2018),

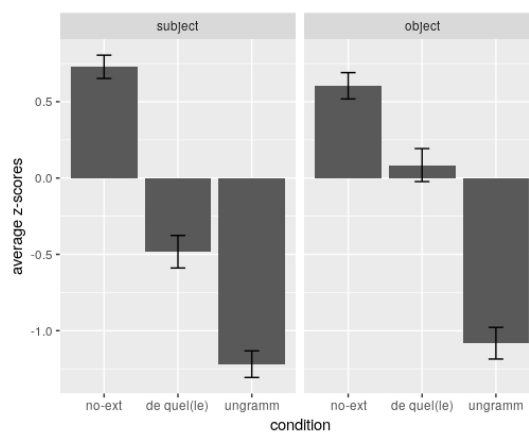


Figure 3: Results of E3

we propose that syntactic theories should be revised in order to take into account the discourse function of the construction. Indeed relativization is a topicalization (Kuno, 1976) whereas *wh*-extraction in direct interrogatives is a focalization (Jackendoff 1972). We therefore expect different results for both constructions: focusing (in a *wh*-question) an element which belongs to the focal domain (a complement of the object) is easier than focusing an element which belongs to the sentence topic (a complement of the subject). Of course, such discourse status are flexible, and some subjects may be more focal, or less topical, and thus make it easier for their complement to be questioned, as in the following attested¹ example:

- (6) De qui l'anniversaire tombe-t-il le 27 ?
 'Of whom does the birthday take place on the 27th?'

On the other hand, relative clauses are subordinate clauses which add a property to an entity (the antecedent) which can have various discourse status in the main clause (in our experiments, the antecedent noun is a complement). They do not put the relativized element into focus. There is thus no discourse clash and no subject penalty.

5 Experiment 4: *où*-extraction out of sentential subjects

We have shown that no subject penalty arises with relative clauses, and proposed that the difficulty found with *wh*-questions come from the discourse status of the construction (questioning means putting the questioned element into focus). This proposal is supported by corpus studies (Abeillé and Winckel, 2018) and experiments on nominal subjects of transitive verbs. What about other subjects? While sentential subjects (7a) are supposed to be (Miller, 2001; Goldberg, 2013), and rare (only 24 in the French Treebank), compared to impersonal variants (7b), infinitival subjects have seldom been discussed (Chaves and Dery, 2018).

- (7) a. ? Que Paul soit parti m'étonne.
 'That Paul left surprises me.'
 b. Ca m'étonne que Paul soit parti.
 'It surprises me that Paul left.'

Infinitival subjects are also rarer than nominal ones (respectively 99 and 26 000 in the French Treebank; Abeillé et al., 2019) and an impersonal variant is possible to:

- (8) a. ? Partir demain m'ennuie.
 'To leave tomorrow bothers me.'
 b. Ca m'ennuie de partir demain.
 'It bothers me to leave tomorrow.'

In Experiment 4, we turn to extraction out of infinitival subjects. We test infinitival subjects with a locative complement and extract this complement using *où* ('where'). We compare extraction out of subjects and out of object, using the impersonal variant. We compare extracted variants with no extraction conditions (coordination), both for subject and object (2x2 design). We have 12 target items and 36 distractors including some ungrammatical ones. It is an acceptability judgement study conducted on internet (Ibex) with 33 participants.

Material for Experiment 4

- (9) a. subj, PP-ext: Il y a une guerre civile dans ce pays, où [aller –] est dangereux en ce moment.
 'There is a civil war in this country, where [to go –] is dangerous right now.'

¹<https://www.quiz.biz/quiz-1027073.html> (Accessed: February 13, 2019)

- b. pred, PP-ext: Il y a une guerre civile dans ce pays, où il est dangereux d’[aller _] en ce moment.
 ‘There is a civil war in this country, where it is dangerous [to go _] right now.’
- c. subj, no-extr: Il y a une guerre civile dans ce pays, et y aller est dangereux en ce moment.
 ‘There is a civil war in this country, and to go there is dangerous right now.’
- d. pred, no-extr: Il y a une guerre civile dans ce pays, et il est dangereux d’y aller en ce moment.
 ‘There is a civil war in this country, and it is dangerous to go there right now.’

We find a tendency for extractions out of subjects (9a) to decrease in acceptability compared to the other three conditions (9b,c,d) but no significant difference. Their rating is very high though (average rating of 7.6). They differ significantly from the ungrammatical items (Figure 4 with z-scores). Surprisingly, there is no significant difference between infinitival subject and impersonal constructions, although the latter is more frequent in corpora.

6 Conclusion

Our results show that relativizing out of subject is either rated better or rated in the same way as relativizing out of objects. This is not true for all extraction, because extracting out of a subject in a wh-question shows a clear decline of acceptability compared to extracting out of an object.

We propose that Ambridge and Goldberg (2008)’s claim that subjects are difficult to extract from because they are background information, only applies to questions. We propose that: (a) the discourse status of the extracted element is a focus in wh-question, not in relative clause; (b) the discourse status of the extracted element must match the discourse status of the noun it is extracted from. We thus predict that wh-questions disfavor extraction of an element out of a subject, which is a default clause topic (Lambrecht, 1994), and that relative clauses don’t. Ross (1967)’s constraint on extraction out of subject phrases only applies to sentential subjects. Our results show no significant interaction effect for extracting out of infinitival subjects, which received good ratings. We conclude that a syntactic constraint banning extraction out of subjects does not hold in French. French not only allows extraction from nominal subjects, but even prefers such extraction, given the right construction and the right relativizer.

References

- Anne Abeillé and Elodie Winckel. 2018. ‘Dont’ and ‘de qui’ relatives in written French, November. Talk at the Conference Grammar and Corpora in Université Paris-Diderot.
- Anne Abeillé, Barbara Hemforth, and Elodie Winckel. 2016. Les relatives en dont du français: études empiriques. In F. Neveu, G. Bergounioux, M.-H. Côté, J.-M. Fournier, L. Hriba, and S. Prévost, editors, *5e Congrès Mondial de Linguistique Française*, volume 27 of *SHS Web of Conferences*.
- Anne Abeillé, Barbara Hemforth, Elodie Winckel, and Edward Gibson. 2018. A construction-conflict explanation of the subject-island constraint, March. Poster at the 31st Annual CUNY Sentence Processing Conference in UC Davis.
- Anne Abeillé, Lionel Clément, and Loïc Liégeois. 2019. Un corpus annoté pour le français : le French Treebank. *TAL Traitement Automatique des Langues*, 60.

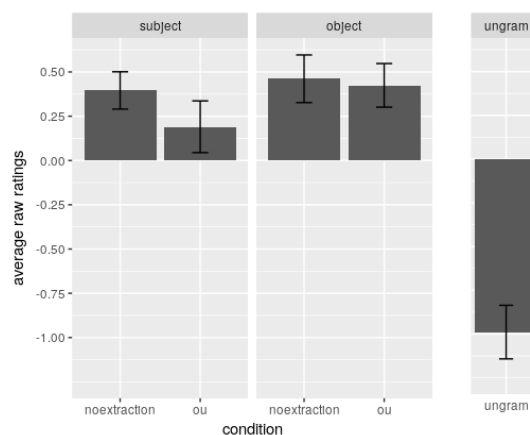


Figure 4: Results of E4

- Ben Ambridge and Adele E. Goldberg. 2008. The island status of clausal complements: evidence in favor of an information structure explanation. *Cognitive Linguistics*, 19(3):349–381.
- Douglas M. Bates. 2010. *lme4: Mixed-effects modeling with R*. Springer.
- Cedric Boeckx. 2012. *Syntactic islands*. Key topics in syntax. Cambridge University Press, Cambridge and New York.
- Rui P. Chaves and Jeruen E. Dery. 2018. Frequency effects in subject islands. *Journal of linguistics*, pages 1–47.
- Noam Chomsky. 1973. Conditions on transformations. In Steven Anderson and Paul Kiparsky, editors, *A festschrift for Morris Halle*, pages 232–285, New York. Winston.
- Noam Chomsky. 1986. *Barriers*, volume 13 of *Linguistic inquiry monographs*. MIT Press, Cambridge, MA.
- Noam Chomsky. 2008. On phrases. In Robert Freidin, David Michaels, Carlos P. Otero, and Maria Luisa Zubizarreta, editors, *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*, pages 133–165, Cambridge, MA. MIT Press.
- Nomi Erteschick-Shir. 2007. *Information Structure: The Syntax-Discourse Interface*. Oxford, oxford university press edition.
- Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28:88–124.
- Danièle Godard. 1988. *La syntaxe des relatives en français*. Paris : Ed. du Centre national de la Recherche Scientifique, Paris.
- Adele Goldberg. 2013. Backgrounded constituents cannot be extracted. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Island Effects*. Cambridge University Press.
- Fabian Heck. 2009. On certain properties of pied-piping. *Linguistic Inquiry*, 40(1):75–111.
- C.-T. James Huang. 1982. *Logical relations in Chinese and the theory of grammar*. Ph.D. thesis, MIT.
- Robert Kluender and Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes*, 8:573–633.
- Ekaterina Kravtchenko, Ming Xiang, and Maria Polinsky. 2009. Are all subject islands created equal? Poster at the 22nd Annual CUNY Sentence Processing Conference in UC Davis.
- Susumu Kuno. 1976. Subject, theme, and the speaker's empathy - a reexamination of relativization phenomena. In Charles N. Li, editor, *Subject and Topic*, pages 417–444, New York, NY. Academic Press.
- Knud Lambrecht. 1994. *Information structure and sentence form : topic, focus, and the mental representations of discourse referents*. Number 71 in Cambridge studies in linguistics. University Press, Cambridge.
- Philip Miller. 2001. Discourse constraints on (non)-extraposition from subject in English. *Linguistics*, 39(4):683–701.
- Maria Polinsky, Carlos G. Gallo, Peter Graff, Ekaterina Kravtchenko, Adam Milton Morgan, and Anne Sturgeon. 2013. Chapter 13. Subject islands are different. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Island Effects*. Cambridge University Press.
- Luigi Rizzi. 1982. *Issues in Italian syntax*, volume v. 11 of *Studies in generative grammar*. Foris Publications, Dordrecht, Holland and Cinnaminson, N.J., U.S.A.
- John Robert Ross. 1967. *Constraints on variables in syntax*. Ph.d thesis, MIT, Cambridge, MA.
- Dominique Sportiche. 1981. Bounding nodes in French. *The Linguistic Review*, 1(2):219–246.
- Jon Sprouse, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language and Linguistic Theory*, 34(1):307–344.
- Arthur Stepanov. 2007. The end of CED? Minimalism and extraction domains. *Syntax*, 10(1):80–126.
- Christine Tellier. 1991. *Licensing theory and French parasitic gaps*, volume v. 26 of *Studies in natural language and linguistic theory*. Kluwer Academic Publishers, Dordrecht, The Netherlands and Boston.

The relation between dependency distance and frequency

Xinying Chen

University of Ostrava, Czech Republic
Xi'an Jiaotong University, China
xy@yuyanxue.net

Kim Gerdes

LPP (CNRS)
Sorbonne Nouvelle, France
kim@gerdes.fr

Abstract

This present pilot study investigates the relationship between dependency distance and frequency based on the analysis of an English dependency treebank. The preliminary result shows that there is a non-linear relation between dependency distance and frequency. This relation between them can be further formalized as a power law function which can be used to predict the distribution of dependency distance in a treebank.

1 Introduction

As a well-discussed norm (Hudson, 1995; Temperley, 2007; Futrell et al., 2015; Liu et al., 2017), dependency distance shows several attractive features for quantitative studies. First, its definition is rather clear. It is the linear distance between a word and its head.¹ Second, it is very easy to quantify. We can simply compute dependency distance as the difference of the word ID and its head's ID in a CoNLL style treebank (Buchholz & Marsi, 2006). These features together with the emergence of large-scale dependency treebanks made dependency distance one of the popular topics in quantitative syntactic studies.

Among various interesting discussions, the most striking finding is probably the dependency distance minimization phenomena. After empirically examining the dependency distance distributions of different human languages and comparing the results with different random baselines, Liu (2008, 2010) found that there is a universal trend of minimizing the dependency distance in human languages. Futrell et al. (2015) conducted a similar study which widened the language range and added one more random baseline. Their results are coherent with Liu's finding. Both Liu (2008) and Futrell et al. (2015) connect this phenomenon with the short-term memory (or working memory) storage of human beings and the least effort principle (Zipf, 1949). Since long dependencies, which have longer distance, occupy more short-term memory storage, they are more difficult or inefficient to process. Therefore, for lowering the processing difficulty and boosting the efficiency of communications, short dependencies are preferable according to the least effort principle.

Initially, the least effort principle was brought up by Zipf for explaining the observed power-law distributions of word frequencies. Later on, similar power-law frequency distributions have been repeatedly observed in various linguistic units, such as letters, phoneme, word length, and etc. (Altmann & Gerlach, 2016). The power law distribution, therefore, has been considered as a universal linguistic law. After investigating the relationships between different word features (such as length vs frequency, frequency vs polysemy, and etc.), people found out an interesting phenomenon. The relations between two highly correlated word features are usually non-linear and can be formulated as a power law function (Köhler, 2002). Köhler (1993) further proposed a word synergetic framework to model the interactions between different word features. This model has proved quite successful also then adapted to syntax features. The first studies mainly focused on the analysis of phrase structure treebanks (Köhler, 2012), which naturally are limited in language types since phrase structure grammar is less suitable for describing free word order languages (Mel'čuk, 1988). As the dependency treebanks are getting dominant, studies based on dependency grammar start to take lead. We can find recent studies discussing the relations between sentence lengths, tree heights, tree widths, and mean dependency distances (Jing & Liu, 2017; Zhang & Liu, 2018; Jiang & Liu, 2015).

¹Hudson's original measures takes two adjacent words to have distance zero. We prefer the alternative definition where $x = y \Leftrightarrow d(x, y) = 0$, i.e. a word has distance zero with itself, making the measure a metric in the mathematical sense.

Knowing that short dependencies are preferable by languages due to the least effort principle and that syntax features behavior similar to word features, we can easily draw our hypotheses:

- *The relation between dependency distance and frequency can be formulated as a non-linear function (probably also a power law function).*

Contrary to above-mentioned studies, our study here is not focusing on mean dependency distances but the distribution of the distance of every single dependency. In the dependency minimization studies or synergetic syntax studies, the observed feature is mean dependency distance per sentence. In a way, these observed dependency distances are treated as a dependent feature of dependency trees. This is a very reasonable choice since the dependency distance is defined as the linear distance between two words in the same sentence. In particular, when the studies discuss other tree-related features such as tree heights and widths, mean dependency distance is a more easily comparable feature than a group of individual dependency distances. However, we believe the value of individual dependency distances is neglected. Individual dependency distances (Liu, 2010; Chen & Gerdes, 2017, 2018) provide more details of the fluctuation than the average which would level-up differences of dependencies in a sentence and it should be given the same attention as the mean dependency distance. Therefore, our study here is trying to pick up the missing detail of previous studies by investigating the relations between individual dependency distances and their frequencies.

The paper is structured as follows. Section 2 describes the data set, the Parallel Universal Dependencies (PUD) English treebank of Universal Dependencies treebanks, and introduces our computing method for dependency distance and frequency. Section 3 presents the empirical results and discussions. Finally, Section 4 presents our conclusions.

2 Material and Methods

Universal Dependencies (UD) is a project of developing a cross-linguistically consistent treebank annotation scheme for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of Stanford dependencies (De Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages while allowing language-specific extensions when necessary. UD is also an open resource which allows for easy replication and validation of the experiments (all Treebank data on its page is fully open and accessible to everyone). For the present paper, we used the PUD English Treebank from the UD 2.3 dataset for our study since English is a rather reasonable choice for a pilot study. Furthermore, PUD is a parallel treebank with a wide range of languages, namely Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish. This makes PUD a good choice for future studies which would further test whether our finding here can be generalized into different human languages. We use the Surface-syntactic UD version of the treebank (Gerdes et al., 2018), which is more suitable for studies in distributional dependency syntax as it corrects the artificially long dependency distances of UD into a more standard syntactic analysis based on distributional criteria (Osborne & Gerdes, 2019).

We first compute the dependency distance for every single dependency in the treebank except the root relation. The dependency distance is computed as the absolute difference between the word ID and its head's word ID. For instance, in Figure 1, there are 4 dependencies. We would take three of them into account except the root dependency. The dependency distances of these three dependencies are: $abs(1-2) = 1$ (for *subj*), $abs(4-2) = 2$ (for *comp*), and $abs(3-4) = 1$ (for *det*).

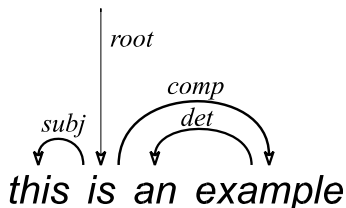


Figure 1: Example dependency tree in SUD analysis.

After computing all the dependency distances of the treebank, we then count the frequencies of each dependency distance, i.e. we count how many dependencies with dependency distance 1 occurred in the treebank, how many dependencies with distance 2 occurred, and so on. We then try to formulate the relation into a non-linear function. We will test different non-linear functions to see which one can predict the empirical data best. In other words, we try to see whether our data can be fitted by the power law function. This result can then either confirm or reject our hypothesis.

We also introduce two random baselines to see whether we can observe similar phenomenon in random dependency trees. Based on the PUD English treebank, we generate two random tree-banks. For the random treebank RT, we just randomly reorder the words of each sentence. For the random treebank PRT, we randomly reorder the words in a way that keeps the sentence’s dependencies projective (non-crossing).

3 Results and Discussion

The PUD English treebank is part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on multilingual parsing (Zeman et al., 2017). There are 1000 sentences in each language. The sentences are taken from the news domain and from Wikipedia. The PUD English treebank contains 21,176 tokens. See Appendix for the frequencies of dependency distances in the treebank.

The scatter plot Figure 2 shows that the relationship between dependency distance and frequency is indeed non-linear.

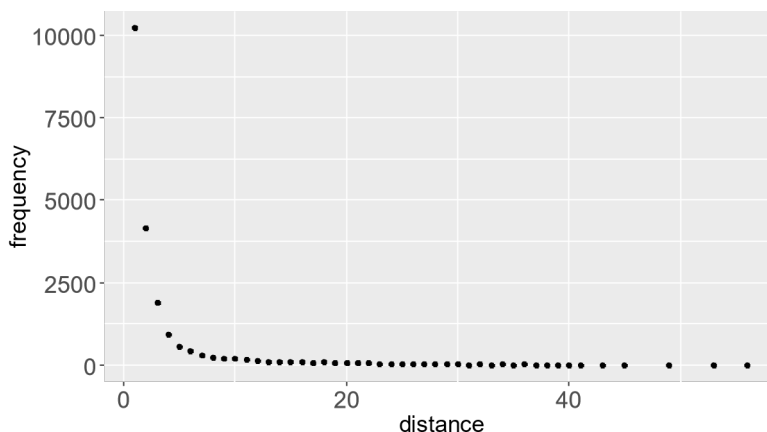


Figure 2: Scatter plot of dependency distance and frequency of PUD English treebank.

Since the observed data points scatter as a L-ish shape, we tried to fit the data to four non-linear functions, namely quadratic, exponent, logarithm, and power law functions. Although there are different ways of measuring the goodness-of-fit (Maćutek & Wimmer, 2013), we choose to use the most common Pearson chi-square goodness-of-fit test to evaluate the fitting results in this study. The formula of the test is defined as

$$R^2 = \sum_{i=1}^n \frac{(f_i - NP_i)^2}{NP_i} \quad (1)$$

with f_i being the observed frequency of the value i , P_i being the expected probability of the value i , n being the number of different data values and N being the sample size. The obtained results of R-squared is presented in Table 2.²

Non-linear Model	Function	R ²
Quadratic	$y=2963.44-206x+3.1x^2$	0.34
Exponent	$\log(y)=7.11-0.16x$	0.92
Logarithm	$y=4100.8-1262\log(x)$	0.49
Power Law	$\log(y)=10.71-2.56\log(x)$	0.91

Table 1: R² of four non-linear models.

The results show that the observed data can indeed be formulated as a power law function. However, it seems that the data also fits an exponent regression very well. This is a very common issue in quantitative linguistic studies (Baixeries et al., 2013). In many situations, both exponent and power-law models can describe the data fluctuation reasonably well. One way to decide which model is better is by adding more observations from other languages. However, this is out of the scope of this pilot study. Another solution can be introducing baselines for comparison, which is our choice in this paper. By comparing the results in Table 1 with the results of two different random treebanks, we try to deliver the answer for this question, which model is better to represent the relation between dependency distance and frequency, exponent or power law?

For the two random English PUD treebank variations, RT and PRT, we replicate the same computation for the frequency and dependency distance, see Appendix. The scatter plots Figure 3 and 4 show that the relations between dependency distance and frequency in RT and PRT are both non-linear.

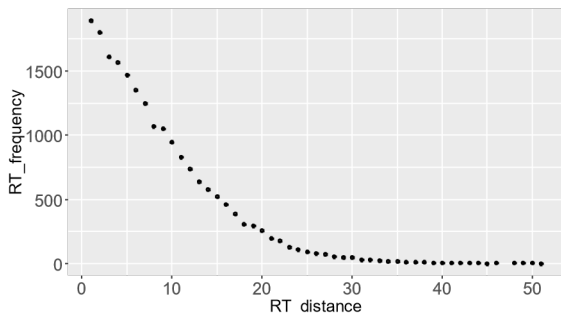


Figure 3: Scatter plot of RT.

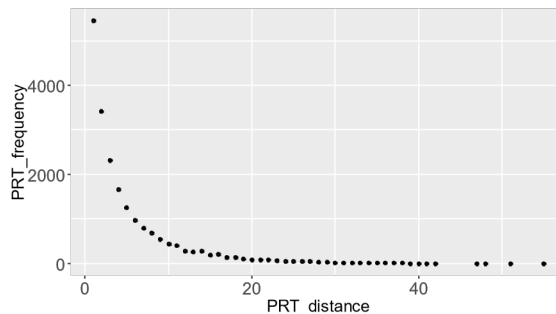


Figure 4: Scatter plot of PRT.

Similarly, we fit the data points to four non-linear models, see Tables 2 and 3 for results. We can see from Table 2 that RT fits to all non-linear models very well except to the power law function, which is very different from the PUD English treebank who fits to power law very well but does not fit to quadratic and exponent models. When we add the projectivity restriction, the fitting results of PRT seems more ‘human language’ like. Similar to the results of PUD in Table 1, PRT fits to exponent and power-law models better. However, the power law fitting result of PUD is clearly more satisfying than the result of PRT.

Non-linear Model	Function	R ²
Quadratic	$y=1883.88-106.28x+1.43x^2$	0.98
Exponent	$\log(y)=8.42-0.17x$	0.98
Logarithm	$y=2220.88-611.66\log(x)$	0.96
Power Law	$\log(y)=11.23-2.37\log(x)$	0.74

Table 2: R² results of RT.

²All parameter values in the models were obtained by R software. The same below.

Non-linear Model	Function	R ²
Quadratic	$y=2551.07-168.63x+2.49x^2$	0.62
Exponent	$\log(y)=7.99-0.17x$	0.97
Logarithm	$y=3258.25-972.05\log(x)$	0.75
Power Law	$\log(y)=11.28-2.55\log(x)$	0.84

Table 3: R² results of PRT.

Beyond considering the projectivity feature of dependency trees that deals with the crossing problem, we would also like to have a closer look at the role of syntax in this question. Our way of addressing this is to exclude less syntactic dependencies from the analysis. The UD/SUD annotation scheme includes predefined dependency structures for some constructions, in particular for MWE and punctuation. The distance of relations such as *fixed*, *compound*, *flat*, and *punct* are not based on distributional criteria of the tokens involved. Therefore, we also tested the results when these dependencies are excluded from our analysis, taking into account only syntactic dependencies (*subj*, *aux*, *cop*, *case*, *mark*, *cc*, *dislocated*, *vocative*, *expl*, *discourse*, *det*, *clf*). See the Appendix for the details. We first tested these three data sets with a linear regression model, and the results are similar to the previous analysis (PUD R²=0.21, RT R²=0.77, PRT R²=0.34). We then repeated the same non-linear regression analysis on these three selected data sets and the results are presented in Table 4.

Syntactic Data Set	Non-linear Model	Function	R ²
PUD English	Quadratic	$y=1216.36-148.07x+3.82x^2$	0.44
	Exponent	$\log(y)=5.84-0.25x$	0.81
	Logarithm	$y=1380.2-523.3\log(x)$	0.56
	Power Law	$\log(y)=8.45-2.53\log(x)$	0.97
RT	Quadratic	$y=434.78-25.62x+0.36x^2$	0.98
	Exponent	$\log(y)=6.78-0.16x$	0.97
	Logarithm	$y=510.91-142.85\log(x)$	0.95
	Power Law	$\log(y)=9.1-2.07\log(x)$	0.74
PRT	Quadratic	$y=656.17-50.02x+0.86x^2$	0.6
	Exponent	$\log(y)=6.27-0.16x$	0.95
	Logarithm	$y=810.18-251.99\log(x)$	0.73
	Power Law	$\log(y)=8.89-2.13\log(x)$	0.89

Table 4: R² results for syntactic dependencies.

Very similar to the results of the previous analysis, PRT is closer to the PUD English results. However, the results with syntactic dependencies demonstrate more clearly that a power law model is the better choice for representing the relation between dependency distance and frequency. First, the original PUD data fits to the power law function best, whereas in the previous analysis we could not easily draw such a conclusion due to the very similar R² values for both power law and exponent models. Secondly, the goodness of the power law model fitting somehow can distinguish the natural PUD data from random baselines.

4 Conclusion

Our results are coherent with our hypothesis that there is indeed a non-linear relation between dependency distance and frequency. Furthermore, this relation can be formulated as a power law function.

