Depling 2019

Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)

Proceedings

27–28 August, 2019
held within the **SyntaxFest** 2019, 26–30 August
Paris, France

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 978-1-950737-63-5

Preface

The Fifth edition of the International Conference on Dependency Linguistics (Depling) follows a biannual series that started in 2011, in Barcelona and continued in Prague (2013), Uppsala (2015), and Pisa (2017). The series responds to the growing need for linguistic meetings dedicated to approaches in syntax, semantics and the lexicon that are centered around dependency structures as a central linguistic notion.

This year's edition is special as Depling is 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 lead 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)
Çagrı 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 Articial Intelligence)
Relja Vulanovic (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 - Inductive biases and language emergence in communicative agents 1
Emmanuel Dupoux

SyntaxFest 2019 Invited talk - Transferring NLP models across languages and domains 2
Barbara Plank

Syntactic dependencies correspond to word pairs with high mutual information 3
Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko and Idan Blank

Reflexives in Czech from a Dependency Perspective 14
Vaclava Kettnerova and Marketa Lopatkova

Coordination of Unlike Grammatical Functions 26
Agnieszka Patejuk and Adam Przepiórkowski

Predicate Catenae: A Dependency Grammar Analysis of It-Clefts 38
Timothy Osborne

Noun Phrases Rooted by Adjectives: A Dependency Grammar Analysis of the Big Mess Construction 49
Timothy Osborne

Cliticization of Serbian Personal Pronouns and Auxiliary Verbs. A Dependency-Based Account . 60
Jasmina Milicevic

The evolution of spatial rationales in Tesnière's stemmas 69
Nicolas Mazziotta

Toward a cognitive dependency grammar of Hungarian 81
András Imrényi

Interpreting and defining connections in a dependency structure 89
Sylvain Kahane

Identifying Grammar Rules for Language Education with Dependency Parsing in German 100
Eleni Metheniti, Pomi Park, Kristina Kolesova and Günter Neumann

How to Parse Low-Resource Languages: Cross-Lingual Parsing, Target Language Annotation, or Both? 112
Ailsa Meechan-Maddon and Joakim Nivre

Word order variation in Mbyá Guaraní 121
Angelika Kiss and Guillaume Thomas

Examining MDD and MHD as Syntactic Complexity Measures with Intermediate Japanese Learner Corpus Data 130
Saeko Komori, Masatoshi Sugiura and Wenping Li

Dependency Parsing as Sequence Labeling with Head-Based Encoding and Multi-Task Learning . 136
Ophélie Lacroix

Towards Deep Universal Dependencies 144
Kira Droganova and Daniel Zeman

Delimiting Adverbial Meanings. A corpus-based comparative study on Czech spatial prepositions and their English equivalents 153

Marie Mikulová, Veronika Kolářová, Jarmila Panevová and Eva Hajičová

A Spanish E-dictionary of Collocations 160
Maria Auxiliadora Barrios Rodriguez and Igor Boguslavsky

Exceptive constructions. A Dependency-based Analysis 168
Mohamed Galal, Sylvain Kahane and Yomna Safwat

Quantitative Analysis on verb valence evolution of Chinese 175
Bingli Liu and Chunshan Xu

Association Metrics in Neural Transition-Based Dependency Parsing 181
Patricia Fischer, Sebastian Pütz and Daniël de Kok

Presenting TWITTIRÒ-UD: An Italian Twitter Treebank in Universal Dependencies 190
Alessandra Teresa Cignarella, Cristina Bosco and Paolo Rosso

Pāninian Syntactico-Semantic Relation Labels 198
Amba Kulkarni and Dipti Sharma

Experiments on human incremental parsing 209
Leonid Mityushin and Leonid Iomdin

Character-level Annotation for Chinese Surface-Syntactic Universal Dependencies 216
Yixuan Li, Gerdes Kim and Dong Chuanming

Invited Talk

Tuesday 27th August 2019

Inductive biases and language emergence in communicative agents

Emmanuel Dupoux
ENS/CNRS/EHESS/INRIA/PSL Research University, Paris

Abstract

Despite spectacular progress in language modeling tasks, neural networks still fall short of the performance of human infants when it comes to learning a language from scarce and noisy data. Such performance presumably stems from human-specific inductive biases in the neural networks sustaining language acquisitions in the child. Here, we use two paradigms to study experimentally such inductive biases in artificial neural networks. The first one relies on iterative learning, where a sequence of agents learn from each other, simulating historical linguistic transmission. We find evidence that sequence to sequence neural models have some of the human inductive biases (like the preference for local dependencies), but lack others (like the preference for non-redundant markers of argument structure). The second paradigm relies on language emergence, where two agents engage in a communicative game. Here we find that sequence to sequence networks lack the preference for efficient communication found in humans, and in fact display an anti-Zipfian law of abbreviation. We conclude that the study of the inductive biases of neural networks is an important topic to improve the data efficiency of current systems.

Short bio

Emmanuel Dupoux directs the Cognitive Machine Learning team at the Ecole Normale Supérieure (ENS) in Paris and INRIA (www.syntheticlearner.com). His education includes a PhD in Cognitive Science (EHESS), a MA in Computer Science (Orsay University) and a BA in Applied Mathematics (Pierre & Marie Curie University, ENS). His research mixes developmental science, cognitive neuroscience, and machine learning, with a focus on the reverse engineering of infant language and cognitive development using unsupervised or weakly supervised learning. He is the recipient of an Advanced ERC grant, the organizer of the Zero Ressource Speech Challenge (2015, 2017, 2019), the Intuitive Physics Benchmark (2019) and led in 2017 a Jelinek Summer Workshop at CMU on multimodal speech learning. He has authored 150 articles in peer reviewed outlets from both cognitive science and language technology.

Invited Talk

Wednesday 28th August 2019

Transferring NLP models across languages and domains

Barbara Plank
IT University of Copenhagen

Abstract

How can we build Natural Language Processing models for new domains and new languages?

In this talk I will survey some recent advances to address this ubiquitous challenge, from cross-lingual transfer to learning models under distant supervision from disparate sources, multitask-learning and data selection.

Short bio

Barbara Plank is Associate Professor in Natural Language Processing at IT University of Copenhagen. She has previously held positions as assistant professor at the University of Groningen and the University of Copenhagen, and a postdoc position at the University of Trento. Her research interests within NLP are broad and include learning under sample selection bias (domain adaptation, transfer learning), learning from beyond the text and multimodal inputs, and in general learning under limited supervision for cross-domain and cross-lingual NLP, applied to a range of applications from author profiling, syntactic language understanding, information extraction and visual question answering.

Barbara is member of the advisory board of the EACL (European Association for Computational Linguistics) and publicity director of the Association for Computational Linguistics.

Syntactic dependencies correspond to word pairs with high mutual information

Richard Futrell[1], **Peng Qian**[2], **Edward Gibson**[2], **Evelina Fedorenko**[2], **and Idan Asher Blank**[3]
[1] Department of Language Science, University of California, Irvine
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[3] Department of Psychology, University of California, Los Angeles
rfutrell@uci.edu, {pqian, egibson, evelina9}@mit.edu, iblank@psych.ucla.edu

Abstract

How is syntactic dependency structure reflected in the statistical distribution of words in corpora? Here we give empirical evidence and theoretical arguments for what we call the Head–Dependent Mutual Information (HDMI) Hypothesis: that syntactic heads and their dependents correspond to word pairs with especially high mutual information, an information-theoretic measure of strength of association. In support of this idea, we estimate mutual information between word pairs in dependencies based on an automatically-parsed corpus of 320 million tokens of English web text, finding that the mutual information between words in dependencies is robustly higher than a controlled baseline consisting of non-dependent word pairs. Next, we give a formal argument which derives the HDMI Hypothesis from a probabilistic interpretation of the postulates of dependency grammar. Our study also provides some useful empirical results about mutual information in corpora: we find that maximum-likelihood estimates of mutual information between raw wordforms are biased even at our large sample size, and we find that there is a general decay of mutual information between part-of-speech tags with distance.

1 Introduction

The field of quantitative syntax requires a way to link the discrete formal structures typically studied in syntax, such as dependency trees, with the probabilistic distributions over wordforms observable in corpora.

Formal syntactic structures are usually taken to define the categorical well-formedness of sentences (Chomsky, 1957), or the latent structures required to derive an interpretation (Heim and Kratzer, 1998). It remains unclear what relationship should obtain between these structures and statistical co-occurrence patterns over linguistic units as one might observe in a corpus. Early work in linguistics tried to use these co-occurrence patterns as the basis on which to define formal syntactic structures, formulating 'discovery procedures' which would enable co-occurrence statistics to be summarized mechanistically using formal syntactic structures (Harris, 1954), but modern generative theories of syntax have eschewed any connection between statistical and syntactic structure (Adger, 2018), and to date it remains unclear whether corpus statistics contain enough information to fully reconstruct syntactic structures as identified by linguists. NLP researchers working on grammar induction and unsupervised parsing have achieved substantial gains in recovering dependency trees on the basis of corpus statistics, but overall accuracy remains modest (Klein and Manning, 2004; Spitkovsky et al., 2012; Le and Zuidema, 2015; Pate and Johnson, 2016; Jiang et al., 2016).

Here we propose a high-level linking hypothesis between dependency structures and co-ocurrence statistics: syntactic dependencies correspond to word pairs with high **mutual information (MI)**, an information-theoretic measure of the strength of covariance between two random variables (Cover and Thomas, 2006). We call this claim the **Head–Dependent Mutual Information (HDMI) Hypothesis**. In doing so we formalize and justify an intuition that has underlain much of the work on grammar induction for over 20 years (de Paiva Alves, 1996; Yuret, 1998; Klein and Manning, 2004). The basic intuition is that MI is a generic measure of strength of covariance, and heads and dependents are those word pairs whose covariance is most strongly constrained by grammatical rules.

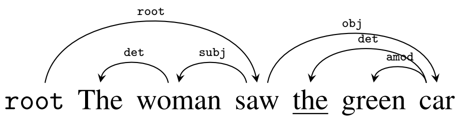


Figure 1: Example dependency tree.

The paper is structured as follows. Sections 2 discusses the methods and dataset we used to measure MI and evaluate the HDMI Hypothesis in corpora; we believe this to be the largest-scale attempt to date to estimate MI between wordforms in natural language text. In Section 3, we present the results of the study, showing that dependencies do identify word pairs with especially high MI as measured in various ways. We also find that mutual information between part-of-speech tags decreases with distance, but we do not observe a similar decay pattern for mutual information between words represented as distributional clusters. Next, in Section 4, we elaborate on the theoretical justification for the HDMI Hypothesis, providing a formal derivation of the hypothesis from an information-theoretic interpretation of the basic postulates of dependency grammar. Section 5 concludes.

2 Measuring Head–Dependent MI in Dependency Corpora

We evaluate the HDMI Hypothesis in a large automatically-parsed corpus of English. To do so, we calculate mutual information between heads and dependents in the corpus. For example, Figure 1 shows an example of a dependency tree. The tree has five **dependency pairs**: ordered pairs of words where the first element is a head and the second is its dependent. The dependency pairs based on this tree are <saw, woman>, <woman, the>, <saw, car>, <car, green>, and <car, the> (excluding the `root` dependency). We calculate Head–Dependent Mutual Information (HDMI) between heads h and dependents d in these pairs:

$$\text{HDMI} = \mathbb{E}\left[\log \frac{p(h,d)}{p(h)p(d)}\right].$$

As a baseline, we also compare HDMI against the mutual information of pairs of words that are not in a direct dependency relationship. See Section 2.2 for details on how these non-dependency word pairs are selected.

For evidence for the HDMI Hypothesis from other languages and hand-parsed corpora, see Futrell and Levy (2017). To our knowledge, the current work is the largest-scale attempt to date to estimate mutual information between words in natural language text and to demonstrate the relationship between dependency and mutual information in a controlled way. The code for our analysis can be found online at `http://github.com/pqian11/mi-hdmi`.

2.1 Estimating Mutual Information

We estimate mutual information using maximum likelihood estimation applied to joint count data over wordforms. The mutual information between wordforms is the true mutual information of interest for our hypothesis, but it is not clear that we can achieve accurate estimates of this quantity due to data sparsity. Therefore we also calculate mutual information between part-of-speech (POS) tags and between distributional clusters, described in more detail below. We include all dependencies except the `root` dependency and those involving wordforms that are not among the top 60,000 most frequent wordforms in the whole corpus.

These experiments also provide data on the convergence of mutual information estimates for wordforms. It is notoriously challenging to estimate information-theoretic quantities such as entropy and mutual information from count data (Miller, 1955; Paninski, 2003; Archer et al., 2013), especially for distributions with long tails, such as wordforms of natural language. Bentz et al. (2017) show that word-level entropy estimates, calculated using maximum likelihood estimation (MLE), converge with around 10^5 tokens of text. But estimating mutual information is more challenging because it requires estimating

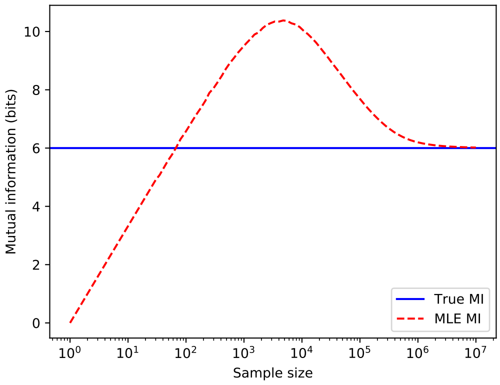


Figure 2: Illustration of bias in MLE estimates of mutual information. The example distribution here is a joint distribution over pairs of bitstrings of length 12, where each pair shares 6 bits, so true mutual information is equal to 6 bits by construction. Empirical MLE estimates of mutual information are shown for various sample sizes. The mutual information estimate initially underestimates the true value, then overestimates it before eventually approaching the true value at around 10^7 samples.

a joint distribution over pairs of words, not just a distribution over single words. It is therefore unknown at what sample size mutual information estimates would converge. Furthermore, while MLE estimates of entropy have a general downward bias, the bias of mutual information is not necessarily downward or upward, as shown in Figure 2. Therefore the MI estimation problem is harder than the entropy estimation problem, because we might not know for some sample size whether we are in an underestimation phase or an overestimation phase.

2.2 Matched non-dependency baseline

We compare the MI of words in dependencies against the MI of words in a **matched non-dependency baseline**. These are word pairs which are not in a direct dependency relationship, and which are matched with the dependency pairs in terms of **displacement**: the linear distance from the head to the dependent and the direction of the dependent with respect to the head, calculated as the linear index of the dependent minus the linear index of the head. For example, given the tree in Figure 1, we might take the non-dependency word pairs <green, the> (displacement -1), <the, saw> (displacement -1), <woman, green> (displacement 3), and <green, saw> (displacement -2). We collect the same number of non-dependency word pairs as dependency word pairs from the corpus. We predict higher MI among the dependency word pairs than among these baseline word pairs.

2.3 Permuted baseline

In order to quantify the magnitude of estimation bias affecting our results, we also compute mutual information for a baseline case where we shuffle the mapping between observed heads and dependents for the entire corpus. In this **permuted baseline**, heads and dependents have analytically zero mutual information: the shuffling process destroys all covariance between heads and dependents within sentences. If our estimation procedures yield any mutual information at all in this case, it can only be due to data sparsity. Therefore the shuffled baseline provides a measure of the strength of the bias affecting our estimates.

2.4 Statistical tests

We wish to statistically compare the MI of dependency word pairs against the MI of the matched non-dependency baseline. To do so, we need some measure of the variance in our MI estimates. Therefore we split our data into 16 equally-sized subsets and calculate MI separately within each subset, and use the standard error of the resulting 16 data points to calculate 95% confidence intervals for each MI estimate. In all figures below, except where otherwise noted, each displayed MI values is the mean of MI values obtained from the 16 subsets. The confidence intervals are too small to be seen. To compare two mean MI estimates statistically, we used two-tailed paired t-tests and report p-values following a False Discovery Rate correction for multiple comparisons (Benjamini and Yekutieli, 2001).

2.5 Dataset

We use the Common Crawl corpus (Buck et al., 2014) of English web text. We filtered the corpus to contain mostly meaningful linguistic utterances and to remove irrelevant web boilerplate text.[1] We parsed and POS-tagged 10% of the filtered corpus using SyntaxNet (Andor et al., 2016). The final dataset used in this paper consists of a total of 320 million tokens of parsed text. SyntaxNet produces function-word-headed dependencies, rather than content-head dependencies, so our results reflect syntactic dependencies rather than semantic dependencies.

2.5.1 POS tags

For MI between POS tags, we use the Penn Treebank POS tags output by SyntaxNet.

POS tags can be interpreted *roughly* as a lower bound on the true MI between full wordforms, because POS tags are mostly a function of individual wordforms. This interpretation is rough because POS tags are to some extent context-dependent. In any case, they can be interpreted as representing the syntactic information present in a word token.

2.5.2 Distributional clusters

Our distributional clusters are derived by spectral clustering from the 300-dimensional GloVe word embedding space trained on 42 billion tokens of the uncased English Common Crawl corpus (Pennington et al., 2014). To generate distributional clusters, we first select the most frequent 60,000 words from a chunk of the whole corpus. After filtering out words that do not have a pretrained embedding in GloVe, we compute the similarity matrix for the remaining 59,998 words and run a spectral clustering algorithm based on the similarity matrix (Pedregosa et al., 2011).

We derive 300 clusters by this method. We found empirically that going above around 300 clusters resulted in many singleton clusters.

We calculate MI between distributional clusters by replacing each word with the index of its cluster and then computing MI by MLE between co-occurrence counts of cluster indices. Because the distributional cluster for a word is a function only of its wordform, the MI between distributional clusters is a true lower bound on MI between full wordforms.

3 Results

3.1 Convergence of MI estimates

Figure 3 shows the convergence of MI estimates for wordforms with increasing sample size, for dependency pairs, matched non-dependency pairs, and the permuted baseline. We see that MI is systematically overestimated at small sample sizes, and that the estimates decrease with increasing sample size. However, even with sample sizes on the order of 10^7 to 10^8 tokens, the estimate does not appear to have converged to a stable value. Furthermore, we see that the permuted baseline, which should ultimately converge to an estimate of zero, still yields a substantial positive MI estimate (0.47 bits) given the full corpus. Overall, we conclude that it is not possible to get an unbiased and stable estimate of MI between wordforms with 10^8 or fewer tokens of text using maximum likelihood estimation. More accurate estimates could come from larger data or from more sophisticated methods of estimating MI.

Now we turn to the convergence of MI estimates based on POS tags and distributional clusters, as shown in Figure 4. Estimates based on POS tags appear to have already converged at 10^5 tokens, and estimates based on distributional clusters appear to converge around 10^7 tokens. Furthermore, the permuted

[1] We filtered out all lines that did not begin with a capital letter and end with punctuation, and all lines containing "copyright", "download", "error", or days of the week or names of months: these lines were overwhelmingly boilerplate text. The filtration process removed about 90% of lines.

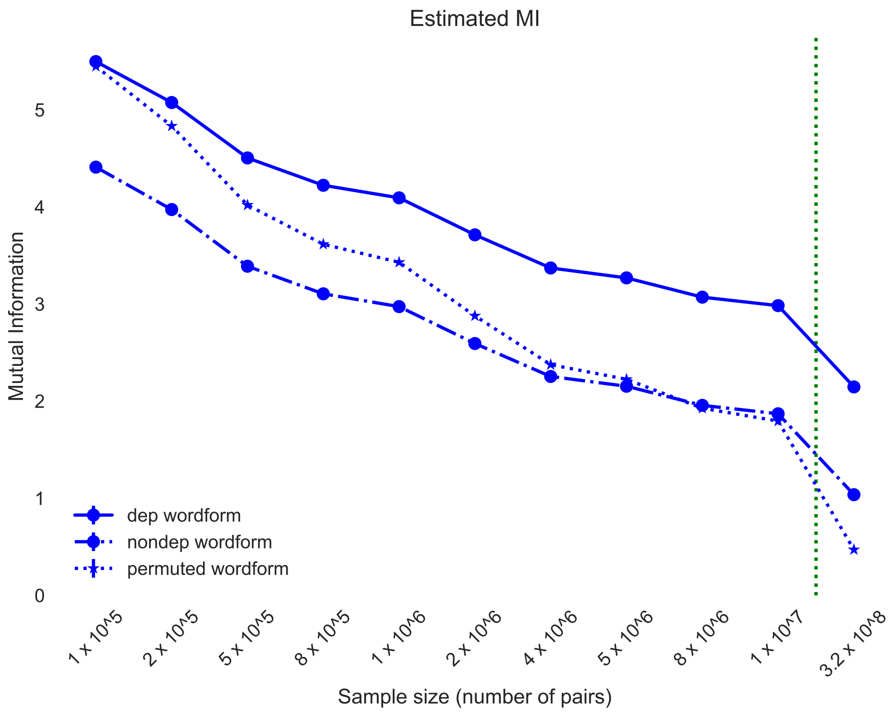


Figure 3: Maximum likelihood-estimated MI by sample size, for wordforms in dependencies (dep), matched non-dependencies (nondep), and the permuted baseline (permuted). The points to the left of the green line are average MI values from 16 subsets of the data. The points to the right of the green line are single point estimates computed from the full corpus.

baseline is near zero for the estimates based on POS tags, and eventually drops to near zero for distributional clusters, an encouraging result that indicates that we have sufficient data to overcome estimation bias due to data sparsity.

3.2 HDMI Hypothesis

Figures 3 and 4 already show that MI in dependencies is higher than in non-dependencies, supporting the HDMI Hypothesis, for POS tags, distributional clusters, and raw wordforms (although the estimation bias for the latter makes the interpretation difficult). For raw wordforms, the difference between dependency MI and baseline MI is significant in all sample sizes at $p < 10^{-16}$.

3.3 Decay with distance

The relationship between mutual information and distance is of theoretical interest beyond the HDMI Hypothesis. Li (1989) and Lin and Tegmark (2017) have reported that mutual information between orthographic letters in natural language text falls off as a power law with distance, but the relationship between mutual information and distance at the level of words has not yet been explored in large corpora. Because of the estimation difficulties observed in Section 3.1, we do not analyze MI between raw wordforms here, but rather only between POS tags and distributional clusters.

We estimate mutual information between POS tags and distributional at different distances. We hold sample size constant for all distances, meaning that we have around 5×10^6 dependency pairs available to estimate MI at each distance. Figure 5 shows the results. We see a clear fall-off of mutual information with distance for POS tags, for both dependencies and non-dependencies. However, we see no fall-off for distributional clusters, indicating that it may be primarily syntactic information that drives high-MI words to be close to each other.

We can also see from Figure 5 that the HDMI Hypothesis holds for both POS tags and distributional clusters in all distances shown (with significance at $p < 10^{-18}$ at all distances shown).

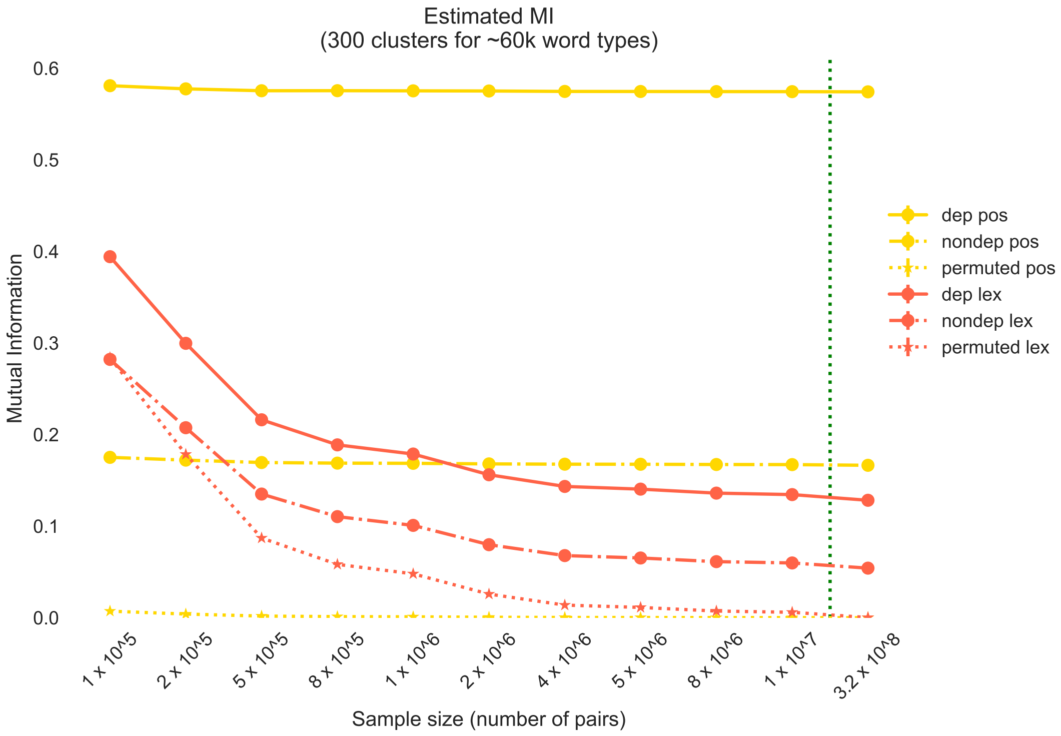


Figure 4: Maximum likelihood-estimated MI by sample size, for wordforms in dependencies (dep), matched non-dependencies (nondep), and the permuted baseline (permuted), based on POS tags (yellow) and distributional clusters (red). Green line as in Figure 3.

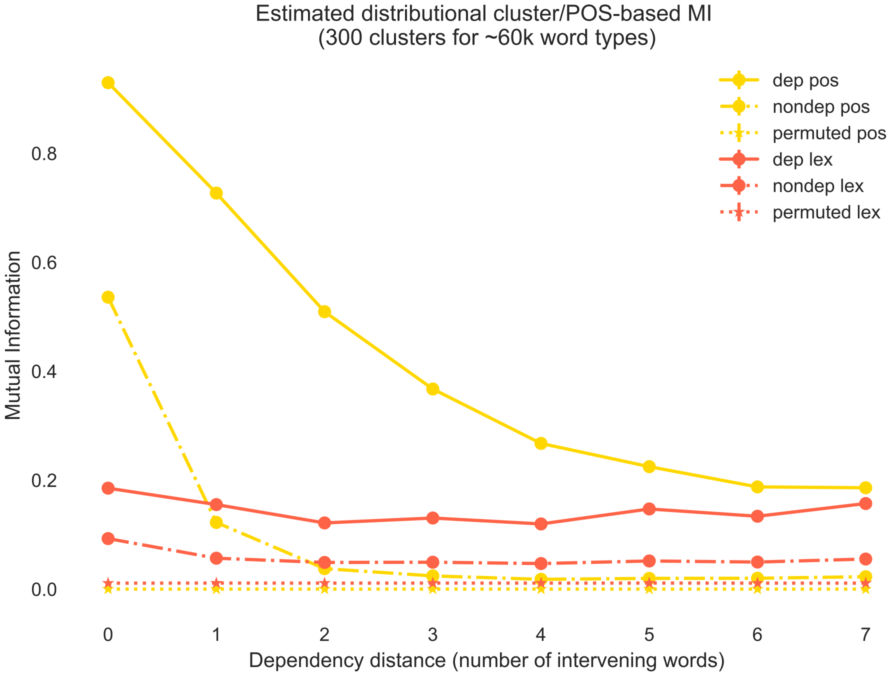


Figure 5: Estimated mutual information between words by number of intervening words (dependency distance), as estimated using POS tags (yellow) and distributional clusters (red). MI for each distance is estimated from ~ 5 million dependency pairs, averaged over 16 subsets of the corpus.

One caveat is in order regarding the interpretation of Figure 5: these results are based on an automatically-parsed corpus, and the more distant dependencies may be less accurately identified by the parser. Long dependencies are known to cause difficulties for shift–reduce parsers such as SyntaxNet (Gulordava and Merlo, 2015). Therefore in our dataset, it may be the case that the longer dependencies are noisier, and their MI will thus regress to the MI of non-dependency word pairs. However, we note that Futrell and Levy (2017) found a similar decay of POS tag MI with distance even in hand-parsed corpora.

4 Theoretical justification

Having established that the HDMI Hypothesis holds empirically in large corpora, we now turn to a theoretical justification for this hypothesis. We propose to view dependency grammar as a method for approximating arbitrary probability distributions over strings. Taking this view, we show that choosing dependency trees to minimize approximation error is equivalent to choosing dependency trees to maximize the head–dependent MI. Therefore the most accurate dependency trees in a linguistic sense will be those with maximal head–dependent MI.

We take the basic postulate of dependency grammar to be that the syntactic well-formedness of a sentence can be fully or mostly characterized in terms of the pairs of head and dependent words in the sentence as identified by some dependency tree (Hudson, 1984, 2010). That is, restrictions on covariance between words in sentences can be stated entirely in terms of the head–dependent pairs forming dependency trees. Given a dependency tree such as the one in Figure 1, all that you would need to know to specify the conditions on what word can go in the underlined position is the identity of the head word—*car*, a noun, licensing a determiner as a dependent. There may also be dependency type labels which are relevant. In the strongest possible formulation of dependency grammar—undoubtedly too strong—the head provides *all* the information you need to specify the possible dependents. More realistically, we can say that the identities of other words in the dependency tree, which are only distantly connected to the underlined word in terms of the dependency structure, play relatively minor roles in the determination of the underlined word.

While dependency grammar was developed to specify categorical well-formedness conditions, we can make a probabilistic generalization and say that the *probability* of a sentence is fully or mostly characterizable in terms of the head and dependent pairs. This assumption is closely related to generative models from the grammar induction literature called **head-outward generative models** (Eisner, 1996; Klein and Manning, 2004), in which the probability of a sentence can be factorized in terms of a dependency tree. Representing a sentence as a sequence of n words $\mathbf{w}_{i=1}^{n}$, and representing the dependency tree for the sentence as a sequence of n heads $\mathbf{t}_{i=1}^{n}$ where t_i gives the head of the ith word w_i, we can factorize the probability of the sentence $\mathbf{w}$ as:

$$p_{\mathbf{t}}(\mathbf{w}_{i=1}^{n}) = \prod_{i=1}^{n} p_{\mathbf{t}}(w_i|t_i).^{2} \tag{1}$$

We propose to view dependency grammar in this sense as a generic method for approximation to arbitrary distributions over sequences, closely related to Chow-Liu trees (Chow and Liu, 1968), which are a general scheme for approximating any joint distribution in terms of only pairwise dependencies.

Any distribution over sequences of symbols could be approximated by Eq. 1 for some set of dependency trees specified by $\mathbf{t}$, to varying degrees of accuracy. Eq. 1 fundamentally expresses an assumption that all the relevant information in the context about the symbol w_i is concentrated in exactly one other symbol t_i—corresponding to the dependency grammar postulate described above.

The independence assumptions of Eq. 1 are obviously too strong for natural language, but surprisingly they provide a reasonable approximation (Eisner, 1996, 1997). It is an interesting scientific question why natural language has the property that it can be well-approximated by such a dependency grammar.

[2] Head-outward generative models differ from our Eq. 1 in that they also put a prior distribution over tree structures $\mathbf{t}$. In contrast, we are only interested in the probability of a string $\mathbf{w}$ given a tree $\mathbf{t}$. One consequence of our formulation is that the *halting probabilities* that appear in Eisner (1996) do not appear in our equations. Since we are not considering prior probabilities on tree structures, our approach in this section is similar to fitting a head-outward generative model by maximum likelihood estimation.

We propose that when linguists are developing dependency grammars and assigning dependency trees to sentences, they are implicitly finding trees $\mathbf{t}$ to make the approximation in Eq. 1 as accurate as possible. That is, for each word, they are choosing the heads that best explain the distribution of each word in the sentence. More formally, they are solving the problem of minimizing the divergence between the dependency approximation in Eq. 1 and the true distribution over sequences of symbols (sentences), which we call p_L. The true distribution over sequences p_L can be written generically as:

$$p_L(\mathbf{w}_{i=1}^n) = \prod_{i=1}^n p_L(w_i|\mathbf{w}_{<i}), \tag{2}$$

where $\mathbf{w}_{<i}$ represents the sequence of symbols up to the ith (non-inclusive).

We now show that minimizing the KL-divergence between the true distribution over sequences p_L (Eq. 2) and the dependency approximation $p_\mathbf{t}$ (Eq. 1) is equivalent to choosing head–dependent pairs that maximize mutual information. This result provides a conceptual link between dependency grammar and information-theoretic statistics observable in corpora.

More formally, let p_L be a conditional probability distribution with support over symbols w_i, called **words**, and a special sentinel symbol which marks the end of a sentence. The distribution p_L generates symbols conditional on a sequence of previous words $\mathbf{w}_{<i}$, called a **context**, also generated by p_L, and starting with a special beginning-of-sentence symbol called `root`. Let $\mathbf{t}$ be a sequence of symbols, called **heads**, where t_i is equal to some word w_j for $j < i$ or to `root`, such that the pairs $\langle t_i, w_i \rangle$ define a dependency graph within each sentence.[3] We hold the distribution p_L to be a fixed target, and we are interested in finding the assignment of heads $\mathbf{t}$ that minimizes the expected per-symbol KL-divergence between the dependency approximation $p_\mathbf{t}$ of L and the true distribution p_L:

$$D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_\mathbf{t}(w_i|t_i)) = \mathbb{E}\left[\log \frac{p_L(w_i|\mathbf{w}_{<i})}{p_\mathbf{t}(w_i|t_i)}\right]. \tag{3}$$

Proposition 1. *The heads* $\mathbf{t}$ *that minimize approximation error (Eq. 3) are given by:*

$$\operatorname*{argmax}_{\mathbf{t}} I[W:T],$$

where W is the distribution over single words generated by p_L, T is the distribution over elements of $\mathbf{t}$, *and $I[W:T]$ gives the mutual information of W and T, called the* ***Head–Dependent Mutual Information*** *(HDMI):*

$$I[W:T] = \mathbb{E}\left[\log \frac{p_\mathbf{t}(w_i|t_i)}{p(w_i)}\right].$$

Proof. We begin by applying Bayes' rule to the numerator of the log probability ratio in Eq. 3:

$$\begin{aligned} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_\mathbf{t}(w_i|t_i)) &= \mathbb{E}\left[\log \frac{p_L(w_i|\mathbf{w}_{<i})}{p_\mathbf{t}(w_i|t_i)}\right] \\ &= \mathbb{E}\left[\log \frac{p(\mathbf{w}_{<i}|w_i)p(w_i)}{p(\mathbf{w}_{<i})p_\mathbf{t}(w_i|t_i)}\right]. \end{aligned} \tag{3}$$

Now we separate the result into two terms:

$$\min_\mathbf{t} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_\mathbf{t}(w_i|t_i)) = \min_\mathbf{t} \mathbb{E}\left[\log \frac{p(w_i)}{p_\mathbf{t}(w_i|t_i)}\right] + \cancel{\mathbb{E}\left[\log \frac{p(\mathbf{w}_{<i}|w_i)}{p(\mathbf{w}_{<i})}\right]}. \tag{4}$$

[3] Our construction includes an assumption that t_i for each word w_i is equal to some previous word $w_{j<i}$. This assumption may appear to entail that our dependency trees are strictly head-initial. However, the assumption is without loss of generality, because the order of the indices in Eq. 2 is arbitrary and does not have to correspond to the linear order of words: different orders simply correspond to different applications of the chain rule for probabilities and will yield the same total probability, as long as the context $\mathbf{w}_{<i}$ is encoded in such a way that the original indices are recoverable. Therefore it is always possible to reassign indices within a sentence such that the dependency graph defined by $\mathbf{t}$ appears to be strictly head-initial, while the value of Eq. 2 will remain the same. Similarly, the second term in Eq. 4 below is also invariant to the choice of indices, because it is equivalent to the average MI between contexts and words. So our result will hold for all tree structures within sentences, be they head-initial, head-final, or mixed within sentences.

The last term in Eq. 4 is the mutual information of words with their contexts under p_L. This quantity (in expectation over words and contexts) is invariant to the choice of **t**, so we can remove it from our minimization objective.

Now using the property that $\log \frac{a}{b} = -\log \frac{b}{a}$, we see that our minimization problem comes out to maximizing the HDMI:

$$\begin{aligned}\min_{\mathbf{t}} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_{\mathbf{t}}(w_i|t_i)) &= \min_{\mathbf{t}} -\mathbb{E}\left[\log \frac{p_{\mathbf{t}}(w_i|t_i)}{p(w_i)}\right] \\ &= \min_{\mathbf{t}} -I[W:T] \\ &= \max_{\mathbf{t}} I[W:T].\end{aligned}$$

□

Proposition 1 means that, if dependency structures are to be interpreted in the sense of Eq. 1, then heads and dependents will be those word pairs with maximal mutual information. This is our proposed theoretical justification for the HDMI Hypothesis.

5 Conclusion

We addressed the question of how syntactic dependency structure is reflected in the statistical covariance structure of words in natural language corpora, from an empirical and theoretical perspective. We advanced a theoretical argument, based on an information-theoretic interpretation of the postulates of dependency grammar, claiming that syntactic heads and dependents should correspond to word pairs with high MI: the HDMI Hypothesis. We reported what we believe is to date the largest-scale attempt to quantify mutual information between words in natural language text as a function of dependency structure, and found empirical support for the HDMI Hypothesis. We also found that MI between raw wordforms cannot be estimated by maximum likelihood estimation without bias even with 320 million tokens of text, and that MI between POS tags falls off with distance, mirroring previous findings about MI between orthographic letters (Li, 1989; Lin and Tegmark, 2017), although we found no fall-off for MI between distributional clusters.