However, the results in Table 1 show that the power-law model is not the only candidate for formulating the relation, and we could also apply an exponential model to it. For figuring out which model is better for representing the relation, we introduce two random baselines. By randomly reordering the words in a sentence, while preserving the words' dependencies, we generate random treebanks: PRT with and RT without the projectivity restriction, in which PRT possesses a more 'natural' structure reproducing more closely the rarity of non-projective relations. We replicate the same analysis on these

two random treebanks and compare the results with the PUD results. We found that we can distinguish the PUD from RT and PRT by looking at the results of power-law fitting. Therefore, we would like to cautiously draw our conclusion here that the power law model is probably a better choice for representing the relation between dependency distance and frequency, a hypothesis that is further strengthened by the results on purely syntactic dependency relations.

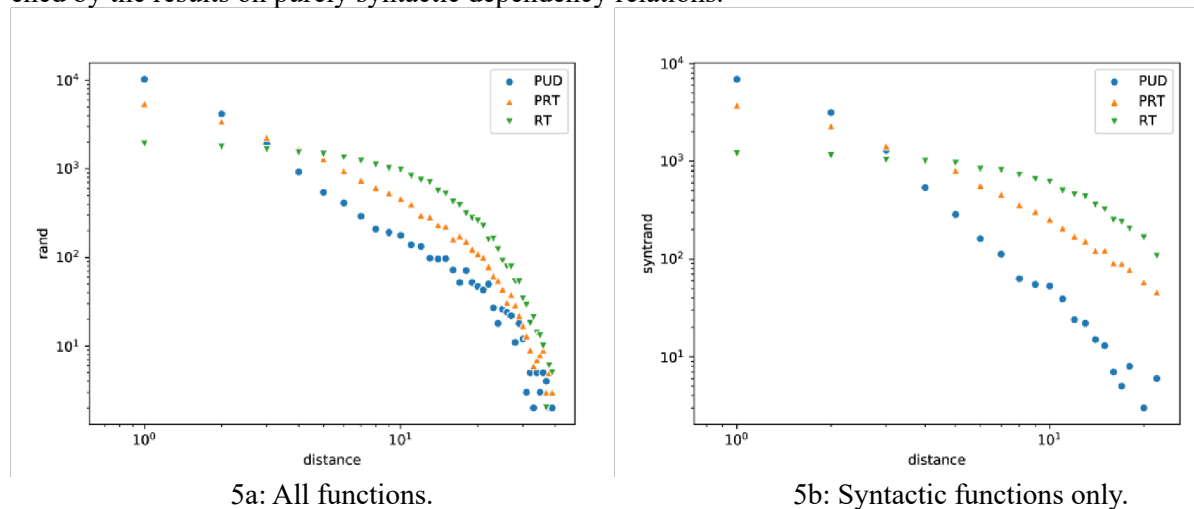


Figure 5: Joint plot of the frequency of dependency distance on a logarithmic scale showing the greater linearity of PUD compared to the random treebanks.

Another interesting phenomenon we can observe from our data is that the projective random data-set has almost as good a fit to a power law function as the syntactically parsed true treebank. Although we need more samples to conduct a statistical significance testing for the difference, it seems that if we compare the natural PUD and the control PRT on the most relevant “syntactic functions only”, for example in the logarithmic presentation of Figure 5b., there is practically no difference between the linearity of PRT and PUD. This shows that projectivity has a major role as the responsible factor for the power-law function of dependency distance. Of course, our conclusion based on this pilot study needs to be tested with more languages in the future. This leads to the open question to actually pinpoint the additional syntactic constraint of PUD, compared to random treebanks, that results in the power law distribution.

We believe the result presented here has several potential applications. We can use the power law model to predict the distribution of dependency distance in a treebank. Since natural language treebanks fit to power law model better than random treebanks, we might even use it as an index for assessing the quality of parse results.

Acknowledgements

This work is supported by the European Union & Ministry of Education of the Czech Republic (No. CZ.02.2.69/0.0/0.0/16_027/0008472) and the National Social Science Fund of China (2018CYY031).

Reference

- Altmann, Eduardo Gabriel, and Martin Gerlach. 2016. Statistical laws in linguistics. In *Creativity and Universality in Language* (pp.7-26). Springer, Cham.
- Baixeries, Jaume, Brita Elvevåg, and Ramon Ferrer-i-Cancho. 2013. The evolution of the exponent of Zipf’s law in language ontogeny. *PLoS one*, 8(3): e53227.
- Buchholz, Sabine and Erwin Marsi. 2006, June. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning* (pp. 149-164). Association for Computational Linguistics.
- Chen, Xinying and Kim Gerdes. 2017. Classifying languages by dependency structure: Typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics* (Depling 2017), Pisa, September. Linköping University Electronic Press.

- Chen, Xinying and Kim Gerdes. 2018. How Do Universal Dependencies Distinguish Language Groups? In *Quantitative Analysis of Dependency Structures*, 72: 277-294.
- De Marneffe, Marie-Catherine and Christopher D. Manning. 2008. The Stanford typed dependency representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.
- Futrell, Richard, 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.
- Gerdes, Kim, et al. 2018, November. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*.
- Hudson, Richard. 1995. Measuring syntactic difficulty. Draft of manuscript, available at <http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>.
- Jiang, Jingyang. 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.
- Jing, Yingqi and Haitao Liu. 2017. Dependency distance motifs in 21 Indoeuropean languages. In *Motifs in Language and Text* (pp.133-150).
- Köhler, Reinhard. 1993. Synergetic linguistics. In *Contributions to Quantitative Linguistics*, pp.41-51. Springer, Dordrecht.
- Köhler, Reinhard. 2002. Power law models in linguistics: Hungarian. *Glottometrics*, 5: 51-61.
- Köhler, Reinhard. 2012. *Quantitative Syntax Analysis*, 65. Walter de Gruyter.
- Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9 (2): 159-191.
- Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120 (6): 1567-1578.
- Liu, Haitao, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21: 171–193.
- Mačutek, Jan and Wimmer, Gejza. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20 (3): 227-240.
- Mel'čuk, Igor Aleksandrovic. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.
- Osborne, Tim and Kim 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. 1-28.
- Petrov, Slav, Dipon Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.
- Temperley David. 2007. Minimization of dependency length in written English. *Cognition*, 105: 300–333.
- Zeman, Daniel. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of LREC*.
- Zeman, Daniel, et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. *CoNLL 2017*.
- Zhang, Hongxin and Haitao Liu. 2018. Interrelations among Dependency Tree Widths, Heights and Sentence Lengths. In *Quantitative Analysis of Dependency Structures*, 72: 31-52.
- Zipf George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Appendix

The table shows the complete dependency distance frequency data from the SUD version of the English PUD treebank. The first three frequency columns take into account all dependency relations of the treebank. The last three frequency columns only count syntactic relations that correspond to actual head-daughter relations, which are the following relations in SUD: *appos*, *clf*, *comp*, *det*, *discourse*, *dislocated*, *expl*, *mod*, *subj*, *vocative*.

Distance	PUD-all	PRT-all	RT-all	PUD-syntactic	PRT-syntactic	RT-syntactic
1	10,236	5,473	1,912	6,866	3,742	1,194
2	4,157	3,438	1,768	3,148	2,285	1,140
3	1,887	2,270	1,646	1,295	1,434	1,021
4	924	1,662	1,532	538	1,050	997
5	544	1,292	1,468	285	809	951
6	412	955	1,335	162	566	829
7	292	747	1,222	112	459	807
8	209	613	1,113	63	360	719
9	192	536	1,009	55	306	650
10	177	462	975	53	255	613
11	139	400	824	39	206	497
12	133	299	741	24	171	454
13	98	287	701	22	152	433
14	96	233	561	15	122	356
15	97	225	521	13	123	319
16	72	162	422	7	91	248
17	52	175	386	5	90	238
18	71	152	312	8	78	203
19	52	124	276	0	66	168
20	47	110	258	3	58	165
21	43	100	226	1	51	147
22	50	79	159	6	46	107
23	27	62	162	1	35	91
24	18	55	122	0	36	66
25	26	44	91	1	21	53
26	24	31	78	1	16	50
27	22	38	78	0	20	48
28	11	29	53	1	15	35
29	18	22	53	0	9	32
30	12	17	34	0	10	20
31	3	13	29	0	7	15
32	5	9	18	0	2	14
33	2	6	21	0	3	9
34	5	7	14	0	3	6
35	3	8	13	0	7	5
36	5	9	10	0	6	4
37	4	3	2	0	0	2
38	2	5	6	0	1	4
39	2	3	5	0	1	3
40	1	2	2	0	2	1
41	1	3	1	0	2	1
42	0	0	3	0	0	2
43	1	3	2	0	2	2
44	0	3	2	0	1	1
45	1	1	3	0	1	1
46	0	0	2	0	0	1
47	0	3	2	0	2	1
48	0	2	0	0	1	0
49	1	0	2	0	0	1
50	0	1	0	0	0	0
53	1	0	1	0	0	0
55	0	1	0	0	0	0
56	1	1	0	0	1	0
57	0	1	0	0	0	0

Full valency and the position of enclitics in the Old Czech

Radek Čech

University of Ostrava, Faculty of Arts
Department of Czech Language
Czech Republic
cechradek@gmail.com

Pavel Kosek

Masaryk University, Faculty of Arts
Department of Czech Language
Czech Republic
kosek@phil.muni.cz

Olga Navrátilová

Masaryk University, Faculty of Arts
Department of Czech Language
Czech Republic
olga@phil.muni.cz

Ján Mačutek

Masaryk University, Faculty of Arts
Department of Czech Language
Czech Republic,
and
Comenius University in Bratislava
Faculty of Mathematic, Physics and Infor-
matics
Department of Applied Mathematics and
Statistics
Slovakia
jmacutek@yahoo.com

Abstract

The paper is focused on the analysis of the relationship between the full valency of the predicate and the position of enclitics in the clause. For this analysis, ones of the oldest Old Czech prose texts were used. We set up the hypothesis - the higher the full valency of the predicate, the lower the probability of the occurrence of the enclitic after the initial phrase of the clause – and test it. The hypothesis was corroborated only for narrative texts. In the case of poetic texts, the hypothesis was rejected.

1 Introduction

Enclitics are language units with a variety of specific grammatical characteristics that have attracted linguists for decades. Despite the fact that a huge number of methods and theoretical approaches were applied to study of this phenomenon, some fundamental questions are still open. Among others, an empirical diachronic description and an explanation of the historical development of enclitics that is based on a larger amount of language material and interpreted from the point of view of the quantitative linguistics remain rather an unexplored field. The reasons are obvious: language material is accessible only with difficulties (in the majority of cases it must be transcribed from a manuscript); an annotation must be performed manually; for the oldest periods, a limited number of texts is available, etc. However, an analysis of the historical development of any language property (or unit) often brings knowledge that substantially enhances an understanding of the phenomena under study. Therefore, in a recent series of papers several properties of enclitics in Old Czech and their historical development were explored (cf. Kosek et al., 2018a, 2018b, 2018c, 2018d), with the aim of obtaining a diachronic perspective of their characteristics. This paper represents a further step in this endeavour. Specifically, we analyse the relationship between the position of the pronominal enclitic in the clause and the so-called full valency (Čech et al., 2010) of the clause predicate. We assume that the full valency (for details, see Section 4) is one of factors which significantly influence the position of the enclitic. Therefore, we set up the following hypothesis:

The higher the full valency of the predicate, the lower the probability of the occurrence of the enclitic after the initial phrase of the clause.

The position of the enclitic immediately after the initial phrase (hereafter, this position will be called the postinitial position, abbreviated as 2P) is considered, according to the Wackernagel's Law (Wackernagel, 1892), the basic position of this unit in the clause, cf. the position of reflexive "sě" in the sentence (1).

(1)

[Co] *sě tobě vidí, Šimone?*

what_{NOM} REFL_{ACC} see_{3.PS.SG.PRAES}

'What is thy opinion, Simon?'

Bible olomoucká (BiblOl) Mathew 17,24

In previous studies (Kosek et al., 2018c; Čech et al., 2019; Kosek et al., 2019) it was shown that there are several factors which move the enclitic to a position which is different from 2P: for instance, the length of the initial phrase, the style, and the impact of the original Latin pretext. All these factors decrease the probability of the occurrence of the enclitic in the 2P position. This study is based on the assumption that the full valency is another factor which should lead to a similar result. Reasons for this assumption are summarized in Section 4 where the notion of full valency is introduced in detail. Statistical methods applied in this paper require certain amount of data to be reliable, therefore, only the most frequent pronominal enclitic in Old Czech, i.e. the enclitic "sě" (accusative reflexive) is analysed. Data from the Olomouc Bible (Bible olomoucká, BiblOl) and Litoměřice-Třeboň Bible (Bible litoměřicko-třeboňská, BiblLitTřeb), which represent the oldest complete Czech Bible translation, are used.

2 Word order of enclitics in Old Czech

For the purpose of this study, we determine two positions of the enclitic (E) in Old Czech:

a) the postinitial position, schematically

[I][E][]*

where symbol [I] represents the initial phrase of the clause and symbol []* represents any consequent syntactic unit(s) of the clause (including the empty unit, i.e. the clause can end with the enclitic). The initial phrase can be represented by one or more words, cf. sentence (1) and (2) respectively.

(2)

[*toho věku*] *sě jemu porodil Isák*

that_{GEN.F.SG} age_{GEN.F.SG} REFL_{ACC} him_{DAT.M.SG} born_{PART.PRET.ACT.NOM.SG.M} Isaac_{NOM.M.SG}

'And as Abraham was a hundred years old, his son Isaac was born to him.'

BiblOl Genesis 21,5

b) non-postinitial positions, schematically

[I][]*[E][]*

cf. sentence (3)

(3)

[*Volanie Sodomských a Gomorrejských*] *rozмноžilo sě jest*

outcry_{NOM.N.SG} sodom_{ADJ.GEN.M.PL} and gomorrha_{ADJ.GEN.M.PL} multiply_{PART.PRET.ACT.N.SG} REFL_{ACC}
be_{AUX.PRET.3.SG}

'The cry of Sodom and Gomorrha is multiplied'

BiblOl Genesis 18,20

3 Language material

For the analysis, some books from the Olomouc Bible (Bible olomoucká, BiblOl) and one book (Acts) from Litoměřice-Třeboň Bible (Bible litoměřicko-třeboňská, BiblLitTřeb) were used. These Bibles originate from the beginning of 15th century (however, it is considered to be copied from missing older translation from 1360, cf. Kyas, 1997; Vintr, 2008) and can be ranked among the oldest Old Czech prose texts (older texts, from the first half of the 14th century, are poetic, and they cannot be used to observe word order characteristics). Since our long-term aim is an analysis of the historical development of the word order characteristics of enclitics, the use of one of the oldest texts seems to be a proper choice - the result of this study can be, afterwards, compared with the results based on later Czech Bible translations.

All the phenomena under the study must be annotated manually, therefore, only eight books from the Bible were analysed. Specifically, four books from the Old Testament and four books from the New Testament were chosen: Genesis (Gen), Isaiah (Is), Job (Job), Ecclesiastes (Ecc), Gospel of St. Matthew (Mt), Gospel of St. Luke (Lk), Acts (Act), and Revelation (Rev).

4 Full valency and word order of enclitics

The notion of full valency (FV) was introduced to linguistics by Čech et al. (2010) and was elaborated by Vincze (2014) and Čech et al. (2015). The FV approach is a reaction to the absence of reliable criteria for distinguishing obligatory arguments (complements) and non-obligatory arguments (optional adjuncts). FV does not distinguish between obligatory arguments and non-obligatory ones. Thus, all directly dependent units of the predicate which occur in the actual language usage comprise its full valency frame.

A higher FV of the predicate means a higher complexity¹ of the clause (at least at this level of the syntactic tree, i.e. at the root of the clause and its direct dependents). We assume that the higher complexity is the factor which increases the probability that the Wackernagel's Law is "violated" for the following reason. The occurrence of the enclitic in the 2P position often means that the enclitic is not in the position adjacent to its syntactically superior word. Further, a more complex the clause structure increases the difficulty of processing the clause structure cognitively, especially when it contains distant dependency relations. Consequently, the tendency to put the enclitic next to its syntactically superior word instead to the 2P position should be positively correlated with the complexity of the clause. Of course, the complexity of the clause could be, in the ideal case, determined from the property of the entire clause structure. However, the character of the language material, which must be annotated manually, forced us to focus exclusively on the FV as the measure of the clause complexity (i.e., only the highest levels of syntactic trees are taken into account). Admittedly, this approach has its limitations and more comprehensive characteristics of the syntactic complexity will have to be applied in future research, but results achieved indicate that the positions of the enclitics are indeed influenced also by syntactic properties of the clause of which they are part.

5 Results

The relationship between the FV and the proportion of enclitics in the 2P position is presented in Table 1 and Figure 1. Here, all data are merged together, i.e. these results represent property of the corpus comprising eight Biblical books (see Section 2). Since some FV sizes do not contain enough instances for a proper evaluation of the data (e.g., there are only four clauses with enclitics for which the FV attains the value of seven), we decided to pool the adjacent bins so that each bin contains at least ten instances. In Tables 2 and 3, the FV size expresses the weighted arithmetic mean, with frequencies being the weights.

¹ In quantitative linguistics it is usual to measure complexity of a syntactic structure as the number of its constituents (e.g. Köhler, 2012, p. 145). For other approaches to syntactic complexity see e.g. Miestamo et al. (2008) or Givón and Shibatami (2009).

FV	2P	non-2P	proportion of 2P
2	2	18	0.1
3	133	75	0.64
4	81	117	0.41
5	47	49	0.49
6.13	14	18	0.44

Table 1: The size of the full valency (FV), number of enclitics in the postinitial (2P) and non-postinitial position (non-2P), and proportion of the 2P in all chosen biblical books.

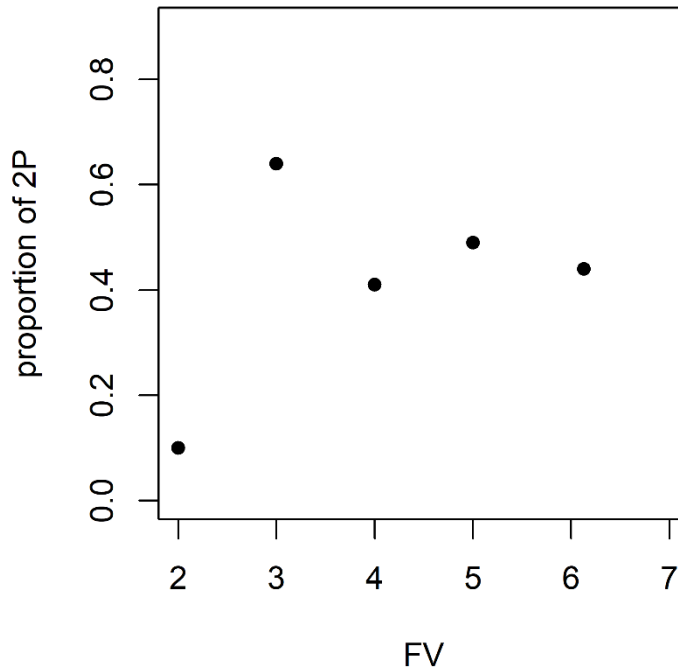


Figure 1: The relationship between the full valency (FV) and the proportion of enclitics in the postinitial (2P) position in all chosen biblical books

It is obvious (see Table 1 and Figure 1) that there is no tendency corresponding to our prediction from Section 1. On the contrary, clauses with the lowest FV have the lowest proportion of the 2P positions of enclitics which is a direct contradiction to the hypothesis. However, it is known that the distribution of particular positions of enclitics is significantly influenced by the style (Kosek et al., 2018c). As Biblical books fundamentally differ with respect to style, we studied also the relationship between the FV and the position of the enclitic separately in individual books. The results are presented in Table 2 and Figure 2.

The analysis of individual books brings rather a different picture. We can see that results from Act, Mt, Lk, and Gen corroborate the hypothesis, while results from Job and Ecc falsify it. As for Is and Rev, there are not enough data for a conclusion. At the first sight, it seems that there are differences between narrative texts (i.e., Act, Mt, Lk, and Gen) and poetic texts (i.e., Job and Ecc). In the case of poetic texts, their specific character can be a reason why the hypothesis is rejected – the author must fulfil some conditions to fit the rules of poetry, which can influence (or violate) the mechanism underlying the hypothesis.

$\mathbf{VF_{Act}}$	$\mathbf{2P_{Act}}$	$\mathbf{non-2P_{Act}}$	$\mathbf{proportion\ of\ 2P_{Act}}$	$\mathbf{VF_{Lk}}$	$\mathbf{2P_{Lk}}$	$\mathbf{non-2P_{Lk}}$	$\mathbf{proportion\ of\ 2P_{Lk}}$
2	2	18	0.1	2.86	20	16	0.56
3	133	75	0.64	4	18	18	0.5
4	81	117	0.41	5.16	8	11	0.42
5	47	49	0.49				
6.13	14	18	0.44				

$\mathbf{VF_{Mt}}$	$\mathbf{2P_{Mt}}$	$\mathbf{non-2P_{Mt}}$	$\mathbf{proportion\ of\ 2P_{Mt}}$	$\mathbf{VF_{Gen}}$	$\mathbf{2P_{Gen}}$	$\mathbf{non-2P_{Gen}}$	$\mathbf{proportion\ of\ 2P_{Gen}}$
2.95	15	6	0.71	2.91	20	12	0.63
4	7	12	0.37	4	13	19	0.41
5.64	5	9	0.36	5.25	11	13	0.46

$\mathbf{VF_{Job}}$	$\mathbf{2P_{Job}}$	$\mathbf{non-2P_{Job}}$	$\mathbf{proportion\ of\ 2P_{Job}}$	$\mathbf{VF_{Ecc}}$	$\mathbf{2P_{Ecc}}$	$\mathbf{non-2P_{Ecc}}$	$\mathbf{proportion\ of\ 2P_{Ecc}}$
2.92	23	15	0.61	2.93	34	27	0.56
4	11	21	0.34	4	12	27	0.31
5.25	11	9	0.55	5	9	2	0.82

Table 2: The size of the full valency (FV), number of enclitics in the postinitial (2P) and non-postinitial position (non-2P), and proportion of the 2P in individual Biblical books.

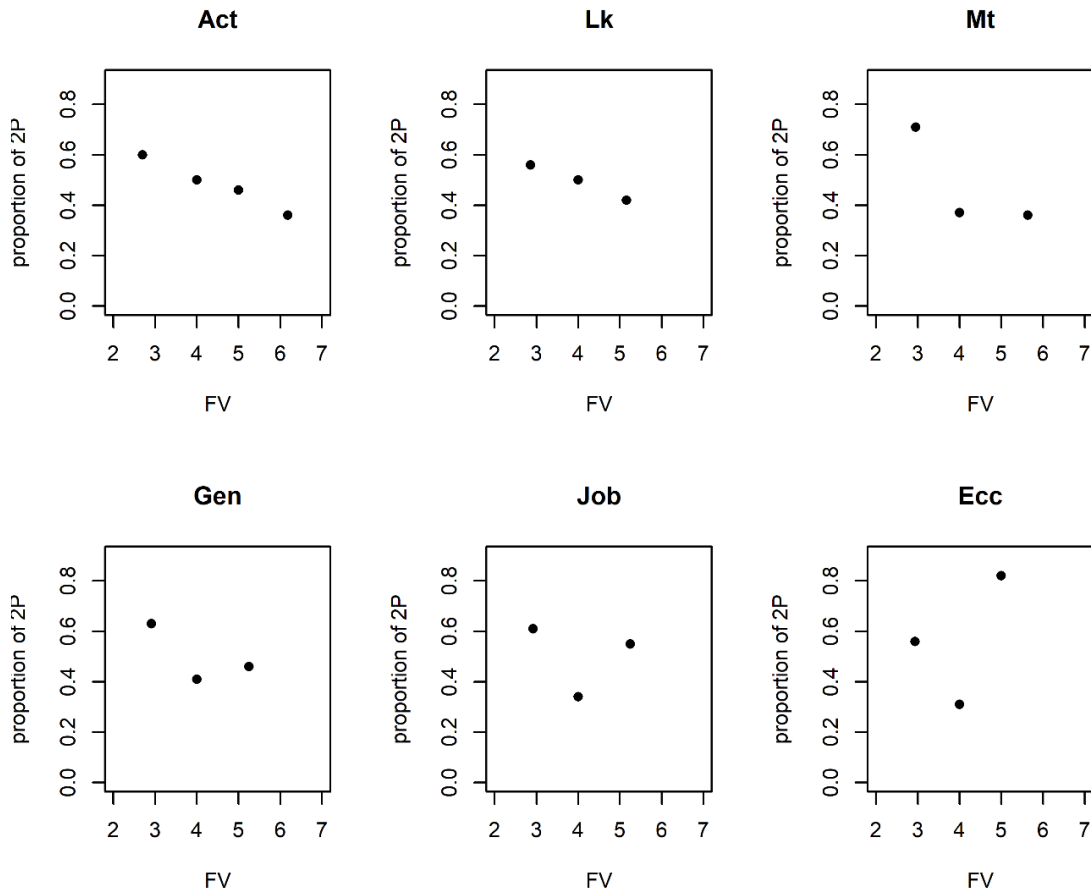


Figure 2: The relationship between the full valency (FV) and the proportion of enclitics in the postinitial (2P) position in individual Biblical books.