Our work establishes a general link between syntactic structure and the statistical properties of texts, joining other work which has established connections between grammatical rules and information-theoretic statistics (Dębowski, 2015). We believe the HDMI Hypothesis can form the basis for improved grammar induction algorithms, by providing a new perspective on the head-outward generative models that have formed the basis of most work in that area. It also provides an intuitive means for comparatively evaluating different theories of dependency grammar (e.g., content-head vs. function-head: Osborne and Gerdes, 2019), in terms of the approximation error induced by different theories according to Eq. 3. In general, we believe the HDMI Hypothesis will also provide a stronger theoretical basis for corpus linguistics by linking the two conceptually independent notions of syntactic and statistical structure.

Acknowledgments

We thank Tim O'Donnell and Roger Levy for many discussions on these topics, and the anonymous reviewers for helpful comments on the paper.

References

Adger, D. (2018). The autonomy of syntax. In Hornstein, N., Lasnik, H., Patel-Grosz, P., and Yang, C., editors, *Syntactic Structures after 60 Years*, pages 153–175.

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based networks. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*, Berlin.

Archer, E., Park, I. M., and Pillow, J. W. (2013). Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19:275–307.

Buck, C., Heafield, K., and Van Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *LREC*.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

Cover, T. M. and Thomas, J. (2006). *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

de Paiva Alves, E. (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.

Dębowski, Ł. (2015). The relaxed Hilberg conjecture: A review and new experimental support. *Journal of Quantitative Linguistics*, 22(4):311–337.

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 340–345.

Eisner, J. M. (1997). An empirical comparison of probability models for dependency grammar. Technical report, IRCS Report 96–11, University of Pennsylvania.

Futrell, R. and Levy, R. (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.

Gulordava, K. and Merlo, P. (2015). Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *CoNLL*, pages 247–257.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23):146–162.

Heim, I. and Kratzer, A. (1998). *Semantics in generative grammar*. Wiley-Blackwell, Malden, MA.

Hudson, R. A. (1984). *Word Grammar*. Blackwell.

Hudson, R. A. (2010). *An introduction to word grammar*. Cambridge University Press.

Jiang, Y., Han, W., and Tu, K. (2016). Unsupervised neural dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 763–771.

Klein, D. and Manning, C. D. (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.

Le, P. and Zuidema, W. (2015). Unsupervised dependency parsing: Let's use supervised parsers. *arXiv preprint arXiv:1504.04666*.

Li, W. (1989). Mutual information functions of natural language texts. Technical report, Santa Fe Institute Working Paper #1989-10-008.

Lin, H. W. and Tegmark, M. (2017). Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299.

Miller, G. A. (1955). Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*, pages 95–100.

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

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Pate, J. K. and Johnson, M. (2016). Grammar induction from (lots of) words alone. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 23–32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. (2012). Three dependency-and-boundary models for grammar induction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*.

Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Reflexives in Czech from a dependency perspective

Václava Kettnerová
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
kettnerova@ufal.mff.cuni.cz

Markéta Lopatková
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
lopatkova@ufal.mff.cuni.cz

Abstract

Reflexives are the source of ambiguity in many languages, including Czech. In this paper, we address Czech reflexives and their description in the dependency-oriented theory, Functional Generative Description. Our primary focus in this paper lies in the reflexives that form analogous syntactic structures as personal pronouns (e.g., *Jan si / jí nevěří.* ‘John does not believe in himself / in her.’). In Czech (similarly as in other Slavic languages), these reflexives encode reflexivity or reciprocity, two closely related phenomena. We offer an in-depth analysis of both these phenomena and propose their description in lexicon and in grammar. Further, we clarify principles underlying ambiguity of reflexive and reciprocal constructions.

1 Introduction

Reflexives appear in a great number of languages. Due to an enormous diversity in their functions, their description represents a tricky task for any syntactic theory. A large number of analyses of reflexives apply methodological principles of a generative syntax, see esp. (Chomsky, 1981; Reinhart and Reuland, 1993; Pollard and Sag, 1992), usually making an effort to provide their unified analysis. Recently, reflexives have been studied in individual languages as well as from a typological perspective, attesting their high ambiguity across languages, see esp. (Faltz, 1985; Geniušienė, 1987; Kemmer, 1993; Frajzyngier and Walker, 2000a; Frajzyngier and Walker, 2000b; König and Kokutani, 2006; Nedjalkov, 2007; König and Gast, 2008; Evans et al., 2011). In this paper, we provide a description of various functions of reflexives in Czech and propose their representation in a dependency-oriented theory, namely in the Functional Generative Description (FGD henceforth) (Sgall et al., 1986; Panevová et al., 2014), with an emphasis put on the distribution of the linguistic information between lexicon and grammar, as two sides of the language description.

In Czech linguistics, reflexives are classified either as a part of verb lemmas or inflectional verb forms, or as the reflexive pronoun. The primary focus in this paper lies in the reflexives representing the reflexive pronoun occurring in reflexive and reciprocal constructions. We offer an in-depth analysis of the deep and surface syntactic structures of these constructions – in FGD, the former one roughly corresponding to the so-called tectogrammatical layer, i.e., a layer of the linguistically structured meaning, while the latter is represented by the so-called analytical layer, see esp. (Sgall et al., 1986; Hajič et al., 2018). We thus provide a comprehensive account of these two related phenomena in verbal as well as nonverbal structures, which allows for generation of well-formed reflexive and reciprocal constructions. We follow and further deepen analysis of the description proposed in (Kettnerová and Lopatková, 2018b; Kettnerová and Lopatková, 2018a) putting under scrutiny other parts of speech than verbs as well.

2 Reflexives in the Functional Generative Description

While the classification of the long forms of the reflexive *sebe*/*sobě*/*sebou* as the reflexive pronoun does not pose any difficulties in Czech linguistics, the status of the reflexive clitics *se*/*si* is rather questionable. Their analysis is heavily dependent on the overall architecture of a linguistic theory within which it is conducted, see esp. (Karlík, 1999; Oliva, 2001; Medová, 2009; Veselý, 2018). In FGD, reflexives are classified according to their function in the language system, i.e., functionally equivalent reflexives take the identical status in the language description, regardless of their clitic or long forms, see esp. (Panevová, 2001; Panevová et al., 2014; Kettnerová et al., 2014). On the basis of their function, reflexives are distinguished into the reflexive pronoun and into the reflexives representing either parts of verb lemmas (often referred to as inherently or derived reflexive verbs, see below), or reflexive inflectional verb forms.

Reflexive Pronoun. In Czech, the reflexive pronoun has the long forms *sebe*/*sobě*/*sebou* or the clitic forms *se*/*si*; the clitic forms are available only in the prepositionless accusative case (*se*) and in the dative case (*si*). Only those reflexive clitics are treated as the reflexive pronoun that can change – depending on their position in a sentence – into long forms when stressed, see also (Komárek et al., 1986). The reflexive pronoun – similarly as non-reflexive pronouns – fills one valency position of a predicate (a verb, a noun, an adjective or an adverb). In Czech, the reflexive pronoun, marking the referential identity between the filled position and another expression, encodes reflexivity (Section 3) or reciprocity (Section 4).

In examples with the verb *věřit* 'to believe' (2), PAT of the verb, see the valency frame in (1), is filled with the reflexive pronoun in the clitic form (2a) or in the long form (2b), respectively, coreferring with *Jan* 'John' in the subject position given by ACT of the verb; in both variants the reflexive encodes reflexivity. Similarly, in examples (3) with the same verb, the reflexive pronoun in the clitic and long form, filling PAT, corefers with ACT of the verb; depending on the context, the reflexive pronoun marks either reflexivity, or reciprocity, see Figure 1a below.

(1) *věřit*^{impf} 'to believe': ACT$_1$ PAT$_{3,dcc}$[1]

(2) a. *Jan si nevěří.*
John REFL$_{clitic.dat}$ not believes
'John does not believe in himself.'

b. *Jan nevěří sobě, věří ale manželce.*
John not believes REFL$_{long.dat}$, believes but wife
'John does not believe in himself but he believes in his wife.'

(3) a. *Lidé ve městě si nevěří.*
people in town REFL$_{clitic.dat}$ not believe
'People in towns do not believe in themselves // in each other.'

b. *Lidé ve městě sobě nevěří.*
people in town REFL$_{long.dat}$ not believe
'People in towns do not believe in themselves // in each other.'

Reflexives in Verb Lemmas. As parts of verb lemmas, only the clitic reflexives *se* and *si* occur (as such they cannot be stressed and they do not fill valency position of a verb). These clitic reflexives appear with reflexive tantum verbs (referred also to as inherently reflexive verbs), see example (4a) with the reflexive *se* as an obligatory part of the verb lemma *blížit se* 'to approach' and (4b) with *si* belonging to the verb lemma *odpočinout si* 'to rest' (Figure 1b). Further, the clitic reflexives serve as derivational means, deriving reflexive verbs (referred also to as derived reflexive verbs) from irreflexive ones; the derivational process can have various semantic and/or syntactic motivations,[2] see examples with the verb *budit* 'to wake' (5a) and with the derived verb *budit se* 'to wake' (5b) (with the reflexive *se* marking decausativity) and examples with the verb *pomáhat* 'to help' (6a) and the derived verb *pomáhat si* 'to help' (6b) (with the reflexive *si* signaling inherent reciprocal meaning).[3]

[1] In valency frames, numbers stand for morphemic cases (1=nom, 2=gen, 3=dat, 4=acc, 6=loc, 7=instr), possibly preceded by required prepositions, dcc stands for dependent content clauses (often referred to as nominal subordinate clauses), and pos represents possessive forms. As it is not relevant for our explanation here, we omit the information on obligatoriness from valency frames.

[2] A detailed analysis of semantic and syntactic functions of the clitic reflexives *se/si* that serve as derivational means, providing an account for a possible difference in the distribution of these two clitics, has not been done for Czech yet. However, such an analysis goes beyond the scope of this paper; from the reflexives representing parts of verb lemmas, only the reflexives in lemmas of inherent reciprocal predicates are considered here in connection with reciprocity, see example (6b) and Section 4.

[3] The clitic reflexives *se* and *si* occur also with the verbal nouns and present participles of verbs that are systematically derived by productive suffixes from verbs with reflexive lemmas; while with the present participles, the clitics are obligatory,

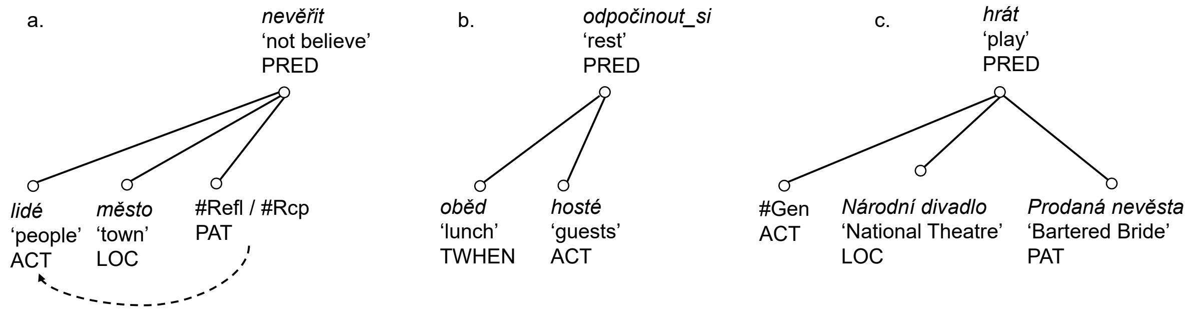


Figure 1: The simplified tectogrammatical trees of sentences (3), (4b) and (7a), respectively. In tree (a), the dashed arrow indicates coreference, pointing from the reflexive pronoun to its antecedent; as ambiguous structures must be distinguished at the tectogramatical layer, this scheme represents – for the sake of brevity – two trees: one with the #Refl lemma and the other with #Rcp lemma, both standing for the reflexive pronoun, distinguishing its function. In tree (b), the reflexive is represented as the part of the verb lemma. In tree (c), the reflexive inflectional verb form is derived on the basis of grammatical rules conditioned by the value of the verbal grammateme 'deagentive' (not displayed), resulting in generalization of ACT (the lemma #Gen). For the annotation principles see esp. (Mikulová et al., 2006).

(4) a. *Horolezci se / *sebe blížili k vrcholu hory.*
mountaineers REFL$_{clitic}$ / REFL$_{long}$ approached to summit of mountain
'Mountaineers were approaching to the summit of the mountain.'

b. *Po obědě si / *sobě hosté odpočinuli.*
after lunch REFL$_{clitic}$ / REFL$_{long}$ guests rested
'The guests had a rest after the lunch.'

(5) a. *Maminka budila děti v sedm hodin.*
'Mother woke children up at seven o'clock.'

b. *Děti se / *sebe budily v sedm hodin.*
children REFL$_{clitic}$ / REFL$_{long}$ woke at seven o'clock
'Children woke up at seven o'clock.'

(6) a. *Jan pomáhal kolegům při práci.*
'John helped his colleagues at work.'

b. *Jan si / *sobě při práci pomáhal s kolegy.*
John REFL$_{clitic}$ / REFL$_{long}$ at work helped with colleagues
'John and colleagues helped at work with each other.'

As for the representation of the clitic reflexives of the given type, they are recorded in the lexicon as parts of relevant lemmas.

Reflexives in Inflectional Verb Forms. With verbs, the clitic reflexive *se* can represent also a part of the reflexive verb form, which is characteristic of marked constructions of the *deagentive* and *dispositional diatheses* (also referred to as middle alternation), see examples (7a) and (7b), respectively. In this case, the clitic reflexive *se* serves as a voice marker, being thus an inflectional means; as such this reflexive does not occupy a valency position of a verb and it cannot be stressed.

The inflectional reflexive verb form brings about a shift of ACT of a verb from the subject position: in case of the deagentive diathesis, the ACT is elided from the surface (7a), see Figure 1c, and in case of the dispositional diatesis, it can be optionally expressed in the dative case (7b).

e.g., *bojící se* 'having fear' (← *bát se* 'to fear') and *stěžující si* 'complaining' (← *stěžovat si* 'to complain'), with the verbal nouns, they are only optional, e.g. *bání (se)* 'fearing' and *stěžování (si)* 'complaining'. In both cases, the presence of the clitic reflexive is considered as evidence of the verbal character of these nouns and participles, see esp. (Dvořak, 2017). These forms are left aside here.

(7) a. *V Národním divadle se / *sebe hrála Prodaná nevěsta.*
in National Theatre REFL_{clitic} / REFL_{long} played Bartered Bride
'The Bartered Bride was played in the National Theatre.'

b. *Koláč se / *sebe (mamince) špatně pekl.*
pie REFL_{clitic} / REFL_{long} (for mother) badly baked
'The pie baked badly (for my mother).

As for the representation of the deagentive and dispositional diatheses, syntactic changes in the surface structure of verbs can be captured by formal rules comprised in the grammar, while the applicability of these diatheses must be recorded in the lexicon as it is given by the lexical meaning of verbs to a great extent and as such it is not derivable from the valency structure of verbs itself.

3 Reflexivity and Its Encoding in Czech

Reflexivity represents language means expressing the fact that two semantic participants of a predicate have a single referent. In Czech linguistics, reflexivity has gained a lot of attention, see esp. (Havránek, 1928; Karlík, 1999; Dočekal, 2008; Medová, 2009; Hudousková, 2009). Within FGD, reflexivity has been studied esp. by Panevová (2001, 2008) and her discussion with Oliva and others (Oliva, 2000; Oliva, 2001; Komárek, 2001; Kettnerová et al., 2014).

In Czech, reflexivity can characterize verbs (8a), nouns (8b), adjectives (9a) and sporadically adverbs (9b) (reflexivity of adverbs are left aside here due to data sparseness). A substantial role in its expression is played by the reflexive pronoun.[4]

(8) a. *Marie se pořád jen lituje.*
'Mary feels sorry for herself all the time.'

b. *Mariina lítost nad sebou*
'Mary's sorrow for herself'

(9) a. *necitlivý k sobě*
'insensitive to herself/himself'

b. *necitlivě k sobě*
'insensitively to herself/himself'

Reflexive constructions can be described as a result of a morphosyntactic operation of reflexivization applied to a valency frame of a predicate. As the applicability of this operation cannot be derived from the valency structure itself, it must be provided with each relevant predicate in the lexicon. However, morphosyntactic patterns underlying reflexivity are so regular that they can be captured in the form of rules contained in the grammar. These patterns are further described below.

Reflexivity in Deep Structures. In the deep syntactic structure of reflexive constructions, the number and type of valency complementations of a predicate are preserved. Moreover, the mapping between semantic participants and valency complemenations[5] remains the same as in non-reflexive constructions, i.e, each semantic participant is mapped onto a single valency complementation.[6] The main difference lies in the fact that in reflexive constructions, two semantic participants refer to a single referent; as a result, the valency complementations involved in reflexivity are linked by a coreferential relation.

[4] In the VALLEX lexicon, reflexivity is captured only with lexical units of verbs that allow the reflexive pronoun in prepositionless dative or accusative – counting only cases where reflexivity affects actants, it is annotated with 578 lexical units of verbs, represented by 690 verb lemmas (if relevant, one lexical unit is represented by lemmas of different aspectual values; moreover, lemmas can have ortographic variants, e.g., *oblékat*^{impf} / *obléci/obléknout*^{pf} 'to put on sth'). In PDT, reflexivity is annotated in 712 instances (only cases affecting actants are counted): 695 in verbal structures, 7 in nominal structures and 10 in adjectival structures, represented by 451 verb lemmas, 8 noun lemmas, and 6 adjective lemmas; however, out of these instances, 171 represent annotation errors: 49 instances are syntactic reciprocals, 16 are lexical reciprocals, 104 are rather reflexive verb lemmas, and 2 are inflectional reflexive verb forms.

[5] Roughly corresponding to semantic actants and deep syntactic actants, respectively, in the Meaning-Text Theory (Mel'čuk, 2004).

[6] Compare with the complex mapping of semantic participants in reciprocal constructions discussed in Section 4.

Let us demonstrate the operation of reflexivization on the verb *uctívat / uctít* 'to respect' and the deverbal noun *úcta* 'respect'. Both these predicates evoke two semantic participants, 'Cognizer' and 'Evaluee', mapped in both cases onto ACT and PAT, respectively, see the valency frames in (10) and (12). The mapping remains the same regardless whether 'Cognizer' and 'Evaluee' refer to different referents or to a single referent; however, in the latter case, the deep syntactic structure of these predicates is characterized by coreference between ACT and PAT, see examples (11) and (13) and their simplified dependency trees in Figure 2a and 2b.[7]

Further, the adjective *uctivý* respectful, derived from the verb *uctívat* 'to respect', is characterized by the same set of semantic participants, Cognizer and Evaluee. However, from these participants, only the latter one can be syntactically structured as a valency complementation of the adjective;[8] this participant is mapped onto PAT, see the valency frame in (14). The participant Cognizer is typically syntactically structured outside adjectival structures, either as the governor of the adjective (15a), or as ACT of the copula verbs *být* 'to be' and *stávat se / stát se* 'to become' with the adjective (15b), see esp. (Boguslavsky, 2003). As a consequence, the coreference relation links PAT of the adjective and either its governor, see Figure 2c, or ACT of copula verbs.[9]

(10) *uctívat*^{impf} */ uctít*^{pf} 'to respect': ACT$_{1}$ PAT$_{4,dcc}$

(11) a. *Tarkovskij začal, tvrdí pisatel, nakonec sám sebe uctívat.*
Tarkovsky began, claims writer, finally alone REFL$_{long.acc}$ respect
'As the writer claims, Tarkovskij finally began to honor himself.'

b. *Tarkovskij se (sám) uctíval.*
Tarkovsky REFL$_{clitic.acc}$ (alone) respect
'Tarkovskij honored himself.'

(12) *úcta* 'respect': ACT$_{2,pos}$ PAT$_{k+3}$

(13) *Tarkovského úcta k sobě*
Tarkovsky's respect to REFL$_{long.dat}$
'Tarkovsky's respect for himself'

(14) *uctivý* 'respectful': PAT$_{k+3}$

(15) a. *člověk uctivý (sám) k sobě*
man respectful (alone) to REFL$_{long.dat}$
'a man respectful to herself/himself'

b. *Člověk je uctivý (sám) k sobě.*
man is respectful (alone) to REFL$_{long.dat}$
'A man is respectful to herself/himself.'

Reflexivity in Surface Structures. Surface positions provided by coreferring valency complementations of a predicate are indicated in the valency frames of the given predicate by morphemic forms. One of these surface position is occupied by *the reflexive pronoun* while the other is filled with *its antecedent*.[10]

The reflexive pronoun can occupy various surface positions, direct or indirect objects (with verbs), attributes (with nouns) and adverbials (with verbs and adjectives). Predominantly, it has the long form, the clitic form of the reflexive pronoun is available only with verbs in the prepositionless dative or accusative

[7] The valency structure of deverbal nouns typically corresponds to the valency structure of their base verbs, see esp. (Kolářová, 2014). In case of primary nouns, valency of verbs with similar meanings should be taken into account, e.g., *láska* 'love' and *milovat* 'to love', see esp. (Piha, 1984).

[8] For specific valency properties of adjectives in Czech see esp. (Panevová, 1998; Panevová et al., 2014).

[9] With deadjectival nouns, one valency complementation – typically ACT – is added to their valency frames that corresponds to the governor of their base adjectives or to ACT in constructions with copula verbs; compare, e.g., the valency frame of the adjective *lhostejný* 'indifferent': PAT$_{k+3,\text{vůči}+3}$ (e.g., *člověk lhostejný k neštěstí*$_{PAT}$ *druhých* 'a man indifferent to others' misery$_{PAT}$' and *Člověk se stane lhostejným k neštěstí*$_{PAT}$ *druhých.* 'A man became indifferent to others' misery$_{PAT}$.') and the frame of the noun *lhostejnost* 'indifference' derived from this adjective: ACT$_{2,pos}$ PAT$_{k+3,\text{vůči}+3}$ (e.g., *lhostejnost člověka*$_{ACT}$ *k druhým*$_{PAT}$ 'man's$_{ACT}$ indifference to others$_{PAT}$').

[10] Further, reflexivity can be optionally emphasized by the expression *sám* 'alone', see examples in (11) and (15).

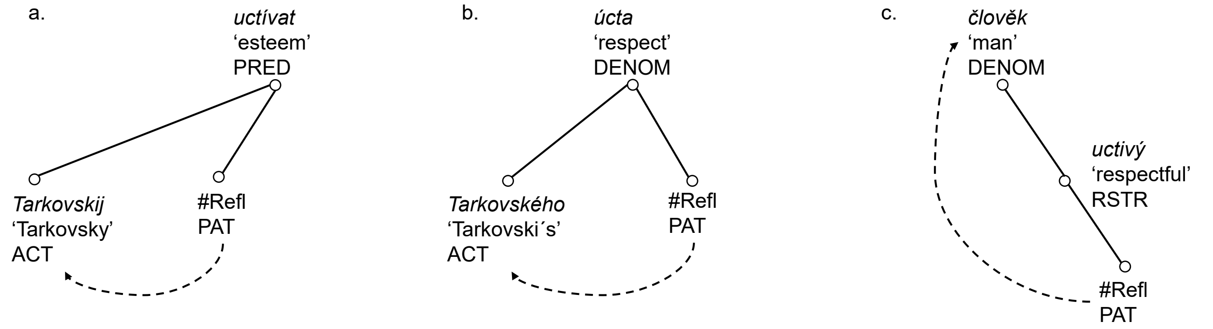


Figure 2: The simplified tectogrammatical trees of examples (11b), (13) and (15a), respectively; the dashed arrow shows coreference.

case, depending on whether it is stressed, or not, compare examples (11a) and (11b). With nouns and adjectives, the clitic forms are not available,[11] only the long forms of the reflexive pronoun are acceptable (Dvořak, 2017).

As for *the antecedent of the reflexive pronoun*, with verbs, it is represented by subject provided by the valency complementation in nominative, typically ACT, see examples with the verb *uctívat / uctít* 'to respect' (11a-b) and its valency frame in (10). With deverbal nouns, the antecedent occupies the attribute position corresponding to subject of their respective base verbs; see the valency frame of the deverbal noun *úcta* 'respect' in (12) and example (13). With adjectives, the antecedent occupies the position of their governors or ACT of copula verbs, being thus external to adjectival structures, see examples (15a-b) with the adjective *uctivý* 'respectful' and its valency frame (14).

4 Reciprocity and Its Encoding in Czech

Reciprocity is understood here as language means expressing a semantic relation of mutuality. In Czech linguistics, reciprocity has not attracted much attention yet; even in summarizing grammars, it is mentioned only marginally, see esp. (Daneš et al., 1987; Grepl and Karlík, 1998). The most elaborated analysis of reciprocity in Czech is provided within FGD, see (Panevová, 1999; Panevová and Mikulová, 2007), being partially reflected in the Prague Dependency Treebank annotation scenario (henceforth PDT) (Hajič et al., 2018).

Reciprocity is characterized by the fact that two (or sporadically three) semantic participants of the situation denoted by a predicate are involved in a mutual relation and this mutual relation is linguistically structured within a single predicate structure. In Czech, verbs (16a), nouns (16b), adjectives (16c), and adverbs (16d) can function as reciprocal predicates (reciprocity of adverbs are left aside here as language data allowing for their analysis are too sparse).

(16) a. *Petr a Pavel se / sebe (vzájemně) střídali ve vyprávění.*
'Peter and Paul changed each other in talking.'

b. *obava přátel o sebe (navzájem)*
'friends' fear for each other'

c. *lhostejní k sobě navzájem*
'indifferent to each other'

d. *daleko od sebe*
'far from each other'

Within reciprocal predicates, two groups can be distinguished: *lexical* and *syntactic reciprocal predicates*. The former group of reciprocal predicates is limited in Czech; these predicates comprise the semantic trait of mutuality in their lexical meaning (e.g., *debatovat* 'to debate', *dohodnout se* 'to agree').

[11] The only exception is represented by verbal nouns systematically derived from verbs.

In contrast, the latter one is broader; it includes predicates that – despite not having the trait of mutuality – allow some of their participants to enter into this relation (e,g., *děkovat* 'to thank', *budit* 'to wake up sb').[12]

For expressing mutuality, syntactic reciprocal predicates make use of the morphosyntactic operation of reciprocalization, applied to their valency frames.[13] This operation can be applied to lexical reciprocal predicates as well, serving, however, a different function: it allows to make the affected semantic participants equal with respect to their participation in the event expressed by a predicate, see esp. (Gleitman et al., 1996).

Similarly as for reflexivity, see Section 3, the applicability of the operation of reciprocalization should be described in the lexicon, as it cannot be determined only on the basis of the valency structure of predicates, while the operation itself is regular enough to be described by rules contained in the grammar.

Reciprocity in Deep Structures. In the deep syntactic structure of reciprocal constructions, the number and type of valency complementations of a predicate are preserved. However, the mapping of semantic participants onto valency complementations is changed: two semantic participants, which – in contrast to reflexivity – refer to distinct referents, are symmetrically mapped onto valency complementations. This complex mapping is then reflected as a coreferential link between the valency complementations involved in reciprocity.

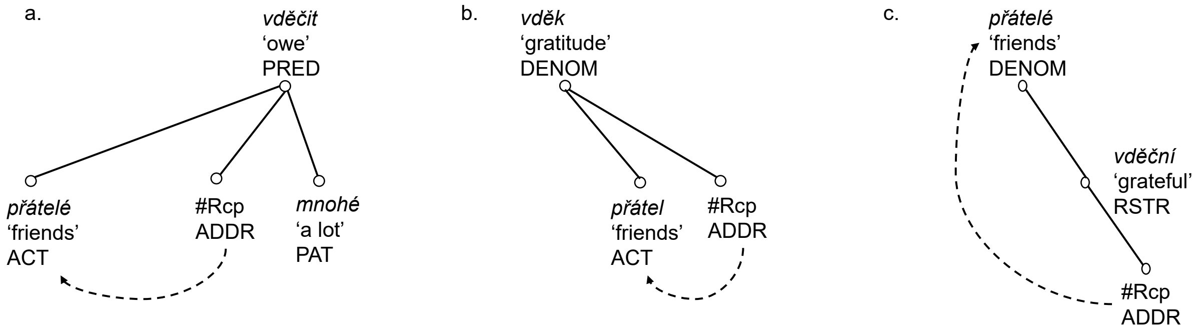


Figure 3: The simplified tectogrammatical trees of examples (18), (20) and (22a), respectively; the dashed arrow shows coreference.

Let us demonstrate the complex mapping of semantic participants onto valency complementations on the syntactic reciprocal predicates from the same derivational family, the verb *vděčit* 'to owe, to be grateful', the noun *vděk* 'gratitude', and the adjective *vděčný* 'grateful'. All these predicates evoke three semantic participants: 'Experiencer', 'Addressee', and 'Reason'. With the verb and the noun, these participants are mapped onto their ACT, ADDR, and PAT, respectively, see the valency frames in (17) and (19). In contrast, with the adjective, only 'Addressee' and 'Reason' can be syntactically structured as its ADDR and PAT, respectively, see the valency frame in (21); 'Experiencer' occurs in the deep structure either as the governor of the adjective, or as ACT of the copula verbs *být* 'to be' and *stávat se / stát se* 'to become' with the adjective (see Section 3 as well).

With these three predicates, participants 'Experiencer' and 'Addressee' can be involved in reciprocity.

[12] In the VALLEX lexicon, there are 241 lexical units of verbs (represented by 319 verb lemmas) indicated as lexical reciprocal verbs. In addition, 1.687 lexical units of verbs are classified there as syntactic reciprocal verbs (`http://quest.ms.mff.cuni.cz/vallex/`). In PDT, however, a vast majority of annotated reciprocal constructions of verbs is formed by lexical reciprocal verbs (411 instances of lexical reciprocal verbs, represented by 133 verb lemmas, out of the overall 439 instances); only in 28 instances, syntactic reciprocal verbs occur, represented by 35 verb lemmas; however, in the manual analysis of reflexive constructions in PDT (see footnote 4), it occurred that 58 other instances of verbal reciprocal structures (49 syntactic reciprocals and 9 lexical reciprocals) were incorrectly annotated as reflexive constructions. In VALLEX, data for nouns and adjectives are not available; in PDT, 558 instances of reciprocity with nouns and 2 instances of reciprocity with adjectives are annotated (plus 7 instances of lexical reciprocity with adjectives were incorrectly annotated as reflexivity). In both data resources, only those cases were counted where reciprocity affects actants.

[13] Conditions of its applicability (esp. semantic homogeneity of semantic participants and their same status with respect to topic-focus articulation) have been described in (Panevová, 1999).

In such a case, with the verb and the noun, both ‘Experiencer’ and ‘Addressee’ are mapped onto ACT and at the same time onto ADDR, see the scheme of this mapping in Figure 4, examples (18) and (20) and their simplified dependency trees in Figure 3a and 3b; with the adjective, the complex mapping involves ADDR from the valency frame of the adjective and either the governor of the adjective, or ACT of the copula verbs *být* ‘to be’ and *stávat se / stát se* ‘to become’ with the given adjective, see the scheme in Figure 4, examples (22a-b) and the simplified tree of example (22a) in Figure 3c.

(17) *vděčit*^{impf} ‘to owe, to be grateful’: ACT$_{\text{nom}}$ ADDR$_{\text{dat}}$ PAT$_{\text{za+acc,dcc}}$

(18) *Přátelé / Němci a Češi si / sobě (vzájemně) vděčili za mnohé.*
friends / Germans and Czechs REFL$_{clitic.dat}$ / REFL$_{long.dat}$ (mutually) owed for a lot
‘Friends / Germans and Czechs owed each other a lot.’

(19) *vděk* ‘gratitude’: ACT$_{\text{gen,pos}}$ ADDR$_{\text{dat,k+dat,vůči+dat}}$ PAT$_{\text{za+acc,dcc}}$

(20) *vděk přátel k sobě (navzájem)*
gratitude of friends to REFL$_{long.dat}$ (mutually)
‘gratitude of friends to each other’

(21) *vděčný* ‘grateful’: ADDR$_{\text{dat,vůči+dat}}$ PAT$_{\text{za+acc,dcc}}$

(22) a. *přátelé vděční sobě (navzájem)*
friends grateful REFL$_{long.dat}$ (mutually)
‘friends grateful to each other’

b. *Přátelé jsou si / sobě (navzájem) vděční.*
friends are REFL$_{clitic.dat}$ / REFL$_{long.dat}$ (mutually) grateful
‘Friends are grateful to each other.’

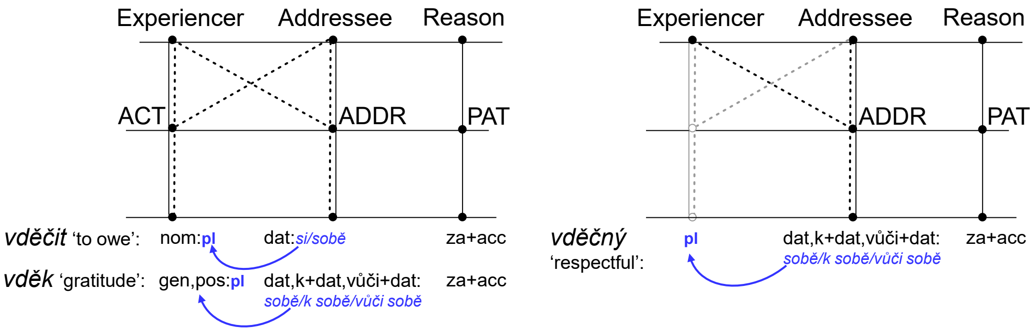


Figure 4: The scheme of the mapping of semantic participants of the verb *vděčit* ‘to owe, to be grateful’, the noun *vděk* ‘gratitude’, and the adjective *vděčný* ‘grateful’ onto valency complementations and surface positions (the solid line depicts unreciprocal structures, the dashed line illustrates reciprocal structures).

Reciprocity in Surface Structures. With reciprocal verbs and nouns, the operation of reciprocalization involves two surface syntactic positions provided by the reciprocalized valency complementations. With reciprocal adjectives, only one surface position provided by the adjectival complementation is available; the second position is typically outside the adjectival structure, given by the governor of adjectives or by ACT of copula verbs.[14]

The syntactically more prominent surface position is pluralized; it can be filled with plural nouns, coordination, see example (18), and collective nouns (e.g., *třída* ‘class’, *družstvo* ‘team’, *posádka* ‘crew’). With verbs, the more prominent position is mostly the position of subject, less frequently the position of

[14] Reciprocity can be optionally emphasized by adverbial modifiers *navzájem, vzájemně* ‘mutually’; in specific cases, the modifiers *spolu* ‘together’ or *mezi sebou* ‘between each other’ can be used as well.

direct object.[15] With nouns, it is represented by the attribute position corresponding to the subject (or direct object) position with their base verbs. With adjectives, the more prominent position is the surface position external to adjectival structure.

For example, with the verb *vděčit* 'to owe, to be grateful' and the noun *vděk* 'gratitude', with which ACT and ADDR are involved in reciprocity, the pluralized more prominent position is given by ACT; this ACT contributes subject to the verbal structure, see example (18), and the corresponding attribute position to the nominal structure (20). In contrast, with the adjective *vděčný* 'grateful', the pluralized position is outside the adjectival structure; typically the governor of the adjective or ACT of a copula verb are pluralized, examples (22a-b).

The less prominent surface position is either deleted from the surface, or if expressed, it can be filled with the reflexive pronoun, or with the expression *jeden druhý* 'each other', both coreferential with the expression in the more prominent position.[16]

The surface expression of the less prominent position is primarily conditioned by (i) morphemic forms of the valency complementation providing the given position and by (ii) a part-of speech of a reciprocal predicate. First, if the valency complementation has the prepositional form *s*+Instr, it is systematically deleted from the surface, regardless of the part-of-speech of its governor; see the valency frame and example of the verb *cítit* 'to sympathize' (23), the frame and the example of the noun *soucit* 'sympathy' (24), and the frame and the example of the adjective *soucitný* 'sympathetic' (25).

(23) a. *cítit*^{impf} 'to sympathize': ACT$_{nom}$ PAT$_{s+instr}$
b. *Lidé spolu v těžkých dobách více cítili.*
people together in difficult times more sympathized
'People sympathized more with each other in difficult times.'

(24) a. *soucit* 'sympathy': ACT$_{gen,pos}$ PAT$_{k+dat,nad+instr,s+instr}$
b. *vzájemný soucit lidí k sobě / nad sebou*
mutual sympathy of people to REFL$_{long.dat}$ / over REFL$_{long.instr}$
'mutual sympathy of people'

(25) a. *soucitný* 'sympathetic': PAT$_{k+dat,nad+instr,s+instr}$
b. *lidé soucitní k sobě / nad sebou navzájem*
people sympathetic to REFL$_{long.dat}$ / over REFL$_{long.instr}$ mutually
'people sympathetic with each other'

Second, if the valency complementation providing the less prominent position has the form of the prepositionless dative or accusative, it can have either the clitic form, or the long form, depending on its position in a sentence and a part of speech of the predicate; with verbs, both forms are available, see example (18), while with nouns and adjectives, only the long forms of the reflexive pronoun are possible, see esp. (Dvořak, 2017).[17]

Last, if the valency complementation giving the less prominent position has other forms than the prepositionless dative or accusative, only the long forms of the reflexive are available, see examples (20), (24b), (25b).

Conclusion

In this paper, we have addressed reflexives in Czech, with an emphasis on the reflexive pronoun. We have proposed their analysis in the Functional Generative Description, supported by data provided in the

[15] In Czech, the direct object position as the more prominent one is mostly involved in reciprocalization with lexical reciprocal verbs. For example, with the lexical reciprocal verb *spojovat / spojit* 'to combine' with the valency frame ACT$_{nom}$ ADDR$_{s+instr}$ PAT$_{acc}$ EFF$_{do+gen,v+acc}$, reciprocalization affects ADDR and PAT, the more prominent position thus being represented by direct object, provided by the accusative PAT (e.g., *Hra spojuje rysy*$_{PAT}$ *komiky a horroru.* 'The play combines comedian and horror features$_{PAT}$.').

[16] The expression *jeden druhý* 'each other', which unambiguously marks reciprocity, is not discussed here (Kettnerová and Lopatková, 2018a).

[17] In constructions with copula verbs, the clitic form of the reflexive pronoun can occur with adjectives as well, see example (22b). However, these constructions require further attention.

VALLEX lexicon and in the syntactically annotated Prague Depencency Treebank. We stress the fact that for their adequate representation both components of the language description – the lexicon and the grammar – must be taken into account.

To conclude, our in-depth analysis of deep and surface syntactic properties of Czech reflexive and reciprocal constructions allows us to explicitly formulate the conditions underlying *ambiguity between reflexivity and reciprocity*, which – to our best knowledge – have not been described yet: (i) the same pair of valency complementations must be affected by reflexivity and reciprocity with a single predicate (as a result, an identical pair of valency complementations are linked by coreference), (ii) the more prominent surface position is represented by the syntactic subject (and the corresponding positions with nouns and adjectives), and (iii) the antecedent of the reflexive pronoun is plural.

Acknowledgements

The research reported in this paper has been supported by the Czech Science Foundation (GAČR), grant No. 18-03984S *Between Reciprocity and Reflexivity: The Case of Czech Reciprocal Constructions* and partially also by the *LINDAT/CLARIN* project No. LM2015071.

This work has been using language resources stored and distributed by the *LINDAT-CLARIN* project of the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2015071.

References

Igor Boguslavsky. 2003. On the Passive and Discontinuous Valency Slots. In *Proceedings of the 1st International Conference on Meaning-Text Theory*, pages 129–138, Paris. École Normale Supérieure.

Noam Chomsky. 1981. *Lectures on Government and Binding.* Foris, Dordrecht.

František Daneš, Miroslav Grepl, and Zdeněk Hlavsa, editors. 1987. *Mluvnice češtiny 3.* Academia, Praha.

Mojmír Dočekal. 2008. *Vázané proměnné v češtině.* Ph.D. thesis, Masaryk University, Faculty of Arts, Brno.

Věra Dvořak. 2017. Dějové substantivum. In Petr Karlík, Marek Nekula, and Jana Pleskalová, editors, *Nový encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha.

Nicholas Evans, Alice Gaby, Stephen C. Levinson, and Asifa Majid, editors. 2011. *Reciprocals and Semantic Typology*, volume 98 of *Typological Studies in Language*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Leonard M. Faltz. 1985. *Reflexivization: a study in universal syntax.* Garland Publishing, New York.

Zygmunt Frajzyngier and Traci Walker, editors. 2000a. *Reciprocals. Forms and Functions*, volume 41 of *Typological Studies in Language*. John Benjamins, AmsterdamPhiladelphia.

Zygmunt Frajzyngier and Traci Walker, editors. 2000b. *Reflexives. Forms and Functions*, volume 40 of *Typological Studies in Language*. John Benjamins, AmsterdamPhiladelphia.

Emma Geniušienė. 1987. *The typology of reflexives.* Mouton de Gruyter, Berlin–New York–Amsterdam.

L. R. Gleitman, H. Gleitman, C. Miller, and R. Ostrin. 1996. Similar, and similar concepts. *Cognition*, 58:321–376.

Miroslav Grepl and Petr Karlík. 1998. *Skladba češtiny.* Votobia, Olomouc.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2018. Prague dependency treebank 3.5. (Dostupné z `http://ufal.mff.cuni.cz/pdt3.5.`).

Bohuslav Havránek. 1928. *Genera verbi v slovanských jazycích I. Nová řada (VIII).* Kr. česká spol. nauk, Praha.

Andrea Hudousková. 2009. Dvě funkce klitiky *se*: různé, a přeci stejné. *Slovo a slovesnost*, 70(4):295–304.

Petr Karlík. 1999. Reflexiva v češtině. In E. Rusínová, editor, *Přednášky a besedy z 32. běhu Letní školy slovanských studií*, pages 44–52. Filozofická fakulta Masarykovy univerzity, Brno.

Suzanne Kemmer. 1993. *The Middle Voice*. John Benjamins, Amsterdam–Philadelphia.

Václava Kettnerová and Markéta Lopatková. 2018a. Lexicographic Potential of the Syntactic Properties of Verbs: The Case of Reciprocity in Czech. In *XVIII EURALEX International Congress, Lexicography in Global Contexts*, pages 685–698, Ljubljana. Ljubljana University Press, Faculty of Arts.

Václava Kettnerová and Markéta Lopatková. 2018b. Mezi reflexivitou a reciprocitou: Poznámky k reflexivním a recipročním konstrukcím vybraných českých sloves. *Prace Filologiczne*, LXXII:131–145.

Václava Kettnerová, Markéta Lopatková, and Jarmila Panevová. 2014. An interplay between valency information and reflexivity. *The Prague Bulletin of Mathematical Linguistics*, 102:105–126.

Veronika Kolářová. 2014. Special valency behavior of Czech deverbal nouns. In Olga Spevak, editor, *Noun Valency*, pages 19–60. John Benjamins, Amsterdam.

Miroslav Komárek, Jan Kořenský, Jan Petr, and Jarmila Veselková, editors. 1986. *Mluvnice češtiny 2*. Academia, Praha.

Miroslav Komárek. 2001. Několik poznámek k reflexi reflexivity reflexiv. *Slovo a slovesnost*, 62(3):207–209.

Ekkehard König and Volker Gast, editors. 2008. *Reciprocals and Reflexives: Theoretical and Typological Explorations*. Mouton de Gruyter, Berlin.

Ekkerhard König and Shigehiro Kokutani. 2006. Towards a typology of reciprocal constructions: Focus on German and Japanese. *Linguistics*, 44:271–302.

Markéta Lopatková, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. 2016. *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.

Lucie Medová. 2009. *Reflexive Clitics in the Slavic and Romance Languages. A Comparative View from an Antipassive Perspective*. Ph.D. thesis, Princeton University, Princeton, NJ, USA.

Igor A. Mel'čuk. 2004. Actants in Semantics and Syntax I. *Linguistics*, 42(1):1–66.

Marie Mikulová, Alevtina Bémová, Jan Haji, Eva Hajiová, Jií Havelka, Veronika Koláová, Lucie Kuová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan tpánek, Zdeka Ureová, Kateina Veselá, and Zdenk abokrtský. 2006. *Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual*. ÚFAL MFF UK, Prague, Czech Rep.

Vladimir P. Nedjalkov, editor. 2007. *Reciprocal Constructions*. John Benjamins, Amsterdam–Philadelphia.

Karel Oliva. 2000. Hovory k sobě/si/sebe/se. In Zdeňka Hladká and Petr Karlík, editors, *Čeština univerzália a specifika. 2. Sborník z konference ve Šlapanicích u Brna 17.-19.11.1999*, pages 161–171, Brno. Masarykova univerzita.

Karel Oliva. 2001. Reflexe reflexivity reflexiv. *Slovo a slovesnost*, 62(3):200–207.

Jarmila Panevová and Marie Mikulová. 2007. On reciprocity. *The Prague Bulletin of Mathematical Linguistics*, 87:27–40.

Jarmila Panevová, Eva Hajičová, Václava Kettnerová, Markéta Lopatková, Marie Mikulová, and Magda Ševčíková. 2014. *Mluvnice současné češtiny 2, Syntax na základě anotovaného korpusu*. Karolinum, Praha.

Jarmila Panevová. 1998. Ještě k teorii valence. *Slovo a slovesnost*, 59(1):1–14.

Jarmila Panevová. 1999. Česká reciproční zájmena a slovesná valence. *Slovo a slovesnost*, 60(4):269–275.

Jarmila Panevová. 2001. Problémy reflexivního zájmena v češtině. In *Přednášky z XLIV. běhu Letní školy slovanských studií*, pages 81–88. UK v Praze, FF, Praha.

Jarmila Panevová. 2008. Problémy se slovanským reflexivem. *Slavia*, 77(1-3):153–163.

Petr Piha. 1984. Case frames of nouns. In Petr Sgall, editor, *Contributions to Functional Syntax, Semantics, and Language Comprehension*, volume 16 of *Linguistic and Literary Studies in Eastern Europe*, pages 225–238. John Benjamins, Amsterdam.

Carl Pollard and Ivan A. Sag. 1992. Anaphors in English and the scope of binding theory. *Linguistic Inquiry*, 23(2):261–303.

Tanya Reinhart and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry*, 24:657–720.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

Vojtěch Veselý. 2018. K slovotvorné funkci reflexivních morfémů *se*, *si*. *Naše řeč*, 101(3):138–157.

Coordination of unlike grammatical functions

Agnieszka Patejuk[1,2] **Adam Przepiórkowski**[1,3,4]
[1]Institute of Computer Science, Polish Academy of Sciences
[2]Faculty of Linguistics, Philology and Phonetics, University of Oxford
[3]Institute of Philosophy, University of Warsaw
[4]Wolfson College, University of Oxford
aep@ipipan.waw.pl adamp@ipipan.waw.pl

Abstract

The aim of this paper is to propose a dependency analysis of coordination of unlike grammatical functions, as witnessed in Slavic and some neighbouring languages (including Romanian, Hungarian and West Armenian). In order to increase the practical impact of the analysis, the proposed representations adhere to Universal Dependencies, a syntactic corpus annotation scheme, though arguments are given for validity of such representations from the theoretical linguistic perspective.

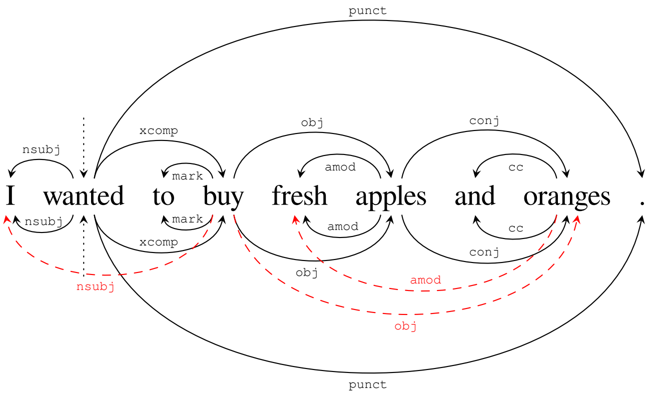
1 Introduction

Coordination is a well-known and long-standing problem for dependency representations of natural language utterances, both in theoretical linguistics and in natural language processing. Representational devices beyond the usual dependency trees are proposed especially for the treatment of coordination in Lucien Tesnière's Dependency Syntax (1959, 2015), Richard Hudson's Word Grammar (1984, 1990, 2010), and Igor Mel'čuk's Meaning–Text Theory (1974, 1988, 2009). Also, the representation of coordination differs widely in different dependency corpora (Popel et al., 2013).

Coordination is also problematic for Universal Dependencies (UD; Nivre et al., 2016; http://universaldependencies.org/). In the current version 2 of the standard, each utterance may be represented by two dependency structures: the basic dependency tree and the enhanced representation, which does not have to be a tree. For example, the two representations of (1) (on one of its interpretations) are shown in (2).[1,2]

(1) I wanted to buy fresh apples and oranges.

(2)



As is clear from the basic dependency tree (above the tokens), coordination is represented in UD as headed by the first conjunct, as in Mel'čuk's Meaning–Text Theory (MTT), but – unlike in that theory – all non-initial conjuncts are `conj` dependents of the initial conjunct, and the coordinating conjunction is a `cc` dependent of the following conjunct. This tree suffers from the usual deficiencies of dependency

[1] This example is based on examples given at http://universaldependencies.org/u/overview/enhanced-syntax.html. All URLs mentioned in this paper were last accessed on 1 April 2019.

[2] In drawing UD representations, the following conventions are adopted in this paper. The basic dependency tree is drawn above the word tokens and the enhanced dependency is drawn below the word tokens. Dependencies which differ between the two representations are drawn as dashed lines in red. The root is marked by a vertical dotted arrow.

trees: it does not represent the fact that *I* is not only the surface subject (`nsubj`; for nominal subject) of the matrix verb *wanted* but also the understood subject of the controlled verb *buy*, or the fact that the adjectival modifier (`amod`) *fresh* is understood here as referring to the whole coordinate structure, *apples and oranges*, rather than just to the first conjunct, *apples*. These deficiencies are corrected in the enhanced dependency structure (below the tokens), where – just as in Lexical Functional Grammar (LFG; Bresnan, 1982, Dalrymple, 2001), Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag, 1987, 1994), and Hudson's Word Grammar (WG) – structure sharing in control constructions is represented explicitly, namely, by the additional `nsubj` dependency. Moreover, the additional `amod` dependency from *oranges* to *fresh* makes it clear that the adjective is shared by the two conjuncts. Finally, the additional `obj` (direct object) dependency from *buy* to *oranges* emphasises the symmetric nature of the two conjuncts with respect to the governing verb *buy*.

One problematic aspect of this representation of coordination, known to the UD community, concerns nested – i.e. immediately embedded – coordination: in the case of three conjuncts, A, B, C, the proposed representation does not distinguish between the flat structure (A,B,C), and the structure in which A and B are conjoined and the resulting coordination is conjoined with C, i.e., $((A,B),C)$.[3] Solutions to this problem are discussed in Przepiórkowski and Patejuk, 2019b. In this paper we deal with another phenomenon problematic for UD, namely, the possibility to coordinate different grammatical functions, as in the attested (3):[4]

(3) [[What]$_{\texttt{obj}}$ and [when]$_{\texttt{advmod}}$] to *eat* to reduce insulin[5]

Such examples violate the overwhelming generalisation that only the same grammatical functions may be coordinated. Normally, languages satisfy this generalisation and attempts to coordinate phrases bearing different grammatical functions result in unacceptability, as in (4)–(5):

(4) *I and an apple have already eaten.
(intended meaning: I have already eaten an apple.)

(5) *I have already eaten an apple and today.
(intended meaning: I have already eaten an apple today.)

The utterance (4) is unacceptable as it involves a coordination of a subject, *I*, and a direct object, *an apple*. Similarly, (5) involves a coordination of a direct object, *an apple*, and an adjunct, *today*.

The assumption that all conjuncts must bear the same relation to the external head is also explicitly made in dependency grammars, e.g. (Hudson, 1984, 225):

> [W]e need to make sure that, in some sense, all the conjuncts in a coordinate structure have the same external relations... If we mix up conflicting external relations, the result is zeugma (e.g. *He came in {(a hurry) and (a taxi)}*, where the conjuncts require conflicting meanings of *in*), or sheer incoherence (e.g. *I ate potatoes and in the kitchen*).

This is also implicitly assumed in constituency- and constraint-based approaches, e.g., in LFG, where – in the f(unctional)-structure – the whole coordinate structure is the value of an attribute such as SUBJ(ect) or OBJ(ect), or belongs to the ADJ(uncts) set, i.e. where all conjuncts bear the same grammatical function.

There are, however, two classes of exceptions – both empirically constrained – to the generalisation that only the same grammatical functions may be coordinated. The first, sylleptic zeugma, is mentioned in the above quote from Hudson, 1984, 225. Such constructions, in which the two conjuncts evoke two different meanings of the head, have a metalinguistic feel and they are easy to distinguish from genuine coordination. We will not deal with zeugma here. The rest of this paper is devoted to the second class of exceptions, illustrated with the English (3). As discussed in §2, such constructions are robust especially in Slavic and they do not evoke different meanings of the head. §3 examines previous dependency approaches to such constructions, while §4 proposes a UD representation. Finally, §5 concludes the paper.

[3] `http://universaldependencies.org/u/dep/conj.html#nested-coordination`

[4] This phenomenon should be carefully distinguished from the coordination of unlike grammatical *categories*, relatively uncontroversial in contemporary linguistics (cf., e.g., Sag et al., 1985, Bayer, 1996, Dalrymple, 2017, but also Bruening and Al Khalaf, 2019 for dissent).

[5] `https://www.dietdoctor.com/what-and-when-to-eat-to-reduce-insulin`

2 Lexico-Semantic Coordination

The phenomenon in question is the so-called lexico-semantic coordination (Sannikov, 1979, 1980), also known as hybrid coordination (Chaves and Paperno, 2007). It occurs mainly in Slavic (incl. Bulgarian, Croatian, Polish and Russian) and in some neighbouring languages (Romanian, Hungarian, West Armenian), as well as – though significantly constrained – in English, French, German, Dutch, Italian and Spanish (Paperno, 2012, Lipták, 2012, Bîlbîie and Gazdik, 2012). In the case of these Germanic and Romance languages, the phenomenon seems to be limited to the coordination of optional *wh*-items (Gračanin-Yüksek, 2007, Lipták, 2012) – e.g., an adjunct and an optional argument – and often occurs in titles, as in (3) above. In the case of the "Slavic sprachbund", the phenomenon is much more robust.

First of all, as discussed in Patejuk and Przepiórkowski, 2012a,b and in Paperno, 2012, in Slavic such constructions are not limited to *wh*-items, although they particularly often involve such items; most of the examples in this paper are of this kind. But apart from *wh*-items, coordination of different grammatical functions may involve negative pronouns (so-called *n*-words; cf., e.g., the Russian (11) and the Polish (16)), certain items expressing existential or universal quantifiers (the latter illustrated in (6) below), and items belonging to a number of other pronominal or quantificational classes.[6] Second, the coordinated items may be obligatory arguments, as in (6) from the National Corpus of Polish (NKJP; Przepiórkowski et al., 2011, 2012), cited here after Patejuk and Przepiórkowski, 2012b, 463.[7]

(6) *Obiecać* można [[wszystko]$_{\texttt{obj}}$ i [wszystkim]$_{\texttt{iobj}}$]. (Polish)
promise.INF may everything.ACC and everyone.DAT
'One may promise everything to everyone.' (NKJP)

Hence, lexico-semantic coordination cannot easily be analysed via 'conjunction reduction', i.e., as some kind of ellipsis; arguments against such an analysis are reviewed in Patejuk, 2015, §5.4.[8] Third, such constructions are textually frequent and often occur in carefully edited texts; there is nothing marginal about them in the languages in which they occur.

Given that lexico-semantic coordination violates the 'same grammatical function' generalisation, it might seem that perhaps it does not involve coordination at all, i.e. that *i* 'and' in (6) and *and* in (3) are not really conjunctions here, but some homophonous elements of a different grammatical class. There are strong arguments against this view. First, in all languages which allow for joining different grammatical functions the joining element has the same form as a conjunction; on the homonymy view, this perfect synchronous correlation is somewhat surprising (even if it may be justified diachronically). Second, and more importantly, as shown in Patejuk and Przepiórkowski, 2012b and in Patejuk, 2015, other typical conjunctions may also occur in such constructions – for instance *lub* 'or', see (7), and *ani* 'nor', see (8).[9] It is worth noting that the conjunction *ani* has a special property – it is an *n*-word, so it requires negation. It retains this property when it combines unlike grammatical functions. As a consequence, the hypothesis that *ani* in (8) is not a conjunction seems like a typical missed generalisation: it combines two items just like conjunctions do, it has the same form as a conjunction, it also has the same properties with respect to negation as that very conjunction. Finally, lexico-semantic coordination may also occur with preconjunctions ('both… and…', 'not only… but also…'), as shown in (9), and it is possible to coordinate more than two items, see (10).

(7) …kto lub czego będzie w Wikipedii szukał. (Polish)
who.NOM or what.GEN will in Wikipedia seek
'…who will seek what in Wikipedia.' (NKJP)

(8) Rząd USA *(nie) ujawnia, kogo ani dlaczego umieścił na liście osób… (Polish)
government USA NEG discloses who.ACC and why put on list people
'The US government does not disclose who and why they put on the list of people…'[10]

[6] See Przepiórkowski and Patejuk, 2014 and, especially, Patejuk, 2015, §5.8 for a comprehensive list of such classes in Polish.

[7] The labels `obj` and `iobj` reflect how this example would be annotated in Polish UD treebanks; cf. Patejuk and Przepiórkowski, 2018.

[8] For this reason we do not refer in this paper to dependency analyses of gapping, non-constituent coordination and the like.

[9] The asterisk before brackets in (8) means that the sentence is ungrammatical if the bracketed material is omitted.

[10] `http://wyborcza.pl/1,76842,15826586,Amerykanie_maja_tajna_liste__nielotow___Trafisz_na.html`

(9) …kiedy wyjawisz nie tylko kto, ale i dlaczego otrzymał awans. (Polish)
when disclose not only who.NOM but and why received promotion
'…when you explain not only who, but also why got promoted.'[11]

(10) Kto, kiedy i dla kogo napisał te wiersze? (Polish)
who.NOM when and for who.GEN wrote these poems
'Who, when and for whom wrote those poems?'[12]

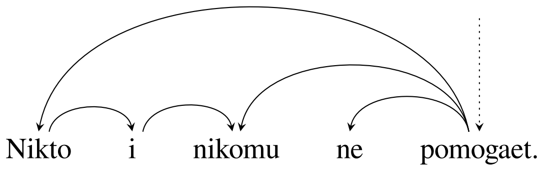
Hence, the combining words should be analysed as true conjunctions, and the phenomenon in question – as true coordination.

3 Theoretical Dependency Approaches

While coordination of unlike grammatical functions has attracted some attention in various linguistic theories, including Transformational Grammar (e.g. Lipták, 2012, Citko and Gračanin-Yüksek, 2013), Categorial Grammar (Paperno, 2012), HPSG (Chaves and Paperno, 2007, Bîlbîie and Gazdik, 2012) and LFG (Gazdik, 2010, Patejuk and Przepiórkowski, 2012a,b, Patejuk, 2015), to the best of our knowledge, it has not been given a serious analysis in dependency theories. Mel'čuk referred to lexico-semantic coordination in a couple of works, but only in passing: in an endnote in Mel'čuk, 1988 (and in Mel'čuk and Pertsov, 1987), then in the main text in Mel'čuk, 2009. Mel'čuk, 1988, 40 states that "There are more complicated cases of double dependency challenging the adequacy of D[ependency]-trees", illustrating this using example (11) together with representation in (12), both based on similar examples in Sannikov, 1979:

(11) Nikto i nikomu ne pomogaet. (Russian)
nobody.NOM and nobody.DAT not helps
'Nobody helps anybody.'

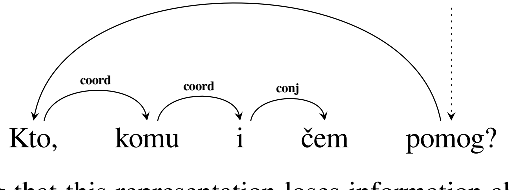
(12)



Mel'čuk, 1988 distances himself from this representation, claiming that "*nikomu* does not depend on *pomogaet* syntactically". Mel'čuk, 2009, 81 returns to lexico-semantic coordination and briefly considers the Russian sentence (13), which would receive a representation like (14) within his Meaning–Text Theory:

(13) Kto, komu i čem pomog? (Russian)
who.NOM who.DAT and what.INS helped
'Who helped whom with what?'

(14)



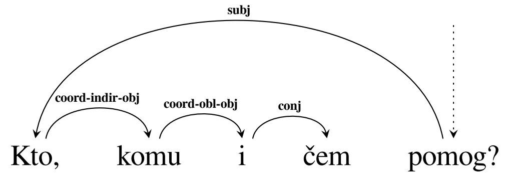
Noting that this representation loses information about grammatical functions of non-initial conjuncts, he proposes to "introduce some special [dependency relations] just for this very special construction: **coord-subj**, **coord-dir-obj**, **coord-indir-obj**, etc." So, presumably, (13) should be represented as (15).[13]

[11] https://www.hbrp.pl/b/dobry-system-awansow-pracownikow-to-awans-dla-calej-firmy/P1C5cBsu1

[12] Danielewiczowa, 1996, 85

[13] It is not clear to us that the two relations assigned to *komu* 'who.DAT' and *čem* 'what.INS' in Meaning–Text Theory are in fact indirect object and oblique object, as indicated in (15), but this does not matter for the argument in the main text.

(15)



Apart from the problem of duplicating many syntactic relations as **coord-**relations, this suggestion is based on the assumption that all conjuncts in lexico-semantic coordination must be dependents of the same head. As demonstrated in Patejuk and Przepiórkowski, 2012b and Patejuk, 2015 on the basis of numerous examples such as the following, this assumption is false:

(16) Nic i nikogo nie może tłumaczyć. (Polish)
nothing.NOM and nobody.GEN NEG can excuse.INF
'Nothing may excuse anybody.' (NKJP)

(17) Czego i ile trzeba dostarczyć organizmowi? (Polish)
what.GEN and how much.ACC should provide.INF organism.DAT
'What – and how much – should one provide one's organism with?'[14]

(18) Jakie i kto może ponieść konsekwencje? (Polish)
what.ADJ.ACC and who.NOM can bear.INF consequences.ACC
'Who can suffer what consequences?'[15]

In (16), *nic* 'nothing.NOM' is the subject of the matrix verb *może* 'may', as well as the understood subject of the controlled verb *tłumaczyć* 'excuse', while *nikogo* 'nobody.GEN' is the direct object of the controlled verb (only). In (17), adopting the common and well-founded assumption in Polish structural and formal linguistics (e.g., Saloni and Świdziński, 2001) that numerals – not nouns – are heads of numeral phrases, the interrogative numeral *ile* 'how much.ACC' is the direct object of *dostarczyć* 'provide, supply', while *czego* 'what.GEN' is a dependent of this numeral.[16] Finally, in (18), *jakie* 'what.ACC' is the adjectival modifier of *konsekwencje* 'consequences', which is the object of *ponieść* 'suffer', which in turn is the infinitival complement of the main verb, *może* 'may', whose subject is *kto* 'who.NOM'.

On Mel'čuk's proposal, the **coord** dependency between conjuncts in (16)–(18) would not only have to encode the grammatical function (**coord-subj**, **coord-dir-obj**, **coord-indir-obj**, etc.), but also information about the actual head of each non-initial conjunct. It is not clear how this information could be encoded within a constrained set of dependency relations (52 in Mel'čuk, 2009). A potential solution, to be considered in more detail below, would be to devise a special labelling convention to account for this phenomenon, where – for each non-initial conjunct – a single dependency label would encode the entire dependency chain to this conjunct. However, this would result in a potentially infinite number of dependency labels.

4 Lexico-Semantic Coordination in UD

4.1 Proposal

How can coordination of unlike grammatical functions be represented in UD? Let us start with the simple (but attested) Polish example (19), where two different dependents of the same head are coordinated: the subject (*kto* 'who.NOM') and the object (*kogo* 'whom.ACC') of the verb *zdradził* 'betrayed'. Since there is no discussion of how to annotate lexico-semantic coordination in UD guidelines, the representation in (20) follows general guidelines related to coordination:

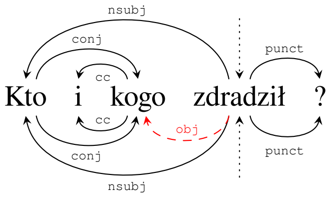
[14] https://vitalia.pl/forum22,446761,0_Czego-i-ile-trzeba-dostarczyc-organizmowi.html

[15] Patejuk, 2015, 99

[16] As discussed in §4.3, in UD the relation between the numeral and the noun is reversed, but the two conjuncts in (17) are still dependents of different heads.

(19) Kto i kogo zdradził? (Polish)
who.NOM and who.ACC betrayed
'Who betrayed whom?'[17]

(20)

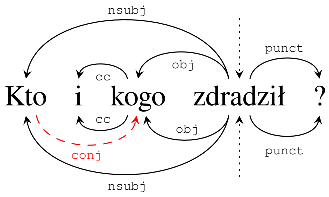


In the basic dependency tree (edges above words), the first conjunct, *kto*, is the subject of *zdradził*, but the information that *kogo* is the object of *zdradził* is lost. This is because *kogo* is annotated as the second conjunct using the `conj` relation – since the basic representation must be a tree, there must not be another incoming relation (object). In effect, the coordination *kto i kogo* 'who.NOM and whom.ACC' is annotated as if it were the subject – which is not true in the case of the second conjunct – and there is no information that *zdradził* has an object. This problem is mitigated in the enhanced dependency representation (edges below words), which lacks such a restriction – the object dependency from *kogo* to *zdradził* is present in the graph, which shows that it is not its subject (despite the basic representation). As a result, appropriate grammatical functions are only provided in the enhanced dependency representation.

Note that it does not seem appropriate to think about the basic representation in (20) as an 'underspecified' version of the more detailed enhanced structure. On such an 'underspecification' view, the basic representation of ordinary coordinated subjects, as in *John and Mary arrived*, would also have to be considered 'underspecified', with the grammatical functions of non-initial conjuncts (here: *Mary*) to be ascertained only upon careful inspection of the enhanced representation. This view is not only questionable conceptually, but also untenable practically: as popular dependency parsers only deal with basic dependency trees – and are unable to learn from or parse with enhanced dependency representations – the lossy and misleading information about grammatical functions at the basic tree level translates into errors in downstream applications, especially those which rely on grammatical functions to extract information about who did what to whom.[18]

For these – conceptual and practical – reasons we propose the alternative basic UD representation in (21); while the enhanced dependency graph is the same as in (20), the basic dependency tree does not include the `conj` dependency between the two conjuncts, i.e. coordination is not fully represented at this level, but the much more important information about grammatical functions is: in (21) *kogo* is identified as the object of *zdradził* at both levels of dependency representation.

(21)



The more theoretical reason for preferring (21) to (20) as a representation of (19) is that coordination plays here a very different role than usual: it does not conjoin phrases which stand in the same syntactic and semantic relation to the head, but rather it joins elements which only have the same information structure status in the sentence. This difference between standard coordination and the coordination of unlike grammatical functions discussed in this section, since it is crucial for the syntactic and semantic interpretation of the sentence, should be represented in the basic dependency tree.

4.2 Potential Alternatives

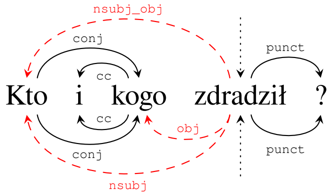
Let us consider a potential alternative solution, inspired by the approach outlined in Mel'čuk, 2009 (see the discussion of (14)), which is aimed at saving the topology of the basic UD representation in (20) by enriching

[17] `http://sliwerski-pedagog.blogspot.com/2018/06/kto-i-kogo-zdradzi.html`

[18] Easiness to extract such relations is an important design goal of UD, as made explicit, e.g., in the following quote: "UD inherits from [Stanford Dependencies] the concern with usefulness for relation extraction [...]" (Silveira and Manning, 2015, 311).

the dependency label from the head to the coordinate structure so that it correctly represents grammatical functions of all conjuncts (instead of suggesting that the entire coordination is the subject), e.g.:

(22)

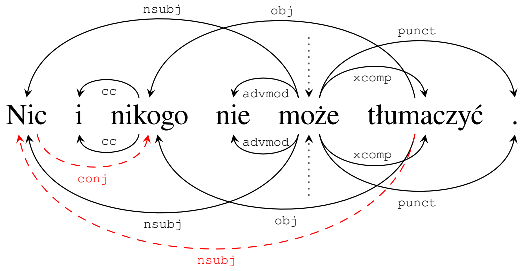


This solution suffers from the same problems as that proposed by Mel'čuk (2009): it greatly multiplies dependency labels (many different numbers and orders of grammatical functions[19] would have to be encoded) and it does not encode information about possibly different heads of particular conjuncts. Consider again (16), repeated below as (23):

(23) Nic i nikogo nie może tłumaczyć. (Polish)
nothing.NOM and nobody.GEN NEG can excuse.INF
'Nothing may excuse anybody.'

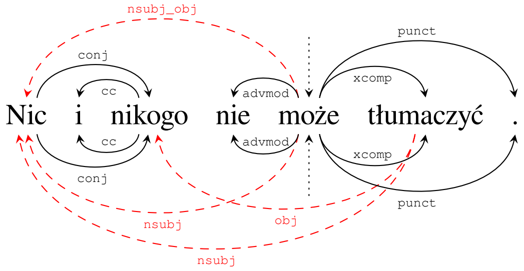
On our proposal, its representation is given in (24) – coordination is not fully represented in the basic tree, but grammatical functions are:

(24)



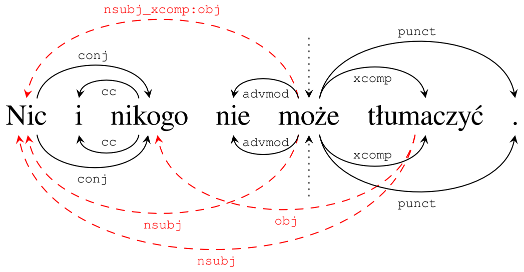
In contrast, the attempt to save the more standard basic representation (the one following from general guidelines) by labelling the dependency from *może* 'can, may' to *nic* 'nothing.NOM' as `nsubj_obj` (instead of `nsubj`) in (25) is misinformative, as it incorrectly suggests that *nikogo* 'nobody.GEN' is the direct object of *może* – rather than the object of *tłumaczyć* 'explain'.

(25)



This problem could be approached in a way analogous to the earlier suggestion on how to modify Mel'čuk's account, namely, by providing – in the basic tree – full paths to non-initial conjuncts, as shown in (26), where the relation targeting *nic* is `nsubj_xcomp:obj` (because *nikogo* is the `obj` of `xcomp`, see (24)):

(26)



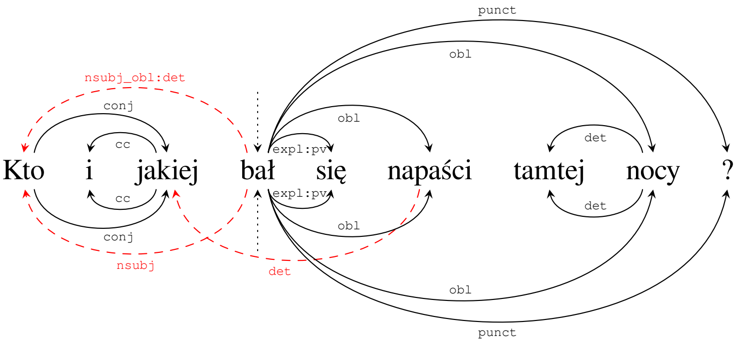
[19] For instance, assuming there are only instances of coordination with 2 conjuncts, each of which has a different grammatical function, this yields $n \times (n-1)$ labels, where n is the number of basic grammatical functions. For 3 conjuncts, there would be $n \times (n-1) \times (n-2)$ labels, and so on.

However, it is clear that such a solution involving full dependency paths as (parts of) dependency labels would further aggravate the issue of the number and complexity of dependency labels.[20] Moreover, in some cases such dependency paths may still be insufficient, e.g. in the case of a predicate with two or more `obl` dependents such that a dependent of one them participates in lexico-semantic coordination, as in (27):

(27) Kto i jakiej bał się napaści tamtej nocy? (Polish)
who.NOM.M and what.ADJ.GEN.F feared.3.SG.M RM assault.GEN.F that.GEN.F night.GEN.F
'Who feared what assault on that night?'

Here, both *napaści* 'assault' and *tamtej nocy* 'that night' are genitive obliques, so the `obl` part of the hypothetical `nsubj_obl:det` dependency from the root verb *bał (się)* 'feared' to *kto* 'who' – the hypothetical head of the lexico-semantic coordinate structure *kto i jakiej* 'who.NOM.M and what.ADJ.GEN.F' – is ambiguous between these two oblique dependents, as shown in (28). Moreover, agreement facts do not help in resolving this ambiguity, because both obliques are feminine, singular, genitive – just like the adjective *jakiej*.[21]

(28)



Hence, we prefer the representation in (24) to hypothetical alternatives shown in (25) and (26) – the proposed solution ensures simple and accurate representation of grammatical functions of coordinated dependents using the standard repertoire of dependency labels, even if coordinated items depend on different heads. This is achieved at the cost of not fully representing the lexico-semantic coordination at the basic level, which is however fully represented in enhanced dependencies.