6 Conclusions

The study brings some important findings. First, even though the hypothesis was falsified when it was tested on both poetic texts and a corpus consisting of eight Biblical books, we do not reject the hypothesis generally. We assume that the poetic character of texts can be interpreted as a border condition which restricts the validity of the hypothesis. Further, it was revealed that mixing texts is another factor that can influence the outcome of hypothesis testing significantly. A mixture of different texts (e.g. with respect to their genre or style) means that particular mechanisms can "fight" each other and, as a consequence, their influence can be weakened (or it can even disappear). Finally, it must be emphasized that this paper is the first attempt to test this hypothesis. Needless to say, further research is necessary in this research field.

Acknowledgements

This study was supported by the project *Development of the Czech pronominal (en)clitics* (GAČR GA17-02545S).

References

- Radek Čech, Petr Pajas, and Ján Mačutek. 2010. Full Valency. Verb Valency without Distinguishing Complements and Adjuncts. *Journal of Quantitative Linguistics*, 17:291-302.
- Radek Čech, Ján Mačutek, and Michaela Koščová. 2015. On the relation between verb full valency and synonymy. In Eva Hajičová and Joachim Nivre (eds.), *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*:68–73. Uppsala University, Uppsala.
- Radek Čech, Pavel Kosek, Olga Navrátilová, and Ján Mačutek. 2019. On the impact of the initial phrase length on the position of enclitics in the Old Czech. *Proceedings of QUALICO 2018* (submitted).
- Talmy Givón and Masayoshi Shibatani. 2009. *Syntactic Complexity: Diachrony, Acquisition, Neuro-Cognition, Evolution*. Benjamins, Amsterdam/Philadelphia.
- Pavel Kosek, Radek Čech, Olga Navrátilová, and Ján Mačutek. 2018a. On the Development of Old Czech (En)clitics. *Glottometrics*, 40:51-62.
- Pavel Kosek, Olga Navrátilová, Radek Čech, and Ján Mačutek. 2018b. Word Order of Reflexive 'sě' in Finite Verb Phrases in the First Edition of the Old Czech Bible Translation. (Part 1). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135(3):177-188.
- Pavel Kosek, Olga Navrátilová, Radek Čech, and Ján Mačutek. 2018c. Word Order of Reflexive 'sě' in Finite Verb Phrases in the First Edition of the Old Czech Bible Translation. (Part 2). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135(3):189-200.
- Pavel Kosek, Olga Navrátilová, and Radek Čech. 2018d. Slovosled staročeských pronominálních enklitik závislých na VF ve staročeské bibli 1. redakce. *SLAVIA časopis pro slovanskou filologii*, 87(1-3):189-204.
- Pavel Kosek, Radek Čech, and Olga Navrátilová. 2019. The influence of the Latin pretext on the word order of pronominal enclitics in the Bible of Olomouc and the Bible of Litoměřice-Třeboň. *Proceeding of the 3rd Diachronic Slavonic Syntax conference DSSL* (submitted).
- Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. De Gruyter, Berlin/Boston.
- Vladimír Kyas. 1997. *Česká Bible v dějinách národního písemnictví*. Praha: Vyšehrad.
- Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson (eds.). 2008. *Language Complexity: Typology, Contact, Change*. Benjamins, Amsterdam/Philadelphia.
- Veronika Vincze. 2014. Valency frames in a Hungarian corpus. *Journal of Quantitative Linguistics*, 21(2):153–176.
- Josef Vintr. 2008. Bible (staroslověnský překlad, české překlady). In Luboš Merhaut et al. (eds.), *Lexikon české literatury* [vol. 4/2: U–Ž; Dodatky A–Ř]: 1882–1887. Academia, Praha.
- Jacob Wackernagel. 1892. Über ein Gesetz der indogermanischen Wortstellung. *Indogermanische Forschungen*, 1(1): 333-436.

Dependency length minimization vs. word order constraints: an empirical study on 55 treebanks

Xiang Yu, Agnieszka Falenska and Jonas Kuhn

Institut für Maschinelle Sprachverarbeitung

University of Stuttgart

firstname.surname@ims.uni-stuttgart.de

Abstract

This paper expands on recent studies of very large treebank collections aiming to find empirical evidence for language universals, specifically for the functionally motivated Dependency Length Minimization (DLM) hypothesis. According to DLM grammars are set up to support the expression of utterances in a way that minimizes the distance between heads and dependents. We construct several incremental baselines that lead from the random free order linearization to the real language by adding various word order constraints. We conduct detailed analyses on 55 treebanks and find that all of the constraints contribute to DLM. We show that DLM on the one hand shapes the regularity and on the other motivates the attested exceptions from canonical word order. The findings contribute to a more fine-grained, differentiated picture of the role of DLM in the interaction of competing constraints on grammar and language use.

1 Motivation and Background

The recent development of comparable dependency treebanks for a considerable number of languages across the typological spectrum (Nivre et al., 2016) has made it possible to address some long-standing hypotheses regarding a functional explanation of linguistic universals.

A number of recent papers (Liu, 2008; Futrell et al., 2015, a.o.) have used evidence from treebanks across languages to address what is arguably the most prominent hypothesis of a functionally motivated universal constraint, the Dependency Length Minimization (DLM) hypothesis, which can be traced back to (Behaghel, 1932).

Phrased as a language typological universal, the DLM hypothesis states that the evolution of languages is driven by the constraint that grammars should allow dependents to be realized as closely as possible to their heads – which is known to reduce the cognitive burden in processing (Gibson, 1998; Gibson, 2000). The Dependency Length (DL) of a sentence is defined as the sum of the distance between the head and dependent of all the dependency arcs in the sentence (see the example in Figure 1).

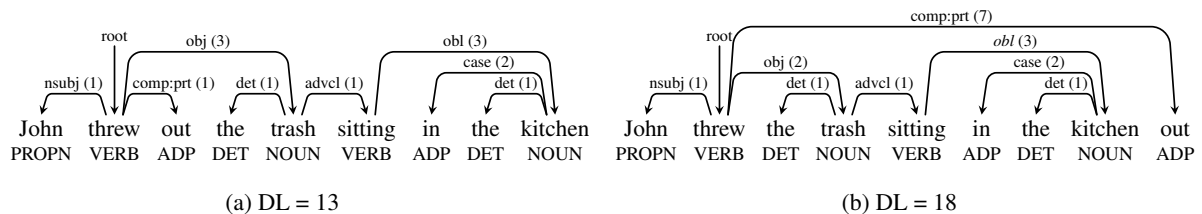


Figure 1: Example dependency trees and their dependency lengths, adapted from Futrell et al. (2015). The tree on the left is preferred since it has shorter DL and lower cognitive burden.

Liu (2008) perform the first cross-treebank study to compare the actual length of dependencies, for 20 languages, against the length that results when the dependency structures are linearized in random ways. The results indicate that languages indeed tend to minimize the dependency distance. Futrell et al. (2015) present a recent expansion of this type of treebank study to a set of 37 languages (which they argue to be the first comprehensive analysis that covers a broad range of typologically diverse

languages), presenting comparisons of the real dependency trees from the treebank with random re-orderings of the same dependency structures. The analysis shows that indeed across all 37 analyzed languages the real DL is significantly shorter than chance. This result corroborates findings from a broad range of empirical studies that are typologically less comprehensive (Gildea and Temperley, 2010; Gulordava and Merlo, 2015; Gulordava et al., 2015). This type of cross-treebank study has prompted a fair number of expansions and discussion regarding the typological implications (e.g. Jiang and Liu (2015), Chen and Gerdes (2018), Temperley and Gildea (2018)). In the present contribution, we go into some detail regarding a question that Futrell et al. (2015) have touched on, namely the relation between (the objective of minimizing) dependency length and language-specific word order constraints (which can also contribute to minimizing the cognitive load in parsing – but may conflict with the DLM objective).

Futrell et al. (2015) are careful to point out their awareness that the type of corpus study they performed makes it hard to distinguish the language typological aspect of the DLM hypothesis on the one hand (which would explain the exclusion of certain logically possible grammatical systems, which go against functional constraints/cognitive processing preferences) from facts about language use, relative to the respective grammatical constraints of a language, on the other.¹ The latter aspect, which is purely a matter of language processing, has also been discussed extensively under the DLM hypothesis (see e.g. Wasow (2002)). Futrell et al. (2015) “do not distinguish between DLM as manifested in grammars and DLM as manifested in language users’ choice of utterances; the task of distinguishing grammar and use in a corpus study is a major outstanding problem in linguistics, which we do not attempt to solve here.”

Besides the methodological question of how one could separate effects from corpus observations, it is worthwhile noting that in the logic of the typological DLM hypothesis, the aspect of language use cannot be completely ignored: if language evolution is indeed driven by these functional constraints, it should favor languages that permit *variation* – so speakers can react to the relative heaviness of constituents in the specific content they want to realize.

A comparison of the real DL with a random re-ordering baseline will necessarily conflate the effect from strict grammatical constraints (which the baseline re-orderings may break arbitrarily) and the relative freedom that any given grammar will leave open within the space of its constraints – and which speakers can exploit to optimize their utterances. Since they are aware of the effects of strict word order constraints that many languages impose, Futrell et al. (2015) present an additional comparison with a *Fixed Word Order Baseline* that is not fully random, but enforces consistent ordering constraints within each sentence. We note that the fact that this baseline chooses a new dependency relation ordering scheme for each sentence does not make it a good candidate for getting closer to the separation of globally fixed word order constraints dictated by the grammar and remaining spaces of free variation – which does seem to play a crucial role for approaching more fine-grained typological generalizations (and which ultimately needs to be clarified before one can claim to have empirical evidence for the DLM hypothesis as manifested in grammars rather than as a cognitive processing preference).

In this paper, we propose alternative baseline realizations of dependency trees that allow us to look more closely at the effect of specific relative ordering constraints in the comparison between random re-ordering realizations and the real treebank sequences along with their DL. We can thus study the manifestations of DLM relative to specific ordering phenomena in isolation. With a differentiated set of baselines, we can identify the DLM effect (1) in the distribution of dependents to both sides of the head; (2) in the direction of each single dependent to its head, and (3) in the ordering of the siblings on the same side of the head. These three phenomena can be seen as different types of word order constraints, where the first one concerns the quantity and balance of dependents, the latter two involve the order of individual dependency patterns, i.e., the combination of part-of-speech and dependency relation of the involved tokens. These word order constraints have been studied in various work (Ferrer-i Cancho, 2008; Ferrer-i Cancho, 2015; Liu, 2010; Gulordava et al., 2015). In this work, we identify all of the constraints

¹Other effects that cannot be captured (as Futrell et al. (2015) note) are differences in inflectional marking (in languages with strong case marking, much of the functional burden from word order is lifted) and the availability of non-projective dependency structures, which are a way of achieving shorter overall dependency length without violating certain hard word order constraints (Ferrer-i Cancho and Gómez-Rodríguez, 2016).

together in a large collection of treebanks, and show that each constraint contributes to the DLM in its unique way. Furthermore, we study their interaction by experimenting alternative word orders deviating from the original data and observe the impact on the dependency length.

We experiment with data instances following the regularities of word order and instances that manifest exceptions separately, and show that the word orders in the real data have shorter DL in both cases. This suggests that DLM is likely not just a result of the fixed word order, since it also happens in the non-canonical word order. Rather, it supports the hypothesis that the DLM influences both the regularities and the exceptions of word order.

2 Experiments on Dependency Length

2.1 Data

We perform our experiments on a selection of 55 treebanks from Universal Dependencies v2.4 (Nivre et al., 2019). The selection consists of training sets from all treebanks with at least 500 sentences (we take sentences with maximum 50 words). Where there are multiple treebanks for one language, we use the largest one to ensure stable and consistent estimation. We remove punctuation from trees and do not consider it when calculating DL, since punctuation biases statistics by introducing many long-distance right dependencies, and such dependencies do not contribute to the meaning of the head.

We consider only projective trees, since the non-projectivity could be interpreted as another way to minimize DL (Ferrer-i Cancho and Gómez-Rodríguez, 2016), which is out of the scope of this paper. Focusing on projective trees allows us to (1) analyze the DL on the subtree level, since the internal ordering of each subtree does not influence other subtrees; and (2) efficiently find the optimal ordering in terms of DL (Gildea and Temperley, 2007).

2.2 Baselines

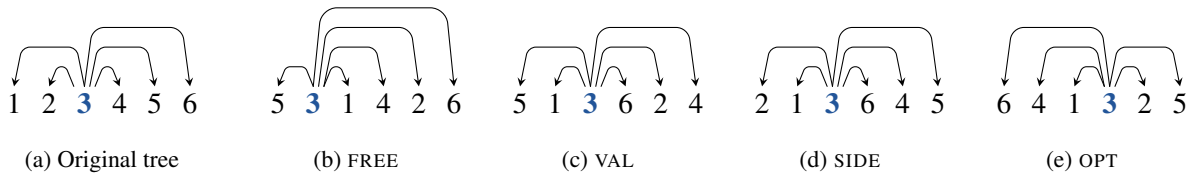


Figure 2: An example tree and results of four baseline linearization methods.

We compare the DL of the observed sentences in the treebanks (**OBS**) with four baselines that generate random linearizations with incremental constraints that leads to the real data:

FREE: free word order baseline, which does not impose any constraints except projectivity to the linearization. It is also used in Futrell et al. (2015).

VAL: same-valency baseline, which ensures that the numbers of left and right dependents of a head in the random subtree are the same as in the real data². In other words, we shuffle the left and right dependents in the original tree separately.

SIDE: same-side baseline, which ensures that the dependents in the random tree stay on the same side of the head as in the real data. This also satisfies the constraints in VAL, and differs from OBS only in the linear arrangement of dependents on each side.³

OPT: optimal baseline, which minimizes the dependency length, as in Gildea and Temperley (2007).

Figure 2 illustrates an example for each baseline, where 2a is the original ordering; 2b shuffles all the tokens in the tree; 2c ensures that there are 2 left dependents and 3 right dependents as in 2a; 2d shuffles the left and right dependents of 2a separately; and 2e is the optimal ordering of the tokens, assuming the label of each dependent also signifies the size of its subtree.

²We use the term *valency* only to describe the number of dependents, not their types.

³SIDE baseline is mentioned but not analyzed in the appendix of Futrell et al. (2015).

2.3 Results

Figure 3 illustrates the average DL of the five described baselines with respect to the sentence length.⁴ From the longest to the shortest are: FREE, VAL, SIDE, OBS, and OPT, and each adjacent pair is clearly separated. We follow with systematic analyses to explain the differences between the baselines. Section 3.2 explains the influence of balanced left and right dependents to the DL (FREE vs. VAL); Section 3.3 illustrates the influence of the head direction as a word order constraint (VAL vs. SIDE); and 3.4 shows the ordering of same-side siblings as another word order constraint (SIDE vs. OBS). In each of the aspects, the constraints from the observed data show clear preference for shorter DL, but does not reach the shortest possible DL due to other constraints, which partly explains the difference between OBS and OPT.

3 Analysis on Word Order Constraints

3.1 Definition and Statistics

Apart from the quantitative distribution of dependents, we study two types of word order constraints conditioned on the dependency pattern, namely **head direction** and **sibling ordering**.

The head direction constraint describes on which side of the head lies the dependent for each dependency pattern. Its dependency pattern is a triple of the universal part-of-speech (UPOS) tag of the head, the UPOS tag of the dependent, and the dependency label. An example dependency pattern from Figure 1 is $\langle \text{VERB}, \text{ADP}, \text{comp:prt} \rangle$, and its value is *right*.

The sibling ordering constraint describes the pairwise precedence relation of the dependents on the same side of the head. It is an approximation of the total order of the siblings, but much simpler and resistant to the data scarcity problem (Gulordava, 2018). Its dependency pattern is a sextuple of the UPOS of the head, the side of the involved dependents, and the UPOS and label of the two dependents. An example is $\langle \text{VERB}, \text{right}, \text{ADP}, \text{comp:prt}, \text{NOUN}, \text{obj} \rangle$, and the value is *left* for Figure 1a and *right* for Figure 1b.

For both dependency patterns, we count the frequency of their values, and use the entropy to measure the **freedom** of that pattern. For example, if a noun appears 50 times on the left of a verb as the subject, and 10 times on the right, then the entropy of $\langle \text{VERB}, \text{NOUN}, \text{obj} \rangle$ is 0.65. We then measure the freedom of the overall word order constraint by taking the average entropy of each single dependency pattern in that constraint type weighted by its frequency in the data. Figure 4 shows all treebanks characterized by the two types of word order freedom. We mark four most common language families and annotate the rest with their language codes. Generally, we do not see correlation of these two types of freedom, which indicates that they characterize different aspects of the word order. Many verb-final languages cluster near the top left corner, since they tend to have strict constraints that all arguments of a verb are uttered before the verb, but the exact ordering of the arguments is very flexible.

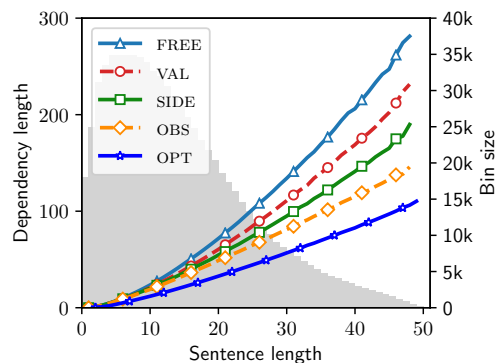


Figure 3: Average DL vs. sentence length across 55 treebanks; grey bars indicate frequency of the sentence length.

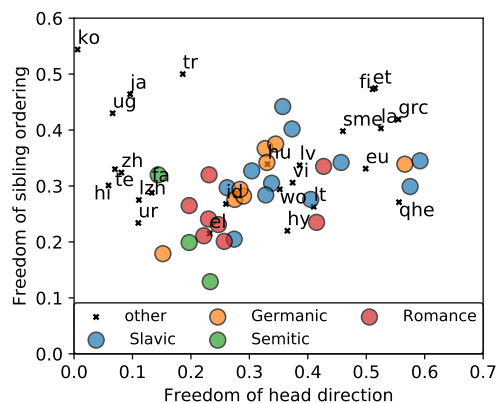


Figure 4: All treebanks characterized by the two types of word order freedom.

⁴The general trend and ranking of the baselines in each individual treebank are consistent with the average.

3.2 Dependent Distribution: FREE vs. VAL

Imposing the valency constraints on the FREE baseline makes the first reduction of DL. It can be easily explained by the fact that DL is shorter when the left and right dependents are more balanced, or in other words, when the head is positioned in the middle of its dependents, as shown in Ferrer-i Cancho (2015). In the FREE baseline, the head is equally likely to be placed in any position of the dependents, while in the VAL baseline, where the number of left and right dependents of each head is the same as from the real tree, it is more likely than chance to have balanced number of dependents on both sides.

To demonstrate this fact, we measure the **imbalance** of a head by the difference of numbers of dependents on both sides divided by the number of dependents. The more balanced a tree is, the smaller the value. We take the averaged imbalance value of all heads with more than one dependent as the imbalance measurement of the whole treebank. We also calculate the expected imbalance of FREE. For a head with n dependents, there are $n + 1$ possible location to insert the head, and the difference of dependents at each location would be: $\{n, n - 2, \dots, 2, 0, 2, \dots, n - 2, n\}$ if n is even, or $\{n, n - 2, \dots, 1, 1, \dots, n - 2, n\}$ if n is odd. The sum of these values is $\left\lfloor \frac{(n+1)^2}{2} \right\rfloor$, normalized by the length n and averaged by the equiprobable locations $n + 1$, thus the expected imbalance value is: $imb(n) = \left\lfloor \frac{(n+1)^2}{2} \right\rfloor / n(n + 1)$.

The measured average imbalance value for VAL is 0.47 (same as OBS) and 0.67 for FREE (also very close to the expectation from the formula), which means that VAL distributes the dependents in a more balanced way. There are only three languages (Telugu (te), Uyghur (ug), and Korean (ko)) where VAL has higher imbalance, mainly because they are verb-final languages, which have very unbalanced verb dependents. This general trend indicates that the real languages tend to have more balanced dependents (in other words, position the head more central) than chance level, thus the reduction of length from FREE to VAL.

Note that in this scenario, we only consider the *number* of dependents as the measurement of balance, which is a simplistic heuristics, while the *subtree size* of the dependents is the more accurate measurement. We explore this factor in Section 3.4.

3.3 Head Direction: VAL vs. SIDE

Next, we consider the length reduction from VAL to SIDE, which puts additional constraints on the head direction based on the real data.

We illustrate the relation between the head direction preference and its effect on DL through two scenarios. The first scenario studies the regularity of head direction: we flip the dependent to the other side of the head if it is on the **majority** side of its dependency pattern based on the statistics of the real data. The second scenario studies the exception of head direction: we flip the dependent if it is on the **minority side**. While flipping the dependent, we keep the order of all other dependents unchanged, and insert the flipped dependent into a position that minimizes the DL. This way, we make sure that the hypothetical flipping is optimal and the comparison to the original order is fair.

All the results are shown in Figure 5, where the y-axis shows the percentage of cases where the original order has shorter DL, i.e., flipping would increase the length, and the x-axis show the freedom of head direction for each language. If a point lies above the 50% line (red line on the plot), it means that the real data in that treebank has shorter DL than by chance.

Figure 5a shows the overall trend of the flipping experiment, where the real data is more likely to have shorter DL than the flipped ones, except for a few verb-final languages (Korean (ko), Japanese (ja), Telugu (te), and Uyghur (ug)), similar to the exceptions in Section 3.2. Since they tend to utter all the arguments before the verb, flipping any arguments would balance the head thus reduce the DL.

Note that by flipping one dependent to the other side, the valency of the head is also changed, therefore it is possible that the reduction of DL is caused by the effect described in Section 3.2. However, we record the imbalance before and after the flipping, and it is almost not changed on average, in other words, the influence from the change of valency is very small for this experiment.

Figure 5b shows the majority flipping scenario, which is very similar to the overall picture, since this scenario has more test cases (by definition), therefore dominates the statistics. The plot shows that the

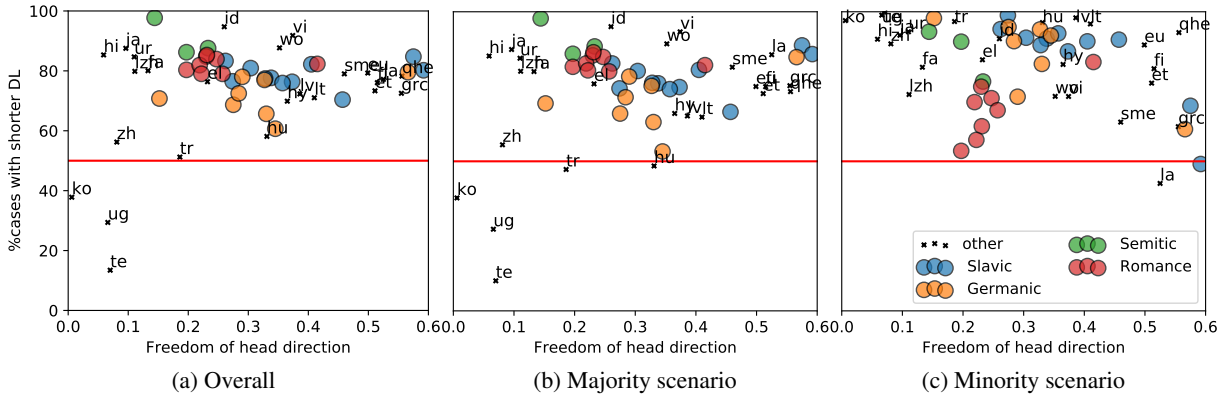


Figure 5: Percentage of trees where the real order has shorter DL than flipping one dependent.

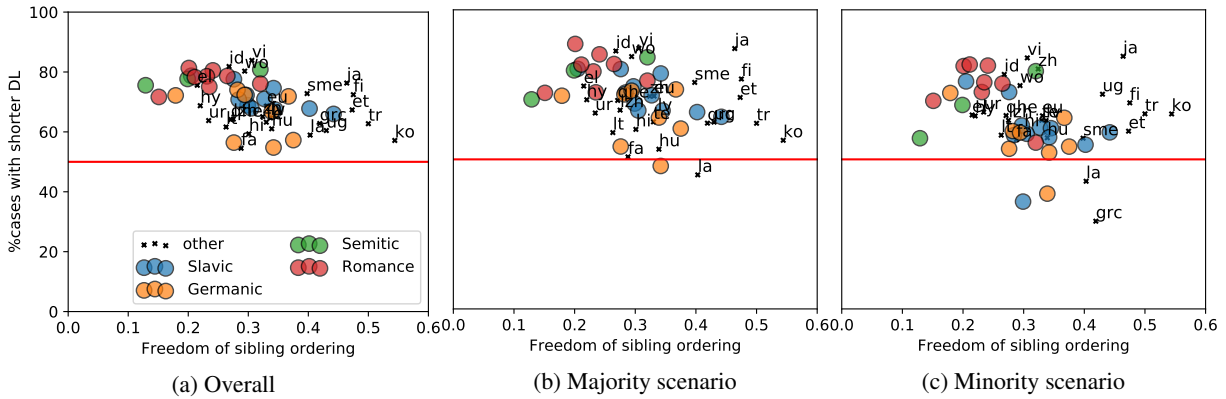


Figure 6: Percentage of trees where the real order has shorter DL than swapping two siblings.

canonical order leads to shorter DL for most of the languages, thus supports the hypothesis that DLM drives the evolution of word order. This then invites the question of why we still use non-canonical order in utterance. We can explain it in terms of DLM by observing the minority scenario, where we “correct” the head direction in the real data if it is not the majority choice. As shown in Figure 5c, for almost all languages, the original minority order has shorter DL than the flipped majority order. More generally, the minority scenario results resemble the mirror image of the majority case, namely the languages with shorter DL in the majority case would have less shorter DL in the minority case.

Assuming the head direction in each instance is realized without influence of DL, then either the majority or the minority scenario should have less than 50% cases with shorter DL. However, the fact that both scenarios have more than 50% positive cases indicates that DLM indeed influences the shaping of the head direction preferences as well as motivating the deviation from such preferences.