4.3 Numeral Phrases: A Challenge for UD

Let us now return to (17), repeated below for convenience as (29), which poses an interesting additional challenge to UD:

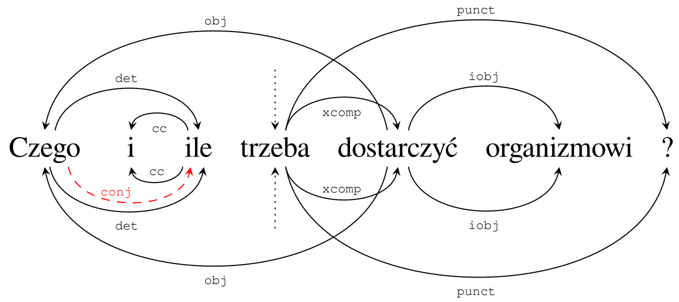
(29) Czego i ile trzeba dostarczyć organizmowi? (Polish)
what.GEN and how much.ACC should provide.INF organism.DAT
'What – and how much – should one provide one's organism with?'

As mentioned above, in Polish – on the standard (non-UD) analysis – the numeral is the head (it receives case marking from the verb), while the accompanying noun is its dependent (it receives case from the numeral). However, following UD guidelines, this dependency relation is reversed: numerals are dependents of nominal heads, so the interrogative numeral *ile* 'how much' is a `det` dependent of *czego* 'what.GEN', which is in turn the direct object of *dostarczyć* 'provide, supply'. One potential problem with the UD representation arises at the level of enhanced dependencies, where *ile* is also a `conj` dependent of *czego*; as shown in (30), there are two different equidirectional dependency relations between these two tokens: `det` and `conj`.

[20] See Schuster et al., 2017, 130–131 for arguments against encoding paths in dependency labels in the context of the UD representation of gapping, the most important of which is that this would introduce an unbounded number of dependency relations.

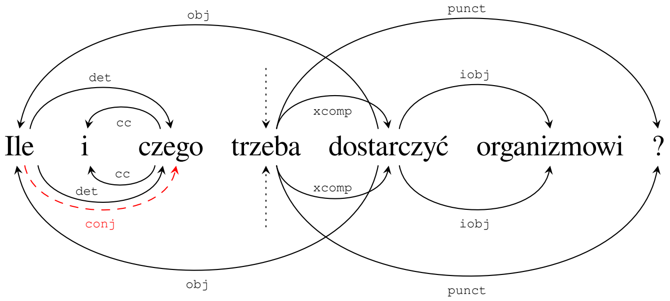
[21] Though relations could be disambiguated by, for instance, adding indices, e.g. `obl1` and `obl2`, but this would further aggravate the problem of number and complexity of labels (resulting in `nsubj_obl1:det`, among others).

(30)



This problem arises regardless of which representation of lexico-semantic coordination – the one proposed here or the one arising from general UD guidelines – is chosen, because the enhanced representations are identical under both. Moreover, this problem is independent of the issue of headedness of Polish numeral phrases: if the UD analysis of numerals were reversed so that the numeral is the head and the noun is the dependent, the problem would resurface in examples such as (29) but with the order of conjuncts reversed (i.e. *Ile i czego trzeba dostarczyć organizmowi?* – also fully acceptable). This is illustrated in (31); the double dependency problem is exactly the same as in (30).

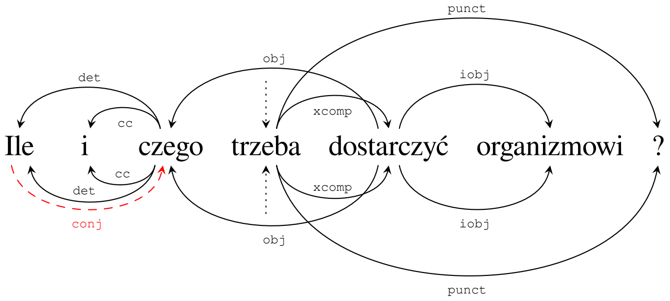
(31)



Rather, the problem stems directly from the UD representation of coordination, which requires that the second conjunct is a conj dependent of the first conjunct: if the second conjunct is independently a non-conj dependent of the first conjunct, the double dependency inevitably arises. Note that this is not a fundamental problem, as – while unmet in UD so far[22] – such double dependencies in the enhanced representation do not seem to violate any deep UD principles, and they could be constrained by well-formedness conditions specifying which dependencies may co-occur this way.

Unfortunately, such constructions also present a more fundamental problem for the standard UD approach to coordination. Let us consider the representation of the sentence from (31) under the current proposal with the usual UD analysis of numeral phrases as headed by nouns:

(32)



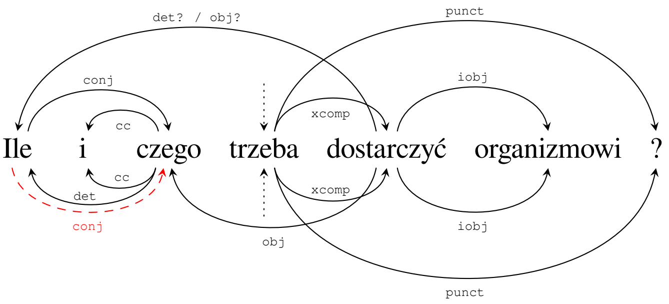
Unlike in (30), there is no problem of two equidirectional relations in (32), but the det edge targeting the first conjunct, *ile* 'how much', originates in the second conjunct, *czego* 'what.GEN', whose incoming edge originates in *dostarczyć* 'provide, supply'. At the level of basic tree lexico-semantic coordination is not represented (there is no conj relation between conjuncts), so the tree is well-formed and does not violate any fundamental UD principles.[23]

[22] However, Przepiórkowski and Patejuk, 2019a discuss a phenomenon which also calls for lifting this ban: Polish particle *się* (sometimes called 'reflexive marker') may have more than one function with respect to the same verb, for instance inherent and impersonal, which also requires two different relations between the same tokens.

[23] The conj relation is present in the enhanced representation, but this representation does not seem to violate any UD rules either: coordination is headed by the first conjunct, this first conjunct has its own a head (even if it is internal to coordination), and

However, an attempt to provide this example with a basic tree including the standard UD representation of coordination fails. Once *czego* is the `conj` dependent of *ile*, the dependency from the verb *dostarczyć* must target *ile*. But, if so, what should be the dependency label on this relation targeting the interrogative numeral? It cannot be `det`, as *ile* is a `det` dependent of *czego*, not of *dostarczyć*; verbs are not supposed to have `det` dependents at all. But it cannot be `obj` either, as this would mean that *ile* stands in the immediate `obj` relation to the verb, and *czego* perhaps does too, depending on the enhanced representation. As the enhanced representation does contain the secondary `obj` dependency from *dostarczyć* to *czego*, the `obj` label in the basic tree would in effect mean that the coordinate structure involves two conjuncts standing in the same `obj` relation to the verb, fully analogous to, for instance, *(Homer) likes donuts and burgers*. This problem is illustrated below.

(33)



So, while the enhanced representation in (33) is the same as in (32), there seems to be no good solution for the basic representation in (33), which assumes that lexico-semantic coordination is represented in the basic tree. Once this assumption is given up, a reasonable representation becomes available, namely, the representation (32) advocated in this paper.

5 Conclusion

In this paper, we presented constructions which violate the principle that only the same grammatical functions may be coordinated – to the best of our knowledge, this is the first comprehensive discussion of such constructions within any dependency framework.

We showed that the few existing suggestions of how lexico-semantic coordination may be analysed in dependency approaches cannot account for the complex data without running into serious problems. Instead, we proposed a UD analysis of such constructions, which represents the vital information about grammatical functions of particular conjuncts on both levels: in basic trees and in enhanced representations. A feature of this representation is that lexico-semantic coordination is fully represented only at the enhanced level, which makes it possible to precisely specify different grammatical functions at the basic level. While non-standard, we believe that – given the very non-standard nature of coordination of unlike grammatical functions – this is an advantage of the proposed representation.

We also demonstrated that the phenomenon of lexico-semantic coordination necessitates giving up the assumption that there is at most one dependency relation from one token to another. (However, on our proposal, such double dependencies only occur at the enhanced level of representation, so they do not violate any deep UD principles.)

We hope that the above considerations will inspire other dependency work on the fascinating topic of coordination of unlike grammatical functions.

Acknowledgements

We thank anonymous SyntaxFest 2019 reviewers and the audience of the 17th International Workshop on Treebanks and Linguistic Theory (University of Oslo, December 2018) for helpful comments. Agnieszka Patejuk gratefully acknowledges the *Mobilność Plus* mobility grant awarded to her by the Polish Ministry of Science and Higher Education.

the `obj` edge to the second conjunct is secondary in a sense (it does not indicate the head of coordination).

References

Samuel Bayer. 1996. The coordination of unlike categories. *Language* 72(3):579–616.

Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. Cambridge, MA: The MIT Press.

Benjamin Bruening and Eman Al Khalaf. 2019. Category mismatches in coordination revisited. *Linguistic Inquiry* Forthcoming.

Gabriela Bîlbîie and Anna Gazdik. 2012. Wh-coordination in Hungarian and Romanian multiple questions. *Empirical Issues in Syntax and Semantics* 9:19–36.

Rui Pedro Chaves and Denis Paperno. 2007. On the Russian hybrid coordination construction. In Stefan Müller, editor, *Proceedings of the HPSG 2007 Conference*, pages 46–64, Stanford, CA: CSLI Publications.

Barbara Citko and Martina Gračanin-Yüksek. 2013. Towards a new typology of coordinated *wh*-questions. *Journal of Linguistics* 49:1–32.

Mary Dalrymple. 2001. *Lexical Functional Grammar*. San Diego, CA: Academic Press.

Mary Dalrymple. 2017. Unlike phrase structure category coordination. In Victoria Rosén and Koenraad De Smedt, editors, *The Very Model of a Modern Linguist*, volume 8 of *Bergen Language and Linguistics Studies*, pages 33–55, Bergen: University of Bergen Library.

Magdalena Danielewiczowa. 1996. *O znaczeniu zdań pytajnych w języku polskim. Charakterystyka struktury tematyczno-rematycznej wypowiedzeń interrogatywnych*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Anna Gazdik. 2010. Multiple questions in French and Hungarian: An LFG account. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'10 Conference*, pages 249–269, Ottawa, Canada: CSLI Publications.

Martina Gračanin-Yüksek. 2007. *About Sharing*. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Richard Hudson. 1984. *Word Grammar*. Oxford: Blackwell.

Richard Hudson. 1990. *English Word Grammar*. Oxford: Blackwell.

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

Anikó Lipták. 2012. Strategies of wh-coordination. *Linguistic Variation* 11:149–188.

Igor Mel'čuk. 1974. *Opyt teorii lingvističeskix modelej «Smysl ⇔ Tekst»*. Moscow: Nauka.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. Albany, NY: The SUNY Press.

Igor Mel'čuk. 2009. Dependency in natural language. In Alain Polguère and Igor Mel'čuk, editors, *Dependency in Linguistic Description*, pages 1–110, Amsterdam: John Benjamins.

Igor Mel'čuk and Nikolaj Pertsov. 1987. *Surface Syntax of English. A Formal Model within the Meaning–Text Framework*. Amsterdam: John Benjamins.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 1659–1666, ELRA, Portorož, Slovenia: European Language Resources Association (ELRA).

Denis Paperno. 2012. *Semantics and Syntax of Non-Standard Coordination*. Ph.D. Thesis, University of California, Los Angeles.

Agnieszka Patejuk. 2015. *Unlike coordination in Polish: an LFG account*. Ph.D. Thesis, Instytut Języka Polskiego PAN, Cracow.

Agnieszka Patejuk and Adam Przepiórkowski. 2012a. A comprehensive analysis of constituent coordination for grammar engineering. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2191–2207, Mumbai, India.

Agnieszka Patejuk and Adam Przepiórkowski. 2012b. Lexico-semantic coordination in Polish. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'12 Conference*, pages 461–478, Stanford, CA: CSLI Publications.

Agnieszka Patejuk and Adam Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically Informed Treebanks of Polish*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Carl Pollard and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. CSLI Lecture Notes, No. 13, Stanford, CA: CSLI Publications.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago, IL: Chicago University Press / CSLI Publications.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zděněk Žabokrtský. 2013. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527, Sofia, Bulgaria.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik. 2011. National Corpus of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 259–263, Poznań, Poland.

Adam Przepiórkowski and Agnieszka Patejuk. 2014. Koordynacja leksykalno-semantyczna w systemie współczesnej polszczyzny (na materiale Narodowego Korpusu Języka Polskiego). *Język Polski* XCIV(2):104–115.

Adam Przepiórkowski and Agnieszka Patejuk. 2019a. From Lexical Functional Grammar to enhanced Universal Dependencies: The UD-LFG treebank of Polish, to appear in *Language Resources and Evaluation* (published online on 4 February 2019).

Adam Przepiórkowski and Agnieszka Patejuk. 2019b. Nested coordination in Universal Dependencies. In *Proceedings of SyntaxFest 2019*.

Ivan A. Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and how to distinguish categories. *Natural Language and Linguistic Theory* 3:117–171.

Zygmunt Saloni and Marek Świdziński. 2001. *Składnia współczesnego języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN, fifth edition.

Vladimir Z. Sannikov. 1979. Sočinitel'nye i sravnitel'nye konstrukcii: ix blizost', ix sintaksičeskoe predstavlenie I. *Wiener Slawistischer Almanach* 4:413–432.

Vladimir Z. Sannikov. 1980. Sočinitel'nye i sravnitel'nye konstrukcii: ix blizost', ix sintaksičeskoe predstavlenie II. *Wiener Slawistischer Almanach* 5:211–242.

Sebastian Schuster, Matthew Lamm, and Christopher D. Manning. 2017. Gapping constructions in Universal Dependencies v2. In Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 123–132, Association for Computational Linguistics, Gothenburg, Sweden.

Natalia Silveira and Christopher Manning. 2015. Does Universal Dependencies need a parsing representation? An investigation of English. In Eva Hajičová and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics (DepLing 2015)*, pages 310–319, Uppsala.

Lucien Tesnière. 1959. *Éléments de Syntaxe Structurale*. Paris: Klincksieck.

Lucien Tesnière. 2015. *Elements of Structural Syntax*. Amsterdam: John Benjamins.

Predicate catenae
A dependency grammar analysis of *it*-clefts

Timothy Osborne
Zhejiang University
Hangzhou, China
tjo3ya@yahoo.com

Abstract

This manuscript proposes a syntactic analysis of *it*-cleft sentences in English in dependency syntax. The connectivity effects of *it*-clefts are addressed in terms of the catena unit. A central claim is that despite the presence of two finite clauses, the matrix predicate of *it*-clefts, which is a catena, reaches into the embedded clause to include the primary predicate residing there. This means that despite the presence of two finite verbs, *it*-clefts are in fact mono-clausal in a central way. Given this essentially mono-clausal status of *it*-clefts, the widely discussed connectivity effects that appear in them are not surprising.

1 Introduction

The connectivity effects of *it*-clefts, pseudoclefts, and specificational copular sentences in general challenge theories of syntax in a central way and have therefore helped give rise to an unending stream of studies into these sentence types over the past few decades (e.g. Akmajian 1970; Gundel 1977; Delahunty 1984, 1986; Heggie 1988; Moro 1997; Heycock & Kroch 1999, Hedberg 2000; Mikkelsen 2004; Reeve 2012; among many others). This contribution demonstrates that a dependency grammar (DG) that acknowledges the *catena* unit (O'Grady 1998; Osborne 2005; Osborne et al. 2012) is in a particularly strong position to account for the connectivity effects of these sentence types. The focus here, however, is on only one of these three sentence types, namely *it*-clefts.

The core phenomenon examined in this manuscript is illustrated with sentence (1):

(1) It was **herself$_1$** that **Jill$_1$** critiqued.

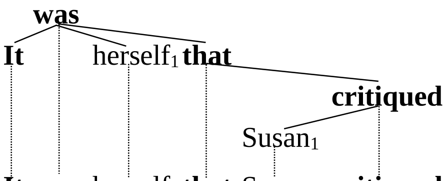
The reading indicated by co-indexation is natural in this case. This is a surprising state of affairs in view of the fact that *herself* appears in the matrix clause, the clause associated with the finite copula *was*, whereas the full noun with which it is co-indexed appears in the embedded clause associated with the finite content verb *critiqued*. Compare sentence (1) with sentence (2) in this regard:

(2) *They told **herself$_1$** that **Jill$_1$** was too critical.

Despite the outward similarity of sentence (2) to sentence (1), sentence (2) is clearly bad. The reflexive pronoun *herself* in the matrix clause cannot take its reference from the full noun *Jill* in the embedded clause. The acceptability contrast across (1) and (2) reveals that *it*-cleft sentences behave uniquely regarding binding patterns. The greater phenomenon is known as *connectivity*. The foregrounded constituent in cleft sentences behaves as though it is "connected" into a simple clause, in the case of (1) the simple clause being *Jill critiqued herself*.

This manuscript demonstrates that a flexible understanding of predicates and their arguments can capture this behavior of *it*-clefts. The central claim is that the matrix predicate in *it*-clefts reaches into the embedded clause to include the main predicate there. The following dependency tree of sentence (1) presents the account in brief:

(3)



a. **It was** herself$_1$ **that** Susan$_1$ **critiqued**.

b. IT WAS THAT CRITIQUED (SUSAN$_1$, HERSELF$_1$)

The words in bold in (3a) form a *catena* and this catena is the matrix predicate of the entire sentence. The arguments of this predicate are *Susan* and *herself*. The predicate-argument analysis of (3a) is given in (3b) according to the convention of predicate-calculus (and using small caps): the predicate is placed on the left and its arguments are listed in parenthesis to the right of the predicate. The key insight concerning this analysis is that the matrix predicate is a catena that includes the expletive *it*, the two finite verbs *was* and *critiqued* as well as the subordinator *that*.

By acknowledging that the matrix predicate is a catena in this manner, it becomes possible to account for connectivity effects in representational terms in surface syntax. Appeals to transformations/derivations that derive *it*-cleft sentences from more basic sentence types (e.g. Akmajian 1970; Pinkham & Hankamer 1975; Emonds 1976; Meinunger 1998; Reeve 2012) and/or appeals to semantic or logical structures (Heycock & Kroch 1999; Lahousse 2009), e.g. Logical Form, are not necessary. Connectivity effects also appear in pseudocleft and specificational copular sentences in general. While the theoretical apparatus developed here can be extended to these related sentence types, no attempt to do so is undertaken here due to length limitations. The manuscript is organized as follows. Section 2 illustrates and discusses connectivity effects in *it*-clefts more extensively. Section 3 provides some background discussion concerning varying notions of predicates. Section 4 establishes that matrix predicates are catenae. Section 5 presents the core analysis of connectivity effects in *it*-clefts in terms of the catena unit. Section 6 draws attention to two additional aspects of *it*-clefts. Section 7 concludes the manuscript.

2 Connectivity effects

The next examples illustrate the effect of Condition A of the traditional binding theory (Chomsky 1981, 1986). Condition A is the requirement in GB (Government and Binding) binding theory that requires a reflexive pronoun to have a local antecedent, roughly a clause-mate, e.g.

Condition A violated

(4) a. *They told **himself$_1$** that **Tom$_1$** was injured.
b. *It surprised **herself$_1$** that **Susan$_1$** won the prize.
c. *Susan asked **himself$_1$** whether **Frank$_1$** would help.

These sentences are robustly ungrammatical because Condition A is violated each time: the reflexive pronoun is not locally c-commanded by an antecedent; that is, *Tom, Susan,* and *Frank* do not locally c-command *himself*, *herself*, and *himself*, respectively. Note that each of these sentences contains two finite clauses, each headed by a finite verb.

It-cleft sentences can have an outward appearance that is similar to sentences (4a-c), yet the presence of the reflexive pronoun is perfectly acceptable (cf. Delahunty 1984: 69; Lahouse 2009; Reeve 2012: 42):

Condition A obviated

(5) a. It was **himself$_1$** that **Tom$_1$** injured.
b. It was **herself$_1$** that **Susan$_1$** surprised.
c. It was **himself$_1$** that **Frank$_1$** asked to help.

The perfect grammaticality of these sentences is unexpected based on the ungrammaticality of sentences (4a-c). Each sentence in both sets is bi-clausal, whereby both clauses are headed by a finite verb. Furthermore, the embedded clauses across the two sets are similar in that they are all introduced by the subordinator *that*. Apparently, some trait of *it*-clefts fundamentally alters the basic binding relationships such that Condition A is obviated.

The situation is the same concerning the other two conditions of the traditional binding theory, that is, *it*-clefts also appear to ignore Conditions B and C. Condition B of the GB binding theory states that a non-reflexive pronoun must be free in its local binding domain, and Condition C of GB binding theory states a fully referential expression, an R-expression, must be free everywhere. To illustrate, each data set now contains three sentences, whereby the a-sentence illustrates the normal situation associated with the binding condition at hand and the b-sentence shows that the cleft sentence ignores this condition. The c-sentences are added to establish a point about mono-clausality:

Condition B
(6) a. They told **him$_1$** that **he$_1$** was injured.
b. *It was **him$_1$** that **he$_1$** injured.
c. ***He$_1$** injured **him$_1$**.

Condition C
(7) a. They told **Tom$_1$** that **he$_1$** was injured.
b. *It was **Tom$_1$** that **he$_1$** injured.
c. ***He$_1$** injured **Tom$_1$**.

Based on the perfect acceptability of the readings in (6a) and (7a), the readings indicated in the *it*-clefts in (6b) and (7b) are unexpectedly unavailable. The c-sentences draw attention to the fact that *it*-clefts behave like mono-clausal sentences in this area despite the fact that *it*-clefts are bi-clausal, containing two finite verbs.

Examples (6-7) suggest an approach to *it*-clefts that derives them from the corresponding non-cleft counterparts – (6b) from (6c) and (7b) from (7c). An important insight in this regard is that the order of the coindexed nominals in each cleft sentence above would match that of the corresponding non-cleft counterpart in which topicalization has occurred:

(8) a. It was **himself$_1$** that **Tom$_1$** injured. = (5a)
b. …but **himself$_1$ Tom$_1$** did injure.

(9) a. *It was **him$_1$** that **he$_1$** injured. = (6b)
b. *…but **him$_1$ he$_1$** did injure.

(10) a. *It was **Tom$_1$** that **he$_1$** injured. = (7b)
b. *…but **Tom$_1$ he$_1$** did injure.

The bolded nominals across each pair match with respect to linear order of appearance and the syntactic function that each fulfills; *himself* each time, *him* each time, and *Tom* each time are all objects of *injured/injure*.

The insight is supported by most so-called *anti-connectivity* effects (cf. Pinkham & Hankamer 1975: 431; Delahunty 1986: 34; Lahousse 2009: 143-145; Reeve 2012: 44). The binding behavior of *it*-clefts does not necessarily match that of the corresponding non-cleft counterpart as illustrated with the following b-sentences. It does, however, match that of the corresponding non-cleft counterpart in which topicalization has occurred as illustrated with the c-sentences:

(11) a. It was **himself$_1$/*him$_1$** that Bill$_1$ asked Sue to wash. (Pinkham & Hankamer 1975: 431)
b. **Bill$_1$** asked Sue to wash ***himself$_1$/ him$_1$**.
c. …but **himself$_1$/*him$_1$** Bill$_1$ did ask Sue to wash.

(12) a. It was **herself$_1$/*her$_1$** that **Sue$_1$** said Bill wants to date.
b. **Sue$_1$** said Bill wants to date ***herself$_1$/her$_1$**.
c. …but **herself$_1$/*her$_1$ Sue$_1$** did say Bill wants to date.

The distribution of pronoun forms in the cleft sentences does not match that of the corresponding non-cleft counterpart in which standard SVO word order obtains (b-sentences). It does, however, match that of the sentences in which OSV order obtains due to topicalization (c-sentences).

To summarize so far, the binding pattern of *it*-cleft sentences can match that of their corresponding non-cleft counterparts in which topicalization has occurred, whereby the foregrounded constituent of the cleft sentence corresponds to the topicalized constituent in the non-cleft counterpart. When the non-cleft counterpart is mono-clausal, the corresponding cleft sentence also behaves as if it is mono-clausal despite the presence of two finite verbs. When the non-cleft sentence is bi-clausal, the foregrounded constituent of the corresponding cleft sentence behaves like a topicalized constituent in the non-cleft counterpart.

The insight established with the examples so far extends to other phenomena, such as to the ambiguities associated with negation and a causal adjunct (13a-c), the distribution of the negative polarity item *any* (14a-c), and ambiguities of quantifier scope (15a-c):

Negation and causal adjunct

(13) a. Frank did **not** leave **because** he had to work. (not > because, not < because)
b. It was **because** he had to work that Frank did **not** leave. (because > not, *because < not)
c. **Because** he had to work, Frank did **not** leave. (because > not, *because < not)

Distribution of NPI *any*

(14) a. Frank did **not** insult **any**one.
b. *It was **any**one that Frank did **not** insult.
c. *…but **any**one Frank did **not** insult.

Ambiguities of quantifier scope

(15) a. **Every boy** kissed **a girl**. (a > every, every > a)
b. It was **a girl** that **every boy** kissed. (cf. Reeve 2012: 42) (a > every, every > a)
c. …but **a girl every boy** did kiss. (a > every, every > a)

The ambiguity of (13a) disappears in the corresponding cleft sentence that foregrounds the causal adjunct (13b), just as it disappears in the corresponding simple sentence that has experienced topicalization of the adjunct (13c).[1] The polarity item *any*- follows its trigger *not* in (14a), but when it precedes it in the corresponding cleft sentence, the sentence is ungrammatical, just as the corresponding simple sentence (14c) is ungrammatical in which the object *anyone* has been topicalized. Concerning examples (15a-c), all three sentences are ambiguous in the same way. The relevant point in this regard is that just as the ambiguity of (15a) is maintained in the cleft sentence (15b), so too it is maintained in the corresponding simple sentence with topicalization (15c).

The examples discussed so far all have the object as the foregrounded constituent in the *it*-cleft. When the subject is foregrounded instead of an object, the *it*-cleft also patterns like the corresponding simple sentence:

Binding (Condition A)

(16) a. It was **Sam$_1$** who hurt **himself$_1$**.
b. **Sam$_1$** hurt **himself$_1$**.

Negation and causal adjunct

(17) a It was Frank who did **not** leave **because** he had to work. (not > because, not < because)
b. Frank did **not** leave **because** he had to work. (not > because, not < because)

Distribution of NPI *any*

(18) a. It was Frank who did **not** insult **any**one.
b. Frank did **not** insult **any**one.

[1] The use of terminology here suggests a transformational approach to syntax, e.g. "foregrounds" and "topicalization". This terminology should be understood in a metaphorical sense and is used in the interest of vivid descriptions that are accessible to a wide audience. The DG espoused here is strictly monostratal in syntax, which means all transformations are rejected that would derive some sentences from other, more basic sentences.

Ambiguities of quantifier scope
(19) a. It was **every** boy that kissed **a** girl. (every > a; every < a)
b. **Every** boy kissed **a** girl. (every > a; every < a)

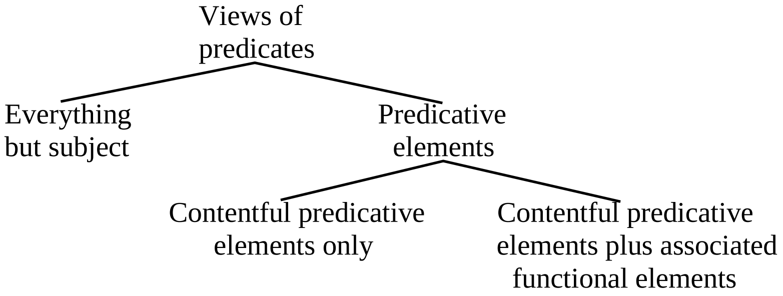
In these cases, foregrounded constituent in the *it*-cleft sentence is the subject. Each time the *it*-cleft sentence, the a-sentence, patterns just like the corresponding simple sentence, the b-sentence. Topicalization in the simple sentence is not needed because the linear order of the bolded constituents is already consistent across the two sentence types.

To summarize the data, *it*-cleft sentences pattern just like the corresponding simple sentences with respect to a number of phenomena of syntax. To ensure completeness of the correspondence, however, one must control for linear order. Doing so necessitates that topicalization occur in the simple sentence if the foregrounded constituent in the corresponding *it*-cleft is a non-subject. This state of affairs suggests strongly that *it*-clefts are in fact mono-clausal in a central respect, despite the appearance of two finite verbs.

3 Predicates

There are two main competing views of what qualifies as a main clause predicate in theories of grammar, a fact that can be verified by a quick check in most dictionaries of linguistic terminology (e.g. *Routledge Dictionary of Grammatical Terms in Linguistics* 1993, p. 213; *Oxford Concise Dictionary of Linguistics* 1997, p. 291), and within one of these views, two distinct sub-views can be discerned. The following diagram gives an overview:

(20)


Views of predicates
Everything but subject
Predicative elements
Contentful predicative elements only
Contentful predicative elements plus associated functional elements

The following sentence is used to illustrate these views of predicates:

(21) Frank has been studying syntax.

The one prominent understanding of predicates takes everything in a simple sentence except the subject as the predicate of the sentence. Hence on this approach, the predicate in (21) is *has been studying syntax*. This understanding of predicates is compatible with traditional phrase structure syntax insofar as the predicate corresponds to the VP of the initial binary division of a sentence S into a subject NP and a predicate VP (S → NP VP).

The main alternative understanding of predicates is inspired by predicate calculus associated above all with Gottlob Frege (1848-1925). A predicate serves to assign a property to an argument or to relate more than one argument to each other. On this approach, the content verb *studying* is deemed as (the core of) the main predicate in sentence (21), and *Frank* and *syntax* are its arguments. Within this alternative approach to predicates, one can discern two sub-views. The one sub-view takes predicates and their arguments as semantic entities that are often manifest as content verbs or adjectives (e.g. Poole 2002: 77-79; Adger 2003: 78-82; Carnie 2013: 57–60); on this sub-view, the matrix predicate in (21) is the content verb *studying* alone. The other sub-view is oriented more toward surface syntax; it takes predicates to consist of at least one main content word plus one or more associated function words. On this sub-view, the matrix predicate in (21) is *has been studying*.

These competing views of predicates are summarized as follows. The matrix predicate on each view appears in bold:

Everything but subject:

(22) a. Frank **has been studying syntax**.

Content predicative word only:

b. Frank has been **studying** syntax.

Content predicative word plus associated function words:

c. Frank **has been studying** syntax.

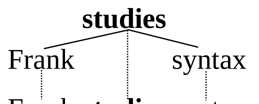
The view of predicates given as (22c) is the one pursued here below. A predicate consists of one or more content words plus any associated function words. Variants of this approach to predicates have been developed in detail (see e.g. Napoli 1989 and Ackermann and Webelhuth 1998). It is also the understanding of predicates that is dominant in the grammars of German (e.g. Helbig and Buscha 1998: 536–543; *Duden* 1984: 567–571). Most importantly, it represents an approach to predicate-argument structures that is particularly congruent with dependency syntax. This congruity is due to the fact that the word combinations that qualify as predicates are catenae in surface syntax, and so are the arguments of these predicates.

4 Predicate catenae

The main insight about predicates and arguments that makes the current account of *it*-clefts possible is that these entities are manifest as catenae in dependency structures. This fact is established and illustrated here using a series of examples, whereby traditional predicate-calculus-style analyses, as first appeared in (3b) above, are included to make the illustrations more concrete.

A catena is a word or a combination of words that are linked together by dependencies (O'Grady 1999; Osborne 2005; Osborne et al. 2012).[2] A typical matrix predicate consists of a content verb and any pure auxiliaries that are present. This fact is illustrated first using the example from above about Frank studying syntax:

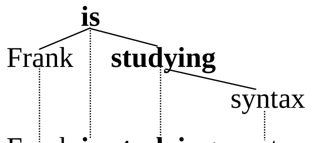
(23)



a. Frank **studies** syntax.

b. STUDIES (FRANK, SYNTAX)

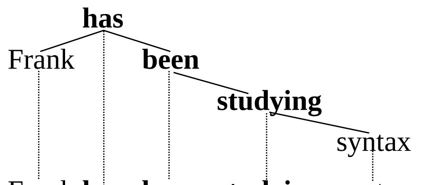
(24)



a. Frank **is studying** syntax.

b. IS STUDYING (FRANK, SYNTAX)

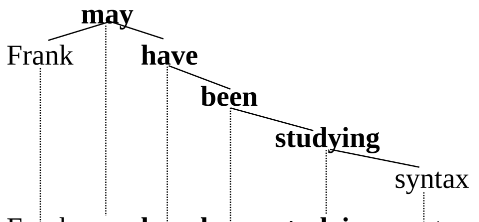
(25)



a. Frank **has been studying** syntax.

b. HAS BEEN STUDYING (FRANK, SYNTAX)

(26)



a. Frank **may have been studying** syntax.

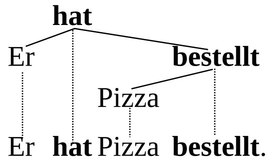
b. MAY HAVE BEEN STUDYING (FRANK, SYNTAX)

Each additional auxiliary verb that appears is easily incorporated into the matrix predicate. On occasion, the words that constitute the matrix predicate are not linearly continuous, a fact illustrated here using two examples from German:

[2] A more formal definition of the catena unit, a set-theoretic one, is given next:

Catena (set-theoretic definition)
Given a dependency tree T, a catena is a set of nodes N in T such that exactly one node in N is not immediately dominated by another node in N.

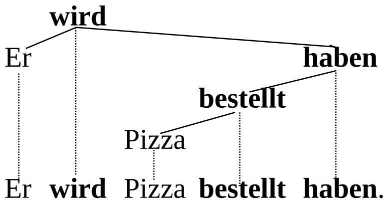
(27)



a. Er **hat** Pizza **bestellt**.
he has pizza ordered
'He ordered pizza.'

b. HAT BESTELLT (ER, PIZZA)

(28)



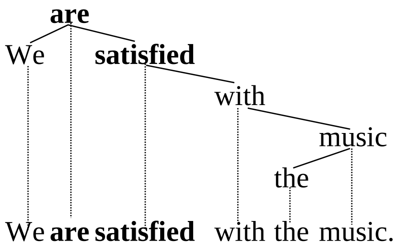
a. Er **wird** Pizza **bestellt** **haben**.
he will pizza ordered have
'He will have ordered pizza.'

b. WIRD BESTELLT HABEN (ER, PIZZA)

Due to the appearance of *Pizza* in these cases, the words that constitute the matrix predicate are not linearly continuous. This fact does not prevent them from forming a catena.

The next examples concern the auxiliary verb *be*. This verb is usually semantically almost empty and hence a pure function word. It forms a predicate with (one of) its post-dependent(s). The next examples involve a predicative adjective and a predicative nominal:

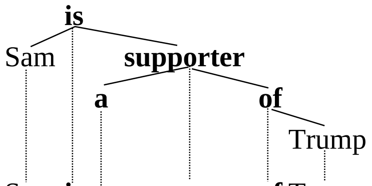
(29)



a. We **are** **satisfied** with the music.

b. ARE SATISFIED (WE, WITH THE MUSIC)

(30)



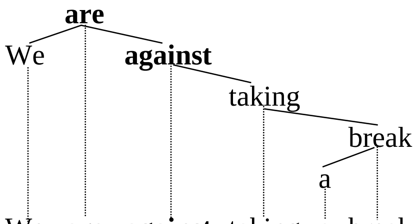
a. Sam **is** **a** **supporter** **of** Trump.

b. IS A SUPPORTER OF (SAM, TRUMP)

The copula *are* in (29) forms the matrix predicate with the predicative adjective *satisfied*, and the copula *is* in (30) forms the matrix predicate with predicative nominal *a supporter of*. Note that there is flexibility concerning the status of the prepositions *with* and *of* in these two examples, that is, concerning their inclusion or exclusion from the matrix predicate. Alternative analyses in this regard might also be plausible: ARE SATISFIED WITH (WE, THE MUSIC) and IS A SUPPORTER (SAM, OF TRUMP). On either analysis each time, the matrix predicate is a catena. Note also that the matrix predicate *is a supporter of* in (30) corresponds to a simple content verb *supports* in the almost synonymous simple sentence *Sam supports Trump*: SUPPORTS (SAM, TRUMP).

The next examples further illustrate the extent to which forms of auxiliary *be* appear in the matrix predicate with whatever occurs as their post-dependent. Prepositions can be directly included in the matrix predicate, whereby the object of the preposition is an argument:

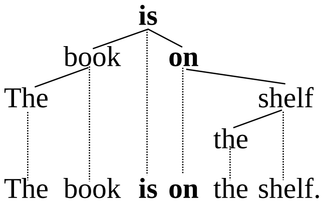
(31)



a. We **are** **against** taking a break.

b. ARE AGAINST (WE, TAKING A BREAK)

(32)



a. The book **is** **on** the shelf.

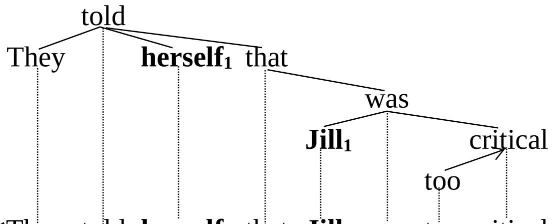
b. IS ON (THE BOOK, THE SHELF)

Examples (29-32) are particularly relevant to the analysis of clefts. They show the manner in which the matrix predicate includes the copula and (part of) a post-dependent of the copula. For cleft sentences, this means that the matrix predicate reaches into the embedded clause.