3.4 Sibling Ordering: SIDE vs. OBS

Finally, we look at the length reduction from SIDE to OBS, which only differ in the ordering of siblings on the same side. Their difference in DL indicates that the ordering of the same-side siblings is also influenced by DLM, as in the example in Figure 1. According to the explanation of DLM, we prefer 1a over 1b because keeping smaller subtrees closer to the head helps reduce DL and cognitive effort.

To verify whether the data supports the claim, we conduct similar experiments as in Section 3.3, where we compare every subtree in the original data to the modification where we swap one pair of dependents on the same side. We also compare two scenarios, majority and minority, where the swapped pair belongs to the regularity and exception of its dependency pattern, respectively.

The results are shown in Figure 6. We notice, that overall as well as in both majority and minority scenarios, the original subtrees has shorter DL than swapping two dependents. This again indicates that given the head direction, the real language tends to arrange the dependents in a way to minimize the

DL, regardless of whether the order is predominant or not. It also supports the similar conclusion as in Section 3.3 that DLM motivates both the regularity and exceptions in the ordering of siblings.

3.5 Technical Notes

It is worth noting that the dependency relation may be an artifact of the treebank design, which does not necessarily reflect the nature of human cognition. Many previous works have analyzed the syntactic and semantic treatment of the UD annotation scheme, cf. de Lhoneux and Nivre (2016), Wisniewski and Lacroix (2017), Osborne and Gerdes (2019). For example, UD tend to make the content word (NOUN) as the head of the function word (ADP), thus the common word order of placing a prepositional phrase *after* the noun it modifies would have longer DL than the opposite case, which is against the DLM hypothesis. In another annotation scheme, e.g. the Penn Treebank (Marcus et al., 1993), where the preposition is the head of the noun, the same word order would support the DLM hypothesis.

Another example of the influence of annotation scheme is the contrast between Korean(ko) and Japanese(ja) in our experiments. These two languages have very similar typological features, but stay rather far away in the plots. One major reason is that the Japanese treebank splits case markers as individual tokens, while the Korean treebank treats them as part of the noun. These idiosyncrasies could significantly change the statistics of the treebanks in terms of DL.

In this work, we do not consider other annotation schemes, e.g. the HamleDT collection (Zeman et al., 2014), nor do we deal with annotation idiosyncrasies within UD. However, we acknowledge that they have certain influence on the analysis of DL. Furthermore, which annotation scheme is closer to human cognition is still an open question.

4 Conclusion

In this paper we have broken down the effect of dependency length minimization step by step, and analyzed its relation to the dependent distribution, head direction, and sibling ordering. The systematic breakdown indicates that natural languages universally show clear preference for shorter dependency length in all three aspects, both in the regularities and in the exceptions. Our findings provide more detailed evidence for the hypothesis that DLM is a universal phenomenon in natural language.

One very interesting direction for future work is the interaction of multiple word order constraints. We have shown that most constraints are locally optimal, i.e., reversing a single constraint would likely increase the dependency length, since it might not be compatible with some other constraints. The grouping effect of constraints with respect to DL might provide explanations to some observations in Greenberg’s linguistic universal (Greenberg, 1963).

In this work, we use dependency patterns extract from the treebank to characterize the word order constraints, which is transparent but maybe not nuanced enough. An alternative way could be to use a statistical linearizer (Bohnet et al., 2010) to model the word order constraints (but the features implicitly related to DL should be carefully disabled), which could serve as a even stronger baseline.

Acknowledgements

This work was in part supported by funding from the Ministry of Science, Research and the Arts of the State of Baden-Württemberg (MWK), within the CLARIN-D research project. We also thank the anonymous reviewers for their feedbacks and ideas for future work.

References

- Otto Behaghel. 1932. *Deutsche Syntax: eine geschichtliche Darstellung. Wortstellung, Periodenbau*. Winter.
- Bernd Bohnet, Leo Wanner, Simon Mille, and Alicia Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 98–106. Association for Computational Linguistics.

- Xinying Chen and Kim Gerdes. 2018. How do universal dependencies distinguish language groups? In Jingyang Jiang and Haitao Liu, editors, *Quantitative Analysis of Dependency Structures*, pages 277–293, Berlin/Boston. De Gruyter Mouton.
- Miryam de Lhoneux and Joakim Nivre. 2016. Should have, would have, could have. investigating verb group representations for parsing with universal dependencies. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 10–19.
- Ramon Ferrer-i Cancho and Carlos Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.
- Ramon Ferrer-i Cancho. 2008. Some word order biases from limited brain resources: A mathematical approach. *Advances in Complex Systems*, 11(03):393–414.
- Ramon Ferrer-i Cancho. 2015. The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change*, 5(1):114–137.
- 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.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.
- 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.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.
- Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.
- 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.
- Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. 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)*, volume 2, pages 477–482.
- Kristina Gulordava. 2018. *Word order variation and dependency length minimisation: a cross-linguistic computational approach*. Ph.D. thesis, University of Geneva.
- 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.
- 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.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*.
- Joakim Nivre, Mitchell Abrams, Željko Agić, et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa: a journal of general linguistics*, 4(1).

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

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI lecture notes. CSLI Publ.

Guillaume Wisniewski and Ophélie Lacroix. 2017. A systematic comparison of syntactic representations of dependency parsing. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 146–152.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

Advantages of the flux-based interpretation of dependency length minimization

Sylvain Kahane

sylvain@kahane.fr

Chunxiao Yan

Modyco, Université Paris Nanterre & CNRS

yanchunxiao@yahoo.fr

Abstract

Dependency length minimization (DLM, also called dependency distance minimization) is studied by many authors and identified as a property of natural languages. In this paper we show that DLM can be interpreted as the flux size minimization and study the advantages of such a view. First it allows us to understand why DLM is cognitively motivated and how it is related to the constraints on the processing of sentences. Second, it opens the door to the definition of a big range of variations of DLM, taking into account other characteristics of the flux such as nested constructions and projectivity.

1 Introduction

The dependency flux between two words in a sentence is the set of dependencies that link a word on the left with a word on the right (Kahane et al., 2017). The size of the flux in an inter-word position is the number of dependencies that cross this position.¹ The flux size of a sentence is the sum of the sizes of the inter-word fluxes.

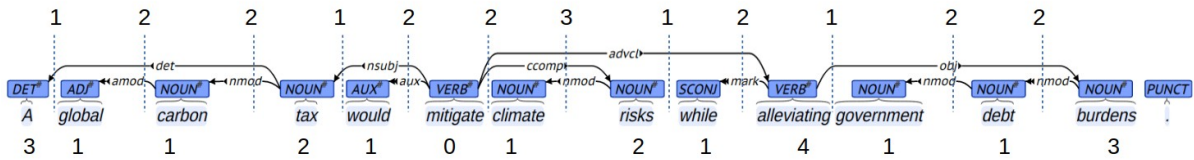


Figure 1. An ordered dependency tree with the flux size on top of each inter-word position and, under each word, the length of the dependency from its governor

On the top line, Figure 1 shows the size of the flux at each inter-word position. In the first position, between *A* and *global*, there is only one dependency crossing ($A <_{\text{det}} \text{tax}$); in the second position, between *global* and *carbon*, there are two dependencies ($A <_{\text{det}} \text{tax}$; $\text{global} <_{\text{amod}} \text{carbon}$).

On the bottom line, Figure 1 shows, for each word, the length of the dependency that links that word to its governor. For example, the first word *A* is linked to its governor *tax* by a dependency of length 3 because this dependency crosses 3 inter-word positions. The dependency length of a sentence is the sum of the lengths of the dependencies of that sentence.

It can be verified, for the sentence in Figure 1, that:

$$\text{Dependency flux size of the sentence} = 1+2+2+1+2+2+3+1+2+2+2+2 = 21$$

$$\text{Dependency length of the sentence} = 3+1+1+2+1+0+1+2+1+2+1+4+1+1+3 = 21$$

It is easy to check that the dependency length is always equal to the dependency flux size. Since the length of a dependency is the number of fluxes on which this dependency belongs, the size of the flux is

¹ The flux in a given position can also be viewed as the set of incomplete dependencies when only one of the two halves of the dependency structure is considered. Gibson (2000) introduces an Incomplete Dependency Hypothesis on the syntactic complexity, but he only considers dependencies of obligatory subcategorized elements.

the sum, on all the dependencies, of the number of fluxes they cross. In other words, these two values are equal to the number of crossings between a dependency and an inter-word position.

Several studies have studied dependency length and shown that natural languages tend to minimize it (Liu, 2008; Futrell et al., 2015). This property is called dependency length minimization (DLM) or dependency distance minimization (dependency lengths can be interpreted as distances between syntactically related words). DLM is correlated with several properties of natural languages. For instance, the fact that dependency structures in natural languages are much less non-projective than in randomly ordered trees can be explained by DLM (Ferrer i Cancho, 2006; Liu, 2008). It is also claimed that DLM is a factor affecting the grammar of languages and word order choices (Gildea and Temperley, 2010; Temperley and Gildea, 2018).

Since the dependency length is equal to the dependency flux size, by trying to minimize the lengths of the dependencies, we also try to minimize the sizes of the inter-word fluxes. This gives us two different views on DLM. The objective of this article is to show that thinking about DLM in terms of flux has several advantages. In section 2, we will show that the interpretation of DLM in terms of flux makes it possible to highlight the cognitive relevance of this constraint. In section 3, we will examine other flux-based constraints related to DLM.

2 Cognitive relevancy of DLM

As we have just seen, DLM corresponds to the minimization of the flux size of the sentence and therefore of all inter-word fluxes. However, since we know that sentences are more or less parsed as fast as they are received by the speakers (Frazier and Fodor, 1978), we can see the flux in a given inter-word position as the information resulting from the portion of the sentence already analyzed that is necessary for its further analysis. In other words, there is an obvious link between the inter-word flux and the working memory of the recipient of an utterance (as well as the producer of the utterance).

The links between syntactic complexity and working memory have often been discussed, starting with Yngve (1960) and Chomsky and Miller (1963). According to Friederici (2011), “the processing of syntactically complex sentences requires some working memory capacity”. The founding work on limitations of working memory is Miller’s (1956), who defended that the span is 7 ± 2 elements; this limitation has been updated between 3 and 5 meaningful items by Cowan’s work (2001). According to Cowan (2010), “Working memory is used in mental tasks, such as language comprehension (for example, retaining ideas from early in a sentence to be combined with ideas later on), problem solving (in arithmetic, carry a digit from the ones to the tens column while remembering the numbers), and planning (determining the best order in which to visit the bank, library, and grocery).” He adds that “There are also processes that can influence how effectively working memory is used. An important example is in the use of attention to fill working memory with items one should be remembering.”

We think that the dependency flux in inter-word positions is a good approximation of what the recipient must remember to parse the rest of the sentence. Of course, it is also possible to make a link between the working memory and DLM if it is interpreted in terms of dependency length: it means that it is cognitively expensive to keep a dependency in working memory for a long time and that the longer a dependency is, the more likely it is to deteriorate in working memory (Gibson, 1998; 2000).

3 DLM-related constraints

DLM is a constraint on the size of the whole flux of a sentence and therefore a particular case of constraints on the complexity of the flux. DLM is neither the only metrics for syntactic complexity (see Lewis (1996) for several constituency-based metrics; Berdicevskis et al. 2018), nor the only metrics on the complexity of the flux and perhaps not the best. We will present other potentially interesting flux-based metrics.

3.1 Constraints on the size of inter-word fluxes

We have seen that the sum of the lengths of the dependencies is equal to the sum of the sizes of the inter-word fluxes. Since there are as many dependencies as there are inter-word positions in a sentence ($n-1$ for a sentence of n words), this means that the average length of the dependencies is equal to the average size of the inter-word fluxes. For the entire UD database (version 2.4, 146 treebanks), this value is equal

to 2.73. But the equality of the average values does not mean that the values of these two variables, dependency length and flux size, are distributed in the same way. We give the two distributions in Figure 2.

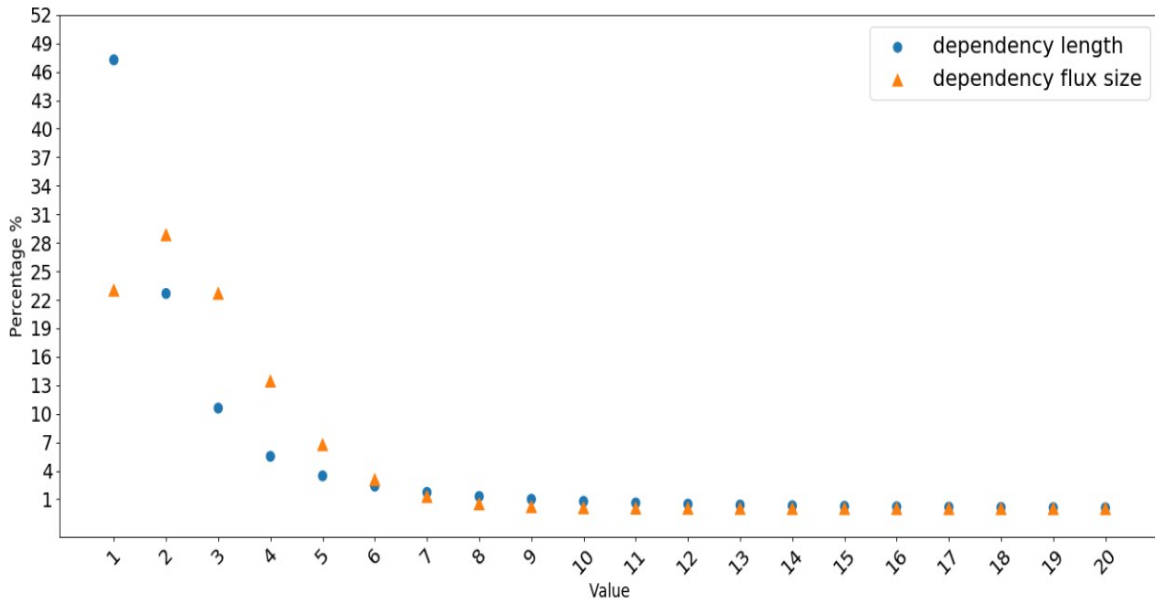


Figure 2. Repartition of dependency lengths vs. flux sizes

For the distribution of dependencies according to their length, we observe that quantities decrease rapidly with lengths, starting with 47% of length 1 dependencies; for the distribution of inter-word fluxes according to their size, we observe a higher quantity of size 2 fluxes than size 1 (29% versus 23%), then a slower decrease at the beginning than for dependency lengths, then much faster. The two curves cross for the value 7. In fact, 99% of the fluxes are of size ≤ 7 , while for the dependency length, it is necessary to reach the value 17 to have more than 99% of dependencies of this length or less (99.97% of fluxes of size ≤ 17 vs. 99.09% of dependency lengths). Said differently, there are 0.91% of dependency lengths ≥ 18 against 0.03% of flux sizes, that is, about 30 times more (see Appendix 1 for more detailed results).

If we look at treebanks separately we see similar results: see table of Appendix 2 which shows the distribution values of 47 treebanks containing more than 100,000 flux positions. Looking at the curves of the percentages of dependency length and flux size, we notice the same crossing between values 1 and 2: The percentage of size 1 fluxes is always lower than the percentage of length 1 dependencies, while the percentage of size 2 fluxes is higher than the percentage of length 2 dependencies. Then, the percentage of fluxes decreases very quickly, and a second crossing is between the value 5 (UD_Finish-FTB), and the value 8 (in 9 treebanks: UD_Urdu-UDTB, UD_Persian-Seraji, UD_Hindi-HDTB, UD_German-HDT, UD_German-GSD, UD_Dutch-Alpino, UD_Chinese-GSD, UD_Arabic-PADT and UD_Japanese-BC-CWJ). After this crossing, the percentage of flux size is lower than the percentage of dependency length and the former decreases much faster than the latter.

Looking at the cumulative percentages from value 2 onwards, the rate of flux size reduction is even sharper, as the crossing is between values 3 (in 4 treebanks: UD_Estonian-EDT, UD_Finnish-FTB, UD_Finnish-TDT, and UD_Polish-PDB) and 5 (in the same 9 treebanks as before). We notice in the treebanks with a crossing at 5, most are the head-final languages like Japanese (UD_Japanese-BCCWJ), German (UD_German-GSD, UD_German-HDT), Dutch (UD_Dutch-Alpino) and Persian (UD_Persian-Seraji), as well as Chinese (UD_Chinese-GSD) which are verb-initial position and head-final for other configurations. An exception is Arabic (UD_Arabic-PADT) which is a typical head-initial language. In treebanks with a crossing at 3, there are head-initial languages such as Finnish (UD_Finnish-FTB and UD_Finnish-TDT), Polish (UD_Polish-PDB) and Estonian (UD_Estonian-EDT). This could result from an asymmetry in the flux in languages according to the position of heads: For head-final languages, the flux size would be less constrained than for head-initial languages. This hypothesis remains to be confirmed by further study.

If we look for which value n we reach the 99% of dependencies of length $\leq n$, we find values ranging from 9 (UD_finish-FTB) to 27 (UD_Arabic-PADT). The same calculation for flux size gives values between 6 (in 12 treebanks: UD_Bulgarian-BTB, UD_Czech-FicTree, UD_Finnish-FTB, UD_French-GSD, UD_Italian-ISDT, UD_Korean-Kaist, UD_Norwegian-Bokmaal, UD_Polish-PDB, UD_Portuguese-GSD, UD_Romanian-RRT, UD_Russian-SynTagRus, and UD_Spanish-GSD) and 11 (UD_Japanese-BCCWJ) (this value is a little exceptional, since the value is equal to 7 for UD_Japanese-BC-CWJ). We also find that variations in dependency length are more sensitive than those in flux size in the different treebanks of a language. For example, for French, UD_French-FTB has 99% length dependencies ≤ 21 , and UD_French-GSD has 99% length dependencies ≤ 16 , while in the case of flux size, the values are 7 and 6 respectively.

If DLM expresses a constraint on the average value of dependency lengths and flux sizes, we see that there is also a fairly strong constraint on the size of each inter-word flux, whereas there is not such a strong constraint on the length of each dependency. For this reason, we postulate that DLM results more on a constraint on flux sizes than on dependency lengths, even if it is not possible to give a precise limit to the size of individual fluxes as Kahane et al. (2017) have already shown.

3.2 Center-embedding and constraints on structured fluxes

Beyond the question of their lengths, the way the dependencies are organized plays an important role in syntactic complexity. In particular, center-embedding structures carry a computational constraint in sentence processing (Chomsky and Miller, 1963; Lewis, 1996; Lewis and Vasishth, 2005). It is important to note that the complexity caused by center-embedding structures cannot be involved in DLM-based constraints. Neurobiological studies have highlighted the independence of memory degradation related to the length of a dependency and the computational aspect expressed by the center-embedding phenomena, which are located in different parts of the brain (Makuuchi et al., 2009).

As shown by Kahane et al. (2017), it is possible to express the constraints on the center-embedding in terms of constraints on the flux, but this requires to consider how all the dependencies belonging to the same flux are structured, by taking into account information about their vertices. Dependencies that share a vertex are referred to as a bouquet, while dependencies that have no common vertex are referred to as disjoint dependencies (Kahane et al., 2017). For example, the flux between *climate* and *risks* in Figure 1 contains 3 dependencies: the dependencies \langle nmod and \rangle ccomp form a bouquet (they share the vertex *risks*), \rangle ccomp and \rangle advcl also (they share the vertex *mitigate*), while \langle nmod and \rangle advcl are disjoint. The flux structure can be represented as shown by the table in Figure 3: vertices on the left of the considered inter-word position give the rows, beginning by the word which is closer to the position, while vertices on the right give the columns, beginning again by the word which is closer to the position. Dependencies which are on the same row or in the same column share a vertex.

	<i>risks</i>	<i>alleviating</i>
<i>climate</i>	\langle nmod	
<i>mitigate</i>	\rangle ccomp	\rangle advcl

Figure 3. Structure of the flux in the position between *climate* and *risks*

The disjoint dependencies correspond to nested constructions. For example, in our example, the dependency between *risks* and *climate* and the dependency between *mitigate* and *alleviating* are disjoint and therefore the unit [*risks,mitigate*] is fully embedded in the unit [*climate,alleviating*].² The number of disjoint dependencies in a flux is very constrained as shown by Kahane et al. (2017): 99.62% of the fluxes in the UD database have less than 3 disjoint dependencies. This suggests that bouquet structures are less constrained than disjoint structures. This is quite predictable if we consider that there are con-

² In the case of a projective tree, the interval between two words connected by a dependency forms a connected portion of the structure and thus a syntactic unit.

straints on working memory and that dependencies in a bouquet share more information than disjoint dependencies.

Note that non-projectivity can also be detected from a structured flux, if we take into account the order in which the vertices of the dependencies of the same flux are located. We plan in our further studies to look more precisely on the distribution of the different possible configurations of the flux.

3.3 Constraints on the potential flux

It must be remarked that we do not really know the flux when processing a sentence incrementally since we do not generally know which words already processed will be linked with a word not yet processed. We call potential flux in a given inter-word position the set of words before the position which are likely to be linked to words after it. See in particular the principles of the transition-based parsing (Nivre, 2003) which consists in keeping all the words already processed and still accessible in the working memory. The largest hypothesis on the potential flux is to consider that all words before the position are accessible. But clearly some words are more likely to have dependents (for instance, only content words can have dependents in UD). It is also possible to make structural hypothesis on the potential flux. We call projective potential flux the set of words accessible while maintaining the projectivity of the analysis. We will limit our study to the projective potential flux even if we are aware that, on one hand, projectivity is far to be an absolute constraint in many languages and, in the other hand, other constraints apply on the potential flux.

Figure 4 shows the value of the projective potential flux of our example. For instance, after *while* three words are accessible: *mitigate*, *risks*, and *while*; words before *mitigate* are not accessible because they are all depending on *mitigate* and *climate* is not accessible because it depends on *risks* and a projective link cannot cover an ancestor.

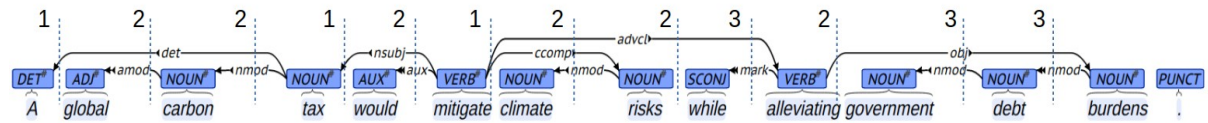


Figure 4. An ordered dependency tree with the projective potential flux of each inter-word position

Figure 5 compares the distribution of the sizes of potential projective fluxes and the sizes of observed fluxes for all UD treebanks (the sizes of observed fluxes have already been given in Figure 2).

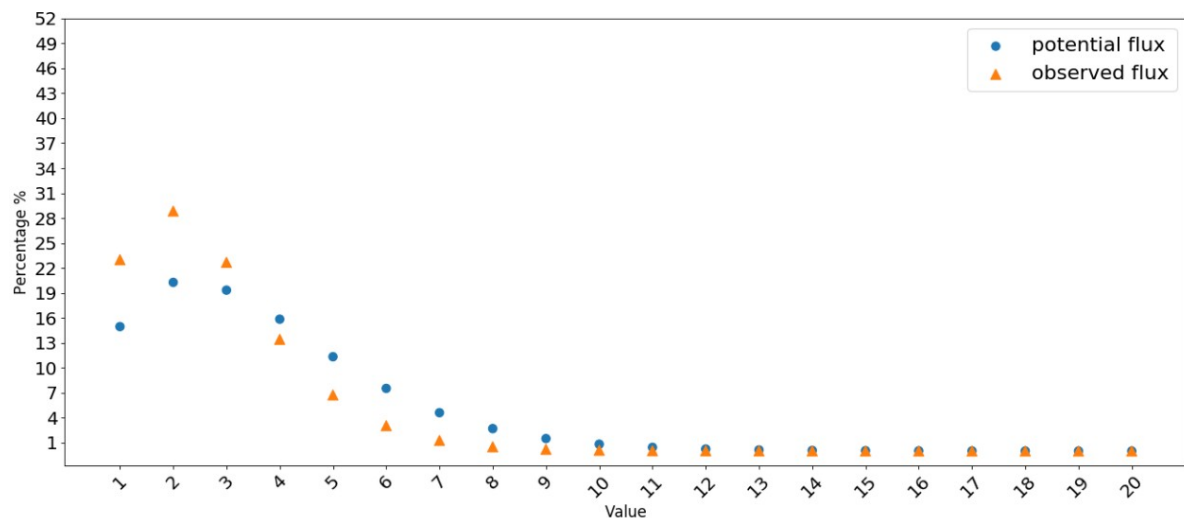


Figure 5. Distribution of projective potential fluxes vs observed fluxes

The distribution of projective potential fluxes is flatter (20% vs. 29% for size 2) with less fluxes with sizes ≤ 3 and more fluxes with sizes ≥ 4 , which means that projective potential fluxes generally have greater size than observed flux. From the size 2, the number of projective potential fluxes decreases but more slowly than the number of observed fluxes. It is necessary to reach size 11 to have more than 99% of the potential fluxes (99.39% of the potential fluxes have a size ≤ 11) while this value is reached with size 7 for the observed fluxes (see Appendix 3 for details).