5 Connectivity accounted for

Many matrix predicates do not reach below the main content verb. This is certainly the case in example (2) above, which is reproduced here as example (33), with the dependency structure and predicate-argument analysis added:

(33)



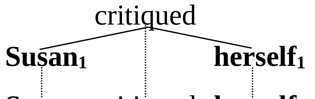
a. *They told **herself$_1$** that **Jill$_1$** was too critical.

b. TOLD (THEY, HERSELF$_1$, THAT JILL$_1$ WAS TOO CRITICAL).

The source of the ungrammaticality in this case is apparent based on the predicate-argument analysis. The reflexive pronoun *herself* fails to find an antecedent at its level of the predicate-argument structure; *Jill* is not its co-argument, but rather is embedded in its co-argument.

The next examples demonstrate that when the reflexive pronoun is licensed, its antecedent is often a co-argument that is ranked higher on the scale of argument functions: SUBJECT > 1ST OBJECT > 2ND OBJECT > OBLIQUE OBJECT.

(34)



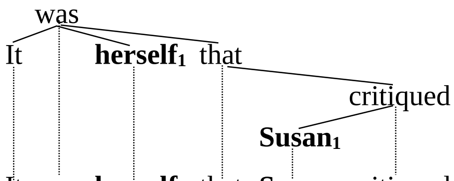
a. **Susan$_1$** critiqued **herself$_1$**.

b. CRITIQUED (SUSAN$_1$, HERSELF$_1$)

The reflexive pronoun *herself* is the object of *critiqued*, and its antecedent is *Susan*, the subject of *critiqued*. Thus, *herself* can appear by virtue of the fact that it finds a more highly ranked co-argument as its antecedent.

The predicate-argument analysis of *it*-clefts is similar. The matrix predicate reaches down from the root copula to include the main predicate in the embedded clause, rendering the foregrounded constituent a co-argument of the argument(s) in the embedded clause. Example (3) is repeated here as (35):

(35)



a. It was **herself$_1$** that **Susan$_1$** critiqued.

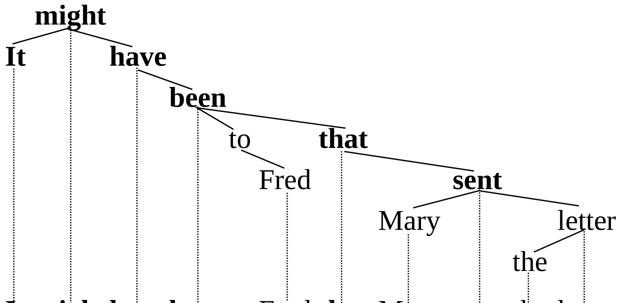
b. IT WAS THAT CRITIQUED (SUSAN$_1$, HERSELF$_1$)

Despite the fact that *herself$_1$* appears in the matrix clause, it can take its reference from the argument in the embedded clause. It can do this because the matrix predicate reaches into the embedded clause in a manner that renders *Susan* and *herself* co-arguments, whereby *Susan*, as a subject, is ranked higher than *herself*, an object. Two key aspects of this analysis are worth restating: first, the copula is a function word and so the matrix predicate necessarily reaches below it to include (part of) a post-dependent, just as in examples (29-32) above; and second, the words constituting the matrix predicate form a catena despite the fact they are discontinuous in the linear dimension and hence do not form a string.

A third aspect of example (35) is tentative: the expletive *It* is included as part of the matrix predicate. Nothing crucial rides on this aspect of the account. An alternative analysis would exclude the expletive *It* from the matrix predicate. The advantage of including it therein is that one is not confronted with the challenge of having to decide how to categorize it: should the expletive be viewed as an argument, an adjunct, or something else?

The next example illustrates the ability of the matrix predicate catena of an *it*-cleft to be very long indeed. The sentence is from Delahunty (1986: 22), whereby the dependency structure and predicate-argument analysis have been added:

(36)



a. **It might have been** to Fred **that** Mary **sent** the letter.

b. IT MIGHT HAVE BEEN THAT SENT (MARY, THE LETTER, TO FRED)

The matrix predicate includes six words, only one of which can be viewed as a full content word, namely *sent*, which is the lowest of the six. We see again that a typical aspect of matrix predicates is the manner in which they reach down from the root of the sentence until they include a full content word.

Example (35) demonstrates how the connectivity associated with binding Condition A is addressed and accommodated in terms of predicate catenae. The same reasoning applies to the other connectivity effects discussed and illustrated in Section 2. These connectivity effects are expected by virtue of the fact that the matrix predicate in an *it*-cleft sentence reaches down to include the main predicate in the embedded clause.

6 Two further aspects

Before concluding this manuscript, two further aspects of the current account are briefly addressed. The first of these concerns the fact that the matrix predicate of *it*-clefts reaches into the embedded clause, but not into the foregrounded constituent. The second concerns the ability of the matrix predicate to include the relative pronoun of the embedded clause.

A widely acknowledged fact about *it*-clefts is that a verb phrase may not be foregrounded, e.g.

(37) a. *It is **blow up some buildings** that you should. (Emonds 1976: 133)
b. *It's **submit her manuscript to Fortune** that Alice did. (cf. McCawley 1998: 66)
c. *It is **(to) apply for special leave** that you must do. (Huddleston and Pullum 2002: 1422)

A related observation is that other predicative elements, such as predicative adjectives and nominals, also cannot be foregrounded:[3]

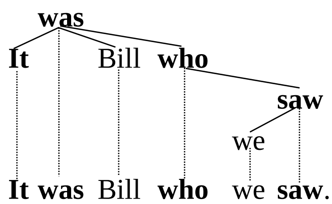
(38) a. *It is **tall** that John is. (Akmajian 1970: 166)
b. *It's **my doctor** that John Smith is. (Heggie 1988: 81)
c. *It is **on the couch** that Frank is.

The ungrammaticality of examples (37-38) is congruent with the current analysis of *it*-clefts. The key trait of *it*-clefts established above is that the matrix predicate necessarily reaches into the embedded clause to include the main predicate that resides there. If there is no main predicate there because that predicate appears instead as (part of) the foregrounded constituent, then the matrix predicate would have to reach into the foregrounded constituent; apparently, it cannot do this. The foregrounded constituent of an *it*-cleft sentence should be an argument or adjunct of the matrix predicate; it cannot include part of the matrix predicate.

The other aspect of *it*-clefts mentioned here concerns the fact that often, the relative pronoun of the embedded cleft clause is included in the matrix predicate, e.g. It was Bill who we saw.

[3] There are some important exceptions to this generalization concerning predicative adjectives. For instance, Heggie (1988: 206) and Reeve (2012: 54-56) observe that if contrastive emphasis is present on the adjective, then predicative adjectives can (at least marginally) be foregrounded, e.g. *A: Her eyes are green. B: No, its BLUE that her eyes are, not GREEN.*

(39)



a. **It was** Bill **who** we **saw**.
b. IT WAS WHO SAW (WE, BILL)

This analysis of the embedded cleft clause, which is a type of relative clause, follows the analysis of relative clauses in Groß and Osborne (2009) and Osborne (2014).[4] The relative pronoun *who* is positioned as the root of the relative clause. In the current context, the relevant aspect of this analysis is that the relative pronoun can be viewed more as a function word than as a content word, so its inclusion in the matrix predicate is consistent with the account above. Consider in this regard that non-subject relative pronouns are often omitted in English, e.g. *the man (who) I know* and that when the relative pronoun is a subject followed by a form of *be*, the two can also be omitted, e.g. *the man (who is) studying syntax*. These observations help reveal that the relative pronoun is non-essential at times, a fact that increases the plausibility of viewing it as a type of function word.

7 Concluding statement

There are of course many aspects of *it*-clefts that have not been addressed above. Hopefully, however, enough of the current approach to *it*-clefts has been presented to convince the reader that such an approach is worth pursuing further. Finally, it is appropriate to state again that the current approach in terms of catenae and predicate-argument structures can be extended to related sentence types, namely to pseudoclefts and specificational copular sentences in general. Connectivity effects also appear in these additional sentence types.

References

Farrell Ackerman and Gert Webelhuth. 1998. *A Theory of Predicates*. CSLI Publications, Stanford, CA.

David Adger 2003. *Core Syntax: A Minimalist Approach.* Oxford University Press, Oxford, UK.

Adrian Akmajian.1970. On deriving cleft sentences from pseudo-cleft sentences. *Linguistic Inquiry*, 1(2):149–168.

Andrew Carnie. 2013. *Syntax: A Generative Introduction*. Wiley-Blackwell, Malden, MA.

Gerald Delahunty. 1984. The analysis of English cleft sentences. *Linguistic Analysis*, 13(2):63–113.

Gerald Delahunty. 1986. *Topics in the Syntax and Semantics of English Cleft Sentences*. Ph.D. thesis, University of California. Reproduced by the Indiana University Linguistics Club.

Duden (Die Grammatik). 1984. Dudenverlag, Mannheim.

Joseph E. Emonds. 1976. *A Transformational Approach to English Syntax: Root, Structure-Preserving, and Local Transformations*. Academic Press, New York.

Thomas Groß and Timothy Osborne 2009. Toward a practical dependency grammar theory of discontinuities. *SKY Journal of* Linguistics, 22:43–90.

Jeanette Gundel. 1977. Where do cleft sentences come from? Language, 53(3): 543–559.

Nancy Hedberg. 2000. On the referential status of clefts. Language, 76(4):891–920.

Lorie Heggie. 1988. *The Syntax of Copular Structures*. Ph.D. dissertation, University of Southern California.

Gerhard Helbig and Joachim Buscha. 1998. *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*, 18th edition. Langenscheidt, Leipzig.

Caroline Heycock and Anthony Kroch. 1999. Pseudo-cleft connectedness: Implications for the LF interface. *Linguistic Inquiry*, 30(3):365–397.

Caroline Heycock and Anthony Kroch. 2002. Topic, focus, and syntactic representations. *Proceedings of WCCFL* 21: 101–125.

Rodney Huddleston and Geoffreey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

[4] Certain aspects of Groß and Osborne's diagrammatic analysis of relative clauses have been suppressed here because they are not directly relevant to the point at hand.

Karen Lahouse. 2009. Specificational sentences and the influence of information structure on (anti-)connectivity effects. *Journal of Linguistics*, 45:139-166.

André Meinunger. 1998. A monoclausal structure for (pseudo) cleft sentences. In: P. N. Tmanji and K. Kusumoto (eds.), Proceedings of NELS, 28:283–297.

Line Mikkelsen. 2005. *Copular Clauses: Specification, Predication and Equation*. [Linguistik Aktuell/Linguistics Today 85]. John Benjamins, Amsterdam.

Andrea Moro.1997. *The Raising of Predicates.* Cambridge University Press, Cambridge, UK.

Donna Jo Napoli. 1989. *Predication theory: A case study for indexing theory*. Cambridge University Press, Cambridge, UK.

William O'Grady. 1998. The syntax of idioms. *Natural Language and Linguistic Theory*, 16:279–312.

Timothy Osborne. 2005. Beyond the constituent: A dependency grammar analysis of chains. *Folia* Linguistica, 39(3–4):251–297.

Timothy Osborne. 2014. Type 2 rising: A contribution to a DG account of discontinuities. In Kim Gerdes, Eva Hajičova and Leo Wanner (eds.), *Dependency Linguistics: Recent Advances in Linguistic Theory using Dependency Structures*, 273–298. John Benjamins, Amsterdam.

Timothy Osborne, Michael Putnam and Thomas Groß 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Oxford Concise Dictionary of Linguistics. 1997. By P.H. Matthews. Oxford University Press, Oxford, UK.

Jessie Pinkham and Jorge Hankamer. 1975. Deep and shallow clefts. In Chicago Linguistics Society Vol. 11, 429-450.

Geoffrey Poole. 2002. *Syntactic Theory*. PALGRAVE, New York.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svatvik. 2010. *A Comprehensive Grammar of the English Language*. Dorling Kindersley (India)/Pearson Education, New Delhi.

Matthew Reeve. 2012. *Clefts and their Relatives*. John Benjamins, Amsterdam.

Routledge Dictionary of Grammatical Terms in Linguistics. 1993. By R. L. Trask. Routledge, London.

Noun phrases rooted by adjectives
A dependency grammar analysis of the Big Mess Construction

Timothy Osborne
Zhejiang University
Hangzhou, China
tjo3ya@yahoo.com

Abstract

The *Big Mess Construction* (BMC) challenges standard assumptions about NP structure in English, e.g. *so big a mess*. Previous accounts of the BMC are couched in phrase structure syntax and most of them take the noun or determiner *a* to be the head of the phrase. In contrast, the current analysis of the BMC is couched in a *dependency grammar* that views the adjective as syntactic root/head of the BMC phrase. The fact that the BMC distributes as an NP, not as an AP, is due to the category changing ability of the degree adverb. This adverb evokes a change in status of its head word from adjective to noun-like category, similar to the manner in which the definite article *the* can cause a change in status of an adjective to a noun, e.g. *the good, the credible*, etc.

1 Introduction

The *Big Mess Construction* (BMC), also discussed under the rubric of *adjectival predeterminers*, defies standard notions about the structure of NPs (and DPs) in English. Some examples of the BMC used to introduce the phenomenon in the literature are next:

(1) a. how serious a problem… (van Eynde 2007: 416)
b. too big a dog… (Zwicky 2007: 113)
c. this delicious a lasagna (Kay and Sag 2012: 229)
d. so prominent a punctuation… (Kim and Sells 2011: 335)
e. so big a part of the present system… (Wood and Vikner 2011: 90)

These phrases have the distribution of NPs, yet the word order each time is unlike that of normal NPs. The adjective precedes the determiner *a,* something which is usually not possible. It becomes possible, however, if a certain type of degree adverb modifies the adjective. The position of the adverb-adjective combination in front of *a* can in fact be obligatory insofar as the more normal NP word order, where the adjective follows the determiner, is blocked, e.g. **a how serious problem…*. The term *Big Mess Construction*, from Berman (1974), is a reference to the example Berman originally discussed and to the challenges to syntactic theory the phenomenon generates.

The BMC is licensed by a limited set of degree adverbs (cf. Huddleston and Pullum 2002: 435; van Eynde 2007: 417; Zwicky 2007: 114; Kim and Sells 2011: 339). This set includes the following members: *as*, *enough*, *how*, *however*, *less*, *more*, *so*, *that*, *this*, and *too*. Similar degree adverbs that modify adjectives fail to license the BMC, e.g. *quite*, *somewhat*, *very*:

(2) a. *big enough a house*, *how big a house*, *so big a house that…*, *that big a house*, *this big a house*, *too big a house*, *as big a house as…*, etc.[1]
b. **quite big a house*, **somewhat big a house*, **very big a house*

[1] Observe that *more* and *less* are not included in these examples. These two licensors of the BMC are unique insofar as they can precede or follow the determiner *a*, e.g. *more difficult a problem* vs. *a more difficult problem* (cf. Huddleston and Pullum 2002:

The particular degree adverbs that license the BMC have some trait that other degree adverbs lack. This trait may be an implication of contrast (Aniya 2016: 8–10), although this matter is not explored in this manuscript.

A distinctive trait of the BMC is the appearance of the indefinite article *a* (cf. Huddleston and Pullum 2002: 435; van Eynde 2007: 416; Kim and Sells 2011: 336). Attempts to construct the BMC fail if *a(n)* is absent:

(3) a. that fun a game
b. *that fun the game

(4) a. too smart a child to…
b. *too smart children to…

The absence of *a* results in ungrammaticality when some other determiner other than *a* appears as in (3b) and also when a determiner is completely absent as in (4b). Data such as these suggest that the presence of the indefinite article *a* is a necessary condition on the occurrence of the BMC.

Another important trait of the BMC is that the preposition *of* appears optionally. The initial examples of the BMC above are given again next, but this time, the preposition *of* appears each time:

(5) a. how serious **of** a problem…
b. too big **of** a dog…
c. this delicious **of** a lasagna
d. so prominent **of** a punctuation…
e. so big of a part **of** the present system…

There is dialectal variation in this area. The appearance of *of* is rare in varieties of British English, but more acceptable in varieties of American English (cf. Kennedy and Merchant 2001: 125 n. 24; Zwicky 2007: 113 n. 1; Kim and Sells 2011: 339–40).

The purpose of this manuscript is to present and defend a novel (and therefore controversial) analysis of the BMC. A survey of existing accounts of the BMC reveals that the noun, the indefinite article *a*, or the preposition *of* is construed as the syntactic head of the phrase:

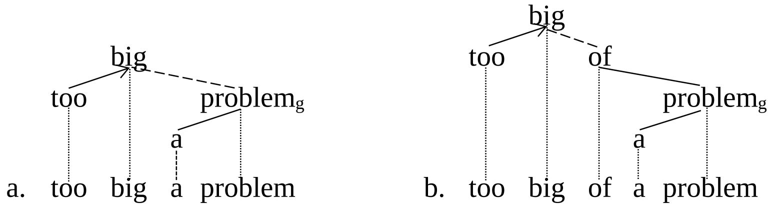
Noun as root/head
Bresnan (1973), van Eynde (2007), Klégr (2010), Kim and Sells (2011), Kay and Sag (2012)

Indefinite article *a* as root/head
Haegeman and Guéron (1999: 420–1), Wood and Vikner (2011)

Preposition *of* as root/head
Kennedy and Merchant (2000)

In contrast to these previous accounts, the claim put forward and defended in this manuscript is that the adjective is in fact the syntactic root/head of the phrase.[2] The structural analysis of the BMC pursued and defended below is illustrated next:

(6)



435). This flexibility is not possible with the other licensors, e.g. *that big a house* vs. **a that big house*. The flexibility of *more* and *less* in this area is not explored in this manuscript.

[2] The term *root* is used here in the DG sense of the hierarchically dominant word in a given phrase. In contrast, the *head* of a given phrase is understood to be the parent of that phrase. The designation *root/head* is intended to accommodate both uses of the terminology, that is, in the current DG as well as in phrase structure syntax more generally.

The appearance of the degree adverb *too* forces a shift in category status, adjective to noun-like word. A similar category shift also takes place in simple phrases such as *the worthy, the pure, the corrupt*, etc., where the appearance of the definite article *the* is enough to shift the category status of the whole from AP to NP. The arrow dependency edge marks *too* as an adjunct. The dashed dependency edge and g-subscript indicate that *rising* is present. *Rising* denotes the particular approach to discontinuities developed by Groß and Osborne (2009), Osborne et al. (2012: 360-366), and Osborne (2014) – more on this below.

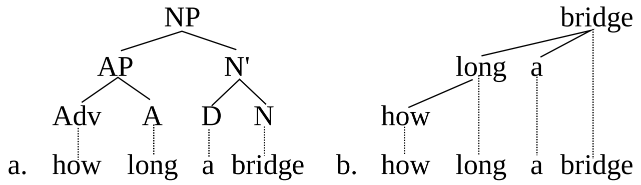
This manuscript is organized as follows. Section 2 considers the existing structural analyses of the BMC, rejecting them all. Section 3 presents the entirely projective DG assumed for the analysis of the BMC. Section 4 briefly considers how it comes to pass that a phrase rooted/headed by an adjective can have the distribution of an NP. Section 5 then presents central traits of the BMC that support the adjective as the root of the phrase. Section 6 gives a concluding statement.

2 Existing analyses of the BMC

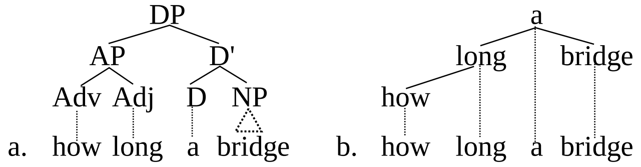
Previous accounts of the BMC are couched in phrase structure syntax. Despite this fact, these earlier accounts are relevant for the current DG approach, and vice versa. This relevance is due to ability to mechanically convert any strictly endocentric phrase structure to the corresponding dependency structure. This is done here now. Each existing phrase structure analysis of the BMC in this section is (if possible) given together with corresponding dependency structure that results from direct translation to dependency.

As mentioned in the introduction, many existing accounts of the BMC view the noun as the head of the BMC phrase (cf. Bresnan 1973: 306; Klégr 2010: 105; van Eynde 2007: 425; Kim and Sells 2011: 353; Kay and Sag 2012: 238) or, on a DP analysis of nominal groups, the indefinite article *a* (Haegeman and Guéron 1999: 420–1; Wood and Vikner 2011: 95). Such accounts produce structural analyses of the BMC along the following lines:

(7)



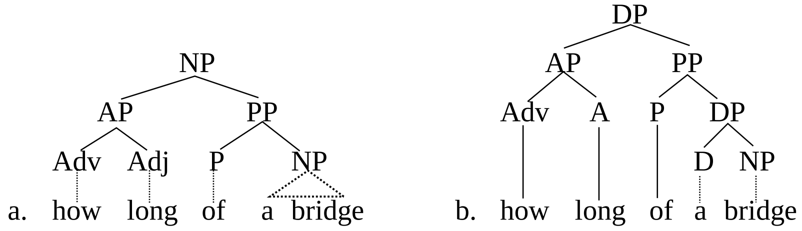
(8)



The analyses given as (7a–b) are those of the NP analysis of nominal groups, and the analyses (8a–b), those of the DP analysis of nominal groups. The b-trees are, again, the corresponding DG structures that result from direct translation (phrase structure → dependency). The named sources certainly vary in the specifics of how they analyze the BMC. From the point of view of the alternative account pursued in this manuscript (adjective as root and couched in DG), however, these differences are minor.

A weakness with the analyses given as (7–8) is the inability to deal with the preposition *of*. When *of* appears in the BMC, it is the head of the PP it introduces just as it is otherwise. This situation essentially makes a necessity an analysis that views the BMC with *of* as an exocentric construction, as illustrated next:

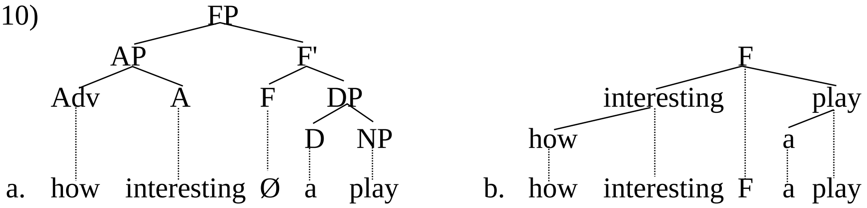
(9)



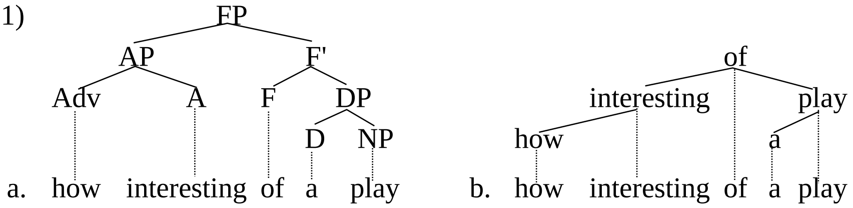
Both of these analyses are exocentric, that is, the root node has a category status that is entirely distinct from that of both of its immediate constituents. DG cannot acknowledge exocentric structures in this manner, and most modern PSGs also avoid exocentric structures as a matter of principle. It is therefore impossible to translate (9a-b) to corresponding dependency structures.

Unlike the accounts just mentioned, Kennedy and Merchant (2000: 124–30) concentrate on the optional appearance of *of* in the BMC (see examples 5a-e) and accommodate it into their analysis of the BMC in a central way that does not result in an exocentric structure. They view the BMC as headed by a functional category that is empty in those instances in which *of* does not appear. When *of* does appear, however, it occupies this head position of the functional category. The analysis they pursue is along the following lines:

(10)



(11)



There are two major drawbacks to this line of analysis given the current DG framework. The first is that in order to accommodate the empty head F shown in (10a), a null node is needed in the corresponding DG analysis, indicated as F in (10b). DGs have in general been reluctant to posit the existence such null elements. The second problem concerns the fact that since the preposition occupies the head position of FP in (11a), the whole is in fact a prepositional phrase, despite being called an FP (functional phrase), and this phrase has the distribution of an NP/DP, not of a PP.

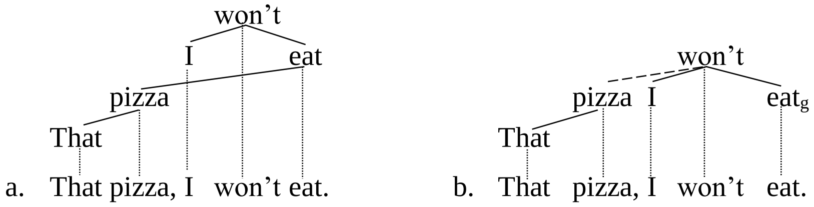
No further attempt is made here to evaluate the analyses given as (7–11) with respect to each other and otherwise. Suffice it to state that the current DG analysis of the BMC is much different, and that an approach that takes the adjective as the root of the phrase is warranted in part due to its ability to address both variants, without or with *of*. The discussion now turns to the DG assumed for addressing the BMC.

3 An entirely projective DG

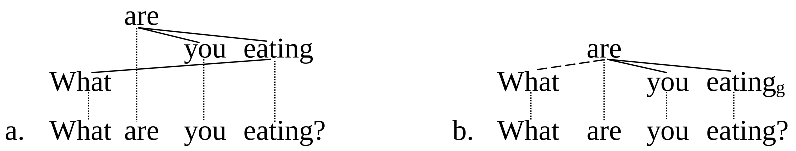
An entirely projective DG is assumed henceforth. Projectivity violations are avoided by attaching the expression in violation of projectivity to a higher word, overcoming the crossing lines in the tree. A number of DGs pursue, or have proposed, this sort of approach to discontinuities (cf. Schubert 1987: 190; Lobin 1993: 31–35; Heringer 1996: 261; Bröker 1999: 55–59, 2003: 294; Eroms and Heringer 2003: 26; Starosta 2003: 276–279; Groß and Osborne 2009; Osborne 2014).

The next examples illustrate how projectivity violations are avoided in the DG assumed henceforth:

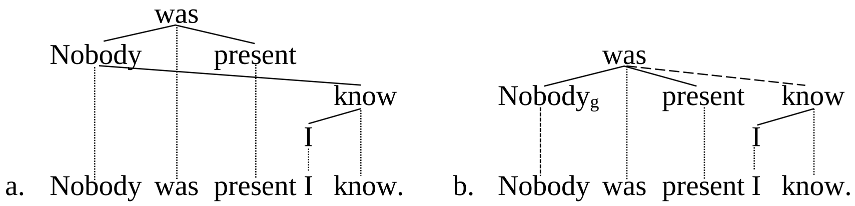
(14)



(15)



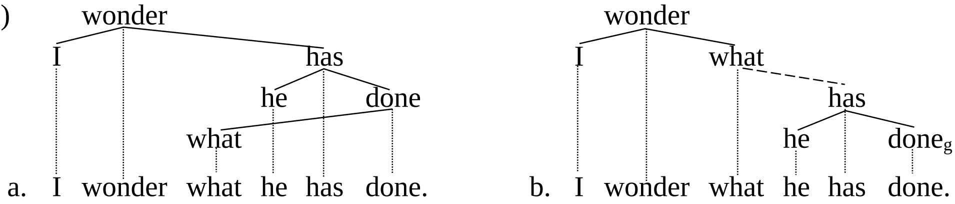
(16)



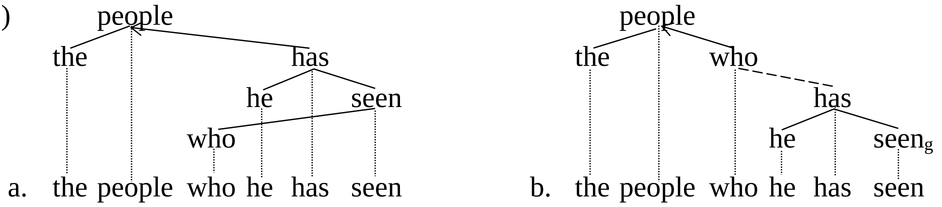
The crossing dependencies in (14a) are due to topicalization, in (15a) to wh-fronting, and in (16a) to extraposition (cf. *Nobody I know was present*). By attaching the constituent in violation of projectivity higher up each time as in the b-trees, the projectivity violation is removed. The dashed dependency edge marks the constituent that has attached higher up, and the g-subscript marks the governor of that constituent.

The current DG extends the sort of analysis illustrated with (14-16) to indirect interrogative and relative clauses, although with an important adjustment. It assumes that in indirect interrogative and relative clauses, the interrogative expression or relative proform is the root of the embedded clause, e.g.

(17)



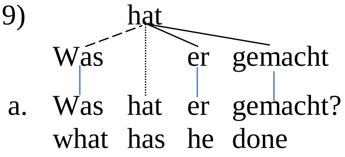
(18)

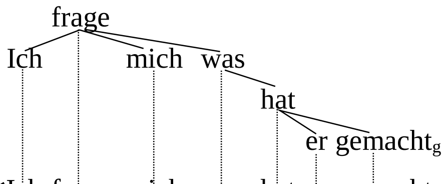


The crossing dependencies in the a-trees are again overcome in the b-trees by attaching the *wh*-element each time to a higher word. In these cases, however, this is done in such a manner that the wh-word becomes the root of the embedded clause. Osborne (2014) motivates the b-analyses in terms of systematic differences in word order across matrix and embedded wh-clauses in English (e.g. *What has he done?, *I wonder what has he done* vs. *I wonder what he has done*).

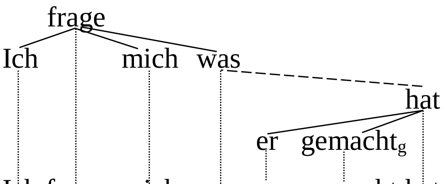
A similar systematic difference in word across matrix and embedded interrogative and relative clauses occurs in German, e.g.

(19)





b. *Ich frage mich, was hat er gemacht.
I ask me what has he done
'I wonder what has he done.'



c. Ich frage mich, was er gemacht hat.
I ask me what he done has
'I wonder what he has done.'

The V2 (verb second) word order of German is maintained in matrix w-clauses in (19a). In the embedded w-clauses in (19b-c), in contrast, VF (verb final) word order becomes necessary, as demonstrated by the ungrammaticality of (19b) in comparison to the grammaticality of (19c). These systematic differences are accommodated as indicated, that is, by establishing a direct link between the w-expression of embedded interrogative clauses and the matrix predicate.

This analysis of embedded interrogative clauses is supported further by the fact that the *wh*-word is linked directly to the preceding predicate, hence the manner in which the matrix predicate (here *wonder* and *sich fragen*) takes an interrogative object valent is accommodated because there is a direct dependency between that matrix predicate and the interrogative word (here *what* or *was*), the latter being the primary marker of an interrogative clause or phrase. This situation is shown in (17b), where *what* is a child of *wonder*, and in (19c), where *was* 'what' is a child of *frage* 'ask'. Extending this sort of analysis to relative clauses as in (18b) is then not a big step.

The approach to discontinuities established in this section is important for the analysis of the BMC. The sort of analysis assumed here for embedded interrogative clauses and relative clauses just sketched is also assumed for the BMC.

4 NP distribution

The BMC has the distribution of an NP, not of an AP. This fact is probably the reason why an analysis like the current one that positions the adjective as the root/head of the phrase has not been proposed until now. The current account must address how it comes to pass that a phrase the root of which is an adjective can have the distribution of an NP. The answer to this question is now offered. This answer is that the degree adverb that licenses the BMC changes the adjective to a noun-like category in a manner similar to how the definite article can change the category status of an adjective to a noun.

Consider the ability of the definite article *the* in the following cases to change the category status of what is normally an adjective:

(20) a. the best and brightest
b. The Good, the Bad, and the Ugly (Title of the 1966 spaghetti western)
c. the wealthy, the poor, the lazy, the insightful, the helpful, etc.

These phrases distribute as NPs, not as APs. The definite article *the* causes a change in status from adjective to noun. In a similar vein, the appearance of the degree adverb in the BMC causes a change in category status, again from adjective to noun-like category. One might object that such cases actually involve noun ellipsis: the head noun is elided and one should therefore not view the adjective in such cases as having taken on the category status of a noun. The problem with this objection is that many of these cases do not allow the appearance of the noun without a shift in meaning. A phrase such as *the wealthy* is distinct in meaning from the phrase *the wealthy people*; the former expresses an abstract trait of what it is to be wealthy that is beyond what the latter expresses.

5 Support for the analysis

The following sections consider some traits of the BMC and establish that these traits are supportive of the current account, that is, that the adjective is in fact the root of the BMC phrase.

5.1 Appositives

There is flexibility in the position of the adverb-adjective combination of the BMC. The combination can also follow the noun (cf. Huddleston and Pullum 2002: 435; van Eynde 2007: 424; Wood and Vikner 2009: 96), e.g.

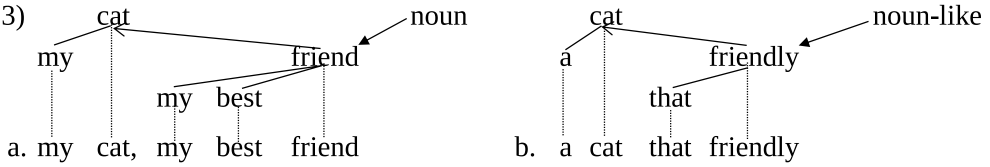
(21) a. that big a bridge
b. a bridge that big (van Eynde 2007: 424)

(22) a. so little altered a house
b. a house so little altered (Wood and Vikner 2009: 96)

Such post-noun positioning is only possible with the adverbs that license the BMC: **a bridge very big*, **a house somewhat altered*. The observation in this area that helps support the current analysis (adjective as root) is that in post-noun position, the adverb-adjective combination appears where appositives appear, and appositives are nouns or noun phrases, not adjectives or adjective phrases.

Compare the following structures, the first containing an appositive NP, and the second the adverb-adjective combination of the BMC, but in post-noun position:

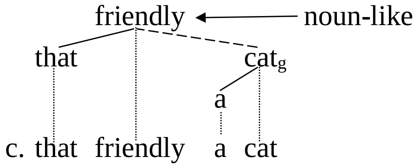
(23)



The adverb-adjective combination *that friendly* in (23b) has the syntactic status of an NP in the same way that the appositive *my best friend* in (23a) is an NP. In both cases, the post-dependent is predicative.

One can extend these insights to instances of the BMC. Since an analysis of *that friendly* as an appositive NP in *a cat that friendly* is plausible, it is also plausible to extend the account to instances of the BMC such as *that friendly a cat*, the string *that friendly* retaining its status as noun-like:

(23)



This demonstrates further that an analysis in terms of apposition is appropriate for examples such as (23b), that is, the adverb-adjective combination has a status that is similar to that of an NP in apposition. Switching to the BMC in (23c), the fact that *that friendly* can appear in a position associated with NPs supports the account here that views *friendly* as the root of the entire phrase *that friendly of a cat*, a phrase that has the distribution of an NP rather than of an AP.

5.2 Converse NPs

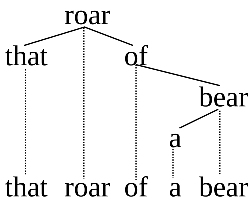
The existence of a related construction in which a noun corresponds to the adjective of the BMC supports the adjective as the root in the BMC. NPs such as *a bear of a guy* are clearly related to the BMC (cf. Bennis et al. 1998; Kennedy and Merchant 2000: 126, Wood and Vikner 2011: 96; Aniya 2016: 3):

(24) a. a bear of a guy (cf. Bennis et al. 1998: 87)
b. a jewel of an island (Wood and Vikner 2011: 96)
c. that idiot of a boy (Aniya 2016: 3)

As with the BMC, the (second) noun must be a singular count noun (e.g. **bears of guys*, **jewels of islands*, **idiots of boys*). These unique NPs are called *converse NPs* here because the canonical hierarchical relationship between modifier and modified is upside down, that is, the modified is a dependent of the modifier.

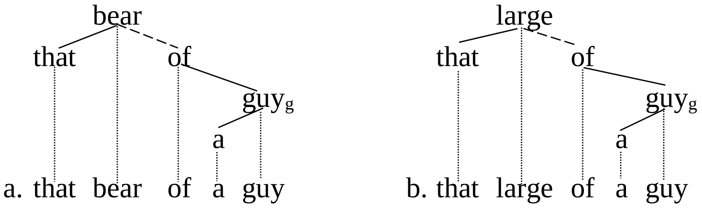
The noun-as-root (or determiner-as-root) analysis is challenged by converse NPs. The difficulty they generate is due in part to the fact that the occurrence of the preposition *of* is obligatory *(*a bear a guy*) and therefore the structure seems to always match the normal hierarchical relationships in NPs containing an *of*-PP. Examine the following analysis of a “normal” NP in (25):

(25)



It seems natural to extend this structural analysis to converse NPs like *that bear of a guy*. The result, then, is that the structural analyses of the two constructions, converse NPs and the BMC, which, again, are clearly related, are closely similar:

(26)



The difference in modifications relations across (26a) and (26b) is captured in the DG analyses in terms of the dashed dependency edge and $_g$-subscript; they indicate that modification relations are the opposite of normal. The modification relations in (26b) are indeed upside down, the modifier *that large* hierarchically dominating what it modifies, i.e. *a guy*.

To summarize the point, positioning the adjective as the root/head of the BMC establishes parallelism in structure across converse NPs and the BMC. If the noun or the determiner were the root/head of the BMC, this parallelism would not obtain.

5.3 Extraposition

An *of*-PP of the BMC can be extraposed in the same manner that the *of*-PP of a normal NP can be extraposed. Such instances of extraposition are most acceptable when the relevant NP is predicative and questioned:

(27) a. Which picture **of your friend** was it?
b. Which picture was it **of your friend**?

(28) a. Which analysis **of that problem** was it?
b. Which analysis was it **of that problem**?

This pattern repeats itself in cases of the BMC:

(29) a. How reliable **of a friend** is he?
b. How reliable is he **of a friend**?

(30) a. How typical **of a politician** is she?
b. How typical is she **of a politician**?

While the a-sentences are perhaps preferable, the b-sentences are passable. What these examples suggest is that the structure of a BMC phrase such as *how reliable of a friend* is similar to the structure of the NP *which picture of your friend*. In both cases, the *of*-PP can be extraposed.

When the *of*-PP is extraposed from an object or subject phrase, acceptability decreases. This reduction in acceptability is, however, consistent across normal NPs and the BMC:

Extrapostion from object NP
(31) a. Which picture **of your friend** do you like?
b. $^?$Which picture do you like **of your friend**?
(32) a. How difficult **of a problem** did you solve?
b. $^{??}$How difficult did you solve **of a problem**?

Extrapostion from subject NP

(33) a. Which picture **of your friend** is best?
b. *Which picture is best **of your friend**?

(34) a. How difficult **of a problem** was given?
b. *How difficult was given **of a problem**?

These examples all demonstrate that the potential to extrapose the of-PP of the BMC is approximately the same as the potential to extrapose the of-PP of a normal NP. This supports the adjective of the BMC as the root of the phrase, since only in this manner is the parallelism in structure achieved across the BMC and normal NPs.

5.4 Left elbows and extraposition within NP

An established fact about the structure of NPs in English (and other languages) is that a pre-modifier of the root/head noun cannot itself be modified by a post-modifier. Osborne (2003) investigates the phenomenon from a DG perspective. He characterizes the relevant constraint as a "ban on left elbows". Some examples of the sort he discusses are illustrated next in the a-, b-, and c-examples:

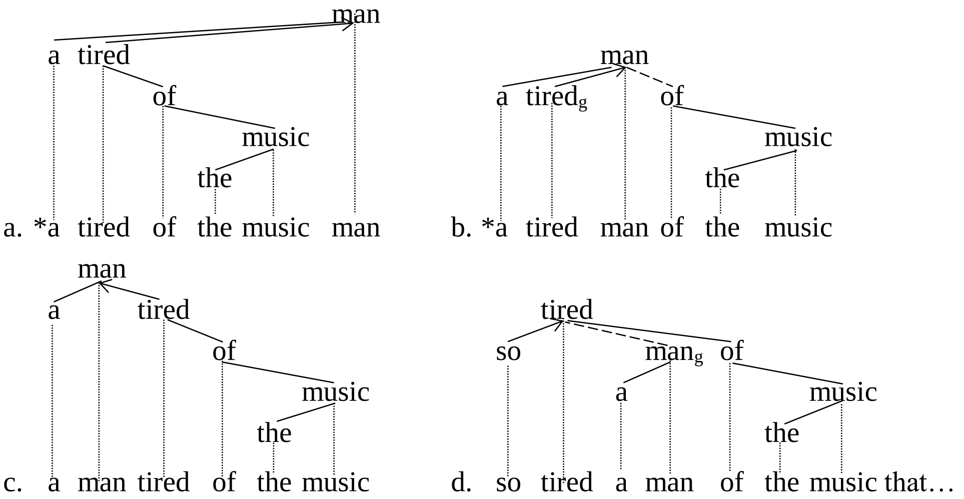
(35) a. *a tired of the music man
b. *a tired man of the music
c. a man tired of the music
d. so tired a man of the music that…

(36) a. *a satisfied with her grade student
b. *a satisfied student with her grade
c. a student satisfied with her grade
d. too satisfied a student with her grade to….

The a-sentences illustrate that the entire AP, *tired of the music* and *satisfied with her grade*, cannot precede the noun that it modifies. The b-sentences show also that the complement of the adjective alone cannot be extraposed to the other side of the noun. In contrast, the NPs are fine if the entire AP is positioned after the noun, as demonstrated with the c-examples. The d-examples are the relevant ones in the context of the BMC. We see there that the complement of the adjective can in fact be separated in the linear dimension from its head adjective by the noun; the BMC allows this.

The following DG structures illustrate how the phenomenon is addressed in the current DG framework:

(37)

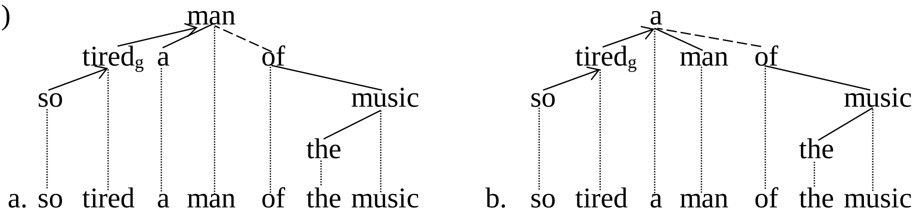


man
a tired
of
music
the
a. *a tired of the music man
man
a tired_g of
music
the
b. *a tired man of the music
man
a tired
of
music
the
c. a man tired of the music
tired
so man_g of
a music
the
d. so tired a man of the music that…

Osborne (2003) addresses the ungrammaticality of examples like (37a) in terms of a ban that blocks a pre-dependent of a noun from itself taking a post-dependent. The constraint is likely due to the grammar's desire to reduce center embedding and thus render NPs easier to produce and process in general. Center-embedding increases dependency distance values and is hence a drag on the efficient production and processing of syntactic structures. The ungrammaticality of (37b) is addressed in terms of the discontinuity: for some reason, the PP complement of an attributive adjective cannot be extraposed. The grammaticality of example (37c) is expected insofar as it violates neither of the previous two constraints. Example (37d), the most relevant one from the perspective of the BMC, also violates neither of the two constraints, for there is no pre-dependent of the noun that itself takes a post-dependent, nor is extraposition of *of the music* present, since the PP *of the music* remains a dependent of the adjective *tired*.

The point to these examples is that the current analysis of the BMC is congruent with both the ban on left elbows in NPs and the inability of extraposition from an attributive adjective to occur within NPs. If, in contrast, the noun or determiner were the root, the resulting structures would not be congruent with these constraints. In particular, both would contradict the ban on extraposition within the NP:

(38)


man
tired$_g$
a
of
so
music
the
a. so tired a man of the music
a
tired$_g$
man
of
so
music
the
b. so tired a man of the music

On these structural analyses, the PP *of the music* has been extraposed within the NP. The same is true of the phrase structures that would result from translation (dependency → phrase structure).

6 Conclusion

This manuscript has presented a novel account of the BMC couched in a DG approach to syntax. It has been argued that the adjective is in fact the syntactic root/head of the BMC. If space had allowed, further sources of support for the analysis could and would have been produced, such as data from gapping and the recursiveness of embedding, that is, one instance of the BMC can be embedded in another, e.g. $^?$*I can't believe that Trump actually wrote that long of that insulting a tweet*.

References

Sosei Aniya. 2016. The Big Mess Construction straightened out. *Studies in Human Sciences Bulletin of the Graduate School of Integrated Arts and Sciences*, Hiroshima University, vol. 11, 1–12.

Hans Bennis, Norbert Corver and Marcel den Dikken. 1998. Predication in nominal phrases. *The Journal of Comparative Germanic Linguistics*, 1:85–117.

Arlene Berman. 1974. *Adjectives and adjective complement constructions in English*. Ph.D. dissertation, Harvard University.

Joan Bresnan. 1973. Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3):275–343.

Norbert Bröker. 1999. *Eine Dependenzgrammatik zur Kopplung heterogener Wissensquellen*. Niemeyer, Tübingen.

Hans-Werner Eroms and Hans Heringer. 2003. Dependenz und lineare Ordnung. In Vilmos Ágel et al. (eds.), *Dependency and Valency: An International Handbook of Contemporary Research*, vol. 1, 247–262. Walter de Gruyter, Berlin.

Christopher Kennedy and Jason Merchant 2000. Attributive comparative deletion. *Natural Language and Linguistic Theory*,18:89–146.

Thomas Groß and Timothy Osborne. 2009. Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics*, 22:43–90.

Liliane Haegeman and Jacqueline Guéron. 1999. *English Grammar: A Generative Introduction*. Blackwell Publishers, Oxford, UK.

Hans Heringer. 1996. *Deutsche Syntax: Dependentiell*. Stauffenburg, Tübingen.

Rodney Huddleston and Geoffrey K. Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Paul Kay and Ivan Sag. 2012. Discontinuous dependencies and complex determiners. In Hans Boas and Ivan Sag (eds.), *Sign-Based Construction Grammar*, 229– 256. CSLI Publications, Stanford, CA.

Jong-Bok Kim and Peter Sells. 2011. The Big Mess Construction: Interactions between the lexicon and constructions. *English Language and Linguistics*, 15(2):335–362.

Aleš Klégr. 2010. Noun phrases with so-adj predeterminers: *So complicated a matter*. In *...for thy speech bewrayeth thee. A Festschrift for Libuše Dušková*, 93–119. Filozofická fakulta, Praha.

Henning Lobin. 1993. *Koordinationssyntax als prozedurales Phänomen*. Series: *Studien zur deutschen Sprache 46*. Gunter Narr Verlag, Tübingen.

Timothy Osborne. 2003. The Left Elbow Constraint. *Studia* Linguistica, 57(3): 233–257.

Timothy Osborne. 2014. Type two rising: A contribution to a DG account of discontinuities. In Kim Gerdes, Eva Hajičova and Leo Wanner (eds.), *Dependency Linguistics: Recent Advances in Linguistic Theory Using Dependency Structures*, 274–298. John Benjamins, Amsterdam.

Timothy Osborne, Michael Putnam and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Klaus Schubert. 1987. *Metataxis: Contrastive Dependency Syntax for Machine Translation*. Foris Publications, Dordrecht.

Stanley Starosta. 2003. Lexicase Grammar. In Ágel et al. (eds.), *Dependency and Valency: An International Handbook of Contemporary Research*, vol. 1, 526–545. Walter de Gruyter, Berlin.

Frank Van Eynde. 2007. The Big Mess Construction. In Stefan Müller (ed.), *Proceedings from the 14th International Conference on Head-Driven Phrase Structure Grammar*, 416–433. CSLI Publications, Stanford, CA.

Johanna Wood and Sten Vikner. 2011. Noun phrase structure and movement: A cross-linguistic comparison of such/sådan/solch and so/så/so. In Petra Sleeman and Harry Perridon (eds.), *The Noun Phrase in Romance and Germanic-Structure, Variation and Change*, 89–109. John Benjamins, Amsterdam.

Arnold Zwicky. 2007. Exceptional degree markers: A puzzle in internal and external syntax. *OSU Working Papers in Linguistics*, 47:111–23.

Cliticization of Serbian Personal Pronouns and Auxiliary Verbs A Dependency-Based Account

Jasmina Milićević
Dalhousie University
Halifax, Canada
jmilicev@dal.ca

Abstract

The paper looks into cliticization of Serbian personal pronouns and auxiliary verbs. *Cliticization* is the operation whereby, in the process of clause construction, a clitic (= unstressed) form of a pronominal/verbal lexeme is chosen, rather than a full (= stressed) form. Cliticization of both pronouns and auxiliaries is obligatory under neutral communicative conditions (i.e., in the absence of contrast or emphasis) and unless specific syntactic/prosodic factors impose the choice of a full form. Under marked communicative conditions, cliticization is precluded. Corresponding rules are proposed within a Meaning-Text dependency framework.

1 Overview of the Problem

Personal pronouns and auxiliary verbs in Serbian (and all other languages stemming from former Serbo-Croatian) have both full (= stressed, tonic) and clitic (= unstressed) forms, the latter being so-called *second-position* clitics (Halpern & Zwicky, eds, 1996). In any sentence featuring pronouns and/or auxiliaries, the choice between full and clitic forms is obligatory, which means that the opposition "tonic ~ clitic" is inflectional in nature.

> The operation whereby the inflectional value (= a grammeme) CLITIC is assigned to a lexical item, in the course of clause synthesis, is called *cliticization*[1].

Roughly speaking, cliticization of both personal pronouns and auxiliary verbs is obligatory under neutral communicative conditions (i.e., in the absence of contrast or emphasis) and unless specific syntactic/prosodic factors impose the choice of a full form. Under marked communicative conditions, cliticization is precluded. It is precisely these conditions that the paper intends to specify.

Here are some preliminary examples of the use of clitic vs. full pronominal and verbal forms; as most examples in the paper, these are taken from the Serbian corpus (*Korpus savremenog srpskog jezika*: www.korpus.matf.bg.ac.rs).

(1) a. *Možda* ***me je*** *Mira podsticala na brbljivost. Gledala* ***me je*** *netremice* ...
lit. 'Maybe me is Mira having.incited on volubililty. [She] having.looked me is intently…'
'Maybe Mira was inciting my volubility. She was looking at me intently…'

b. *No, bilo kako bilo, prepoznao ga* ***jeste***.
lit. 'But, be it as it may, having.recognised him [he] is.'
'But, be it as it may, he did recognize him.'

c. *Ali nije gledala* ***njega****, gledala je* ***mene***.
lit. 'But [she] is.not having.looked him, having.looked [she] is me.'
'But she wasn't looking at him, she was looking at me.'

Example (1a) illustrates a communicatively unmarked context, where clitic forms are used by default and the corresponding full forms would be inappropriate; we see here instances of the accusative 1p pronominal clitic, **me** 'me', and the 3sg past tense auxiliary clitic, **je** 'is'. In sentence (1b), a full form of the past tense auxiliary is used contrastively—to insist that the fact of recognizing did take place; note also a marked word order, with the auxiliary clause-final. The corresponding clitic auxiliary is possible here if the contrast is expressed lexically: […] <u>*zaista*</u> *ga* ***je***

[1] The term *cliticization* has at least another two usages that I do not subscribe to: 1) a diachronic process of becoming a clitic; 2) the operation of attachment of a clitic to its host.

prepoznao ‘[…] really him is [he] having.recognized’. Finally, the use of full personal pronouns **njega** ‘him’ and **mene** ‘me’ in sentence (1c) is warranted by the contrastive focus they bear; in this type of context clitic forms are excluded.

While some other aspects of clitic behavior, in particular their linear placement, have been extensively researched, cliticization (in the sense intended here) has received less attention. Kayne (1975) is a seminal study of cliticization in French, which has served as a springboard for work on this phenomenon in other languages. A discussion of cliticization in Slavic languages can be found, for instance, in Dimitrova-Vulčanova (1999) and Franks (1998 and 2010); the most complete existing account of the cliticization in Serbian/Croatian is the one in Browne (1975: 276-282). Some aspects of the problem were addressed in Progovac (2005: 126-136), Mrazovac, 2009: 364-366), and (in a different perspective) Caink 2000; Peti-Stantić (2017 and 2018) reports on some recent research on the topic on Croatian data.

Cliticization is theoretically interesting because it involves the interplay of the syntactic and communicative (a.k.a. information) structures in sentence production and is linked to other important phenomena such as subject ellipsis and conjunction reduction.

In the remaining part of this Section, I provide some basic facts about Serbian lexical items susceptible of undergoing cliticization (1.1) and describe the essentials of the theoretical framework adopted (1.2). Conditions under which the cliticization of personal pronouns and auxiliaries occurs are informally characterized in Section 2; their formal description, in terms of rules belonging to a Meaning-Text linguistic model, is offered in Section 3; Section 4 is reserved for a conclusion.

1.1 Full and clitic forms of personal pronouns and auxiliaries

As indicated above, cliticizable lexical items in Serbian include personal pronouns and auxiliary verbs.[2] The paradigms of three personal pronouns and two auxiliary verbs follow;[3] the stressed vowel (in the full forms) is boldfaced; tonal accents are not shown.

	JA ‘I’		ON ‘he’		VI you [PL] ’	
	TONIC	CLITIC	TONIC	CLITIC	TONIC	CLITIC
NOM	**ja**	——	on	——	v**i**	——
ACC/GEN	m**e**ne	me	nj**e**ga	ga	v**a**s	vas
DAT	m**e**ni	mi	nj**e**mu	mu	v**a**ma	vam
INSTR	mn**o**m(e)	——	nj**i**m(e)	—	v**a**ma	——
LOC	m**e**ni	——	nj**e**mu	——	v**a**ma	——
VOC	——	——	——	——	v**i**	——

BITI ‘be’ in the present, past tense aux.				
	SG		PL	
	TONIC	CLITIC	TONIC	CLITIC
1	j**e**sam	sam	j**e**smo	smo
2	j**e**si	si	j**e**ste	ste
3	j**e**ste	je	j**e**su	su
HTETI lit. ‘want’ in the present, future tense aux.				
1	h**o**ću	ću	h**o**ćemo	ćemo
2	h**o**ćeš	ćeš	h**o**ćete	ćete
3	h**o**će	će	h**o**će	će

Table 1: Full and clitic forms of some personal pronouns and auxiliary verbs

Pronominal clitic forms exist in the accusative, genitive and dative. The nominative, i.e., subject, pronouns are never cliticized; they are dropped in neutral communicative conditions (Serbian is a PRO-Drop language). Oblique case personal pronouns, whether full or clitic, function as objects of verbs, nouns and adjectives.

The auxiliary BITI ‘be’ has the forms identical to that of the copula and the locative verbs; all three verbs exhibit identical behavior with respect to cliticization and linear placement. A finite auxiliary, whether full or clitic, is the head of its clause (Milićević, 2009b) and the top node of the corresponding dependency tree (see immediately below).

1.2 The Framework

Within a Meaning-Text linguistic model, a semantically-driven, dependency-based, synthesis-oriented stratificational model (Mel’čuk, 2016: 41-85), cliticization happens in the transition between the Surface-Syntactic Representation [SSyntR] and the Deep-Morphological Representation [DMorphR] of a clause. Formally, the basic structure of the SSyntR is a (linearly non-ordered) dependency tree; that of the DMorphR is a (fully ordered) string.

[2] In addition, the interrogative conjunction DA LI has a clitic form, $\text{LI}_{\text{INTERR}}$ (homophonous with the emphatic particle $\text{LI}_{\text{EMPHATIC}}$, with no corresponding full form); it will not be considered in this paper.

[3] There is a third auxiliary, BITI in the aorist tense, used to construct the conditional mood forms; it is currently undergoing grammaticalization and becoming a particle, just like its cognate in Russian.

Cliticization is part of the operation of *morphologization*, whereby lexemes in the SSyntS are assigned syntactic inflectional values. Two other major operations—*linearization* and *prosodization* of the SSyntS—are part of this transition, which is guided in an essential way by the communicative structure (Mel'čuk, 2001) of the clause under synthesis.

During linearization, all lexemes of the clause that have been assigned the grammeme CLITIC (including auxiliary verbs) are gathered in a *clitic cluster* and linearly positioned together, according to special linearization rules (Milićević, 2009a)—not with respect to their governors, but with respect to a *host*. The clitics are by default positioned after the first available host, which means that they often "land" clause-second (whence their name).

Full pronominal forms obey the same linearization rules as full-fledged nominal complements; their linear positioning is normal in that it is done taking their governor(s) as the reference point. A full finite auxiliary is the reference point for the linearization of all other clause elements, just as a finite lexical verb is.

Since our dependency trees are not linearly ordered, for two (or more) clauses containing items that differ only along the "tonic ~ clitic" opposition, the basic dependency structures are identical; their respective communicative structures are different, and so are, of course, their DMorphSs. As an illustration, the corresponding structures for sentences in (2) are given in Figure 1; an underlying question [in square brackets] is supplied for each sentence, providing a minimal communicative context in which it can felicitously be uttered.

(2) a. [Did you tell him?]
Rekao ***sam mu***.
'Having.told [I] am to.him.' = 'I told him.'

b. [Who did you tell?]
Rekao ***sam njemu***.
'To.him [I] am having.told.' = 'It's to him that I told.'

c. [Why didn't you tell him?]
Jesam mu *rekao*.
'[I] am to.him having.told.' = 'I did tell him.'

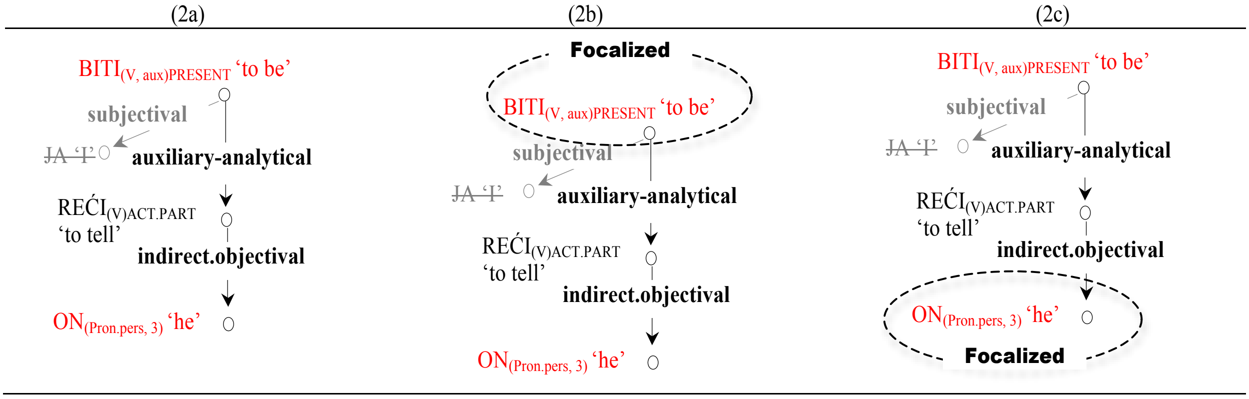


Figure 1: SSyntSs of sentences in (2) with communicative information specified

(2a)	REĆI$_{\text{ACT.PAST.PART, SG, MASC}}$ [BITI$_{\text{PRES, }\underline{\textbf{CLIT}}\text{, 1, SG}}$ ON$_{\underline{\textbf{CLIT}}\text{, SG, MASC}}$]
(2b)	REĆI $_{\text{ACT.PAST.PART, SG, MASC}}$ [BITI $_{\underline{\text{CLIT}}\text{, 1, SG}}$] ON $_{\underline{\textbf{FULL}}\text{, SG, MASC}}$
(2c)	BITI $_{\text{PRES, }\underline{\textbf{FULL}}\text{, 1, SG}}$ [ON$_{\underline{\textbf{CLIT}}\text{, SG, MASC}}$] REĆI$_{\text{ACT.PAST.PART, SG, MASC}}$

Figure 2: DMorphSs of sentences in (2)

Remarks:

1) In all the structures in Figure 1, the pronominal subject is slated for deletion since it is communicatively unmarked; it gets deleted in the transition towards the morphological string.
2) The branches of the dependency tree are labeled with language-specific Surface-Syntactic Relations [SSyntRels]; for more on these and the whole Meaning-Text dependency framework, see Polguère & Mel'čuk, eds (2009).
3) Linearizations other than those shown in Figure 2 are possible, without modifying the tonicity status of the auxiliary and the personal pronoun.

In the SSyntS of (2a), the auxiliary BITI 'to be' and the pronoun ON 'he' bear no marked communicative values and neither of them appears within a syntactic configuration which does not allow for cliticization (for instance, in coordination or as the only word in a clause); therefore, they are both assigned the grammeme CLITIC, which appears in the DMorphS of (2a).

The communicative value **Focalized**, assigned to the pronoun ON 'he' in the SSyntS of (2b), marks it as logically prominent with respect to some contextual information (cf. the corresponding underlying question); it is this communicative marking that triggers the assignment of the grammeme TONIC to the pronoun in the transition towards the morphological string. An analogous situation obtains with the auxiliary BITI 'to be' in the structures underlying (2c).

This architecture of the Meaning-Text Model determines the form of cliticization rules: they are transition rules, operating between (fragments of) SSyntRs and DMorphRs of utterances and having as conditions the communicative load and syntactic/prosodic environment of the items whose tonicity status they specify.

2 Factors Relevant for Cliticization of Personal Pronouns and Auxiliary Verbs

The use of clitic vs. full forms of pronouns and auxiliaries is determined both by communicative factors and syntactic/prosodic ones. Three cases can be distinguished.

Case 1
A full form of a PRON/$V_{(Aux)}$ is freely chosen to express a value of a *communicative opposition* (Mel'čuk 2001: 93-258):

- The value **Focalized** (the marked value of the Focalization opposition) or/and the value **Emphatic** (the marked value of the Emphasis opposition).

(3) a. *Nije pričao* ***meni****, već drugovima.*
'He was not telling [this] to me, but to [his] friends.'

b. *Šta će* ***meni*** *filozofija!* ***Meni*** *se živi, voli,* ***meni*** *se hoće sreće.*
'What FUT.1SG to.me philosophy! To.me REFL lives, loves, to.me REFL wants of.happiness.' =
'What do I need philosophy for? I want to live, to love, I want happiness.'

(4) a. *U tom smislu zaista* ***jesam*** *spreman da se izvinim i gospodinu Cvetkoviću.*
'In that sense I really AM ready to apologize also to Mr. Cvetković.'

b. *Kad bi mu rekla da ga voli, on bi joj odgovarao: E,* ***jesi*** *teška guska!*
'When she would tell him that she loved him, he would answer: Well, you ARE a silly goose.'

We see focalized items in (3a) and (4a); those in (3b) and (4b) are emphatic.
Cf. also (1b) and (1c).

- The rhematic focus

(5) [Kome kažeš? 'To whom are you saying (that)?']
a. ***Njemu****.*
'To.him.'
b. *Kažem* ***njemu*** / $^{\#}$***mu****.*
'I.am.saying'

A pronoun used as an answer to a WH-question carries the rhematic focus and must appear in a full form. This holds not only when it is clause-initial/the only word in the clause (this environment being unavailable for an enclitic for prosodic reasons), as in (5a), but also when it appears clause-internally, as in (5b), where *Kažem mu*, otherwise a fully grammatical sentence, is inappropriate.

Case 2

A full form of a PRON/V$_{(\mathrm{Aux})}$ is imposed by syntactic/prosodic factors (rather than freely chosen to express some communicative opposition values).

1) The word order constraints are such that a PRON/V$_{(\mathrm{Aux})}$ must be/preferably is clause-initial or follows an internal prosodic break (i.e., it finds itself in a linear position unavailable for an enclitic).

(6) a. *On deluje pošteno.* ***Njemu*** *se veruje i on je sad najpopularniji ministar u vladi.*
'He seems honest. To.him REFL trusts = He is trusted and he is now the most popular minister in government.'

b. [*Da li* ***je*** *slika kod vas?* 'Is (the) picture with you?']

(i) ***Jeste****.*
lit. 'Is.' = 'Yes, it is.'

(ii) *Da, kod nas* ***je*** */ *jeste.*
lit. 'Yes, with us'

c. *Salinitet, ili slanoća,* ***jeste*** */ *je količina soli u morskoj vodi.*
'Salinity, or saltiness, is the quantity of salt in sea water'.

The pronoun in (6a) preferably appears in the clause-initial position (because it functions as a semantic theme within a thematic progression sequence) and is therefore full; however, it could have been used in the corresponding clitic form clause-internally ([…] *Veruje* ***mu*** *se i sada je najpopularniji …*).

A full form of the auxiliary is standardly used as an elliptic (only-word) affirmative answer to a YES/NO question, as shown in (6b-i).[4] When not clause-initial, as in (6b-ii), a V$_{(\mathrm{Aux})}$ must appear in a clitic form. This contrasts with the behavior of personal pronouns in the same syntactic environment; cf. (5b).

In (6c), a full form of the auxiliary BITI 'be' is used because it follows an internal prosodic break (marked by a comma in writing).[5]

2) Coordination

A pronoun used in coordination (with another pronoun or with a noun) must be full, as illustrated in (7); however, this restriction does not hold for the auxiliaries, as shown in (8b) and (8c).[6]

(7) a. *Mada bih, u tom slučaju, lišio i* ***nju*** *i* ***njega*** *dubokog, radosnog uzbuđenja.*
'Although I would, in that case, deprive both her and him of deep, gleeful excitement.'

b. *Pričala je uz kafu,* ***meni*** *i mojoj* ***supruzi****, na kakve je sve prepreke na Ketedri nailazila.*
'She was telling over coffee, to me and my wife, about different obstacles she was facing at the Department.'

(8) a. ***Je*** *li on član kluba ili* ***nije****?*
'Is INTERR he member of.club or not.is?' = 'Is he or not a club member?'

b. *Ne zanima me to, ali* ***jesam*** *i* ***biću*** *patriota.*
'This doesn't interest me, but I am and will be a patriot.'

c. *Bio* ***sam*** *i* ***jesam*** *potpuno svestan svojih postupaka.*
'Having.been am and am = I was and still am completely aware or my actions.'

[4] Negative forms of auxiliaries can of course be used in negative answers, but they are always full, so the question of their tonicity status does not arise.

[5] This rule is often transgressed in journalistic and informal styles.

[6] Note that ***je*** 'is' in (8a) is a full form; in Serbian, it is used only in questions formed by means of the interrogative particle LI, but in Croatian it can also appear in answers to such questions.

3) Prepositions and conjunctions

(9) a. *Mislim na*$_{\text{(Prep)}}$ ***nju***.
‘I am thinking of her.’

b. *Sviđa mi se to. A*$_{\text{(Conj)}}$ ***vama****?*
‘Likes to.me REFL that = This is likable to me. And to.you?’

c. *I baš zato što je to istina cela stvar i*$_{\text{(Conj)}}$ ***jeste*** *tako smešna!*
‘And precisely because this is true the whole thing and is so funny = is so funny in the first place.’

Pronouns as propositional objects must appear in a full form.[7] No communicative load is attached to the full form; to express focalization, prosody is used (symbolized by capitalization in our examples): *Mislim na NJU* ‘It is of her that I am thinking’.
Some “focalizing” conjunctions impose the use of a full form a PRON/V$_{\text{(Aux)}}$.

4) Presence of a specific dependent [for pronouns only]

A pronoun governing a restrictive modifier (*baš* ‘precisely’, *samo* ‘only’, *jedino* ‘uniquely’, *isključivo* ‘exclusively’, …) must appear in a full form. (Again, we could say that such a modifier has a focalizing effect, and that this triggers the assignment of the grammeme TONIC to the pronoun.)

(10) a. *Zašto baš* ***tebi****?*
‘Why precisely to.you?’

b. *Može samo* ***meni*** *nešto da se desi.*
‘Can only to.me something that$_{\text{(Conj)}}$ REFL happens’ = ‘Something can happen only to.me.’

5) Presence of a specific co-dependent [for pronouns only]

(11) a. *Predstavi me/nas* ***njemu.***
‘Introduce me/us to.him’

b. *Predstavi *mu me/nas* vs. *Predstavi mu ga*.
‘Introduce to.him me/us.’ ‘to.him him’

If a dative and a 1/2p accusative pronoun cooccur, one of them must appear in the full form; cliticizing both pronouns leads to ungrammaticality. The incompatibility of dative – accusative clitic sequences is known in other Slavic languages, for instance Bulgarian (Franks 1998: 85), as well as in Romance languages (Miller & Monachesi 2003: 87ff).

Case 3
A clitic form of a PRON/V$_{\text{(Aux)}}$ is chosen by default, i.e., if no communicative load is attached to it and no syntactic/prosodic factors are present which preclude cliticization.

(12) a. *Na vreme* ***ću vas*** *obavestiti.*
‘On time FUT.1SG you to.notify.’ = ‘I will notify you in time.’

b. *Da* ***sam*** *znala, ne* ***bih vam*** *nista rekla.*
‘That$_{\text{(Conj)}}$ [I] am having.known, not [I] would to.you nothing having.said.’ = ‘Had I known, I wouldn’t have told you anything.’

[7] Except if the stress is shifted to the proposition; cf. *Na te mislim kada zora sviće* ‘Of you I think when the dawn breaks’ (a line from a popular song); this kind of stress shift is (in most cases) optional and stylistically marked as poetic, dated or regional.

c. *Teren u Podgorici* ***je*** *bio veoma težak za igru ali* ***smo mu*** *se prilagodili i ostvarili cilj.*
‘Field in Podgorica is being.been very difficult for play but [we] are to.it REFL having.adapted and having.reached goal.’ = ‘The football field in P. was very difficult to play on, but we adapted to it and reached the goal.’

b. *Poznata* ***mi je*** *ta priča. Znam da* ***ti je*** *poznata.*
‘Known to.me is that story. [I] know that$_{(Conj)}$ to.you is known.’ = ‘I know the story. I know that you do.’

Cf. also (1a).

Remark:
For some pronouns and auxiliaries appearing in set expressions, tonicity value is fixed; for instance: [TONIC] *Što se* L$_{(Pers.pron)GEN,\ \mathbf{FULL}}$ *tiče*; ‘As for L’ [marks a Focalized Theme]; *Teško <Blago>* L$_{(Pers.pron)DAT,\ \mathbf{FULL}}$ ‘Woe/Joy to L’; *Što jes(te) jes(te)* Lit. ‘What is is’ = ‘This is uncontestable’; etc. [CLITIC] *Eto* ***ti ga*** *sad!* ≈ ‘What’s this, all of a sudden’ [marks surprise and disapproval]; *Šta* ***(ti) ga*** *znam* ‘What do I know’; etc.

As shown above, in most cases, clitic and full forms of personal pronouns are in complementary distribution, and so are clitic and full forms of auxiliaries. There are two types of situation where this does not hold.

1) In some unmarked contexts, either a clitic or a full form is possible without any perceptible communicative difference: *Čini* ***mi***$_{\mathbf{CLITIC}}$ *se da …* ‘[It] seems to.me REFL that$_{(Conj)}$ …’ <***Meni***$_{\mathbf{FULL}}$ *se čini da…* > ‘To.me [it]seem REFL that$_{(Conj)}$ …’

2) In some neutralizing contexts, the communicative load carried by a full form is also expressed by another clause element; thus, sentence *Stvarno* ***jeste***$_{\mathbf{FULL}}$ *tako* ‘Really [it] IS like.that’, in which the adverb STVARNO ‘really’ provides a neutralizing context, allows for a paraphrase making use of the corresponding clitic form of the auxiliary: *Stvarno* ***je***$_{\mathbf{CLITIC}}$ *tako* ‘Really [it] is like.that’. Also, interchangeability of a full and a clitic form is possible if the communicative load carried by a full form can alternatively be expressed by a lexical mean: *Jeste*$_{\mathbf{FULL}}$ *tako* <*Stvarno* ***je***$_{\mathbf{CLITIC}}$ *tako*>.

3 Cliticization Rules for Personal Pronouns and Auxiliary Verbs

To account for the fact described in Section 2, two cliticization rules are needed, one for the pronouns and another one for the auxiliaries; they are given in Figures 3 and 4, respectively. (Shaded areas in the left-hand side of a rule indicate the context of its application. Both rules are a “short hand” for several more specific rules.)

SSynt-level DMorph-level

L$_{(Pron.pers)}$ ⇔ L$_{(Pron.pers)CLIT}$ | L is NOT 1) communicatively marked
2) placed clause-initially or after a clause-internal prosodic break
3) the governing member of the **coordinative** SSyntRel
4) the governing member of the **restrictive** SSyntRel
5) the dependent member of the **prepositional** or **conjunctional** SSyntRel

Figure 3: Cliticization rule for personal pronouns

According to this rule, the cliticization of personal pronouns will take place in all cases except those illustrated in (1c), (2b), (3), (5), (6a), (7) (9a/b) and (10). As for the case illustrated in (11), it will be taken care of by filter rules presiding over the construction of the clitic cluster (Milićević, 2007: 109-114).

SSynt-level DMorph-level

L$_{(V,\ Aux)}$ ⇔ L$_{(V,\ Aux)CLIT}$ | L is NOT 1) communicatively marked
2) placed clause-initially or after a clause-internal prosodic break
3) the dependent member of the **conjunctional** SSyntRel

Figure 4: Cliticization rule for auxiliary verbs

This rule will allow for the cliticization of auxiliary verbs in all cases except those illustrated in (1b), (2c), (4), (6b-i), (6c), (8b/c) and (9c).

4 Conclusion

The use of clitic forms of Serbian personal pronouns and auxiliary verbs is the default case, while using tonic forms requires additional conditions. Tonic forms are either freely chosen to express marked values of communicative oppositions or are imposed by specific syntactic configurations/prosodic environments. This is in line with the conclusions of Peti-Stantić (2018) for Croatian; cf.: "Short, clitic forms [in Croatian] are the first (and the only) choice in informationally neutral contexts".

Tonic forms are more prominent morphologically and syntactically: unlike clitics, which are deficient, stress-lacking wordforms, they are full-fledged wordforms and full-fledged sentence elements, less restricted in their linear positioning. Thus, being tonic is a sort of a promotion. It is not surprising, then, that tonic forms appear under more involved communicative/syntactic conditions.

To what extent are the conditions that license cliticization similar cross-linguistically? Are the factors identified above for Serbian 2P clitics applicable to clitics of other types? I would expect communicative factors to be more generally applicable than syntactic factors, but this has yet to be determined on a large enough sample of languages.