It is interesting to note that the projective potential flux is not the same for head-initial and head-final dependencies. If the governor is before its dependent, they are both accessible for further projective dependencies. But if the dependent is before the governor, only the governor is accessible for further projective dependencies. Consequently, we decided to compare the distribution of the sizes of the projective potential fluxes of two head-initial languages (Arabic and Irish) with those of two head-final languages (Japanese and German) (Figure 6).³

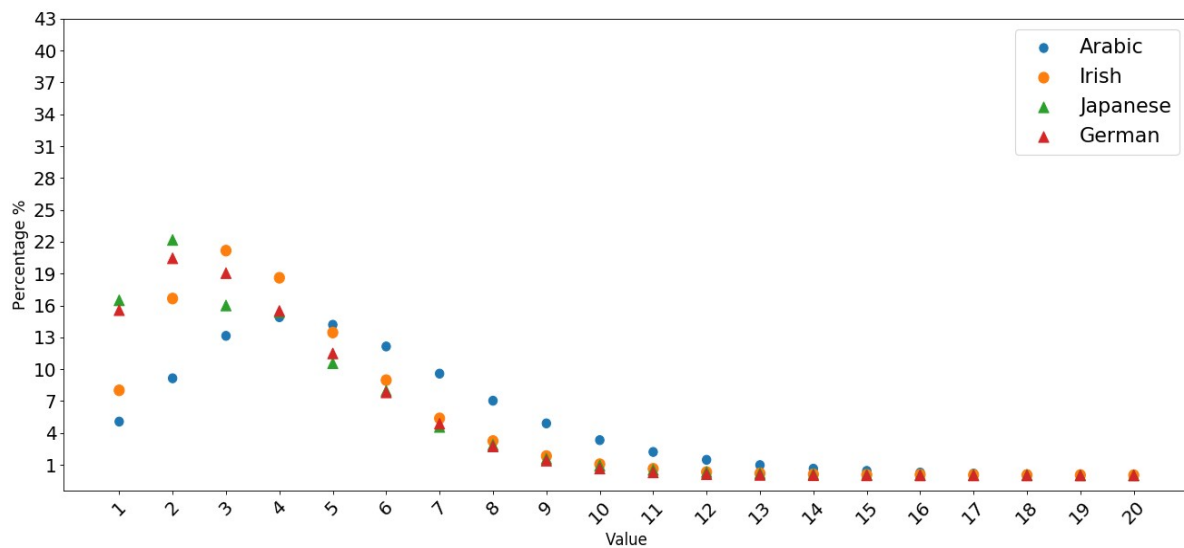


Figure 6. Comparison of projective potential flux sizes for head-initial and head-final languages

For head-final languages, the distribution of projective potential flux sizes is similar to the general distribution presented in Figure 5: Projective potential fluxes with size 2 are the most numerous, 22% for Japanese and 20.5% for German. Then the percentage reduces with the increase in size, the two final-headed languages have a percentage fairly close for size 7. More than 99% of projective potential fluxes have a size ≤ 10 for Japanese and ≤ 11 for German.

As expected, the distribution of projective potential flux sizes is different for head-initial languages: Projective potential fluxes with size 3 (Irish) and size 4 (Arabic) are the most numerous (21% for Irish and 15% for Arabic). Compared to head-final languages, not only does the percentage of projective potential flux size in head-initial languages increase more slowly (than for head-final languages) to reach the most represented size, but also it decreases more slowly afterwards. Thus, the distribution of head-initial languages is flatter than for head-final language and Arabic is particularly flat, which means that there are much more projective potential fluxes with greater sizes. From size 8 onwards, the distribution of Irish is very close to that of the two head-final languages. But in the case of Arabic, the distribution only approaches the other three from size 15. If we look at the cumulative percentage, more than 99% of projective potential flux have a size ≤ 11 for Irish and ≤ 15 for Arabic.

This difference in the distribution of projective potential fluxes for head-initial and head-final languages could have some consequences. Figure 7 shows the observed flux sizes for the same four languages.

³ German is V2 (verb second position) in main clauses and head-final in subordinated clauses.

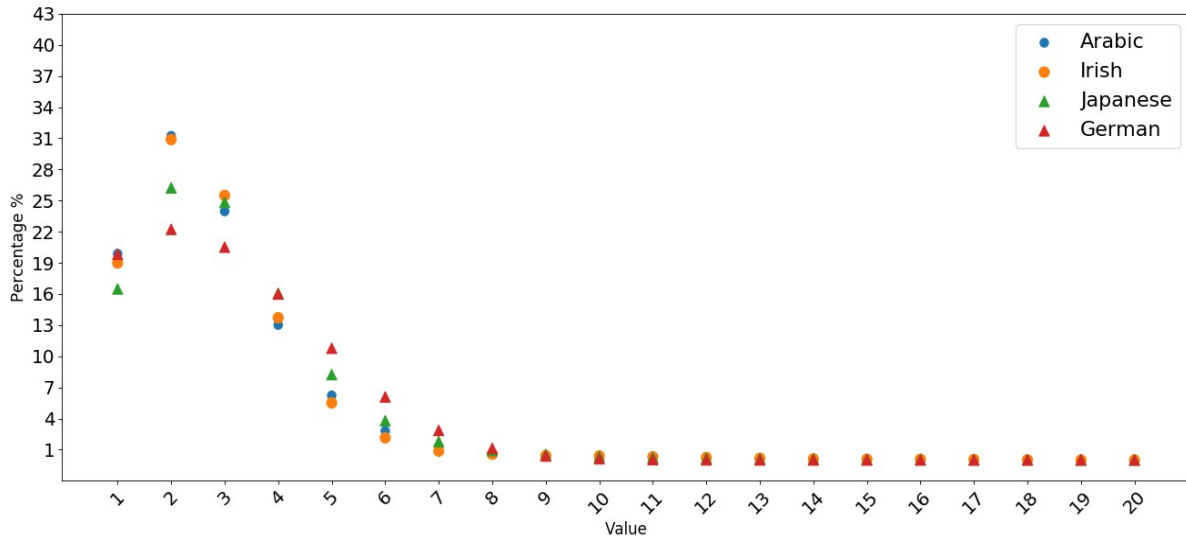


Figure 7. Comparison of observed flux sizes for head-initial and head-final languages

We have already noted that in UD treebanks, projective potential fluxes tend to have larger sizes than those of observed fluxes (Figure 5), but this trend is accentuated in head-initial languages: Projective potential fluxes with small sizes are much less numerous than for observed fluxes: For Arabic, 5%, 9%, and 13% of projective potential fluxes with sizes 1, 2, and 3 compared to 20%, 31%, 24% of observed fluxes. For Irish, 8%, 17% and 21% of projective potential fluxes with sizes 1, 2, and 3 compared to 19%, 31%, 25.5% of observed fluxes. Consequently, 73% of the potential fluxes with a size ≥ 4 compared to 25% of the observed fluxes for Arabic, and 54% compared to 25% for Irish.

It could seem contradictory that the projective potential flux is larger in head-initial languages (than in head-final languages), while the observed flux is significantly smaller (we have more fluxes with small sizes). It may be the result of a phenomenon of compensation: A larger potential flux increases the complexity of the (human, as well as automatic) parsing and a smaller observed flux would compensate for this. We do not have a better explanation for the time being and we leave this question open for further studies based on a deeper analysis of the data.

4 Conclusion

We have shown that dependency length minimization (DLM) is also a property of inter-word dependency fluxes. Such a view allows us to reformulate many assumptions on DLM. For instance, Gildea and Temperley (2018) remark that the idea that languages tend to place closely related words close together can be expressed as DLM. But as DLM comes down to reduce the flux and therefore the working memory, this can be reformulated by saying that the idea that languages tend to place closely related words close together can be expressed as a reduction of working memory.

We hope that this article will motivate studies on constraints on dependency flux, which are not limited to DLM. In particular, we believe that the constraints on the flux are far to be limited to its average size and that the structure of the flux plays an important role in its complexity. We have in particular shown an asymmetry between head-initial and head-final languages concerning the flux that could be related to the different potential flux in these two kind of languages.

Acknowledgments

We would like to thank our three reviewers for valuable remarks.

References

- Berdicevskis Aleksandrs et al. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 8-17.
- Ramon Ferrer i Cancho. 2006. Why do syntactic links not cross?. *EPL (Europhysics Letters)*, 76, 1228-1234.
- Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1), 87-114.
- Nelson Cowan. 2010. The magical mystery four: How is working memory capacity limited, and why?. *Current directions in psychological science*, 19(1), 51-57.
- Angela D Friederici. 2011. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4), 1357-1392.
- Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model, *Cognition*, 6(4), 291-325.
- 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.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1-76.
- Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000, 95-126.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length?. *Cognitive Science*, 34(2), 286-310.
- Sylvain Kahane, Alexis Nasr and Owen Rambow. 1998. Pseudo-projectivity: a polynomially parsable non-projective dependency grammar. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics* (Vol. 1).
- Sylvain Kahane, Chunxiao Yan, and Marie-Amélie Botalla. 2017. What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. In *4th international conference on Dependency Linguistics (Depling)* (pp. 73-82).
- Richard L. Lewis. 1996. Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of psycholinguistic research*, 25(1), 93-115.
- Richard L. Lewis and Shrawan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3), 375-419.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159-191.
- Michiru Makuuchi et al. 2009. Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, 106(20), 8362-8367.
- George A. Miller and Noam Chomsky. 1963. Finitary models of language users.
- Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the Eighth International Conference on Parsing Technologies*, 149-160.
- Joakim Nivre. 2006. Constraints on non-projective dependency parsing. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Daniel D. K. Sleator and Davy Temperley. 1995. Parsing English with a link grammar. *arXiv preprint cmp-lg/9508004*.
- David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: typological/cognitive universal?. *Annual Review of Linguistics*, 4, 67-80.
- Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5), 444-466.

Appendix 1. Results of distribution of dependency lengths and flux sizes ≤ 20

n	% DL = n	% DL \leq n	% FS = n	% FS \leq n
0	0	0	0	0
1	47.26	47.26	23	23
2	22.66	69.92	28.84	51.84
3	10.59	80.51	22.68	74.51
4	5.52	86.03	13.43	87.94
5	3.46	89.49	6.74	94.69
6	2.38	91.87	3.05	97.74
7	1.72	93.59	1.28	99.02
8	1.3	94.89	0.51	99.53
9	1	95.89	0.21	99.74
10	0.78	96.66	0.09	99.83
11	0.61	97.28	0.05	99.88
12	0.49	97.76	0.03	99.91
13	0.39	98.15	0.02	99.93
14	0.31	98.46	0.01	99.95
15	0.25	98.71	0.01	99.96
16	0.21	98.92	0.01	99.96
17	0.17	99.09	0.01	99.97
18	0.14	99.22	0	99.97
19	0.11	99.34	0	99.98
20	0.09	99.43	0	99.98

Appendix 2. Results of distribution of dependency lengths (DLs) and flux sizes (FSs) in 47 treebanks

	99% DLs \leq n	99% Fss \leq n	Crossing value %DL > %FS	Crossing value %FS > %DL	Crossing value cum- mul % FS > cumul % DL
UD_Ancient_Greek-PROIEL	17	8	7	2	4
UD_Ancient_Greek-Perseus	16	8	7*	2*	4
UD_Arabic-NYUAD	24	9	7	2	4
UD_Arabic-PADT	27	9	8	2	5
UD_Bulgarian-BTB	12	6	6	2	4
UD_Catalan-AnCora	19	7	7	2	4
UD_Chinese-GSD	22	9	8	2	5
UD_Croatian-SET	15	7	7	2	4
UD_Czech-CAC	15	7	6	2	4
UD_Czech-FicTree	12	6	6	2	4
UD_Czech-PDT	14	7	6	2*	4
UD_Dutch-Alpino	17	8	8	2	5
UD_English-EWT	15	7	7	2	4
UD_Estonian-EDT	12	7	6	2	3
UD_Finnish-FTB	9	6	5	2	3
UD_Finnish-TDT	12	7	6	2	3
UD_French-FTB	21	7	7	2	4
UD_French-GSD	16	6	6	2	4
UD_Galician-CTG	17	7	7	2	4
UD_German-GSD	17	8	8	2	5
UD_German-HDT	17	8	8	2	5
UD_Hebrew-HTB	18	7	6	2	4
UD_Hindi-HDTB	21	8	8	2	5
UD_Italian-ISDT	16	6	6	2	4
UD_Italian-PoSTWITA	17	7	7	2	4
UD_Italian-VIT	19	7	7	2	4
UD_Japanese-BCCWJ	26	11	8	2	5
UD_Japanese-GSD	21	7	7	2	4
UD_Korean-Kaist	14	6	6	2	4
UD_Latin-ITTB	15	7	6	2	4
UD_Latin-PROIEL	17	8	7	2	4
UD_Latvian-LVTB	13	7	6	2	4

UD_Norwegian-Bokmaal	13	6	6	2	4
UD_Norwegian-Nynorsk	14	7	6	2	4
UD_Old_French-SRCMF	12	7	6	2	4
UD_Old_Russian-TOROT	14	8	6	2	4
UD_Persian-Seraji	24	9	8	2	5
UD_Polish-PDB	13	6	6	2	3
UD_Portuguese-Bosque	17	7	7	2	4
UD_Portuguese-GSD	17	6	6	2	4
UD_Romanian-Nonstandard	15	7	6	2	4
UD_Romanian-RRT	15	6	6	2	4
UD_Russian-SynTagRus	14	6	6	2	4
UD_Slovenian-SSJ	14	7	7	2	4
UD_Spanish-AnCora	18	7	7	2	4
UD_Spanish-GSD	17	6	6	2	4
UD_Urdu-UDTB	24	9	8	2	5

* There are several crosses, but they are after the value 40 which are rather unrepresentative.

Appendix 3. Results of distribution of projective potential flux ≤ 20

n	% Projective potential flux = n	% Projective potential flux $\leq n$
0	0	0
1	14.96	14.96
2	20.27	35.23
3	19.34	54.57
4	15.85	70.42
5	11.34	81.76
6	7.53	89.29
7	4.61	93.9
8	2.7	96.6
9	1.51	98.11
10	0.83	98.94
11	0.45	99.39
12	0.25	99.64
13	0.14	99.78
14	0.08	99.86
15	0.05	99.91
16	0.03	99.94
17	0.02	99.96
18	0.01	99.97
19	0.01	99.98
20	0.01	99.99

Length of non-projective sentences: A pilot study using a Czech UD treebank

Ján Mačutek

Comenius University in Bratislava
Faculty of Mathematics, Physics and
Informatics
Department of Applied Mathematics
and Statistics
Slovakia
jmacutek@yahoo.com

Radek Čech

University of Ostrava
Faculty of Arts
Department of Czech Language
Czech Republic
cechradek@gmail.com

Jiří Milička

Charles University in Prague
Faculty of Arts
Institute of Comparative Linguistics, and
Institute of the Czech National Corpus
Czech Republic
jiri@milicka.cz

Abstract

Lengths (in words) of projective and non-projective sentences from a Czech UD dependency treebank are compared. It is shown that non-projective sentences are significantly longer (in addition, the same result was obtained in this study also for Arabic, Polish, Russian, and Slovak). The hyperpascal distribution, which was suggested as the model for frequency distribution of sentence length measured in words, fits well the data from both projective and non-projective sentences; however, its parameters attain different values for the two groups. Proportions of non-projective sentences in the treebanks used are presented, together with a discussion on factors which can influence them.

1 Introduction

Non-projectivity of syntactic dependency trees belongs to research topics which are of interest for a relatively wide spectrum of scholars. From the theoretical linguistics point of view, non-projectivity opens many questions related to the structure of natural language (e.g. Hajičová et al., 2004; Kuhlman and Nivre, 2006; Miletic and Urieli, 2017), while in the area of natural language processing it is relevant with respect to parsing (e.g. Gómez-Rodríguez and Nivre, 2013). In addition, non-projectivity can be understood as a violation of one of the dominant rules of the dependency grammar, namely, that a “dependent must appear in a sentence immediately adjacent to its head except that the two may be separated by dependent(s) of either words. This rule is applied recursively, so that if the inserted dependent has a dependent of its own, the latter may in turn be inserted between its own head and *the head’s head*” (Ninio, 2017).¹ This rule has the decisive impact on a transfer from the two-dimensional tree-structure to the linear phonetic structure and seems to be closely connected to the so-called dependency distance

¹ Strict requirements on projectivity of dependency trees appeared much earlier, see e.g. Hays (1964).

minimization, which is, in turn, related to cognitive requirements of language users (cf. Liu et al., 2017; Ninio, 2017).²

Although several papers on theoretical aspects of non-projective syntactic dependency trees were published in recent past (see e.g. Ferrer-i-Cancho, 2017; Ferrer-i-Cancho et al., 2018, and references therein), it seems that no empirical study was dedicated to properties of sentences which, according to the dependency syntax formalism, are represented by non-projective trees. For a better understanding of the phenomenon of non-projectivity, it would be useful to compare properties of projective and non-projective sentences, and to investigate their relations to properties of other language units. In this paper, which can be considered a pilot study in this area, we therefore focus on the comparison of two basic aspects.

First, sentence length (throughout the paper, measured in the number of words which the sentence consists of) in these two groups will be compared. Theoretical considerations without an empirical analysis could lead to ambiguous conclusions here. On the one hand, non-projective trees could appear more often as representations of longer sentences, because longer sentences offer more possibilities to “play” with word order, and, consequently, to display this property. On the other hand, both an increasing sentence length and the appearance of non-projectivity increase the cognitive processing difficulty of a sentence, so one cannot a priori exclude the possibility that the two phenomena could compete, and that, as a result of their competition, length of non-projective sentences would not be allowed to increase too much (although, obviously, such sentences must contain at least three words). This apparent dilemma was, however, solved already. The chance that a crossing appears in a sentence (i.e., that the sentence is non-projective) increases with the increasing mean dependency distance in the sentence (Jiang and Liu, 2015; Ferrer-i-Cancho and Gómez-Rodríguez, 2016). Next, the mean dependency length tends to increase with the increasing sentence length (Ferrer-i-Cancho and Liu, 2014; Jiang and Liu, 2015). It follows that the longer sentence, the more likely it is non-projective (this hypothesis has been corroborated also empirically by Ferrer-i-Cancho et al., 2018). Ferrer-i-Cancho (2017) provides another indirect support – in random trees, the number of crossings increases with the growing number of vertices (i.e. if words in a sentence were ordered randomly, longer sentences would, again, have a higher chance to be non-projective).

Second, we will compare the frequency distributions of sentence lengths from both groups. The question is whether the same probability distribution can serve as a model in both cases; and, if the answer is positive, whether parameters of the distribution can distinguish the two groups. It can be expected that, for projective sentences, we will be able to fit the data by a special case of the very general model derived by Wimmer and Altmann (2005); in addition to being general and thus fitting well most of linguistic data, the model has also its linguistic background and its parameters are interpretable in terms of the Zipfian equilibrium of requirements of “speaker” and “hearer” (cf. Zipf, 1949). In addition, Best (2005) already suggested some of its special cases specifically as models for sentence length, one of them for sentence length measured in the number of words. As non-projective trees can be, in a way, considered an anomaly, the model for frequency distribution of their length is much more questionable.

2 Language material and methodology

For the analysis, the Czech-PDT UD treebank is used. This treebank is based on the Prague Dependency Treebank 3.0 (Bejček et al., 2013), it consists of Czech journalistic texts from 1990s. Specifically, we used a training file named `cs_pdt-ud-train.conllu` from the Lindat Clarin repository (<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2895>). Headings, titles, indication of a place where an article was written etc., i.e. all units which do not, in fact, represent a sentence, were removed. All units of this kind share one common property, namely, the absence of punctuation (full stop, question mark, exclamation mark), in Czech-PDT UD annotation schema. Consequently, we used this feature of the annotation to identify them.

In the study, 35,213 sentences were analyzed in total. First, we determined non-projective trees as follows. In each tree, we need to find out whether there is a word whose children's edges are crossed by its parent's edge. For illustration, consider sentence (1)

² One arrives at the same conclusion - that language users prefer shorter dependency distances and thus avoid non-projective sentences - if one starts with the cognitive requirements and takes into account the least effort principle (cf. Zipf, 1949), i.e. without specific assumptions on grammar (Ferrer-i-Cancho, 2016; Ferrer-i-Cancho and Gómez-Rodríguez, 2016; Gómez-Rodríguez and Ferrer-i-Cancho, 2017).

(1) *Do Prahy měl přijet ráno*
 to Prague be supposed_{PRET 3 SG.} to come morning

‘He was supposed to come to Prague in the morning’

and examine its words one by one.

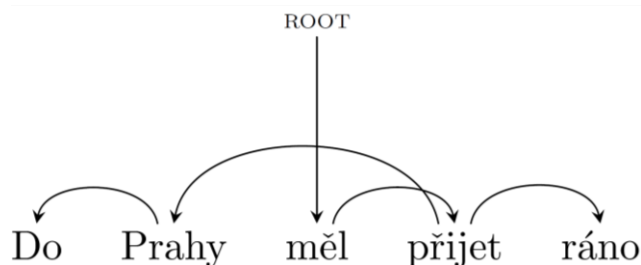


Figure 1. Syntactic relations between words in sentence (1) based on the Universal Dependencies annotation scheme. The root of the sentence is the word “*měl*”.

For each word, we look at the list of its children: does the list form an uninterrupted sequence within the sentence? If yes, the sentence is projective; if not, is the interruption caused only by the word itself (i.e., by the parent of the children under consideration)? If yes, the sentence is projective; if not, i.e. if there is also at least one other word which splits the sequence of the children, the sentence is non-projective. In sentence (1), the root, i.e. the word “*měl*”, has only one child, namely, the word “*přijet*”. Then, the word “*přijet*” has children “*Prahy*”, and “*ráno*”. The sequence of these words is interrupted not only by their parent word “*přijet*”, but also by the word “*měl*”. Thus, the sentence is non-projective (see Figure 1). This algorithm can be described by the following pseudocode (Word stands for the examined word, ID for its index, AllChildren is a zero-based sequential list of its children):

```

Word.Projectivity ← IsProjective;
d ← 0;
for i ← 1 to Word.AllChildren.Count - 1 do
  if (Word.AllChildren[0].ID + i + d ≠ Word.AllChildren[i].ID) then
    if (Word.ID = Word.AllChildren[0].ID + i + d) then
      Increment(d)
    else
      Word.Projectivity ← IsNonProjective;
  
```

The source code that was used can be found at <http://milicka.cz/kestazeni/nonprojective1.zip>, the function TWord.IsProjective is placed in the UDParser unit.

3 Results

Before we present results on sentence length, we shortly address the issue of proportions of non-projective trees in the treebank used. According to our analysis, non-projective trees form 8.04% of the sample. This proportion is smaller than findings presented by Havelka (2007, p. 614, Table 1) who reported 23.15% of non-projective trees in the Prague Dependency Treebank. Given that Havelka (2007) and this paper use the same treebank, but that the former study used the PDT annotation scheme while we use the UD annotation, the difference in results seems to be a consequence of using different annotation schemes. This topic deserves a more detailed study (e.g. comparing sentences which are projective according to one annotation scheme but non-projective according to the other).³

³ Havelka (2007, p. 614, Table 1) found the following proportions of non-projective trees: 11.16% in Arabic (out of the total of 1,460 trees), 5.38% in Bulgarian (12,823 trees), 23.15% in Czech (72,703 trees), 15.63% in Danish (5,190 trees),

Our results on sentence length confirm the findings presented earlier by Ferrer-i-Cancho et al. (2018). Non-projective sentences in the Czech treebank we used are significantly longer (with the 95% confidence interval for the mean being $\langle 21.33; 21.93 \rangle$) than projective ones (the 95% confidence interval for the mean is $\langle 16.04; 16.19 \rangle$), see also Figure 2.⁴

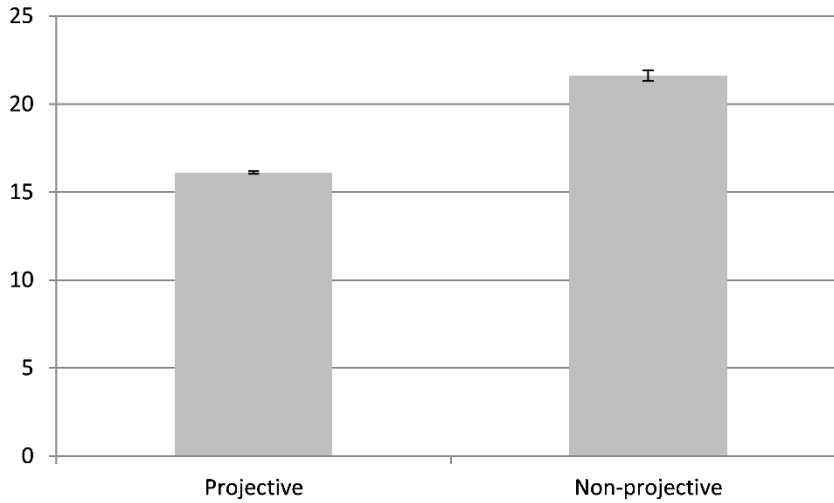


Figure 2. Length of projective and non-projective sentences in the Czech treebank (with 95% confidence intervals).

Basic descriptive statistics (which allow to formulate some – admittedly very tentative – conjectures on the comparison of lengths of projective and non-projective sentences) can be found in Table 1.

	projective	non-projective
mean	16.25	21.52
standard deviation	8.46	10.16
skewness	1.01	1.40
relative entropy	0.80	0.80

Table 1. Basic statistics on length of projective and non-projective sentences in the Czech treebank.

It is interesting that while lengths of non-projective sentences seem to be more dispersed (they achieve a higher standard deviation) and their frequency distribution more skewed, they do not differ from pro-

36.44% in Dutch (13,349 trees), 27.75% in German (39,216 trees), 5.29% in Japanese (17,044 trees), 18.94% in Portuguese (9,071 trees), 22.16% in Slovene (1,534 trees), 1.72% in Spanish (3,306 trees), 9.77% in Swedish (11,042 trees), and 11.60% in Turkish (4,997 trees). They vary quite a lot, and especially the difference between the proportions in languages so similar as Portuguese and Spanish is striking. Although we focus on the Czech treebank in this paper, we ran preliminary analyses on the proportions of non-projective sentences in several other languages using the UD annotation, with results as follows: 1.90% in Arabic (out of the total of 999 sentences), 8.04% in Czech (35,213 sentences), 0.23% in Polish (13,748 sentences), 4.81% in Russian (48,176 sentences), and 1.80% in Slovak (7817 sentences). The proportions are much lower both in Arabic and in Czech (i.e. in the two languages for which we can directly compare our results with the ones by Havelka, 2007). In addition to different annotation schemes, the differences can be caused also by the treatment of the treebanks (“[w]e take the data as is”, Havelka, 2007, p. 612, vs. our approach described in Section 2 – we removed headings, titles etc., and analyzed only proper sentences). Yet another possible source of differences cannot be neglected, namely, the sentences themselves and the text which they form. The influence of text type/genre (e.g., written vs. spoken language; or, within written texts, e.g. belletristic prose, journalistic texts, scientific papers, etc.) and author on dependency syntax (in general, including non-projectivity) is a topic which, although touched in several papers (Hollingsworth, 2012; Wang and Liu, 2017; Yan and Liu, 2017; Mehler et al., 2018; Wang and Yan, 2018), is waiting for a systematic analysis.

⁴ Non-projective sentences are significantly longer also in Arabic, Polish, Russian, and Slovak treebanks (cf. a short discussion on proportions of non-projective sentences at the beginning of Section 3). All these treebanks were processed in the same way as the Czech one, i.e. only proper sentences (as opposed to titles, headings etc.) were taken into account.

jective ones with respect to their relative entropies. Again, the question whether these observations represent a general tendency or whether they are specific for the Czech language (or even for this particular dependency syntax formalism) can be answered only after a more comprehensive analyses of this and related phenomena.

Best (2005) claims that frequencies of sentence lengths measured in words can be modelled by the hyperpascal distribution (cf. Wimmer and Altmann, 1999, pp. 279-281), with

$$P_x = \frac{\binom{k+x-1-s}{x-s}}{\binom{m+x-1-s}{x-s}} q^{x-s} P_0,$$

where $x = s, s + 1, s + 2, \dots$ are sentence lengths; s is the shift of the distribution (in this context, the length of the shortest sentence observed); k , m , and q are free parameters.⁵ He also provides a theoretical substantiation of the model. Frequencies of lengths of sentences represented by projective and non-projective trees were fitted by this distribution; however, extreme outliers (the two longest projective sentences, consisting of 78 and 162 words, and the longest non-projective sentence, with 119 words) were removed⁶ before the numerical procedures for the fit were performed. After the removal of the outliers, the longest sentence consists of 76 words in case of projective sentences and 91 words in case of non-projective ones.⁷

The results of the fit are presented in Figure 3 and Table 2. Full data (also for Arabic, Polish, Russian, and Slovak) can be found at http://www.cechradek.cz/data/2019_Macutec_et_al._Nonprojectivity_Length_proportions.zip.

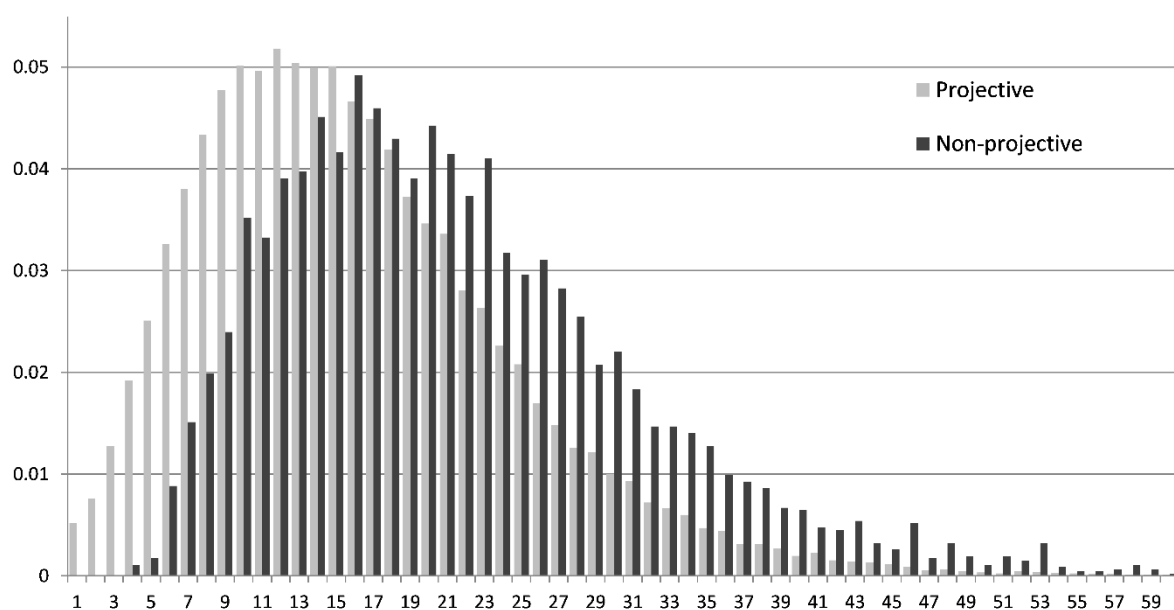


Figure 3. Relative frequencies of lengths of projective (black) and non-projective (grey) sentences in the Czech treebank.

⁵ The hyperpascal distribution has three free parameters - k , m , and q . The value of P_0 is uniquely determined by the other parameters (cf. Wimmer and Altmann, 1999, p. 280).

⁶ The usual boxplot-based rules for detection of outliers (i.e., outliers are values below $q_1 - 1.5IQR$ or above $q_3 + 1.5IQR$, with q_1 and q_3 being the first and the third quartile, respectively; $IQR = q_3 - q_1$ is the interquartile range, cf. Tukey, 1977) indicate too many outliers for highly skewed distributions such as ours. While there are more sophisticated versions of boxplot available for such data (e.g. the one suggested by Bruffaerts et al., 2014), in this paper we use, as a rule of thumb, a boxplot with much wider whiskers defined by $q_1 - 5IQR$ and $q_3 + 5IQR$.

⁷ We remind that lengths of projective and non-projective sentences were compared (see Table 1 and Figure 2) using all sentences, i.e. prior to the removal of the outliers.

It is well-known that, in terms of the p-value, the chi-square goodness of fit test rejects practically all null hypotheses if the sample size is large enough.⁸ In linguistics, it became standard to evaluate the goodness of fit of a model using the so-called discrepancy coefficient $C = \chi^2/N$, where χ^2 is the value of the test statistic in the chi-square goodness of fit test,⁹ and N is the sample size. As a rule of thumb, $C \leq 0.02$ indicates a good fit; a “more tolerant” version of the rule accepts a good fit of a model if $C \leq 0.05$ (cf. Mačutek and Wimmer, 2013, where also other possibilities how to avoid the problem of large samples are mentioned). Parameters were estimated by the minimum χ^2 method (cf. Hsiao, 2006).

	projective	non-projective
k	9.14	1.66
m	3.84	0.20
q	0.74	0.87
s	1	5
N	32379	2831
C	0.0073	0.0384

Table 2. Fitting the hyperpascal distribution to frequency distribution of length of projective and non-projective sentences in the Czech treebank.

Values of parameters k and m for projective and non-projective sentences are quite far from each other (we postpone testing and attempts to interpret both the parameter values and their differences until data from more languages are available). It means that the two frequency distributions differ in their shape, not only in the shift to the right represented by the increase of parameter s . The relatively worse (but still acceptable) fit of the hyperpascal distribution to length frequencies of non-projective sentences can be explained by their smaller number, and perhaps also by the fact that they can be considered, in a way, an anomaly, and it cannot be a priori excluded that their properties (among them their length) can differ from the “normal” (i.e. projective) sentences.

4 Conclusion and perspectives

Our results provide a further empirical corroboration of the hypothesis that non-projective sentences are longer than projective ones. Moreover, we show that frequency distribution of sentence length can be fitted by the same model in the two groups, albeit with different parameter values.

In addition to results, the paper also opens several questions. First, proportions of non-projective sentences vary not only across languages, but they depend also on the annotation scheme (such as PDT or UD), and probably on genre and author of a text as well. A systematic study, e.g. one where three out of the four “variables” under consideration (i.e., language, annotation scheme, genre, author) are fixed and the influence of the fourth one is investigated, is necessary before this problem can be at least partially solved.

Second, while word is one of reasonable units in which sentence length can be measured, it is not the only one possible – on the contrary, quantitative approaches to language modelling prefer immediate “neighbours” in the hierarchy of language units (cf. e.g. Köhler, 2012). Sentence length measured in the number of clauses could reveal other properties of non-projective sentences (and their differences from projective ones).

Third, as we suppose that “[n]o property of things or linguistic entities is isolated; each of them is in at least one relation to the other properties of the same thing, or those of other things” (Altmann, 1993; cf. also Köhler, 2005, who tries to build a general language theory which encompasses different language units, their properties and their interrelations and mutual influences), neither is sentence length. The Menzerath-Altmann law (in general cf. Cramer, 2005) predicts that longer sentences should consist of shorter clauses (cf. Köhler, 1982; Heups, 1983, Teupenhayn and Altmann, 1984; the law seems to be

⁸ Browne and Cudeck (1993) wrote that “... goodness-of-fit tests are often more a reflection on the size of the sample than on the adequacy of the model”. This problem is not specific to goodness-of-fit tests only, one encounters it whenever a statistical test with a fixed level of significance is used (cf. e.g. Kunte and Gore, 1992).

⁹ $\chi^2 = \sum_{i=s}^L \frac{(f_i - NP_i)^2}{NP_i}$, where f_i is the observed frequency of sentences with length i , NP_i is the frequency of sentences with length i predicted by the model, and L is the length of the longest sentence observed.

valid also within the dependency syntax formalism - see Mačutek et al., 2017, where results on the relation between lengths of clauses and phrases, i.e. one level lower, are presented). The question is whether the non-projective sentences “obey” this law; if yes, whether the parameters in the mathematical formulation of the law reflect the difference between them and projective sentences (we allow ourselves to formulate the hypothesis that the decrease of clause length for longer non-projective sentences will be steeper, which could compensate for their higher cognitive processing difficulty).

Acknowledgements

Supported by research projects VEGA 2/0054/18 (J. Mačutek) and ERDF CZ.02.1.01/0.0/0.0/16_019/0000734 “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (J. Milička).

References

- Gabriel Altmann. 1993. Science and linguistics. In Reinhard Köhler and Burghard B. Rieger (eds.), *Contributions to Quantitative Linguistics*, pp. 3-10. Kluwer, Dordrecht.
- Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikušlová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3>.
- Karl-Heinz Best. 2005. Satzlänge. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 298-304. de Gruyter, Berlin / New York.
- Michael W. Browne and Robert Cudeck. 1993. Alternative ways of assessing model fit. In Kenneth A. Bollen and J. Scott Long (eds.), *Testing Structural Equation Models*, pp. 136-161. SAGE, Newbury Park (CA).
- Christopher Bruffaerts, Vincenzo Verardi, and Catherine Vermandele. 2014. A generalized boxplot for skewed and heavy-tailed distributions. *Statistics and Probability Letters*, 95:110-117.
- Irene M. Cramer. 2005. Das Menzeratsche Gesetz. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 659-688. de Gruyter, Berlin / New York.
- Ramon Ferrer-i-Cancho. 2016. Non-crossing dependencies: Least effort, not grammar. In Alexander Mehler, Andy Lücking, Sven Banisch, Philippe Blanchard, and Barbara Frank-Job (eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pp. 203-234. Springer, Berlin / Heidelberg.
- Ramon Ferrer-i-Cancho. 2017. Random crossings in dependency trees. *Glottometrics*, 37:1-12.
- Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320-328.
- 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 and Haitao Liu. 2014. The risks of mixing dependency lengths from sentences of different length. *Glottotheory*, 5(2):143-155.
- 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 and Joakim Nivre. 2013. Divisible transition systems and multiplanar dependency parsing. *Computational Linguistics*, 39(4):799-845.
- Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 81:5-22.
- Jiří Havelka. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 608-615. ACL.
- David G. Hays. 1964. Dependency theory: A formalism and some observations. *Language*, 40(4):511-525.

- Gabriela Heups. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In Reinhard Köhler and Joachim Boy (eds.), *Glottometrika 5*, pp. 113-133. Brockmeyer, Bochum.
- Charles Hollingsworth. 2012. Using dependency-based annotations for authorship identification. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (eds.), *Text, Speech, and Dialogue*, pp. 314-319. Springer, Cham.
- Cheng Hsiao. 2006. Minimum chi-square. In Samuel Kotz and Norman L. Johnson (eds.), *Encyclopedia of Statistical Sciences, Vol. 7*, pp. 4812-4817. Wiley, Hoboken (NJ).
- 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 treebank. *Language Sciences*, 50:93-104.
- Reinhard Köhler. 1982. Das Menzerathsche Gesetz auf Satzebene. In Werner Lehfeldt and Udo Strauss (eds.), *Glottometrika 4*, pp. 103-113. Brockmeyer, Bochum.
- Reinhard Köhler. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 760-775. de Gruyter, Berlin / New York.
- Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. de Gruyter, Berlin / Boston.
- Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL 2006*, pp. 507-514. ACL.
- Sudhakar Kunte and Anil P. Gore. 1992. The paradox of large samples. *Current Science*, 62:393-395.
- 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.
- Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni and Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 100-107. Linköping University Electronic Press, Linköping.
- 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.
- Alexander Mehler, Wahed Hemati, Tolga Uslu, and Andy Lücking. 2018. A multidimensional model of syntactic dependency trees for authorship attribution. In Jingyang Jiang and Haitao Liu (eds.), *Quantitative Analysis of Dependency Structures*, pp. 315-347. de Gruyter, Berlin / Boston.
- Aleksandra Miletic and Assaf Urieli. 2017. Non-projectivity in Serbian: Analysis of formal and linguistic properties. In Simonetta Montemagni and Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 135-144. Linköping University Electronic Press, Linköping.
- Anat Ninio. 2017. Projectivity is the mathematical code of syntax. Comment on “Dependency distance: A new perspective on syntactic patterns in natural languages” by Haitao Liu et al. *Physics of Life Reviews*, 21:215-217.
- Regina Teupenhayn and Gabriel Altmann. 1984. Clause length and Menzerath’s law. In Joachim Boy and Reinhard Köhler (eds.), *Glottometrika 6*, pp. 127-138. Brockmeyer, Bochum.
- John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading (MA).
- Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135-147.
- Yaqin Wang and Jianwei Yan. 2018. A quantitative analysis on literary genre *essay*’s syntactic features. In Jingyang Jiang and Haitao Liu (eds.), *Quantitative Analysis of Dependency Structures*, pp. 295-314. de Gruyter, Berlin / Boston.
- Gejza Wimmer and Gabriel Altmann. 1999. *Thesaurus of Univariate Discrete Probability Distributions*. Stamm, Essen.
- Gejza Wimmer and Gabriel Altmann. 2005. Unified derivation of some linguistic laws. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 791-807. de Gruyter, Berlin / New York.
- Jianwei Yan and Siqi Liu. 2017. The distribution of dependency relations in *Great Expectations* and *Jane Eyre*. *Glottometrics*, 37:13-33.
- George K. Zipf. 1949. *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*. Addison-Wesley, Cambridge (MA).

Gradient constraints on the use of Estonian possessive reflexives

Suzanne Lesage

Université de Paris, LLF, CNRS

suzanne.lesage.broyelle@gmail.com

Olivier Bonami

Université de Paris, LLF, CNRS

olivier.bonami@univ-paris-diderot.fr

Abstract

We report on a corpus study of the use of reflexive vs. nonreflexive possessives in Estonian sentences headed by verbs taking an allative argument. We parsed the Estonian National Corpus using UDPipe trained with the Estonian Dependency Corpus, extracted relevant data automatically, eliminated false positives and annotated the data by hand. This allowed us to document effects of grammatical functions, word order and person on the choice of a reflexive vs. non-reflexive, using generalized linear mixed models. We hypothesize that the documented effects are due to the combined effects of grammatical relations, information structure, and ambiguity avoidance.

1 Introduction

Estonian allows two ways of referring to the possessor of a noun: adnominal genitive pronouns, which agree in person and number with their antecedent (1), and two reflexive forms *oma* and *enda*, which do not agree (2). In the remainder of this paper we call relevant uses of genitive pronouns NONREFLEXIVE POSSESSIVES, and relevant uses of *oma* REFLEXIVE POSSESSIVES. We leave aside *enda* for brevity.

- (1) a. Peeter vii-s mind_i minu_i vanema-te juurde.
Peeter.NOM lead-PST 1SG.PART 1SG.GEN parent-PL.GEN at
Peeter led me to my parents' place.
- b. Peeter_j vii-s Jaani_j tema_{j/*i} vanema-te juurde.
Peeter.NOM lead-PST JaanSG.PART 3SG.GEN parent-PL.GEN at
Peeter_i led Jaan_j to his_{j/*i} parents' place.
- (2) a. Ma_i vii-si-n Jaan-i_j oma_{i/*j} vanema-te juurde.
1SG.NOM lead-PST-1SG Jaan-GEN POSS.REFL parent-PL.GEN at
I led Jaan to my parents' place.
- b. Peeter_j vii-s Jaani_j oma_{i/*j} vanema-te juurde.
Peeter.NOM lead-PST JaanSG.PART 3SG.GEN parent-PL.GEN at
Peeter_i led Jaan_j to his_{i/*j} parents' place.

In canonical constructions, reflexive possessives are bound by the local subject (2), while nonreflexives can either be bound by a local non-subject, as in (1), or be locally unbound. The complementary distribution between reflexive and nonreflexive possessives collapses in some constructions, notably when the head verb has a noncanonical argument structures. Of particular interest here are bivalent verbs taking a subject and an allative argument. Most of these verbs, including those whose use is illustrated in (3), are psych verbs expressing the stimulus as a subject and the experiencer as an allative. As illustrated below, with such verbs, both reflexive and nonreflexive possessives can bind either the subject or the allative argument.

- (3) a. Mu-lle meeldi-vad kassi-d_i nende_j/oma_i iseloomu pärast.
1SG-ALL please-3PL.PRS cat-PL.NOM 3PL.GEN/REFL/POSS temper-GEN because
'I like cats because of their temper'.

- b. Lille-de-le_i sobi-b taeva-st alla sadanud vesi oma;/nende_i pehmuse
 flower-PL-ALL be.suitable-3SG.PRS sky-ELA. adown fallen water.NOM POSS.REFL /3PL
 tõttu väga hästi.
 fragility because very well
 'Rain water is suitable to flowers because of their fragility.'

(Lesage, accepted) reports the results of two psycholinguistic experiments on this noncanonical construction, showing *inter alia* that: (i) Speakers do not exhibit a categorical preference for reflexives being bound by the surface subject (resp. nonreflexives being bound by the allative argument); (ii) Binding preferences are modulated by word order, with reflexives showing a preference for an initial antecedent irrespective of its grammatical function. In the present paper, we set out to explore whether these results from comprehension experiments are confirmed in production, on the basis of a corpus study. We first train a dependency treebank on a large web corpus to help select relevant examples, which were then all validated by hand. We then annotate the examples for various syntactic and semantic properties, and run a number of logistic regression models to establish which factors influence the choice of a reflexive or nonreflexive form of the possessive. Finally we hypothesize that the observed preferences for possessive choice follow from syntactic and pragmatic constraints.

2 Data collection and annotation

The main challenge for our study is that the combination of factors we set out to investigate is too rare for data to be easily available: as is usual in languages without articles, there is no mandatory overt expression of possession in Estonian, which makes possessive forms comparatively infrequent; in addition, the construction of interest is found only with a handful of verbs.

For this reason, we relied on resources from the Universal Dependencies community to parse a large web corpus and use it for initial data selection. Specifically, we trained UDPipe (Straka and Straková, 2017) on the Estonian UD v2.4 treebank, the Universal Dependencies version of the Estonian Dependency Treebank (Muischnek et al., 2014). We then used this to parse the 1.1 billion token Estonian National Corpus (Kallas and Koppel, 2018). We relied on the morphological, POS and dependency annotation to select all sentences satisfying the following criteria:

- The sentence contains a token *v* of one of eight verbs taking an allative argument: *meeldima* 'please', *sobima* or *kõlbama* 'be suitable for', *meenuma* 'come to one's mind', *võimaldama* 'make possible', *kuuluma* 'belong', *jätkuma* 'be enough', *maitsuma* 'please by its taste'.
- The sentence contains a token *p* of a reflexive or nonreflexive possessive.
- The possessive word *p* is the possessor of some noun that has *v* on its head path – that is, the noun is a direct or indirect dependent of *v*.
- The verb *v* has an allative dependent.
- The person-number features expressed on *p*, if any, are compatible with the person-number features expressed either on the verb (if any), the subject (if it is overt), or the allative dependent.

This search allowed us to retrieve 5,593 candidate examples of a use of a possessive referring to the surface subject or allative argument of a verb in the relevant construction. We then sorted through the examples by hand to eliminate the numerous false positives due to parsing errors, other uses of the form *oma*, and/or possessives with antecedents other than the two co-arguments of the verb under examination. This narrowed down the dataset to 1,307 sentences. We then classified these examples in 5 groups as indicated in Table 1. Note that what we call the surface subject is that argument which may trigger agreement on the verb. This argument, when overtly expressed, is either in the nominative or partitive case, depending on factors orthogonal to our concerns. By design, all our examples include an allative argument,¹ whereas the subject is sometimes unexpressed. Direct objects are rare in our corpus, since only the verbs *meenutama* 'remind' and *võimaldama* 'make possible' takes a direct object. There are various interesting cases where the possessive is embedded within a direct dependent of the verb; we

¹In principle, the grammar allows for another allative dependent with the status of an adjunct, but no such case is found in our data.

Type of relation	Count
Surface subject	415
Allative argument	285
Direct object	86
Other oblique dependent	366
Embedded within a dependent	155

Table 1: Syntactic relation between the possessed noun and the head verb of the antecedent.

leave these examples aside for purposes of this paper, as they do not form a uniform class and there is not enough data for a more fine-grained classification to be informative. Hence we will focus on the 1,152 sentences corresponding to the first four row in Table 1. Since the number of possessed direct objects is low, and all cases where the possessed noun is neither the subject nor the allative argument are structurally similar, we grouped together possessed noun under ‘Direct object’ and ‘Other oblique dependent’ under a single value ‘other’, with 452 data points.

Each example was then annotated using a combination of information collected from dependency parses and manual work. We annotated the following:

- The type of possessive (reflexive or nonreflexive)
- The grammatical function of the antecedent (surface subject or allative argument).
- The grammatical function of the possessed noun.
- The person and number, and animacy of each argument.
- The volitional involvement in the event of the participant realized as the subject.
- The relative order of the two arguments and the relative order of the possessive and its antecedent.

3 Results

Following (Bresnan et al., 2007) and many other studies of binary alternatives in corpus data, we fitted mixed effects logit models to our data, using the *lme4* (Bates et al., 2014) and *lmerTest* (Kuznetsova et al., 2017) R packages. The dependent variable was the type of *possessive*. All candidate models treated the identity of the verb as a random effect. As for fixed effects, two important independent variables of interest here are the grammatical function of the possessed noun and that of the antecedent. It is important to note however that the values of these two variables are not independent: if the possessed noun is the allative argument (resp. subject), by design, the antecedent can only be the subject (resp. the allative argument); if the possessed noun is another dependent of the verb, then it can take either the subject or the allative argument as its antecedent. Because of this, we combined the two variables into one, whose values are noted $f1 \rightarrow f2$, where $f1$ is the function of the possessed noun and $f2$ that of the antecedent. Figure 1 shows the proportions of use of a possessive reflexive for each pair of grammatical functions. As the figure highlights, it is not obvious that differences between all 4 levels are statistically significant. Hence we used forward difference coding of the 4 values of that variable to be able to assess the significance of differences between adjacent levels.

We examined various combinations of this variable with other fixed effects, and report only on the best fit. The model parameters are shown in Table 2.

As the table indicates, we found a significant effect of the combined choice of a function for the possessed noun and a function for the antecedent for all pairs of adjacent conditions except *other* \rightarrow *sbj* and *all* \rightarrow *sbj*. Overall, situations where the antecedent is the surface subject favor using the reflexive form, whereas situations where the antecedent is the allative argument favor the nonreflexive. Note that the clearest effect opposes allative vs. subject antecedents, but that there is still a significant difference among examples with allative antecedents depending on the function of the possessed noun.

Non-third person antecedents comparatively disfavor using the reflexive. Contrary to our expectations, none of our various schemes for integrating an effect of animacy or volitional involvement turned out to be significant². We tried various ways of taking into account word order. A binary variable indicating

²It would have been relevant to observe the role of case marking of the subject, reflecting the definiteness among other features,

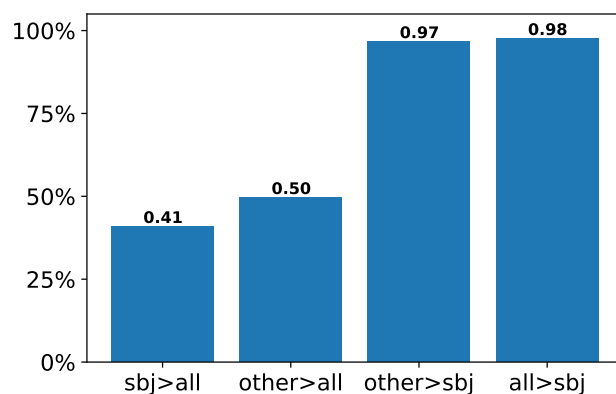


Figure 1: Proportion of use of the reflexive possessive (vs. nonreflexive possessive) for each combination of grammatical functions of the possessed noun and antecedent.