Given the fact that in some cases a full form of a pronoun/auxiliary is selected freely, to express a communicative opposition, we could ask whether tonicity is really (or only) a syntactic inflectional category. It looks like in this case a syntactic inflectional category has been "enlisted" to express some semantic/communicative information. This situation is similar to gender conversion as a means of expressing some derivational meanings (e.g., in Spanish) or a change of nominal class in order to express plurality (e.g., in Bantu languages), where a syntactic feature is pressed into service for word formation or inflection purposes.

Acknowledgments

I am grateful to Wayles Browne and Igor Mel'čuk for their comments on a pre-final version of this paper.

References

Wayles Browne. 1975. Serbo-Croatian Clitics for English-speaking Learners. *Kontrastivna analiza engleskog i hrvatskog ili srpskog jezika,* Zagreb: Institut za lingvistiku Filozofskog fakulteta; 105-134. [Reprinted in *Journal of Slavic Linguistics*, 12(1-2), 249-283, 2004.]

Andrew Caink. 2000. Full Form Auxiliaries in Serbian/Croatian/Bosnian. In: T. H. King & I. Sekerina, eds, *Formal Approaches to Slavic Linguistics, 8. The Philadelphia Meeting*. Ann Arbor: Michigan Slavic Publications; 61-77.

Mila Dimitrova-Vulčanova. 1999. Clitics in the Slavic Languages. In: H. van Riemsdijk ed., *Clitics in the Languages of Europe*. Berlin/New York: Mouton de Gruyter; 83-123.

Steven Franks. 1998. Clitics in Slavic. Position paper. Workshop "Comparative Slavic Morphosyntax". Spencer, Indiana, 5-7 June 1998; pp. 96.

Steven Franks. 2010. Clitics in Slavic. *Glossos*, 10; 1-157.

Aaron Halpern and Arnold Zwicky. 1996. *Approaching Second: Second Position Clitics and Related Phenomena*. Stanford (CA): CSLI.

Richard Kayne. 1975. *French Syntax. The Transformational Cycle*. Cambridge (MA): MIT Press.

Korpus savremenog srpskog jezika: www.korpus.matf.bg.ac.rs

Igor Mel'čuk. 2001. *Communicative Organization in Natural Language*. Amsterdam/Philadelphia: John Benjamins.

Igor Mel'čuk. 2016. *Language. From Meaning to Text*. Moscow/Boston: Academic Studies Press.

Phillip Miller and Paola Monachesi. 2003. Les pronoms clitiques dans les langues romanes. In: D. Godard, ed., *Les langues romanes, problèmes de la phrase simple*. Paris: CNRS Éditions.

Jasmina Milićević. 2007. Co-occurrence of Serbian Second-Position Clitics: Syntactic and Morphonological Constraints. In: N. Nathout and F. Montermini, eds, *Morphologie à Toulouse. Actes du colloque international de morphologie 4^e Décembrettes*. München: Lincom Europa; 99-121.

Jasmina Milićević. 2009a. Linear Placement of Serbian Clitics. A Description within a Dependency Based Approach. A. Polguère and I. Mel'čuk, eds; 235-276.

Jasmina Milićević. 2009b. Serbian Auxiliary Verbs—Syntactic Heads or Dependents? W. Cichocki, ed. *Proceedings of the 31st Annual Conference of the Atlantic Provinces Linguistic Association*, Fredericton, November 2-3, 2007; 43-53.

Pavica Mrazovac. 2009. *Gramatika srpskog jezika za strance*. Sremski Karlovci/Novi Sad: Izdavačka knjižarnica Zorana Stojadonovića.

Anita Peti-Stantić. 2017. On Wackernagel Ten Years Later: Personal Pronouns in Information Structure. Talk given during 12th Annual Meeting of the Slavic Linguistic Society. Ljubljana, September 23, 2017.

Anita Peti-Stantić. 2018. Klitike naše svagdašnje: Pun ti je kufer? I meni. PPT presentation, Zagreb Linguistic Circle, January 3, 2018.

Alain Polguère and Igor Mel'čuk, eds. 2009. *Dependency in Natural Language*. Amsterdam/Philadelphia: John Benjamins.

Ljiljana Progovac. 2005. *Syntax of Serbian: Clausal Architecture*. Bloomington: Slavica Publishers.

The evolution of spatial rationales in Tesnière’s stemmas

Nicolas Mazziotta
Centre de linguistique française, générale et romane (UR Traverses)
Université de Liège
nicolas.mazziotta@uliege.be

Abstract

This paper investigates the evolution of the spatial rationales of Tesnière’s syntactic diagrams (*stemma*). I show that the conventions change from his first attempts to model complete sentences up to the classical stemma he uses in his *Elements of structural syntax* (1959). From mostly symbolic representations of hierarchy (directed arrows from the dependent to the governor), he shifts to a more configurational one (connected dependents are placed below the governor).

1 Introduction

The constant use of diagrams is one of the most famous characteristics of Tesnière’s *Éléments de syntaxe structurale* (Tesnière, 1959) (henceforth, *Elements*). These diagrams, known as *stemmas* in this work, are visual representations of the syntactic analysis of sentences, i.e. words and the syntactic relations between them. In this paper, I will focus on the comparison between Tesnière’s first stemma – observed in his drafts and in (Tesnière, 1934a) – and the « classical » ones he draws in his *Petite grammaire russe* (Tesnière, 1934b, 157, 162 and 164) as well as in his *Esquisse d’une syntaxe structurale* (Tesnière, 1953) and his *Elements*. From the early stemmas to the classical ones, Tenière changed the graphical elements and the configurational rules at work to achieve what he thought was a better representation of his analysis – his pedagogical concerns were prominent (Tesnière, 2015, Chapters 276-277).

Before I delve into the details of the stemmas, I would like to introduce my perspective and my motivation. My perspective is mostly a semiotic one: the description of the mechanics of diagrams as bidimensional graphical formalisms. Semioticians of peircian obedience (Stjernfelt, 2007) describe diagrams as complex *icons*, i.e. signs that share structural characteristics with what they represent.[1] For instance, a map is an icon of the territory it represents, because areas on the map and symbols used are placed in accordance with the location of elements of interest of the territory (Bertin, 2005). This paper will often deal with a recurrent issue in the description of diagrams: the evaluation of the relevance of the observed graphical elements and the way they are laid out on the plane. What is incidental? What does, indeed, qualify as an icon, and is genuinely diagrammatical? For instance, the thickness of the strokes in all the stemmas reproduced in this paper is incidental. It does not represent anything at the conceptual level: some may be thinner than others with no consequences on information.

My motivations are epistemological and methodological. By emphasizing the evolutions of graphical rationales proposed by a major historical author in the field, I intend to draw attention to the fact that most syntacticians apply formal conventions to encode analyses in graphical inscriptions. These conventions actually constraint what can be expressed.

In this historical survey, I will proceed from the classical diagrams of the *Elements* to the earliest ones, mainly because whereas the formers are sometimes remembered, the latters are almost completely forgotten. I will introduce Tesnière’s conceptual rationales as well as the theoretical foundations of my analysis alongside the analysis itself. In the conclusion, I will show that the characteristics of the alternative types of stemmas are still relevant to solve specific problems in linguistics.

[1] In this conception, which I will not investigate here, diagrams are tools that allow for discovering novel knowledge about what they represent (Stjernfelt, 2007, 99-105).

2 Classical stemmas

The foundational idea in the *Elements* is that words are not the only elements of the sentence, but that there exist syntactic relations that can also be qualified as *elements*:

> a sentence of the type *Alfred speaks* is not composed of just **two** elements, *Alfred* and *speaks*, but rather of **three** elements, the first being *Alfred*, the second *speaks*, and the third the connection that unites them, without which there would be no sentence. To say that a sentence of the type *Alfred speaks* consists of only two elements is to analyze it in a superficial manner, purely morphologically, while neglecting the essential aspect that is the syntactic link (Tesnière, 2015, Chapter 1, § 5, emphasis from the author)

In this section, I first describe how these basic elements are represented in the graphical medium (2.1) and how they combine on each axis of the plane (2.2). To do so, I formulate discursive descriptions of the diagrams. The last subsection deals with more complex configurations (2.3).

2.1 Graphical entities

In accordance with his epistemological stance, Tesnière draws stemmas that consist of arrangements of discrete graphical items. In semiotics, the Groupe μ (1992) proposed the concept of *graphical entities* (or *entities*, for short). Entities are *Gestahlt*, i.e. forms that, from a cognitive perspective, can be identified and described as single objects; e.g.: a dot, an arrow, a face, a car, etc. (henceforth, entities will be noted using a fixed-width font). They do not need to have a meaning to qualify as entities; e.g. an `arrow` can be recognized as such without knowing how to interpret it. This characteristic is especially important in learning procedures: one can recognize `strokes` in fig. 1 (Tesnière, 1959, Chapter 3, § 8) before being instructed about how to understand them – although our background knowledge provides us with very good insight.

(1) Alfred frappe Bernard
 'Alfred hits Bernard'

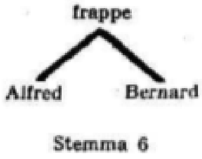


Figure 1: Classical stemma of (1)

In the stemmas of the *Elements*, such as fig. 1, entities represent linguistic signs ("words") as well as relations. In the terminology I will use henceforth, words and relations are analytical concepts that are *reified* (Kahane and Mazziotta, 2015b; Mazziotta, 2016b) in the diagram: entities are used to represent them in a discrete way on the graphical substratum.[2] The basic inventory of graphical entities in these diagrams is thus:

1. Words at use in the sentence are reified by entities that can be called `words`, i.e. a graphical image of the linguistic units.

2. Relations are reified by `strokes`.

[2] The history of syntactic diagrams demonstrates that it is possible to conceive diagrams that do not reify relations. See, e.g. the diagrams by Clark or Reed and Kellogg (Brittain, 1973; Mazziotta, 2016a).

2.2 Configurational rules and super-entities

For them to work as tools for the linguists, stemmas have to be organized in accordance with a specific syntactic analysis.[3]

> "le stemma note toujours une correspondance point par point avec les opérations qu'il est censé représenter" ['elements in the stemma always strictly correspond to the operations it is supposed to represent'] (Samain, 1995, 131, my translation)

Therefore, to fully understand the rationales of the stemmas, one must achieve a description of the rules that govern the combination of entities in correspondence with Tesnière's syntactic epistemology (i.e. the represented "operations"). To do so, one needs to remember the basic theoretical rationales of tesnierian syntax.

Syntactic rationales. There are two major kinds of syntactic relations that are reified in classical stemmas. The first one is *connection* (Tesnière, 2015, Part 1). A connection between two words is hierarchized and asymmetrical. It corresponds to a subordination relation and it is very close to the modern concept of syntactic dependency. Connections in (2) are illustrated in fig. 2 (Tesnière, 1959, Chapter 3, § 8). The second type of relation occurs between words that share the same grammatical function, i.e. coordinations and appositions. It is called *junctions* in Tesnière's terminology. The junction between *Alfred* and *Bernard* in (3) is illustrated in fig. 3 (Tesnière, 1959, Chapter 134, § 4).

(2) Mon vieil ami chante cette fort jolie chanson
'My old friend sings this very nice song'

(3) Alfred et Bernard tombent
'Alfred and Bernard fall'

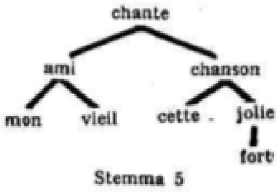


Figure 2: Classical stemma of (2): connections

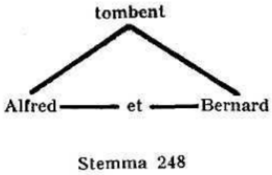


Figure 3: Classical stemma of (3): junction

`Words` and `strokes` are graphically arranged according to rules that distinguish between connection and junction. As I will explain, the most important rules governing the spatialization (their spatial organization) of the classical stemmas can be described by focusing on the behaviour of entities on the vertical axis.

Vertical axis. Each extremity of a single `stroke` is close to a `word`. This corresponds to a syntactic relation between words. The distinction between the two types of relations is expressed by the relative vertical coordinates of the two `words`:

[3] See also (Petitot, 1995).

- Words that are linked together by a connection are reified by words that do not have the same vertical coordinates. The one that is located higher on the plane corresponds to the governor; the lower one to the dependent.
- Words that are bound by a junction are reified by words that have the same vertical coordinate.
- The corollary of these first two rules is that the topmost word represents the root of the syntactic hierarchy.

Such configurations of words and stroke are super-entities that can be interpreted as wholes (*Gestahlt*). Therefore, a connection is a super-entity consisting in two words with different vertical coordinate, connected by a stroke. Similarly a root is the topmost word of the stemma.

Two important additions to these rules are necessary. Firstly, in the case of junctions, the description remains unsatisfactory so far. The junctor *et* is not, from a linguistic perspective, coordinated with either of the conjuncts. The second rule I just stated is not sufficient. Furthermore, there are stemmas in the *Elements* (e.g. see stemma 266) where junction occurs without any junctor (asyndesis). In this case, a single stroke is drawn between the words that reify the conjuncts. To solve this, one must posit that there can be two alternative reifications of junction: (i) the entity reifying the junction is a single stroke and; (ii) the entity reifying the junction is a stroke interrupted by a word.[4] From a cognitive perspective, both entities are junctions, i.e. graphical representations of junctions (fig. 4)[5].

Alfred —— Bernard Alfred —— et —— Bernard

Figure 4: Variety of the junction

Secondly, words that are at the same level of distance of the root have the same vertical coordinate. This convention is implicit in the *Elements*, and probably due to the requirement of printing a visually pleasing book.[6] The cognitive consequence of co-dependents consistently having the same vertical coordinates is that syntactic distance between a word and its descendents is iconically represented, and can be grasped in a single glimpse.

Horizontal axis. The semiotic values of the horizontal coordinates of the words is not similar to the values of the vertical coordinates.

First of all, the horizontal axis plays a major ergonomic role. For the sake of readability, entities may not overlap.[7] In accordance with configurational rules and with this ergonomic constraint, the following are simple corollaries:

- Junction strokes are horizontal.
- Connection strokes are vertical or slanted.

Despite being corollaries, these properties have major cognitive consequences: the slope of a stroke can be used to instantly tell apart the two types of entities in structures that look like trees.[8]

Tesnière insists that the stemma is a representation of the syntactic structure (or *structural order*), rather than the *linear order* of the words (Tesnière, 2015, Chapter 4 sqq.). The horizontal positionning of the words that represent co-dependents of the same verb actually corresponds to a distinction between the kinds of dependencies between them: the subject, the object and the oblique complement are placed, in this order, before the adjuncts, with no respect to the linear order – fig. 5 depicts (4). Such a convention only affects the dependents of the verb.

[4] "If the junction is marked by a junctive, the junction line will be constituted by two parts. The junctive appears between these segments." (Tesnière, 2015, Chapter 136, § 3) See also (Mazziotta, 2014, 145-146).

[5] The first stemma is reconstructed to provide simple comparison material. The second one is extracted from fig. 3

[6] Tesnière hardly ever suggests that a distinction between the vertical coordinates of co-dependent could be meaningful – see stemma 296, Chapter 169, §§ 19-20 for a unique discussion on the subject.

[7] In sciences, diagrams may display some tolerance with respect to this constraint (e.g. scatterplots often have overlapping elements).

[8] Although stemmas are not trees from a mathematical perspective (Kahane and Osborne, 2015; Mazziotta, 2014).

(4) Marie vous rendra sûrement votre livre demain
Mary you.DATIVE will give back certainly your book tomorrow
‘Mary will certainly give you back your book tomorrow’

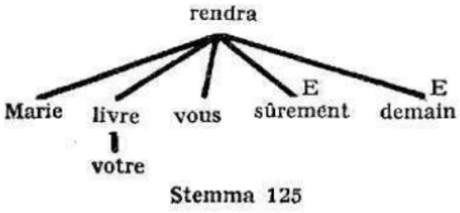


Figure 5: Horizontal order in classical stemmas. See *infra* (2.3) about the letter “E” occurring next to connection `strokes`.

Aside from that, horizontal coordinates have hardly any structural values: the only other constraint is ergonomic (entities cannot overlap). By default, the `words` are horizontally arranged according to the linear order of the sentence, but since this convention is not bound to the theoretical foundations, it remains mostly incidental.[9]

2.3 Complex configurations

The main configurational rules between `words` and `strokes` are somewhat minimalistic, but as Tesnière wants to refine the analysis, he adds complexity to the system in two ways: by using entities that look like `words` as labels and by introducing a special entity to encode a special syntactic operation: *transfer*.

Labels. The fact that relations are reified make it possible to make visual statements about them.[10] Some strokes have `labels`. Letters or numbers that are placed in the direct proximity of the `stroke`. For instance, in fig. 5, the `label` “E” qualifies a `stroke` as a representation of an adjunct relation (Fr. *circonstant* in Tesnière’s terminology). In many cases, `words` and `labels` share the same components (`letters`): only their positions in the stemma can help distinguishing between them.

Transfer. Tesnière also uses another complex graphical entity to represent an operation that he calls *transfer* (Fr. *translation*) (Tesnière, 2015, Part 3). In Tesnière’s model, words have a natural syntactic potential that corresponds to the word classes they belong to; e.g., an adjective naturally depends on a noun and a noun naturally depends on a verb. By the means of grammatical markers such as case endings, prepositions and conjunctions, words can be transferred from one class to another in order to depend on a word belonging to an incompatible class. For instance, a word like *Peter* is a noun, that cannot depend on another noun, unless it is transferred, by the genitive case, to become an adjective. Tesnière gives (5) as an example.

(5) le livre d’ Alfred
the book of Alfred
‘Alfred’s book’

Hence **le livre Alfred vs. le livre d’Alfred.* The transfer relation is reified by a `stylized T`, such as the one in fig. 6.

Tesnière describes two types of transfer (Tesnière, 2015, Part 3). The ones similar to the aforementionned example *d’Alfred*, and the ones that imply the subordination of a clause, i.e. a structure governed by a finite verb. The former are “first-degree transfers” and the latter are “second-degree transfers”.

[9] Scholars of different disciplines have suggested to use the horizontal axis exclusively to encode linear order (Ihm and Lecerf, 1963; Bertin, 2005; Groß, 1992).

[10] See note 2 about systems that do not reify relations.

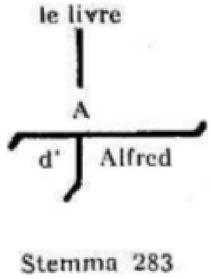


Figure 6: Classical stemma of (5): transfer

The `stylized T` is a complex super-entity consisting of many arranged subentities. From a cognitive perspective, this entity is perceived as a whole (*Gestahlt*) and the description of its parts is only relevant with respect to configurational conventions:

- The `word` representing the transferred word is placed on below the `horizontal bar` of the `stylized T` on the side where the lower part of the `stylized T` is slanted.
- The `word` representing the grammatical means used to transfer the word is placed on the other side.
- A `label`, identifying the resultig word class,[11] is placed on top of the `stylized T`
- The `horizontal bar` is doubled in order to represent second-degree transfer.

The super-entity formed by the `stylized T` and the two aforementionned `words` behaves like a `word` with respect to all configurational rules.

3 Early stemmas

In this section, I study early stemmas in comparison with classical ones, in order to emphasize the contrasts between two different diagrammatic conventions of similar analyses.

The epistemological grounding and the main theoretical choices characterizing Tesnière's early stemmas are similar to the ones of the classical stemmas: relations are reified (Section 2) and connection and junction are distinguished (2.2). The first handcrafted stemma is found in the correspondence between Tesnière and Fernand Mossé in 1932.[12] The first – and to my knowledge only – printed early stemmas appear in "Comment construire une syntaxe" ['How to build a syntax'] (Tesnière, 1934a). Only two stemmas of this kind have been published, and we have yet to find other drafts using the same conventions in Tesnière's archive (BnF NAF 28026).

Tesnière (1934a, 225) draws the stemma of (6) – fig. 7.[13]

(6) De même qu'on voit un grand fleuve qui retient encore, coulant dans la plaine, cette force violente et impétueuse qu'il avait acquise aux montagnes d'où il tire son origine; ainsi cette vertu céleste, qui est contenue dans les écrits de saint Paul, même dans cette simplicité de style, conserve toute la vigueur qu'elle apporte du ciel d'où elle descend. (Bossuet, *Panégyrique de saint Paul*)
'As we see a large river that still retains, running across the plain, this violent and impetuous strength it had gained in the mountains it originates from; similarily, this celestial virtue found in the scriptures of saint Paul, even when the style is simple, keeps all the vigor it brings from the heaven it comes down from.' (my translation)

I will now review the graphical entities of this early stemma (3.1) as well as their configurational rules (3.2). The last part of this section will focus on transfer (3.3).

[11] Tesnière uses the following labels: "I" for "verb", "O" for "noun", "A" for "adjective" and "E" for "adverb" (Tesnière, 2015, Chapter 33).

[12] See (Mazziotta and Kahane, Forthcoming). Tesnière has written a letter on the matter on the 26th of July 1932 and the box at the BnF (NAF 28026) contains an early stemma analyzing the Latin sentence that candidates of the French *baccalauréat* had to translate.

[13] Swiggers (1994, 215) provides a copy of the stemma, but neither this paper nor the original publication are easily accessible. There are several errors in this early stemma (probably made by the publisher). *Un* should connect with *fleuve* (not with *grand*), the direction of the arrow connecting *elle* to *apporte* should be inverted, and *vigueur* should connect with *apporte* (and not *elle*).

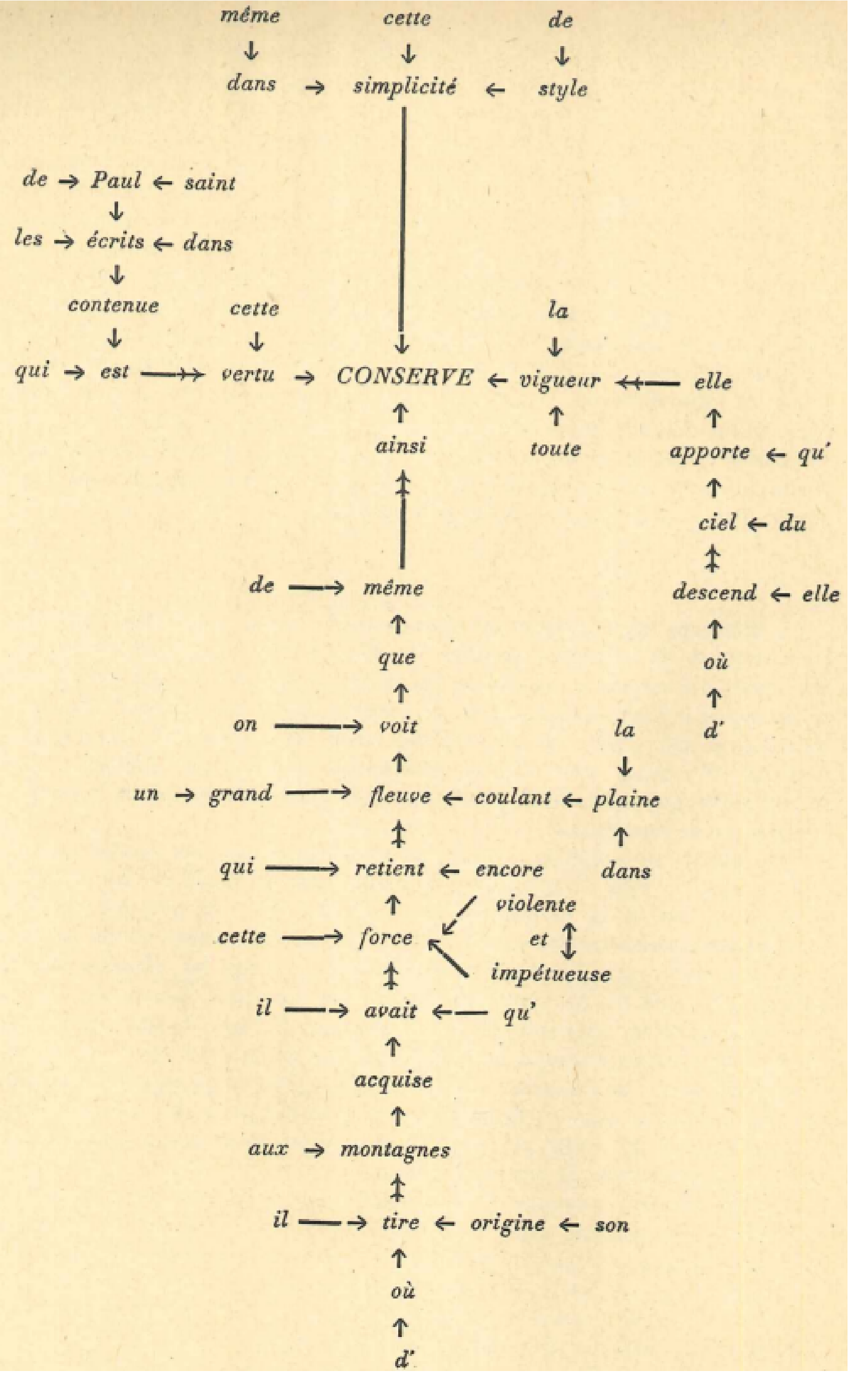


Figure 7: Early stemma (see note 13 for corrections)

3.1 Graphical entities

The graphical entities at use in the early stemmas correspond to the words and the relations represented in the classical stemmas:

- `Words` reify words of the sentence in a similar manner as they do in classical stemmas (2.1). The `word` that corresponds to the root is capitalized.
- Three kinds of `arrows` reify syntactic relations. The internal structure of `arrows` is worth considering. The discontinuities in the `stroke` composing some `arrows` are incidental, but `arrow heads` are similar to labels that identify different types of `arrows`, and, consequently, of relations:
  - `simple arrows` "→" correspond to connections;

- `double arrows` "↔" correspond to junctions;[14]
- `two-headed arrows` "↠" correspond to a special type of connection (see 3.3).[15]

It is already clear that early stemmas use symbolic conventions to distinguish between different types of relations. By contrast, classical stemmas use simple `strokes` that need to be spatialized in order to be identified as the reification of specific relations. Configurational rules are set accordingly.

3.2 Configurational rules

The configurational conventions between `words` and `arrows` of all kinds are:

- Similarily to `strokes` in classical stemmas, `arrows` connect `words` that appear at both extremities.
- `Single arrows` and `two-headed arrows` express a hierarchy. The `word` representing the governor is placed near the `arrow head` and the one representing the dependent at the other extremity.

As in classical stemmas, coordination also deserves a closer look. There is no hierarchy between the words and the first of the two aforementionned rules is sufficient to describe the behavior of the entities reifying conjuncts. The specific convention is that the `word` that reifies the coordinator is used similarily as `labels` in classical stemmas: it is placed beside the `dual arrow` (*et* in fig. 8).

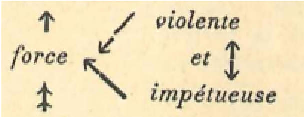


Figure 8: Junction in early stemmas

The recursive application of the configurational rules in early stemmas leads to a *gravitational* repesentation of the sentence – Tesnière uses the French term *gravitation* to describe the relationships in his system (Tesnière, 1934a, 224). Words that depend on a governor are represented by `words` surrounding the `word` corresponding to the governor. Both axes of the plane are used simultaneously: it is not possible to identify a specific function for one, that the other would not share. Furthermore, the two corollaries identified in classical stemmas (namely that junctions are represented by `strokes` that look horizontal and connections by `strokes` that look vertical or slanted) cannot be used to easily locate different kinds of relations. No configurational contrast can encode this difference, which is expressed by symbolic means: the kind of `arrow` at use.

Another major consequence of this behavior is the way syntactic distance is made visible (2.2). As a first approximation, it might seem that graphical distance iconically represents this syntactic distance. However, the graphical distance between words is never sufficient to express the syntactic hiearchy: it must be supplemented by `arrows`, the length of which is not relevant. For instance, in fig. 7, the arrow between *simplicité* and *CONSERVE* is longer than both arrows linking *même* to *voit* through *que*. Graphical distance may incidentally correspond to syntactic distance, but only the count of `arrows` implied and their directions are relevant.

Additionally, the slopes of the `arrows` do not correspond to anything in the syntactic analysis. The `arrow` between *vigueur* and *CONSERVE* is orthogonal to the one between the latter *simplicité*. This contrast has no value in the diagrammatic system, as stated by Tesnière:

> Ses subordonnés directs sont placés devant, derrière, au dessus ou au dessous, peu importe. ['Its direct subordinates are placed in front, behind, below or on top [of the governor]; it does not matter.'] (*Letter to Mossé*, 26 Jul. 1932; BnF NAF 28026)

The direct consequence of this free placement is that the `word` corresponding to the root element of the sentence cannot be identified by its positionning in the diagram alone: the central position of *CONSERVE*

[14] Note that in the first draft of an early stemma, Tesnière uses a simple `stroke` instead (BnF NAF 28026, B42, 148B).

[15] In his drafts, Tesnière uses a `crossed-out arrow` rather than a `two-headed arrow`: "Nous l'indiquons par une flèche barrée [. . .]" [' We note this by the means of a crossed-out arrow'] (Tesnière, 1934a, 228, note 1).

is also incidental. Therefore, the root is identified by the means of a symbolic convention (the use of capital letters). Configurational rules would suffice to identify it, not in a cognitive-efficient way, but rather by evaluating the direction of each `arrow` in the diagram.

3.3 Representation of transfer

Although Tesnière had already elaborated the concept of transfer by the time he published his first stemmas (Tesnière, 1934a, 227-228), the entities and the configurational rules do not encode transfers in a straightforward way.

> [I]l faut, à côté des régissants et des subordonnés de toute sorte, prévoir une place pour les subordonnants, c'est-à-dire pour les éléments qui, n'étant eux-mêmes ni régissants ni subordonnés, ont pour mission de marquer la subordination des autres éléments. Cette réserve faite, toute phrase peut être représentée par un stemma qui indique la hiérarchie de ses connexions. ['Aside from governors and subordinates of any kind, there must be a place subordinators, i.e. elements that are neither governors nor dependents, but that make subordination possible for other elements. Apart from that, any sentence can be depicted as a stemma that represents the hierarchy of its connections.'] (Tesnière, 1934a, 225, my translation)

Fig. 9 is the fragment of the stemma that depicts the analysis of "cette vertu qui est contenue dans les écrits de saint Paul" ['this virtue found in the scriptures of saint Paul']. From a linguistic point of view, the construction *de saint Paul* is similar to *d'Alfred*: both are PPs, and both are analyzed as transfers by Tesnière.

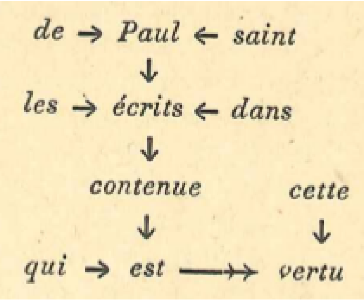


Figure 9: Dependents and translatives in early stemmas

However, there is no difference between a dependent and a translative in stemmas such as fig. 9: *de* (translative) and *saint* (adjective) are both connected to *Paul* by the means of "→".

Only second-degree transfers can be identified, since a special type of `arrow` "↠"[16] connects a subordinate finite verbs to their governors.[17] However no convention can tell apart the translative from any other dependent: in fig. 9, the translative *qui* 'who' has the same satus as the dependent *contenue* 'contained, found'.

4 Conclusion

Classical and early stemmas use two different sets of diagrammatical entities and rules of spatialization. Tab. 1 summarizes the comparison. Both systems have in common that words are reified by `words` and that relations are reified by specific line-like entities drawn from one `word` to another. Classical stemmas confer a greater importance to *configurational* means of representing syntax, whereas earlier ones favor *symbolic* means. This has cognitive consequences. The conventions of the classical stemmas allow for a straightforward identification of key elements of the analysis by the means of geometric properties:

- the root of the stemma is reified by the topmost `word`;

[16] See footnote 15 about this entity.

[17] See footnote 13 about the errors in the stemma.

	Classical stemmas	Early stemmas
word	`word` (2.1)	`word` (3.1)
relation	`stroke` (2.1)	`arrow` (3.1)
relation type	relative position and slope (2.2)	`arrow head` (3.1 and 3.2)
syntactic distance	vertical distance (2.2)	arrow count (3.2)
transfer	`stylized T` (2.3)	(hardly present, 3.3)

Table 1: Classical *vs.* early stemmas

- dependents of the same level are reified by `words` on the same horizontal line;
- the distinction between connection and junction corresponds to the slope of the `strokes`.

These straightforward arrangements are not available in early stemmas, which use symbolic conventions, such as the use of capitals for the root, the use of different arrow heads, etc.

Furthermore, surveying the evolution of diagramming systems actually helps understanding issues that are still relevant today. Tesnière's `stylized T` and flexibility in the choice of diagrams are two illustrations of the link between diagram use, theoretical debate, and efficiency of expression. As exposed in this paper, Tesnière tried to elaborate minimalistic conventions, but he somewhat failed in the case of transfer. The `stylized T` remains the most idiosyncratic entity he uses, and I have not heard of any colleague using it to make diagrams. Nevertheless, it is striking that the invention of such bizarre entities is actually possible without breaking the rest of the system. The `stylized T` behaves like a `word`, and it solves the distinction problem between dependent and translatives (3.3). By positionning translatives in a specific way on the stemma, Tesnière brings his own answer to the problem of function words in dependency syntax[18] – a problem that still cannot be solved in a consensual way (Kahane and Mazziotta, 2015a; Osborne and Gerdes, 2019).

The choice between configurational and symbolic means to express components of syntactic analysis is still a constant issue in modern dependency linguistics. Careful linguists try to select the diagrammatic conventions that suits their demonstration. For instance, Mel'čuk usually uses a configurational system (fig. 11a),[19] but he favors symbolic conventions to represent dependencies when he wants to evaluate projectivity (fig. 10). By doing so, he makes it possible to visualize crossing `arrows` (Mel'čuk, 1988, 37).[20]

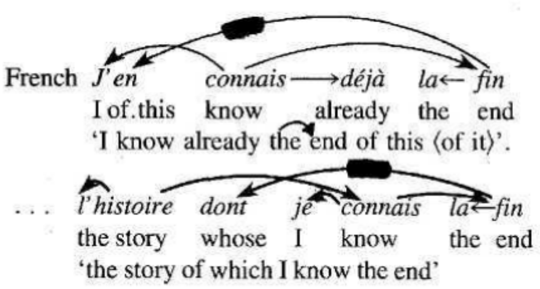

French J'en connais déjà la fin
I of.this know already the end
'I know already the end of this (of it)'.
... l'histoire dont je connais la fin
the story whose I know the end
'the story of which I know the end'

Figure 10: Symbolic conventions express projectivity violations

Symbolic devices are convenient, since not relying on spatialization gives more freedom for geometric arrangement. For instance, as illustrated by fig. 11 (Mel'čuk and Iordanskaja, 2015, 26 and 34), the authors' usual way to draw a dependency tree combines the main configurational principles of classical

[18] Although such an answer may actually be seen as a constituency-based solution (Osborne, 2013). For a presentation of the possible structural interpretations of transfer, see (Kahane and Osborne, 2015, l-lx).

[19] The use of `arrows` (the convention is the opposite of Tesnière's) is redundant, because the vertical arrangement already expresses the direction of dependencies.

[20] Of course, this is not the only way to visualize projectivity violation; e.g. the early diagrams by Ihm and Lecerf (1963, 10) duplicate the `words` and align them on a projection `stroke`, which is still a common practice.

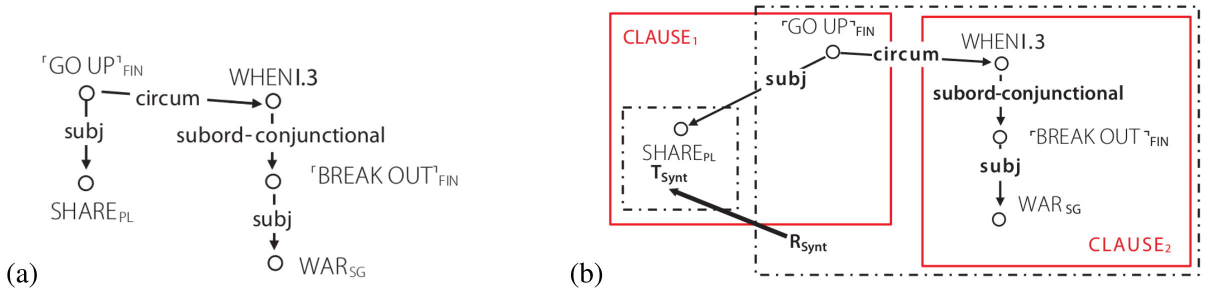


Figure 11: Expanding a diagrammatic system

stemmas with arrows and labels interrupting them. When adding boxes and arrows reifying the communicative structure to a dependency tree, they actually move the first dependent of *go up* further to the left, with absolutely no consequence on the meaning of the diagram, thus leaving space for adding a new box.

In his stemmas, Tesnière had to cope with some theoretical issues that still find concurrent solutions in modern linguistics. Sometimes these various solutions are incompatible, because they acknowledge different views of syntax. Sometimes diagrammatic flexibility is a way to achieve a better visualization of the reasoning. Differences may be incidental from a formal point of view, but they are of utmost importance from a cognitive perspective.