	Estimate	Std. Error	z value	p-value	
(Intercept)	0.005898	0.582356	0.010	0.991919	
sbj->all vs.other->all	-1.181075	0.274058	-4.310	1.64e-05	***
other->sbj vs.all->sbj	-2.271744	0.426328	-5.329	9.90e-08	***
other->sbj vs.all->sbj	-0.612193	1.331760	-0.460	0.645741	
person=1	-2.230623	0.296334	-7.527	5.18e-14	***
person=2	-1.308191	0.485689	-2.693	0.007071	**
order=ant_first	0.998497	0.262998	3.797	0.000147	***

Table 2: Parameters of GLMM modelling the proportion of use of a reflexive possessive.

whether the antecedent is realized before the other dependents turned out to be most relevant and lead to a significant effect: possessives that follow their antecedent are comparatively more likely to be reflexive. Finally, no significant interaction is documented for any combination of our dependent variables.

4 Discussion

Our model confirms that binding constraints on possessives are not categorical: in the constructions under examination, *oma* is often bound by a nonsubject argument, and nonreflexive possessives may (although they rarely are) be bound by the subject. We thus are in the familiar situation where one and the same constraint (reflexives tend to be bound by the subject) that is categorical in some language/some part of the grammar is gradient in another language/another part of the grammar (Bresnan et al., 2001; Sorace and Keller, 2005). More importantly, binding preferences are modulated by the dependency configuration relating the possessive and its antecedent, their relative word order, and their shared person feature.

4.1 Grammatical functions

The most obvious effect is that subject antecedents show a higher preference for reflexive possessives than allative antecedents. This is just a gradient reflex of the well-known observation that relative obliqueness constrains binding (Pollard and Sag, 1992): reflexive dependents of a verb tend to be used when they are bound by a less oblique dependent, where obliqueness may be characterized using a hierarchy such as the one in (4). In the case of reflexive possessives, we submit that relative obliqueness of the possessed noun and the antecedent likewise constrains binding of the possessive.

- (4) Subject < Direct object < Oblique argument < Adjunct

More subtle are the differences among contexts with allative antecedents (*sbj->all* vs. *other->all*). A likely reason for the observed difference has to do with how far one stands from a subject as suggested by one reviewer, but in the construction under scrutiny, the subject is mostly nominative. We found few examples with a partitive subject that we had to remove for a more homogeneous data set.

prototypical reflexive binding situation. As we observed in the introduction, Estonian possessive reflexives tend to be bound by the subject of the local clause they occur in, which is by definition the least oblique dependent of the verb. In the *sbj->a11* condition, we are departing maximally from that situation. Not only is the antecedent not a subject, it is also strongly *more oblique* than the possessed noun. Hence we have a strong expectation that a nonreflexive rather than a reflexive possessive be used. In the *other->a11* condition, the situation is different. Remember that, in this condition, the possessed noun is either a nonargument oblique or a direct object; obliques are more common, and make up 65% of the data. It follows that, in a clear majority of examples, the antecedent is more oblique than the possessed noun. Hence, on average, only the strongest expectation that the antecedent be a subject, but not the weaker expectation that it be less oblique than the possessed noun, is violated in the *other->a11* condition. Closer examination confirms that oblique possessed nouns are indeed driving the difference between the *sbj->a11* and the *other->a11* condition: whereas the proportion of reflexives is 60% for oblique possessed nouns, it drops to 22% for direct object possessed nouns.

4.2 Word order

We turn briefly to the effects of word order. As noted above, our model shows that, all other things being equal, possessives preceding their antecedents are less likely to be expressed as a reflexive than possessives that follow their antecedent.

This generalization is likely to be linked to information structure, given the tight link between word order and information structure in Estonian. The main relevant generalization here is that the dependent of a main clause verb that is realized first in linear order strongly tends to be topical (Lindström, 2005; Tael, 1988). (Bickel, 2004) suggests that, in Himalayan languages, reflexives are topic-oriented rather than subject-oriented: the reflexive tends to take the topic as its antecedent, which will coincide with the subject in most situations, but is likely not to in sentences with an experiencer expressed as an oblique. Our data supports the idea that Estonian reflexives are *both* subject-oriented and topic-oriented. While a subject antecedent favors the use of a reflexive, a topical (and hence initial) antecedent also favors such a use. Hence, where topicality and obliqueness do not align, we expect that conflicting constraints on the use of reflexives will lead to a somewhat balanced distribution of reflexives and nonreflexives.

This hypothesis helps explain the striking fact, apparent in Figure 1, that proportions of use of a reflexive reach much more extreme values when the antecedent is the subject than when it is the allative argument. Note that, unlike what happens in the canonical transitive construction, where subjects overwhelmingly precede objects, verbs with an allative argument tolerate much more easily realization of that argument before the subject (Metslang, 2013). In our data, this is true in 48% of the cases where the subject is overt. Importantly, antecedents tend to precede possessives: this is the case for 70% of allative antecedents and 84% of subject antecedents. Hence, when the antecedent is a subject, both obliqueness and topicality (as manifested in word order) favor the choice of a reflexive possessive, leading to a very high proportion of reflexives. Where the antecedent is an allative though, more often than not, obliqueness and topicality pose conflicting constraints on the choice of the possessive form: the obliqueness relation between possessive and antecedent favors a nonreflexive, while the topicality of the antecedent favors a reflexive. We conjecture that this is why, while nonreflexives are more common, reflexives are still a relevant option in most cases where the antecedent is the allative argument.

4.3 Person

We finally turn to the effect of person. As noted above, first and second person antecedents comparatively disfavor the use of a reflexive possessive. We submit that this may be due to speakers optimizing their speech for ambiguity avoidance, in accordance with Grice's maxim of manner (Grice, 1975).

To see how this plays out, let us reason first about cases in which the antecedent and possessed noun are co-arguments—that is, the *sbj->a11* and *a11->sbj* conditions. Example (5a) exhibits a situation where the antecedent is first person. In this situation, neither choice of pronoun form leads to ambiguity: reflexive *oma* has to corefer with the allative argument, as it is the only other referring expression in the local clause; but nonreflexive *minu* does not carry any ambiguity either, because it is explicitly 1st person singular. Now consider (5b). Using reflexive *oma* again does not lead to ambiguity. However,

if the speaker were to choose instead nonreflexive *tema*, this would lead to ambiguity between a local antecedent (namely the allative argument) and an extra-sentential antecedent. This line of reasoning should push a rational speaker to comparatively favor the use of a reflexive with third person antecedents as compared to first and second person antecedents.

- (5) a. Mu-lle_i meeldi-b minu/oma_i naine.
 1SG-ALL please-3SG.PRS 1SG.GEN/POSS.REFL wife.NOM
 ‘I like my wife.’
- b. Peetri-le_i meeldi-b tema_{i/j}/oma_i naine.
 Peeter-ALL please-3SG.PRS 3SG.GEN/POSS.REFL wife.NOM
 ‘Peeter likes his wife.’

Note that exactly the same reasoning is valid, *mutatis mutandis*, in the all→subj condition. Overall then, when the possessive, possessed noun and antecedent are the only three referential expressions in the clause, pragmatic reasoning predicts a higher proportion of use of the reflexive with third person antecedents. A model identical to that above but trained on only the subj→all and all→subj conditions does confirm that this prediction is borne out.

If we now turn to the remaining other→all and other→subj conditions, things are less clear, both conceptually and empirically. Conceptually, ambiguity avoidance does not make sharp predictions in such configurations. Table 3 lists all relevant configurations of person of the two co-arguments of the verb, and indicates what the ambiguity potential is depending on the choice of a possessive form; here local ambiguity is ambiguity with a clause-local antecedent, while global ambiguity is ambiguity with an extra-sentential antecedent. The clear prediction here is that, all other things being equal, reflexives should be rarest where both arguments are nonthird person (because using a nonreflexive avoids local ambiguity) and most frequent where both arguments are third person (because using a reflexive avoids global ambiguity). Where the other two situations should stand between these two extremes is unclear, in the absence of a hypothesis on the relative costs of local and global ambiguity.

Person of antecedent	Person of other arg.	Possessive type	Local ambiguity	Global ambiguity	# of observations	% of reflexives
non-3rd	non-3rd	non-reflexive	no	no	4	33%
non-3rd	non-3rd	reflexive	yes	no	2	
non-3rd	3rd	non-reflexive	no	no	43	51%
non-3rd	3rd	reflexive	yes	no	45	
3rd	non-3rd	non-reflexive	no	yes	4	95%
3rd	non-3rd	reflexive	yes	no	79	
3rd	3rd	non-reflexive	yes	yes	51	81%
3rd	3rd	reflexive	yes	no	224	

Table 3: Potential ambiguity of possessives in sentences with two candidate antecedents.

Empirical results are inconclusive. The raw counts in Table 3 clearly indicate that there is not enough data to conclude anything in the first condition, and proportions go against predictions when comparing the two last conditions. Be that as it may, a GLMM predicting possessive type on the basis of person configurations as indicated in Table 3 and grammatical function of the antecedent revealed no significant effect of person on this subset of the data. This in itself does not invalidate the idea that ambiguity avoidance constrains the choice of possessive forms: it could be that the preferences at play here are too small to be documented on a dataset of this size, or that some other factors counteract the effects of ambiguity avoidance; but it may also be that our hypothesis does not hold, and that the person effect documented in the co-argument conditions is due to some other factor.

5 Conclusion

In this paper we have used corpus evidence to explore constraints on the choice of reflexive vs. non-reflexive forms of possessives in one particular construction of Estonian. Our main empirical findings

are (i) that such these constraints are not categorical, and (ii) that separate influences of relative obliqueness, word order, and person can be documented. These results are in line with previous observations in comprehension experiments. We explored two separate but complementary lines of explanation for these findings: an interplay of grammatical relations and information structure on the one hand, and an influence of ambiguity avoidance.

On a methodological level, this paper highlights how useful the availability of dependency treebanks and parsing resources is for the linguistic study of rare syntactic phenomena in understudied languages.

References

- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Balthasar Bickel. 2004. The syntax of experiencers in the Himalayas. In Peri Bhaskararao and Karumuri V. Subbarao, editors, *Non-nominative Subjects*, volume 1, pages 77–112. John Benjamins, Amsterdam.
- Joan Bresnan, Shipra Dingare, and Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*, pages 13–32.
- Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kramer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Sciences, Amsterdam.
- Paul Grice. 1975. Logic and conversation. In Donald Davidson and Gilbert H. Harman, editors, *The logic of grammar*. Dickenson, Ensino.
- Jelena Kallas and Kristina. Koppel. 2018. Eesti keele ühendkorpus 2017.
- Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. Imertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).
- Suzanne Lesage. accepted. Liage du réfléchi possessif en estonien : une approche expérimentale. *Études finno-ougriennes*.
- Liina Lindström. 2005. *Finiitverbi asend lauses: sõnajärg ja seda mõjutavad tegurid suulises eesti keeles*, volume 16. Tartu Ülikooli kirjastus.
- Helena Metslang. 2013. Coding and behaviour of Estonian subjects. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 4(2):217–293.
- Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. 2014. Estonian dependency treebank and its annotation scheme. In *Proceedings of the 13th Workshop on Treebanks and Linguistic Theories*.
- Carl Pollard and Ivan A. Sag. 1992. Anaphors in English and the scope of Binding Theory. *Linguistic Inquiry*, pages 261–303.
- Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115(1497–1524).
- Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.
- Kaja Tael. 1988. Infostruktuur ja lauseliigendus. *Keel ja Kirjandus*, pages 133–143.

What can we learn from natural and artificial dependency trees

Chunxiao Yan

Modyco

Université Paris Nanterre

CNRS - France

yanchunxiao@yahoo.fr

Marine Courtin

LPP

Université Paris 3 Sorbonne Nouvelle

CNRS - France

marine.courtin@sorbonne-nouvelle.fr

Abstract

This paper is centered around two main contributions : the first one consists in introducing several procedures for generating random dependency trees with constraints; we later use these artificial trees to compare their properties with the properties of natural trees (i.e trees extracted from treebanks) and analyze the relationships between these properties in natural and artificial settings in order to find out which relationships are formally constrained and which are linguistically motivated. We take into consideration five metrics: tree length, height, maximum arity, mean dependency distance and mean flux weight, and also look into the distribution of local configurations of nodes. This analysis is based on UD treebanks (version 2.3, Nivre et al. 2018) for four languages: Chinese, English, French and Japanese.

1 Introduction

We are interested in looking at the linguistic constraints on syntactic dependency trees to understand what makes certain structures plausible while others are not so plausible. To effectively do this kind of work, we need to observe natural trees (syntactic trees that are the results of linguistic analysis) to see what this population looks like. Similar work has been done for example by Jiang and Liu (2015) on the relation between sentence length, dependency distance and dependency direction. But observing natural trees only has its limits : we cannot see what is special about them and their properties, and we cannot distinguish the effects of the various constraints that affect them. We can only observe the structures that are the result of all these constraints and their interactions. On the other hand, if we start from a blank canvas, randomly generated trees, and incrementally add constraints on these trees, we might be able to study one by one the effects of each constraint, and to progressively add them to get closer to natural trees. Using artificially generated trees can also be insightful to determine which constraints are formally motivated (they are a result of the mathematical structure of the tree) and which constraints are linguistically or cognitively motivated. Research in the line of Gildea and Temperley (2009) who have used random and optimal linearisations to study dependency length and its varying degrees of minimization can help us to discover constraints that would be helpful to explain why we only find a small subset of all potential trees in syntactic analyses on real data.

Our objective is therefore twofold: first we want to see how different properties of syntactic dependency trees correlate, in particular properties that are related to syntactic complexity such as height, mean dependency distance and mean flux weight, then we want to find out if these properties can allow us to distinguish between artificial dependency trees (trees that have been manipulated using random components and constraints), and dependency trees from real data.

2 Looking into the properties of syntactic dependency trees

2.1 Features

In this work we use the five following metrics to analyze the properties of dependency trees:

Feature name	Description
Length	Number of nodes in the tree
Height	Number of edges between the root and its deepest leaf
Maximum arity	Maximum number of dependents of a node
Mean Dependency Distance (MDD)	Two nodes in a dependency relation are at distance 1 if they are neighbours, distance 2 if there is a node between them etc. For every tree, we look at the mean of those dependency distances. See (Liu 2008) for more information about this property.
Mean flux weight	Mean of the number of concomitant disjoint dependencies (Kahane et al. 2017; see comments below)

Table 1: Tree-based metrics

We chose these properties because we believe that they all interfere in linearization strategies, that is how words are ordered in sentences, and the effects of those linearisation strategies. Recently, there have been many quantitative works (Futrell et al., 2015; Liu 2008) that have focussed on dependency length and its minimization across many natural languages. In complement to these linear properties we also use “flux weight”, a metrics proposed by Kahane et al. (2017) which captures the level of nestedness of a syntactic construction (the more nested the construction is, the higher its weight in terms of dependency flux). In their paper, they claim the existence of a universal upper bound for flux weight, as they have found it to be to 5 for 70 treebanks in 50 languages.

In addition to these tree-based metrics, we propose to look at local configurations using the linearised dependency trees. To look at these configurations, we extract and compare the proportion of all potential configurations of bigrams (two successive nodes) and trigrams (three successive nodes). For bigrams, we have three possible configurations: $a \rightarrow b$ which indicates that a and b are linked with a relation on the right, $a \leftarrow b$ which indicates that a and b are linked with a relation on the left, and $a \diamond b$, which indicates that a and b are not linked by a dependency. For trigram configurations (a, b, c), the possibility is much wider and we obtain 25 possible configurations. There are projective configurations like: $a \rightarrow b \rightarrow c$, $(a \rightarrow b) \& (a \rightarrow c)$, $(a \rightarrow c)$ and $(b \leftarrow c)$, but also non-projective cases like: $a \leftarrow c$ and $b \rightarrow c$.

2.2 Hypotheses

In this section, we describe some of our hypotheses concerning the relationship between our selected properties. First, we expect to find that tree length is positively correlated with other properties. As the number of nodes increases, the number of possible trees increases including more complex trees with longer dependencies (which would increase MDD) and more nestedness (which would result in a higher mean flux weight). The relationship with maximum arity is less clear, as there could be an upper limit, which would make the relation between both of these properties non-linear. We are also particularly interested in the relationship between mean dependency distance and mean flux weight. An increase in nestedness is likely to result in more descendents being placed between a governor and its direct dependents, which would mean an overall increase in mean dependency distance.

For local configurations, we know that in natural trees, most of the dependencies occur between neighbours, see for example Liu (2008), the proportion varying depending on the language. It will be interesting to see how much that is still the case in the different random treebanks, depending on the added constraints.

For trigrams of nodes we are interested in the distribution of four groups of configurations that represent four different linearization strategies: “chain” subtrees that introduce more height in the dependency tree in with both dependents in the same direction, “balanced” subtrees that alternate dependents on both sides of the governor, “zigzag” subtrees which are similar to chains but with the second dependent going in the opposite direction as the first one, and “bouquet” subtrees where the two dependents are linked to the same governor (see examples in Figure A1 in the Appendix). If one group of configurations is preferred in natural trees compared to artificial ones, it could indicate that there exists some linguistic and/or cognitive constraints that make the configuration more likely to appear. We are also interested in the hypothesis advanced by Temperley (2008) who proposes that languages that strongly favor head-initial or head-final de-

dependencies will still tend to have some short phrases depending on the opposite direction, which could constitute a way of limiting dependency distances.

3 Random tree generation with constraints

In this section we will look at random dependency tree generation with constraints. We distinguish two different steps in the dependency tree generation process : the generation of the unordered structure, and the generation of the linearisation of the nodes. Throughout this generation process, we limited ourselves to projective trees. In order to compare the properties of natural and random trees we used 3 different tree generating algorithm, to which we assign the following names : original random (1), original optimal (2) and random random (3).

The first algorithm “original random” samples an unordered dependency structure from a treebank (i.e the original structure), and generates a random projective linearisation for it:

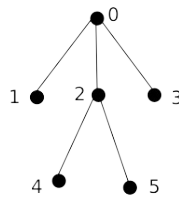


Figure 1: unordered tree

1. We start the linearisation at the root.
2. Then, we select its dependent nodes [1,2,3] and randomly order them, which gives us [2,1,3].
3. We select their direction at random, which gives us [“left”, “left”, “right”], and the linearisation steps [0], [20], [120], [1203].
4. We repeat steps 1 through 2 until every node has been linearized, which gives us (for example) [124503].

The second algorithm “original optimal” also samples an unordered dependency structure from a treebank, but instead of generating a simple projective linearisation, we add a second constraint to minimize dependency distances inside the linearised dependency tree. The idea comes from Temperley (2008): to minimize dependency distances in a projective setting, dependents of a governor should be linearized alternately on opposing sides of the governor, with the smallest dependent nodes (i.e those that have the smallest number of direct and indirect descendants) linearized first. Using the same structure unordered tree as in fig. 1 we described the procedure below:

1. We start the linearisation at the root.
2. Then, we select its dependent node [1,2,3] and order them in order of their decreasing number of descendant nodes, which gives us [1,3,2].
3. We select a first direction at random, for example “left”, and order these nodes alternating between left and right, which gives us these linearisation steps [0], [10], [103], [2103].
4. We repeat steps 1 through 2 until every node has been linearized, which gives us for example [425103].

The third algorithm “random random” is the only one to implement two random steps : first generate a completely random structure, then linearize it following the same procedure as in algorithm 1). The unordered structure generation step is described in fig 2.

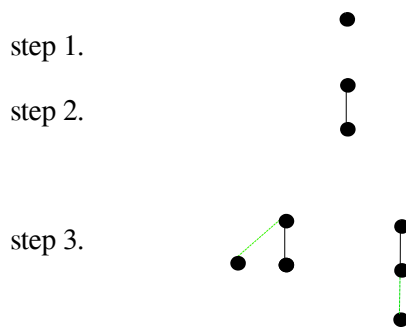


Figure 2: Random tree generation

1. We start the generation process with a single node
2. We introduce a new node and randomly draw its governor. For now, since there is only one potential governor, the edge has a probability of 1.
3. We introduce a new node and randomly draw its governor. There are two potential governors which gives us a probability 0.5 of drawing the node 0 and the same probability for the node 1. These potential edges are drawn in green on the graph.
4. We repeat this last step until all nodes have been drawn and attached to their governor.¹

These tree generation algorithms are only some of the many possible algorithms that could be implemented, but they give us tools to analyze how different generation strategies will affect the properties of the generated trees, as we incorporate more and more constraints into the two generation steps. They are also easily extensible, for example during the linearisation process we could introduce a probability of creating a head-final edge, to produce trees that resemble more the trees of a head-final language like Japanese. For the unordered structure generation, we could introduce a constraint to limit length, arity or height. We need to distinguish constraints that happen during the unordered structure generation step and constraints that have to do with linearisation, like constraints on dependency distances and on flux weights.

One question that still remains concerns the ordering of the two steps : unordered structure generation and linearisation generation. So far we have only implemented the full generation starting with the generation of the unordered structure and then moving on to the linearisation, which is a synthesis approach as described in Meaning-Text-Theory (Mel'čuk 1998), but it would be interesting to go in the analysis direction, starting with a sequence of nodes, and then randomly producing a structure for it. This could allow us to see how generation algorithms impact the distribution of trees, especially as we add constraints into the generation. We could then see if one type of random generation (synthesis vs analysis) produces structures that resemble natural dependency structures more, or if they introduce biases towards some types of structures.

4 Results and discussion

4.1 Correlation between properties

For each pair of properties presented in section 2.1 we measured the pearson correlation coefficient to find out the extent to which the relationship between these variables can be linearly captured. We looked into these results for the different natural treebanks (“original”) and the artificial ones (“original random”, “random random” and “original optimal”). Tables presenting the full results are showcased in tables 1-4 in appendix, with rankings for the correlation between parentheses.

Based on these results, we notice that mean dependency distance and mean flux weight are overall the most correlated properties with values ranging from 0.70 (jp_pud, “original”) to 0.95 (fr_partut, “original optimal”). This can be explained by the fact that mean flux weight increases as the number of disjoint flux increases, which in turn tends to create longer dependencies than structure with few disjoint flux. An interesting observation about this correlation is that it is intensified in all the artificial treebanks, and is the strongest in the “original optimal” version. Introducing a dependency distance minimization constraint will

¹ Note that this algorithm gives us a uniform probability on derivations, but that some derived trees are more probable than others, for example if the length of the tree is 4 we only have 1 derivation to obtain a tree of height 4, and 2 derivations to obtain a tree with 2 dependent on the root and 1 on one of these dependents.

favour shorter dependencies, which provides less opportunities for configurations that introduce disjoint flux. Therefore the mean flux weight will also decrease.

If we look at the correlation between length and height, we find that it is strong in original structures (0.78 correlation) as well as in the random ones (0.71 correlation in “random random”, which is the only format in which the height of the tree is affected by the manipulation). This means that the relationship between these two properties is not motivated by linguistic factors only. From a mathematical point of view, longer sentences have the potential to introduce more hierarchy which increases the height. Thus, there is a correlation between these two properties regardless of whether the structure is natural or random. Zhang and Liu (2018) have proposed that the relationship between these two properties in natural treebanks of English and Chinese can be described by a power law. Further examination could tell us if it is also the case for randomly generated trees, or if the relationship is better modelled by another type of function.

We also find quite strong correlations between mean dependency distance and height in the artificial treebanks (0.76, 0.79, 0.72 respectively for “original random”, “original optimal” and “random random”) while this correlation is less important for the natural treebanks (0.46). It is quite interesting that the correlation decreases in the original trees. Our interpretation is that perhaps there is a more complex relationship at play between height and mean dependency distance in real data that cannot be linearly captured, and this complex relationship would be altered by the random components when generating the various artificial trees, especially as we relinearize the nodes.

4.2 Distribution of configurations

In this section we look at the distribution of local syntactic configurations by extracting trigrams and looking at their dependency relations. First we look at the non-linearized configurations : $a \rightarrow b \rightarrow c$ and $b \leftarrow a \rightarrow c$, to analyze the differences in local structures between natural and randomly generated trees. Then we analyze the distribution of the four different groups presented in section 2.2, and how this distribution is impacted by language and the type of treebank (natural and artificial). We will discuss here a few points and present the full results in appendix.

In fig 3, we can see the distribution of non-linearized configurations for one example language, French. For the “random random” trees, we have 45% of $b \leftarrow a \rightarrow c$ configurations and 55% of $a \rightarrow b \rightarrow c$ configurations. For all other treebank types, the first configuration is by far the most frequent one. This will likely have some repercussions in the distribution of linearized configurations.

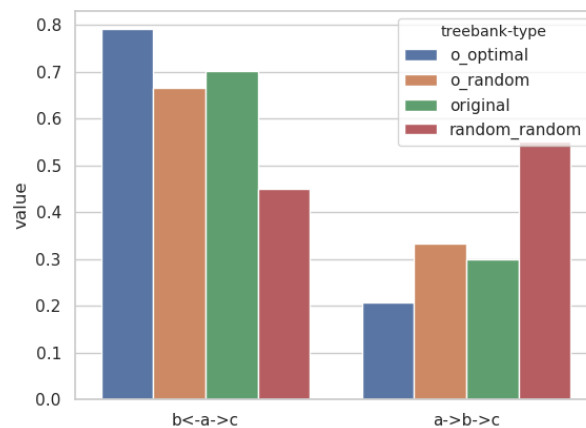


Figure 3: Non-linearized trigram configurations distribution for French

We also observe that the results are fairly similar across all 4 languages, with “original optimal” showing the most unequal distribution (80%-20% respectively for $b \leftarrow a \rightarrow c$ and $a \rightarrow b \rightarrow c$ configurations), followed by “original” and “original random” (around 60%-40%, although there is some variation depending on the language). One possible explanation for favouring $b \leftarrow a \rightarrow c$ could be that it helps minimizing dependency distances, since it can lead to “balanced” configurations which are the optimal way to arrange dependents without introducing longer dependencies. If that is the case, we will see a high proportion of “balanced” configurations when we look more in detail at how these configurations are linearized. Another line of explanation could be that having too many $a \rightarrow b \rightarrow c$ configurations introduces too much height in the trees, which could be a factor of complexity that natural languages try to avoid whenever possible. Differences

between “original optimal”, “original random” and “original” can be explained by the linearization process: the optimal trees tend to favour shorter dependencies, which means that a higher percentage of triplets of nodes will all be connex, while non-optimal trees will sometimes linearize the nodes further away, thus excluding them from the extraction of triplets. It would be interesting to see if the distribution is similar when we look at all configurations of triplets and not just at local ones.

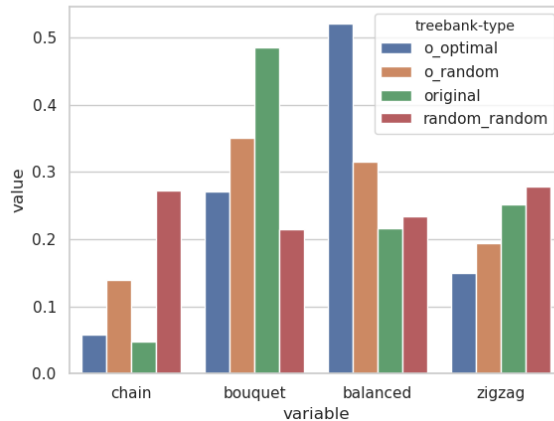


Figure 4: Trigram configurations distribution for French

We then go on to look at these configurations once they have been subdivided according to the classification proposed in section 2.2. Note that the configurations “bouquet” and “balanced” are a result of the $b \leftarrow a \rightarrow c$ configurations and that $a \rightarrow b \rightarrow c$ will produce either “chain” or “zigzag”. We show the distribution for French in fig 4. First we comment the results that are stable across languages: “random random” trees have a slight preference for “chain” and “zigzag” as a result of the preference for $b \leftarrow a \rightarrow c$ configurations, but inside each group (“chain” and “zigzag” / “bouquet” and “balanced”) the distribution is equally divided. The “original optimal” trees have a very marked preference for “balanced” which is to be expected because alternatively ordering dependents of a governor is the preferred strategy to minimize dependency length. Next we find “zigzag” configurations, followed by “bouquet” and very few “chain”. Contrary to the potential explanation we advanced for the high frequency of $b \leftarrow a \rightarrow c$ configurations, “balanced” configurations are not particularly frequent in the original trees (23% in Chinese, 14% in English, 21% in French and 27% in Japanese), especially when compared to the “bouquet” configurations (37%, 52%, 48%, 30% respectively). Bouquet configurations are much more frequent in the natural trees than in the artificial ones. We have yet to find a satisfactory explanation for this. Even if we know that some arbitrary choices in the UD annotation scheme inflate the percentage of bouquet (*conj*, *fixed* and *flat* relations are always encoded as a bouquet), this does not seem sufficient to explain the difference with the other configurations. We also remark that, if we were to use a schema with functional heads most of these “bouquet” configurations would become “zigzagz” or “chain”, so we could potentially find an explanation by investigating there. For the optimal model, the bouquet is not an optimal strategy to minimize dependency distances, so the bouquet configuration will, of course, be less critical in the optimal model.

Compared to the other languages, Japanese has an interestingly high percentage of “zigzag” configurations. This can be partly explained by the segmentation used in the Japanese treebanks. The particles and agglutinated markers (for polarity, aspect, politeness...) have been annotated as separate tokens, which often creates many dependents on a single governor. A lot of these dependencies fall outside the trigram windows and are excluded from our analysis. Japanese being a head-final language, the configurations captured will often contain a head-final dependency (*obj*, *acl*, *nmod*...) and a marker of the dependent, which means that it will often fall into the “zigzag” bin. Nonetheless “bouquet” are still quite frequent as a governor often has several marks, and “balanced” capture nominal modifiers or compounds, and their case or topic marker.

5 Conclusion

In this paper we introduced several ways to generate artificial syntactic dependency trees and proposed to use those trees as a way of looking into the structural and linguistic constraints on syntactic structures for 4 different languages. We propose to incrementally add constraints on these artificial trees to observe the ef-

facts these constraints produce and how they interact with each other. We limited ourselves to generating projective trees, which we now realize was a very strong constraint that strongly restricts the types of structures available, and therefore the variations of the different observed properties, and think that it would be interesting to also look at the result when allowing non-projective edges.

To expand on this work we would also like to see how the observed properties and the relations between them are affected by the annotation scheme, in particular contrasting schemas where content words are governors (as is the case in UD) and schemas where function words are governors (for example using the SUD schema proposed by Gerdes et al. (2018)), as it will have an impact on height, dependency distances, and the types of configurations that can be extracted from the treebanks.

In the present paper, we have looked at local syntactic configurations through the extraction of sequences of nodes (pairs and triplets). However these configurations are not representative of all configurations inside the trees, as some syntactic relations are more likely to appear in more global configurations. In the future, we plan on looking at these larger configurations by extracting subtrees and analyzing their distribution. We also intend on digging deeper into the analysis of the present data, and propose predictive models that could help us clarify the relationship (whether they be linear or not) between the different features in order to build a more solid basis to verify our hypotheses and propose explanations for the observations we made.

References

- 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. <https://doi.org/10.1073/pnas.1502134112>
- Daniel Gildea and David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2), 286–310. <https://doi.org/10.1111/j.1551-6709.2009.01073.x>
- Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2), 159–191. <https://doi.org/10.17791/jcs.2008.9.2.159>
- 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 (2015: 93-104).
- 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 (Depling2017)* (pp.73-82).
- Igor Mel'čuk. 1998. *Dependency syntax: Theory and Practice*. SUNY press.
- Joakim Nivre, Mitchell Abrams, Željko Agić et al. 2018. Universal Dependencies 2.3, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-2895>.
- David Temperley. 2008. Dependency-length minimization in natural and artificial languages*. *Journal of Quantitative Linguistics*, 15(3), 256–282. <https://doi.org/10.1080/09296170802159512>
- Hongxin Zhang and Haitao Liu. 2018. Interrelations among Dependency Tree Widths, Heights and Sentence Lengths. In: Jingyang Jiang & Haitao Liu (eds.). *Quantitative Analysis of Dependency Structures*, Berlin/Boston: DE GRUYTER MOUTON. pp. 31-52.

Appendix

original treebanks

	height_length	arity_length	arity_height	mdd_length	mdd_height	mdd_arity	mfw_length	mfw_height	mfw_arity	mfw_mdd
en_lines	0.77 (1)	0.62 (6)	0.37 (10)	0.68 (4)	0.44 (8)	0.72 (3)	0.65 (5)	0.49 (7)	0.42 (9)	0.77 (2)
en_gum	0.8 (2)	0.68 (4)	0.5 (10)	0.68 (5)	0.53 (9)	0.79 (3)	0.64 (6)	0.57 (7)	0.56 (8)	0.83 (1)
en_esl	0.71 (2)	0.57 (6)	0.2 (10)	0.62 (5)	0.23 (9)	0.65 (3)	0.62 (4)	0.33 (7)	0.32 (8)	0.72 (1)
en_partut	0.76 (2)	0.55 (6)	0.3 (10)	0.57 (5)	0.31 (9)	0.6 (3)	0.58 (4)	0.37 (7)	0.33 (8)	0.78 (1)
en_ewt	0.82 (3)	0.72 (4)	0.61 (10)	0.7 (5)	0.64 (9)	0.85 (2)	0.65 (8)	0.66 (7)	0.68 (6)	0.87 (1)
en_pud	0.65 (2)	0.44 (6)	0.09 (9)	0.49 (3)	0.08 (10)	0.48 (5)	0.48 (4)	0.19 (7)	0.13 (8)	0.72 (1)
fr_gsd	0.73 (2)	0.52 (6)	0.22 (10)	0.61 (5)	0.28 (9)	0.65 (3)	0.63 (4)	0.37 (7)	0.32 (8)	0.74 (1)
fr_sequoia	0.81 (3)	0.69 (4)	0.56 (9)	0.68 (5)	0.54 (10)	0.82 (2)	0.67 (6)	0.59 (8)	0.63 (7)	0.83 (1)
fr_spoken	0.84 (1)	0.59 (6)	0.42 (10)	0.61 (5)	0.44 (9)	0.78 (2)	0.67 (4)	0.56 (7)	0.46 (8)	0.71 (3)
fr_partut	0.79 (1)	0.54 (6)	0.35 (10)	0.61 (5)	0.37 (8)	0.67 (3)	0.65 (4)	0.47 (7)	0.37 (9)	0.77 (2)
fr_pud	0.64 (2)	0.47 (6)	0.1 (10)	0.58 (3)	0.17 (9)	0.54 (5)	0.56 (4)	0.29 (7)	0.21 (8)	0.77 (1)
zh_cfl	0.76 (2)	0.61 (5)	0.37 (10)	0.58 (7)	0.52 (8)	0.63 (4)	0.6 (6)	0.66 (3)	0.42 (9)	0.78 (1)
zh_gsd	0.58 (6)	0.56 (7)	0.14 (10)	0.61 (4)	0.39 (8)	0.65 (2)	0.63 (3)	0.6 (5)	0.27 (9)	0.74 (1)
zh_pud	0.57 (3)	0.53 (6)	0.07 (10)	0.56 (5)	0.41 (8)	0.52 (7)	0.56 (4)	0.59 (2)	0.19 (9)	0.77 (1)
zh_hk	0.83 (2)	0.73 (6)	0.56 (10)	0.79 (4)	0.72 (7)	0.79 (3)	0.69 (8)	0.74 (5)	0.61 (9)	0.86 (1)
jp_pud	0.58 (4)	0.47 (6)	0.0 (10)	0.59 (3)	0.13 (9)	0.59 (2)	0.54 (5)	0.36 (7)	0.17 (8)	0.7 (1)
jp_gsd	0.74 (2)	0.61 (6)	0.31 (10)	0.69 (4)	0.37 (9)	0.7 (3)	0.66 (5)	0.48 (7)	0.4 (8)	0.8 (1)
jp_modern	0.85 (1)	0.62 (4)	0.46 (9)	0.58 (6)	0.44 (10)	0.72 (3)	0.61 (5)	0.58 (7)	0.5 (8)	0.81 (2)

original_optimal

	height_length	arity_length	arity_height	mdd_length	mdd_height	mdd_arity	mfw_length	mfw_height	mfw_arity	mfw_mdd
jp_gsd	0.74 (4)	0.61 (7)	0.31 (9)	0.76 (3)	0.74 (5)	0.57 (8)	0.68 (6)	0.79 (2)	0.26 (10)	0.89 (1)
jp_pud	0.58 (5)	0.47 (7)	0.01 (10)	0.64 (3)	0.59 (4)	0.36 (8)	0.57 (6)	0.71 (2)	0.05 (9)	0.88 (1)
jp_modern	0.85 (4)	0.62 (7)	0.46 (9)	0.81 (5)	0.86 (3)	0.6 (8)	0.79 (6)	0.88 (2)	0.41 (10)	0.94 (1)
en_esl	0.71 (4)	0.57 (7)	0.2 (9)	0.73 (3)	0.7 (5)	0.48 (8)	0.63 (6)	0.75 (2)	0.16 (10)	0.89 (1)
en_ewt	0.82 (4)	0.72 (7)	0.61 (10)	0.8 (6)	0.84 (2)	0.8 (5)	0.69 (8)	0.82 (3)	0.63 (9)	0.93 (1)
en_partut	0.76 (4)	0.55 (7)	0.3 (9)	0.76 (5)	0.77 (3)	0.47 (8)	0.71 (6)	0.8 (2)	0.27 (10)	0.94 (1)
en_lines	0.77 (4)	0.62 (8)	0.37 (9)	0.79 (3)	0.77 (5)	0.63 (7)	0.72 (6)	0.8 (2)	0.35 (10)	0.9 (1)
en_gum	0.8 (4)	0.68 (8)	0.5 (9)	0.79 (5)	0.81 (3)	0.7 (7)	0.7 (6)	0.81 (2)	0.49 (10)	0.92 (1)
en_pud	0.65 (4)	0.44 (7)	0.08 (9)	0.68 (3)	0.62 (5)	0.35 (8)	0.61 (6)	0.69 (2)	0.07 (10)	0.89 (1)
zh_gsd	0.58 (5)	0.56 (6)	0.13 (10)	0.72 (2)	0.55 (7)	0.53 (8)	0.65 (4)	0.65 (3)	0.21 (9)	0.86 (1)
zh_cfl	0.76 (2)	0.61 (8)	0.38 (10)	0.75 (4)	0.69 (6)	0.68 (7)	0.7 (5)	0.75 (3)	0.39 (9)	0.87 (1)
zh_hk	0.83 (2)	0.73 (6)	0.57 (9)	0.81 (4)	0.79 (5)	0.81 (3)	0.62 (8)	0.73 (7)	0.55 (10)	0.87 (1)
zh_pud	0.58 (5)	0.53 (7)	0.08 (9)	0.65 (3)	0.6 (4)	0.39 (8)	0.54 (6)	0.68 (2)	0.07 (10)	0.86 (1)
fr_sequoia	0.81 (4)	0.69 (8)	0.56 (10)	0.8 (5)	0.83 (2)	0.72 (7)	0.73 (6)	0.83 (3)	0.56 (9)	0.94 (1)
fr_gsd	0.73 (4)	0.52 (7)	0.22 (10)	0.74 (3)	0.73 (5)	0.44 (8)	0.69 (6)	0.78 (2)	0.22 (9)	0.92 (1)
fr_partut	0.78 (4)	0.54 (7)	0.35 (9)	0.75 (5)	0.81 (3)	0.47 (8)	0.71 (6)	0.82 (2)	0.31 (10)	0.95 (1)
fr_spoken	0.84 (3)	0.59 (8)	0.42 (9)	0.82 (4)	0.81 (5)	0.67 (7)	0.77 (6)	0.84 (2)	0.32 (10)	0.88 (1)
fr_pud	0.64 (5)	0.47 (7)	0.1 (10)	0.68 (3)	0.65 (4)	0.36 (8)	0.63 (6)	0.72 (2)	0.14 (9)	0.92 (1)

original_random

	height_length	arity_length	arity_height	mdd_length	mdd_height	mdd_arity	mfw_length	mfw_height	mfw_arity	mfw_mdd
fr_pud	0.64 (3)	0.47 (7)	0.1 (10)	0.63 (4)	0.62 (5)	0.37 (8)	0.61 (6)	0.71 (2)	0.2 (9)	0.85 (1)
fr_spoken	0.84 (3)	0.59 (8)	0.42 (9)	0.82 (4)	0.8 (6)	0.6 (7)	0.81 (5)	0.86 (2)	0.4 (10)	0.9 (1)
fr_partut	0.79 (4)	0.54 (8)	0.36 (10)	0.78 (5)	0.79 (3)	0.55 (7)	0.76 (6)	0.85 (2)	0.41 (9)	0.92 (1)
fr_gsd	0.73 (3)	0.52 (7)	0.22 (10)	0.71 (4)	0.69 (6)	0.45 (8)	0.7 (5)	0.78 (2)	0.26 (9)	0.88 (1)
fr_sequoia	0.81 (4)	0.69 (8)	0.56 (10)	0.81 (5)	0.82 (3)	0.69 (7)	0.78 (6)	0.86 (2)	0.59 (9)	0.93 (1)
jp_gsd	0.74 (4)	0.61 (7)	0.31 (10)	0.75 (3)	0.71 (6)	0.57 (8)	0.73 (5)	0.8 (2)	0.37 (9)	0.87 (1)
jp_modern	0.85 (4)	0.62 (7)	0.46 (10)	0.85 (3)	0.83 (6)	0.6 (8)	0.84 (5)	0.87 (2)	0.47 (9)	0.93 (1)
jp_pud	0.58 (4)	0.47 (7)	0.0 (10)	0.6 (3)	0.54 (6)	0.39 (8)	0.57 (5)	0.71 (2)	0.07 (9)	0.78 (1)
en_esl	0.71 (3)	0.57 (7)	0.2 (10)	0.69 (4)	0.64 (6)	0.44 (8)	0.65 (5)	0.74 (2)	0.23 (9)	0.85 (1)
en_pud	0.65 (3)	0.44 (7)	0.08 (10)	0.6 (5)	0.62 (4)	0.33 (8)	0.6 (6)	0.71 (2)	0.13 (9)	0.85 (1)
en_partut	0.76 (3)	0.55 (7)	0.3 (10)	0.73 (6)	0.74 (4)	0.48 (8)	0.73 (5)	0.81 (2)	0.33 (9)	0.9 (1)
en_gum	0.8 (3)	0.68 (7)	0.5 (10)	0.79 (4)	0.77 (5)	0.67 (8)	0.76 (6)	0.82 (2)	0.55 (9)	0.9 (1)
en_ewt	0.82 (4)	0.72 (8)	0.61 (10)	0.81 (5)	0.82 (3)	0.76 (7)	0.77 (6)	0.86 (2)	0.67 (9)	0.92 (1)
en_lines	0.77 (4)	0.62 (7)	0.37 (10)	0.77 (3)	0.74 (6)	0.58 (8)	0.75 (5)	0.8 (2)	0.42 (9)	0.9 (1)
zh_gsd	0.58 (5)	0.56 (6)	0.13 (10)	0.66 (2)	0.48 (8)	0.52 (7)	0.64 (3)	0.63 (4)	0.23 (9)	0.8 (1)
zh_hk	0.83 (2)	0.73 (8)	0.56 (10)	0.82 (3)	0.79 (5)	0.75 (6)	0.74 (7)	0.8 (4)	0.6 (9)	0.89 (1)
zh_cfl	0.76 (3)	0.61 (8)	0.37 (10)	0.75 (4)	0.67 (6)	0.63 (7)	0.71 (5)	0.76 (2)	0.42 (9)	0.85 (1)
zh_pud	0.57 (5)	0.53 (6)	0.07 (10)	0.63 (2)	0.47 (7)	0.45 (8)	0.59 (4)	0.62 (3)	0.18 (9)	0.81 (1)

random_random

	height_length	arity_length	arity_height	mdd_length	mdd_height	mdd_arity	mfw_length	mfw_height	mfw_arity	mfw_mdd
zh_gsd	0.61 (5)	0.53 (7)	0.14 (10)	0.65 (3)	0.55 (6)	0.42 (8)	0.64 (4)	0.65 (2)	0.23 (9)	0.85 (1)
zh_hk	0.8 (3)	0.75 (7)	0.58 (10)	0.8 (4)	0.79 (5)	0.73 (8)	0.75 (6)	0.81 (2)	0.63 (9)	0.91 (1)
zh_pud	0.57 (5)	0.54 (6)	0.14 (10)	0.62 (3)	0.54 (7)	0.44 (8)	0.61 (4)	0.62 (2)	0.25 (9)	0.85 (1)
zh_cfl	0.72 (5)	0.6 (8)	0.37 (10)	0.77 (2)	0.68 (6)	0.64 (7)	0.75 (4)	0.76 (3)	0.47 (9)	0.88 (1)
fr_pud	0.57 (4)	0.51 (7)	0.11 (10)	0.59 (3)	0.52 (6)	0.41 (8)	0.57 (5)	0.61 (2)	0.19 (9)	0.83 (1)
fr_partut	0.68 (6)	0.65 (7)	0.42 (10)	0.72 (3)	0.68 (5)	0.61 (8)	0.7 (4)	0.75 (2)	0.48 (9)	0.89 (1)
fr_spoken	0.75 (5)	0.68 (7)	0.49 (10)	0.76 (3)	0.74 (6)	0.66 (8)	0.76 (4)	0.79 (2)	0.52 (9)	0.9 (1)
fr_sequoia	0.75 (5)	0.71 (8)	0.6 (10)	0.77 (4)	0.79 (3)	0.73 (7)	0.75 (6)	0.83 (2)	0.63 (9)	0.92 (1)
fr_gsd	0.63 (5)	0.58 (7)	0.23 (10)	0.68 (3)	0.59 (6)	0.5 (8)	0.66 (4)	0.68 (2)	0.31 (9)	0.86 (1)
en_lines	0.71 (5)	0.65 (7)	0.4 (10)	0.74 (3)	0.68 (6)	0.6 (8)	0.72 (4)	0.75 (2)	0.45 (9)	0.89 (1)
en_pud	0.58 (5)	0.52 (6)	0.08 (10)	0.61 (3)	0.51 (7)	0.4 (8)	0.6 (4)	0.64 (2)	0.18 (9)	0.82 (1)
en_partut	0.65 (5)	0.59 (7)	0.28 (10)	0.66 (3)	0.61 (6)	0.52 (8)	0.66 (4)	0.72 (2)	0.36 (9)	0.86 (1)
en_ewt	0.77 (6)	0.75 (7)	0.66 (10)	0.79 (4)	0.82 (3)	0.77 (5)	0.74 (8)	0.85 (2)	0.7 (9)	0.93 (1)
en_esl	0.63 (5)	0.57 (6)	0.17 (10)	0.67 (2)	0.57 (7)	0.47 (8)	0.65 (4)	0.66 (3)	0.27 (9)	0.85 (1)
en_gum	0.73 (6)	0.71 (7)	0.53 (10)	0.77 (3)	0.75 (4)	0.69 (8)	0.74 (5)	0.8 (2)	0.59 (9)	0.91 (1)
jp_gsd	0.69 (5)	0.64 (7)	0.37 (10)	0.73 (3)	0.66 (6)	0.59 (8)	0.72 (4)	0.74 (2)	0.44 (9)	0.89 (1)
jp_pud	0.52 (5)	0.5 (6)	0.06 (10)	0.56 (3)	0.48 (7)	0.36 (8)	0.53 (4)	0.6 (2)	0.13 (9)	0.82 (1)
jp_modern	0.7 (5)	0.69 (7)	0.46 (10)	0.72 (3)	0.69 (6)	0.67 (8)	0.7 (4)	0.78 (2)	0.54 (9)	0.9 (1)

Examples of 4 types of trigram configurations

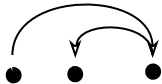
1. Balanced



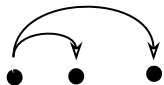
2. Chain



3. Zigzag

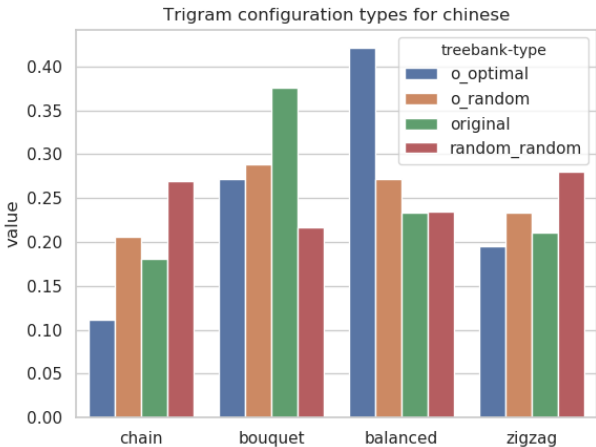
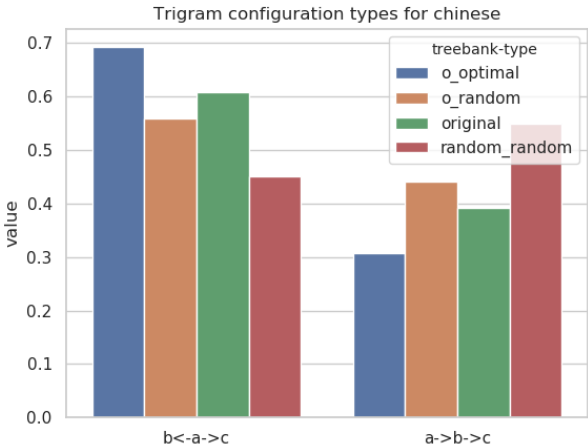


4. Bouquet

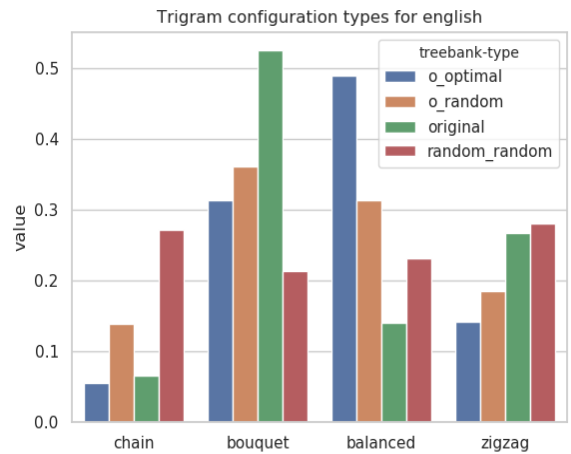
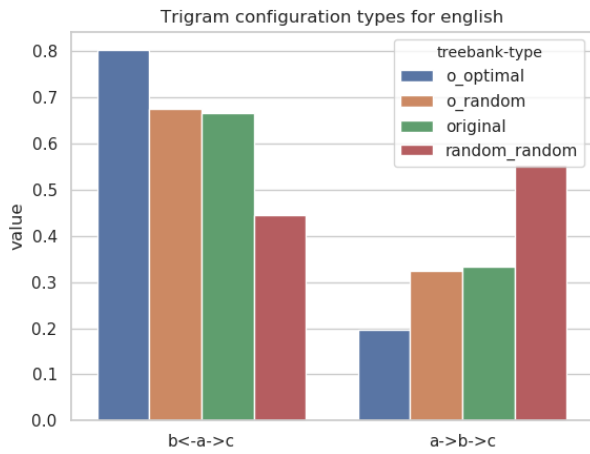


Trigrams configurations by type

Chinese



English



Japanese