Acknowledgements

I would like to thank Sylvain Kahane, Jean-Christophe Vanhalle and anonymous reviewers of the Depling comitee for their suggestions. I would also like to thank Jacques François and Lene Schøsler, who discussed preliminary versions of this paper.

References

Jacques Bertin. 2005. *Sémiologie graphique: Les diagrammes – Les réseaux – Les cartes*. EHESS, Paris.

R. C. Brittain. 1973. *A Critical History of Systems of Sentence Diagramming in English*. University of Texas [Unpublished Ph. D. dissertation].

Groupe μ. 1992. *Traité du signe visuel*. Seuil, Paris.

Thomas Groß. 1992. Konstruktive Stemmatologie. *Papiere zur Linguistik*, 42(2):115–139.

P. Ihm and Y. Lecerf. 1963. *Éléments pour une grammaire générale des langues projectives*. European Atomic Energy Community, Joint Nuclear Research Center, Ispra Establishment (Italy), Scientific Data Processing Center.

Sylvain Kahane and Nicolas Mazziotta. 2015a. Dependency-based analyses for function words – introducing the polygraphic approach. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 181–190, Uppsala, Sweden, August. Uppsala University, Uppsala, Sweden.

Sylvain Kahane and Nicolas Mazziotta. 2015b. Syntactic polygraphs. A formalism extending both constituency and dependency. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 152–164. Association for Computational Linguistics.

Sylvain Kahane and Timothy Osborne. 2015. Translators' introduction. In *Elements of structural syntax* (Tesnière, 2015), pages xxix–lxxiv. Translation of (Tesnière, 1959).

Françoise Madray-Lesigne and Jeannine Richard-Zappella, editors. 1995. *Lucien Tesnière aujourd'hui. Actes du colloque international du CNRS URA 1164 – Université de Rouen 16 - 17 - 18 Novembre 1992*. Peeters, Louvain and Paris.

Nicolas Mazziotta and Sylvain Kahane. Forthcoming. L'émergence de la syntaxe structurale de Lucien Tesnière. In Cécile Mathieu and Valentina Bisconti, editors, *Entre vie et théorie: La biographie des linguistes dans l'histoire des sciences du langage*. Lambert Lucas, Limoges.

Nicolas Mazziotta. 2014. Nature et structure des relations syntaxiques dans le modèle de Lucien Tesnière. *Modèles linguistiques*, 69:123–152.

Nicolas Mazziotta. 2016a. Drawing syntax before syntactic trees. Stephen Watkins Clark's sentence diagrams (1847). *Historiographia Linguistica*, 43(3):301–342.

Nicolas Mazziotta. 2016b. *Représenter la connaissance en linguistique. Observations sur l'édition de matériaux et sur l'analyse syntaxique. Mémoire de synthèse en vue de l'obtention de l'Habilitation à diriger des recherches*. Université Paris Ouest, Nanterre – La Défense. `http://hdl.handle.net/2268/204408`.

Igor Mel'čuk and Lidija Iordanskaja. 2015. Ordering of simple clauses in an English complex sentence. *Rhema*, 4:17–59.

Igor Mel'čuk. 1988. *Dependency syntax: theory and practice*. State University of New York, Albany.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa*, 4(1)(17). `http://doi.org/10.5334/gjgl.537`.

Timothy Osborne. 2013. A look at Tesnière's *Éléments* through the lens of modern syntactic theory. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 262–271. Charles University in Prague, Matfyzpress, Prague, Czech Republic, Prague, Czech Republic, August.

Jean Petitot. 1995. Approche morphodynamique de l'iconicité des stemmas. des connexions tesnièriennes aux images-schèmes des grammaires cognitives. In Madray-Lesigne and Richard-Zappella (Madray-Lesigne and Richard-Zappella, 1995), pages 105–112.

Didier Samain. 1995. Le graphe et l'icône. Remarques sur la logique du schématisme chez lucien tesnière. In Madray-Lesigne and Richard-Zappella (Madray-Lesigne and Richard-Zappella, 1995), pages 129–135.

Frederik Stjernfelt. 2007. *Diagrammatology: An investigation on the borderlines of phenomenology, ontology, and semiotics*. Springer, Dordrecht.

Pierre Swiggers. 1994. Aux débuts de la syntaxe structurale: Tesnière et la construction d'une syntaxe. In *Actes du colloque international: Lucien Tesnière. Linguiste européen et slovène (1893-1993). Ljubljana, 18-20 novembre 1993*, pages 209–219. Faculté des Lettres et Philosohie de l'Université de Ljubljana, Ljubljana.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Lucien Tesnière. 1934a. Comment construire une syntaxe. *Bulletin de la Faculté des Lettres de Strasbourg*, 7:219–229.

Lucien Tesnière. 1934b. *Petite grammaire russe*. Didier, Paris.

Lucien Tesnière. 1953. *Esquisse d'une syntaxe structurale*. Klincksieck, Paris.

Lucien Tesnière. 2015. *Elements of structural syntax,* translated by Timothy Osborne and Sylvain Kahane. Benjamins, Amsterdam/Philadelphia. Translation of (Tesnière, 1959).

Toward a cognitive dependency grammar of Hungarian

András Imrényi
Department of Hungarian Linguistics
ELTE, Budapest
imrenyi.andras@gmail.com

Abstract

The paper presents the key tenets of a novel approach to the structure of Hungarian clauses that combines aspects of cognitive linguistics and dependency grammar. Clauses are given a multigraph description (as in XDG), with separate semantic graphs dedicated to frame semantic relations (S1), speech function (S2) and contextualization (S3). These stand in symbolic association with formal dimensions pertaining to morphology (F1), word order (F2) and prosody (F3). It is shown that the finite verb, or a catena of elements including the verb, is central for both S1 and S2, functioning as a 'clause within the clause'. Further, the clause is shown to manifest modularity, whereby a single node in one dimension may correspond to a catena of interconnected elements in another.

1 Introduction

The goal of this paper is to present the broad outlines of a novel approach to the structure of Hungarian clauses. This approach is inspired by two major sources, cognitive linguistics (CL) and dependency grammar (DG). In line with CL, syntax is regarded as the study of learned pairings of meaning and form in the sentence, with a lexicon-syntax continuum hosting signs (constructions) of varying degrees of complexity and schematicity (cf. Goldberg, 1995; Langacker, 1987, 2005, 2008; Diessel, 2015). In line with DG, the internal structure of multiword constructions is described by reference to syntagmatic rather than part/whole relations.

This combination of ideas entails a focus on form-meaning correspondences that hold between semantic relations and associated formal devices. In fact, under a Langackerian interpretation of cognitive linguistics, familiar concepts of DG such as subject and object can no longer be regarded as theoretical primitives. Rather, they must be reduced to symbolic associations between phonologically relevant properties (e.g. word order or case morphology) and aspects of clausal meaning.[1] Crucially, meaning from a CL perspective is taken to comprise all facets of conceptualization (mental processing), including such factors of construal as perspective, focusing, and specificity (cf. Langacker, 2008).

Work along these lines has produced a comprehensive description of Hungarian clause structure (Imrényi, 2017a) and the rudiments of a theory, or research program, that may be called cognitive dependency grammar (CDG).[2] In this paper, I focus on the following aspects of the novel account: i. the nature of the three semantic dimensions (S1, S2, S3), and the rationale for positing them (Section 2), ii. the dual role of the clausal core (a catena that serves as 'clause within the clause') in S1 and S2 (3.1), iii. contextualizing relations (S3) and iv. cross-dimensional mappings (3.2).

The proposal stems from an intention to describe Hungarian in its own right, and focuses on basic conceptual matters. Thus, (in-depth) cross-linguistic application and practical implementation are beyond the scope of the paper. In terms of the theoretical landscape, my aim is to show that insights from cognitive and functional approaches (including Halliday's Systemic-Functional Grammar) can be fruitfully integrated into DG-oriented theorizing. In a Hungarian context, I seek to develop an alternative to mainstream generative accounts (e.g. É. Kiss, 2002), offering a new set of conceptually defined categories as well as a new way of raising the basic questions about sentence structure.[3] However, the proposal is at an early stage of its development, to be seen as a new beginning rather than a full-fledged framework.

[1] As Langacker (2008: 6) puts it, "all valid grammatical constructs are symbolic, hence reducible to form-meaning pairings".

[2] CDG is different from Hudson's Word Grammar (WG) (Hudson, 1990, 2007, 2010), which is also CL-oriented and dependency-based, by more closely following a Langackerian conception of cognitive linguistics (Langacker, 1987, 2008).

[3] For example, instead of the notion of "focus", I work with the concept of "overriding" (see Section 3.1), and "topic" is treated as a subtype of "contextualizers" (3.2). Instead of asking questions such as what position a constituent "occupies" or "moves into", both the function and the form (e.g. linearization) of elements are defined in relational terms, with respect to other elements in the syntagmatic chain.

2 Beyond Tesnière's drama metaphor: three dimensions of clausal meaning

As Tesnière famously states, "the verbal node […] is a theatrical performance. Like a drama, it obligatorily involves a process and most often actors and circumstances" (Tesnière, 2015 [1966]: 97). Under these assumptions, the sentence *Alfred gives the book to Charles* can be semantically analysed by saying that the verb designates the process of giving, with the three dependents expressing the actors (or participants) associated with it. Although many practitioners of DG make a strict separation between syntax and semantics, it is hard to escape the view that the drama metaphor underlies all DG analyses in which a finite lexical verb serves as the root node of the sentence. Concomitantly, at least in prototypical cases, an analysis in terms of subject, object, etc. closely corresponds to a semantic account that treats the referents as participants of the designated process.

Like all metaphors, however, the drama metaphor also has limitations; it does not capture all aspects of clausal meaning. One issue can be illustrated by a sentence such as *What does Alfred give to Charles?*, in which *what* has a dual role to play: on the one hand, it pertains to a participant (the thing which is given), and on the other, it endows the sentence with the function of a wh-question. Therefore, a DG representation that only treats it as direct object misses something important about its function and behaviour. Additionally, adverbs such as *unfortunately* or *probably* do not designate any participant or circumstance of the onstage process (the theatrical performance); rather, they indicate the speaker's evaluation/assessment of the foregrounded information.

In my research on Hungarian, I have found it useful to distinguish between three semantic dimensions, in a way that is consonant with Halliday's (1994: 35) notions of 'clause as representation', 'clause as exchange' and 'clause as message' (see also Langacker, 2001, 2015). The dimensions, represented as semantic graphs, address the following questions.

S1: What grounded process is designated by the clause?[4] What are its participants and circumstances?

S2: What is the speech function of the clause? Is the speaker stating the existence (occurrence) of the grounded process, or does the clause serve a different purpose?

S3: How does the speaker contextualize the message to facilitate its efficient processing and intended interpretation?

Under these assumptions, the dual role of *what* can be captured by analysing it in both S1 and S2, whereas elements like *unfortunately* and *probably* are assigned exclusively to S3.

In contrast with Halliday, I use dependency structures (semantic graphs) to represent these complementary aspects of meaning, and link them to dimensions of linguistic form (grounded in phonological space) including morphology/segmental content (F1), word order (F2) and prosody (F3). Thus, the proposal is very close to the formalist framework of Extensible Dependency Grammar (XDG), whose key tenet is the following:

> An XDG grammar allows the characterisation of linguistic structure along several *dimensions* of description. Each dimension contains a separate graph, but all these graphs share the same set of nodes. Lexicon entries synchronise dimensions by specifying the properties of a node on all dimensions at once. (Debusmann et al., 2004: 2)

However, my approach departs from XDG by having a CL-orientation and also by lifting the constraint that all graphs must share the same set of nodes. In the next section, I give a brief illustrative discussion of the proposal.

3 The CDG approach to Hungarian

3.1 The clausal core and its dual function. The analysis of S1 and S2

A finite lexical verb has two basic roles in clausal semantics. Firstly, it designates a grounded instance of a process type (e.g. an instance of buying), which is at the centre of the theatrical performance.[5] By virtue of this, it can and often must be accompanied by dependents designating participants and circumstances. Secondly, the finite verb, whether a lexical or auxiliary verb, also marks the illocutionary force and polarity (in short, the speech function) of the sentence, at least by default.[6] Here are two formulations of this insight.

[4] The notion of grounded process is taken over from Langacker (2008). 'Process' is understood highly generally to encompass actions, states, etc., the key criterion being that they unfold in time and are processed by sequential scanning (Langacker, 2008: 111). 'Grounding' is interpreted as the operation whereby an instance of a type (here, a process type) is situated with respect to the ground, defined by Langacker as involving "the speech event, its participants (speaker and hearer), their interaction, and the immediate circumstances (notably, the time and place of speaking)" (Langacker, 2008: 259).

[5] Note that additional elements may also crucially contribute to the function of designating the grounded process (e.g. in the case of idioms and light verb constructions), and elements other than the verb may also serve in this capacity in languages like Hungarian.

[6] For proposals treating illocutionary force and polarity together, as aspects of an integrated system, see Croft (1994: 466) and Langacker (2009: 232).

> I said that the main use of the Verb is to convey affirmation, because we will see below that it is also used to convey other movements of our soul; such as *to desire, to pray, to command,* etc. (Arnault and Nicole, 1662: 90; quoted by Kahane, to appear).

> [The meaning of finite verbs] includes what is called an illocutionary force which guides the listener; so if I say to you *Bill has died*, you know that this is a new property of Bill that I am inviting you to add to your memory. [...] Similarly, the finite verbs in *Has Bill died?* and *Remember me!* each carry the illocutionary force for a question and a command. (Hudson, 2010: 264)

Illocutionary force is generally assumed to characterize entire clauses (or indeed utterances) rather than the finite verb itself. However, its linking to the finite verb is highly motivated in Hungarian, which even allows this unit to serve as a full-fledged positive declarative sentence under appropriate circumstances (with sufficient contextual support).

(1) a. A Disney meg-vette a 21st Century Foxot.
the Disney.NOM PREV-bought.3SG.DEF the 21st Century Fox.ACC
'Disney bought 21st Century Fox.'
b. A: A Disney megvette a 21st Century Foxot? *'Did Disney buy 21st Century Fox?'*
B: Igen, megvette. *'Yes, they [*lit. *he/she/it] bought it.'*

In (1a), the finite verb *megvette* 'PREV.bought.3SG.DEF' evokes the frame of a commercial transaction (cf. Fillmore 1982), necessarily involving a Buyer and some Purchased Goods.[7] Owing to this frame-evoking role, the finite verb is central for the organization of S1. However, its function is more complex than this: it also expresses a statement that an instance of buying took place, which makes it central for S2 as well. Since Hungarian finite verbs have a way of marking not only the person and number of the subject but also the definiteness (contextual recoverability) of the object referent (marked by DEF in the glosses), speaker B employs *megvette* 'he/she bought it' without any dependents in her reduced answer in (1b). In this context, the verb is sufficient to convey the same message as the elaborate clause in (1a).

I use the term **clausal core** to refer to a minimal unit in the clause which expresses the same grounded process (e.g. a grounded instance of buying) as the clause as a whole.[8] In Hungarian, the clausal core need not be coextensive with the finite verb; rather, it may also be a multiword catena of elements.[9] This is the case with routinized, more or less idiomatic expressions such as *feleségül vesz* 'marry [a woman]' (lit. 'take as wife') and *figyelembe vesz* 'take into consideration', where the frame is evoked by a multiword unit. Additionally, there are constructions (e.g. those involving auxiliaries) where the evoking of a process type and the grounding of an instance of that type are effected by separate words.

The clausal core of a neutral positive declarative clause such as (1a) is characterized more specifically as a **proto-statement**. A proto-statement has the dual function of 1) designating the grounded process that is also expressed by the clause as a whole and 2) expressing a statement to the effect that this process exists (existed, will exist) at some time and in some mental space. It is thus a schematic clause, and (1a) includes the proto-statement *megvette* as a 'clause within the clause'. The notion of mental space (cf. Fauconnier, 1985) is necessary because of auxiliaries such as *kell* 'must', *lehet* 'may' and *akar* 'want', which may also appear inside a clausal core along with the infinitival form of a main verb (e.g. *meg akarja venni* 'he/she wants to buy it'). Auxiliaries allow speakers to talk about the existence of a process in the world of necessary or possible actions, somebody's intentions, etc. rather than the Reality Space (the world of actual occurrences). I assume that the proto-statement function is linked to the clausal core as an unmarked default value.

Having discussed the three semantic dimensions and the dual role of the clausal core, let us turn to the analysis of example (2).

(2) Ki vette meg a 21st Century Foxot?
who.NOM bought.3SG.DEF PREV the 21st Century Fox.ACC
'*Who bought 21st Century Fox?'*

[7] Preverb+verb combinations are lexicalized units (much like English phrasal verbs) that often have a partially or fully idiomatic meaning. A preverb (glossed as PREV) such as *meg* immediately precedes the verb stem by default; however, we will see later that in certain constructions, this default order gets reversed.

[8] Cf. the notion of existential core in Langacker (2012).

[9] For the notion of catenae, see Osborne and Gross (2012).

In (2), the clausal core is *vette meg*, which evokes the frame of buying, and designates the same grounded process as the clause as a whole. S1 is a graph consisting of the frame semantic (thematic) relations Agent and Patient (more specifically, Buyer and Purchased Goods). S1 is symbolically associated with an F1 dimension which highlights relevant morphological properties. In particular, the nominative case of *ki* 'who.NOM' and the accusative case *-(V)t* of *21st Century Foxot* '21st Century Fox.ACC' make it clear which company acted as the Buyer and which one assumed the role of Purchased Goods.[10] In the diagram, the semantic and formal representations are separated by a horizontal line. Dotted lines are used to mark correspondences between elements of the two.

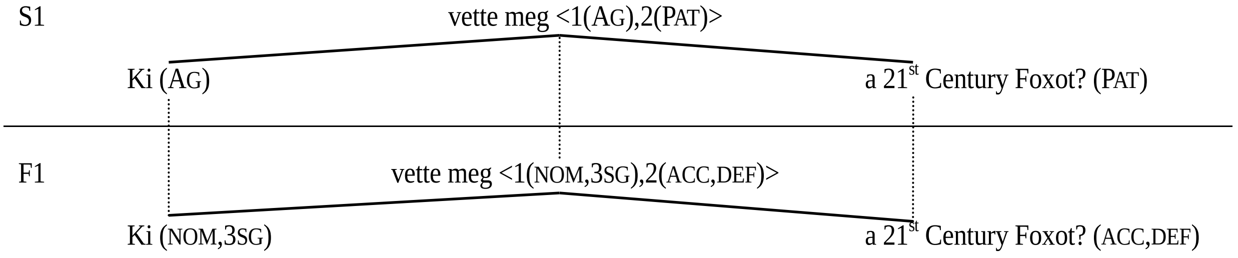

S1
vette meg <1(AG),2(PAT)>
Ki (AG)
a 21st Century Foxot? (PAT)
F1
vette meg <1(NOM,3SG),2(ACC,DEF)>
Ki (NOM,3SG)
a 21st Century Foxot? (ACC,DEF)

Figure 1. An illustration of S1 and F1.

In the description of Hungarian, I see no pressing need for making reference to grammatical functions (e.g. subject, object). For example, subjecthood can be reduced to a set of **construction-specific mappings** between thematic roles (e.g. Agent, Patient, depending on the construction) and morphological properties (nominative case, person-and-number agreement with the verb). This approach draws on Brassai's following insight:

> the thing denoted by the nominative is the actor in the plot of active verbs, the sufferer in that of passive verbs, and it is in a particular state in the plot of middle verbs. The generalization cannot be taken any further, hence the true [semantic] interpretation cannot be considered completely successful. (Brassai, 2011 [1864]: 199, my translation)

From the perspective of construction grammar, Brassai had no reason to be dissatisfied. Taking the position that constructions are basic and the categories and relations in them derivative (cf. Croft, 2001: 46), we can say that it is up to particular constructions to determine what role is associated with the nominative dependent. In active construals of a transitive event, this role will be the Agent/Experiencer, in passive constructions, the Patient/Theme, etc. More specifically, knowing the verb *megvesz* 'buy' involves knowing the fact that its nominative dependent is associated with the Agent thematic role. Needless to say, the notion of subject is harder to eliminate in an account of English, where e.g. weather verbs pose further challenges. However, no such issues arise in Hungarian, hence I follow Brassai's general approach.

The careful reader must have noted that while (1a) includes the verb *megvesz* 'buy' in its default preverb+verb order (*megvette*), the opposite linearization is found in (2) (*vette meg*). To account for this, we need to take a look at S2 and the formal dimensions with which it is symbolically associated, namely F2 (for word order) and F3 (for prosody).

Recall that the clausal core is also central for S2 owing to its ability to determine, at least by default, the speech function of the clause as a whole. For example, the declarative illocutionary force and positive polarity of (1a) are "projected" or "inherited" from the proto-statement *megvette,* which also serves to state the occurrence of an instance of buying.

The key step toward an S2 analysis of (2) is that a proto-statement (a positive declarative clausal core) is also materially present in it. In particular, it includes all of the segmental content (*meg* and *vette*) which is used by default, in preverb+verb order, to state that an instance of buying took place. What is special about (2) is that the default positive declarative function of the clausal core is not "projected to" or "inherited by" the clause as a whole, whose speech function is to inquire about the identity of a participant. This can be captured in S2 by saying that *ki* 'who.NOM' acts as an **overrider** (OVR) of the core's default speech function; it stands in a relation of overriding with the core. On the

[10] More generally, the form of a word (in contrast with other forms in the same paradigm) has a cuing role for the recognition of its referent's function. In some paradigms such as that of nouns with a first or second person possessive marker, the usual *-(V)t* ending of accusatives can go missing (e.g. *A fiam szereti a lányod[at]* 'the son.my loves the daughter.your', 'My son loves your daughter'). In examples like this, we can say that the base form (e.g. *lányod*) is linked to both the nominative and the accusative slots of the case paradigm, and it is up to word order (preverbal dependents are more likely to denote agents/experiencers), world knowledge, etc. to provide disambiguation. In other words, case is more properly interpreted as a value within a paradigm rather than as something which is necessarily manifested in specific segmental content.

formal side, this is coded by the overriding of default linearization, i.e. inversion (F2) and the destressing of the clasual core (F3).

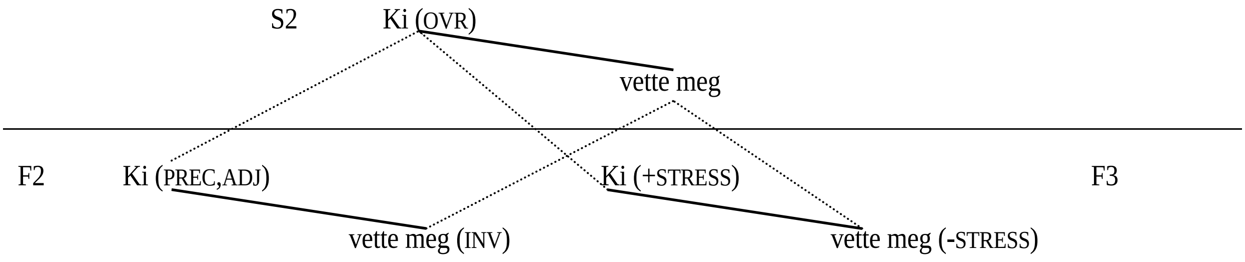


Figure 2. An illustration of S2, F2 and F3.

In S2, the dominant element is *ki*. As an overrider, it imposes its speech function on the structure as a whole. Since *a 21st Century Foxot* does not modify the type of speech function which is associated with the clause, it is absent from S2. In F2, it is signalled that the interrogative pronoun precedes (PREC) and is adjacent (ADJ) to *vette meg*, with the latter also displaying preverb-verb inversion (INV). Finally, F3 marks relationships of relative prosodic prominence. The label +STRESS and -STRESS are to be interpreted as ‘more stressed’ and ‘less stressed’, respectively.

The concept of overriding is useful for the description of immediately preverbal, inversion-triggering elements in Hungarian because it is suitable for capturing a seemingly highly heterogeneous set of elements displaying the formal properties just mentioned. In particular, preverb-verb inversion occurs not only after interrogative pronouns but also after the negative particle *nem* ‘not’ (e.g. *A Disney nem vette meg a 21st Century Foxot* ‘Disney did not buy 21st Century Fox’) and after so-called identificational foci (e.g. *A DISNEY vette meg a 21st Century Foxot* ‘It was Disney who bought 21st Century Fox).[11] It holds for all of these constructions that the default speech function of the clausal core (that of stating the existence of a grounded process) cannot prevail, as the clause as a whole serves a fundamentally different function.

3.2 Contextualizing relations (S3). Cross-dimensional mappings

Finally, let us see an illustration of the third semantic dimension, which represents contextualizing relations. Someone supporting the independence of 21st Century Fox might opt for the following construal of the transaction.

(3)	A Disney	sajnos	állítólag	meg-vette	a 21st Century Foxot.
	the Disney.NOM	unfortunately	allegedly	PREV-bought.3SG.DEF	the 21st C. Fox.ACC

‘Disney unfortunately allegedly bought 21st Century Fox.’

Before we turn to the details of the analysis, it is worth discussing how it relates to previous proposals in the literature.

As a first approximation, S3 is meant to describe what other theories call topic/comment or theme/rheme articulation. In Halliday’s approach, which goes back to the Prague School (Garvin, 1964; Firbas, 1992),

> The Theme is the element that serves as the point of departure of the message; it is that which locates and orients the clause within its context. The speaker chooses the Theme as his or her point of departure to guide the addressee in developing an interpretation of the message […]. The remainder of the message, the part in which the Theme is developed, is called […] the Rheme (Halliday, 2014: 89).

In the history of Hungarian linguistics, Sámuel Brassai had offered a similar account, and may be credited with the discovery of information structure, preceding Gabelentz (1868) by several years. Brassai used the term inchoative, derived from the Latin verb *inchoare* ‘begin’, to name sentence-initial elements preceding the main part of the sentence (conveying new information) for which the speaker wants to prepare the listener. He defined the function of inchoatives as follows: they “prepare the ground in the listener’s mind for comprehending the meaning of the sentence, in other words they have an attention-directing, preparatory role, linking up the mental operations of the listener with those of the speaker” (Brassai, 2011 [1860]: 54, my translation).

[11] For the notion of identificational focus, see É. Kiss (1998). The usual strategy of É. Kiss is to define the function associated with the so-called focus position (Spec-FP) based on identificational focus only (cf. É. Kiss, 2002: 78), even though other elements occupying the same position are not equally compatible with this definition. Overriding is proposed here as a more schematic notion that applies in the same way to varied types of preverbal, inversion-triggering elements. Additionally, it may also inform the description of English (cf. Imrényi, 2017b: 309).

Halliday notes that what other linguists call topic represents only one subtype of theme, the "topical theme" (Halliday, 2014: 89). (And in the same way, Brassai's category of inchoatives is much broader than that of topics.) However, given that *topic* and *theme* are synonyms in present-day English, this terminology seems rather confusing. Therefore, in line with Halliday's formulation that the theme "locates and orients the clause within its context", I work with the notion of **contextualization** instead. More precisely, the phenomenon is captured in CDG by the notion of contextualizing relations, constituting the third semantic dimension (S3).

In example (3), the "comment" or "rheme" is expressed by *megvette a 21st Century Foxot* 'bought 21st Century Fox'. This is the contextualized part of the clausal network, with which three elements (contextualizers) stand in a contextualizing relationship, namely *a Disney* 'Disney.NOM', *sajnos* 'unfortunately' and *állítólag* 'allegedly'.

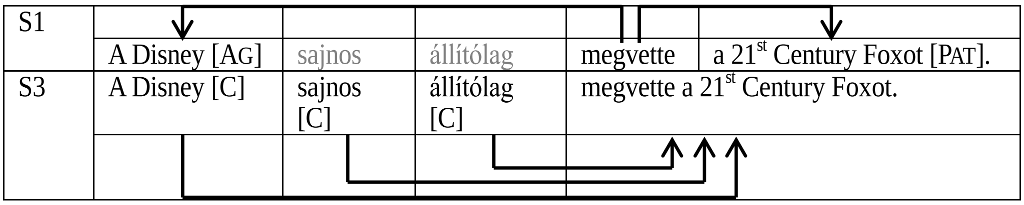
Gumperz (1982: 131) defines contextualization as the process by which discourse participants "foreground or make relevant certain aspects of background knowledge and underplay others", using the term 'contextualization cue' to refer to linguistic signals for the situated understanding of socio-cultural meaning (see also Verschueren, 1999: 112). However, my understanding of contextualization is mostly informed by Halliday's following remark: "the message begins with »let me tell you how this fits in«, and/or »let me tell you what I think about this«" (Halliday, 2014: 109). Drawing on this insight, I suggest that the dual role of contextualizers is to facilitate the efficient processing of the foregrounded information and/or to signal the speaker's intended interpretation.

In (3), starting the sentence with *a Disney* 'Disney.NOM' is optimal as it is a fundamental aspect of our knowledge that processes are anchored to participants (for anchoring, see Langacker, 2012); this is how we memorize and retrieve them. I regard topic as a subtype of contextualizers that stands in an aboutness relation with the contextualized message. A topic offers a natural point of departure for the speaker to activate a body of knowledge and also gives an early signal to the listener as to where she can integrate the new information.

While topics primarily aid the efficiency of processing, the use of *sajnos* 'unfortunately' and *állítólag* 'allegedly' is motivated by the need for speakers to signal their intended interpretation of the message. Expressing an evaluative stance is often important because speakers tend to be engaged in attempts at influencing how their listeners perceive and interpret the world, they may want to express empathy, etc. *Sajnos* contributes significantly to the interpretative context of the contextualized part (the "rheme") in (3) by signalling the transaction's negative evaluation by the speaker. Finally, the use of *állítólag* 'allegedly' is motivated by the cooperative principle (Grice, 1975). Since the speaker of (3) is unable to assume full personal responsibility for the validity of the message, she avoids potentially misleading the listener by using an evidential expression to indicate the fact that her information comes from others.

One advantage of the concept of contextualization is that it readily accounts for cases when a contextualizer appears at the right periphery. For example, the linearization *A Disney sajnos megvette a 21st Century Foxot állítólag* is also grammatical in Hungarian. In this variant, *állítólag* 'allegedly' supplies retroactive contextualization (cf. Verschueren, 1999: 112). By the same token, left-dislocation and right-disloction of a referential expression (e.g. *Messi, he's a brilliant player* vs. *He's a brilliant player, Messi [is]*) can both by subsumed under the analysis.

In the remainder of this section, I offer a simplified representation of S3 in the context of cross-dimensional mappings. Ignoring the formal dimensions for the sake of simplicity, we can analyse example (3) as follows.



S1					
	A Disney [AG]	sajnos	állítólag	megvette	a 21st Century Foxot [PAT].
S3	A Disney [C]	sajnos [C]	állítólag [C]	megvette a 21st Century Foxot.	

Figure 3. An analysis of (3) in S1 and S3.

As the analysis shows, S3 contains three contextualizing relations (C) aiding the processing and intended interpretation of the foregrounded information.[12] In S1, *sajnos* and *állítólag* play no role, which is marked by grey colour. The fact that elements need not participate in all of the dimensions motivates the following formulation of the principle of **connectedness** (cf. Mel'čuk, 1988: 23): Each element must be linked to at least one other element, in at least one dimension.

Moreover, on the basis that *megvette a 21st Century Foxot* represents a single node in S3 but a combination of interconnected elements in S1, the following principle seems justified: A catena (connected subgraph) of one dimension may function as a single node in another. This may be seen as an example of **network modularity** (Newman, 2006).

[12] In this paper, I cannot go into a detailed discussion of the various types of contextualizers. I only mention the following for orientation: a) anchoring to a person or thing (a subtype of which is the topic or aboutness function), b) situating a process in space or time, c) epistemic modality and evidentiality, d) attitude, e) inter-clausal relations (e.g. the marking of serial order). The ultimate source is Brassai (2011 [1860]: 52–54), listing and illustrating no fewer than 18 subtypes of inchoative.

This principle also holds for the mapping between S1 and S2. For example, *melyik cég* 'which firm.NOM' is a single node of S2 (as an overrider) in *Melyik cég vette meg a 21st Century Foxot?* 'Which firm bought 21st Century Fox?' but it is a combination of two connected elements, i.e. a multiword catena, in S1.

4 Summary and conclusions

The goal of this paper has been to give a concise presentation of a new approach to Hungarian and the rudiments of a theory, or research program, that has emerged from it. Cognitive dependency grammar (CDG) is envisaged as the study of form-meaning correspondences in multiple dimensions, each of which takes the form of a graph. The idea of a multigraph description is shared with XDG; however, the content of the dimensions is closer to Halliday's Systemic Functional Grammar. Three semantic dimensions have been introduced for frame semantic relations (S1), speech function (S2) and contextualization (S3), linked to one or more of the formal representations F1 (morphology), F2 (word order) and F3 (prosody). Highlights of the proposal include the idea that subjecthood may be reduced in languages like Hungarian to a set of construction-specific mappings between thematic roles and morphological properties; the concept of overriding for describing a varied class of inversion-triggering elements; and the notion of contextualization subsuming topics along with other expressions aiding the efficient processing and/or intended interpretation of a message.

Acknowledgments

The research behind this paper was supported by a postdoctoral fellowship awarded by the National Research, Development and Innovation Office (NKFIH) of Hungary, under project number PD 120934. Further support was received from NKFIH under project number K 129040. The valuable comments of three anonymous reviewers are also gratefully acknowledged.

References

Arnault, Antoine, and Pierre Nicole. 1662. *La logique ou l'Art de penser*, 1st ed. Jean Guignart, Charles Savreux, & Jean de Lavnay, Paris. [Eng. transl. *Logic; or, The Art of Thinking*, translated by Several Hands. 1st ed. H. Sawbridge, London, 1685.]

Brassai, Sámuel. 2011 [1860–1888]. *A magyar mondat.* [The Hungarian sentence.] Texts selected by László Elekfi and Ferenc Kiefer. Tinta, Budapest.

Croft, William. 1994. Speech act classification, language typology and cognition. In Tsohatzidis, Savas L. (ed.), *Foundations of speech act theory: Philosophical and linguistic perspectives*. Routledge, London & New York. 460–477.

Croft, William. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford University Press, Oxford.

Debusmann, Ralph, Denys Duchier, Alexander Koller, Marco Kuhlmann, Gert Smolka and Stefan Thater 2004. A Relational Syntax-Semantics Interface Based on Dependency Grammar. *Proceedings of the 20th international conference on Computational Linguistics*. Geneva, Switzerland. http://acl.ldc.upenn.edu/coling2004/MAIN/pdf/26-753.pdf.

Diessel, Holger. 2015. Usage-based construction grammar. In: Ewa Dabrowska and Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*. Mouton de Gruyger, Berlin. 295–321.

É. Kiss, Katalin. 1998. Identificational focus versus information focus. *Language* 74:245–273.

É. Kiss, Katalin. 2002. *The syntax of Hungarian*. Cambridge University Press, Cambride.

Fauconnier, Gilles. 1985. *Mental spaces: Aspects of meaning construction in natural language.* MIT Press, Cambridge, MA.

Fillmore, Charles J. 1982. Frame semantics. In: The Linguistic Society of Korea (eds.): *Linguistics in the Morning Calm*. Hanshin, Seoul. 111–137.

Firbas, Jan. 1992. *Functional sentence perspective in written and spoken communication.* Cambridge University Press, Cambridge.

Gabelenz, Georg von der. 1868. Ideen zu einer vergleichenden Syntax: Wort- und Satzstellung. *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* 6:376–384.

Garvin, Paul (ed.). 1964. *A Prague School reader on esthetics, literary structure, and style.* Georgetown University Press, Washington, DC.

Goldberg, Adele. 1995. *Constructions: a Construction Grammar approach to argument structure*. University of Chicago Press, Chicago.

Grice, Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics, 3: Speech acts*. Academic Press, New York. 41–58.

Gumperz, John. 1982. *Discourse strategies*. Cambridge University Press, Cambridge.

Halliday, M.A.K. 1994. *An introduction to Functional Grammar*. 2nd edition. Arnold, London.

Halliday, M.A.K. 2014. *Halliday's introduction to Functional Grammar.* 4th edition. Revised by Christian Matthiessen. Routledge, London & New York.

Hudson, Richard. 1990. *English Word Grammar*. Blackwell, Oxford.

Hudson, Richard. 2007. *Language networks. The new Word Grammar.* Oxford University Press, Oxford.

Hudson, Richard. 2010. *An introduction to Word Grammar*. Cambridge University Press, Cambridge.

Imrényi, András. 2017a. Az elemi mondat viszonyhálózata. In: Tolcsvai Nagy Gábor (ed.), *Nyelvtan.* [Grammar.] Osiris, Budapest. 664–760.

Imrényi, András. 2017b. Form-meaning correspondences in multiple dimensions: The structure of Hungarian finite clauses. *Cognitive Linguistics* 28(2):287–319.

Kahane, Sylvain. To appear. *How dependency syntax appeared in the French Encyclopedia: from Buffier (1709) to Beauzée (1765).*

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar, vol. 1. Theoretical prerequisites*. Stanford University Press, Stanford.

Langacker, Ronald W. 2001a. Discourse in Cognitive Grammar. *Cognitive Linguistics* 12:143–188.

Langacker, Ronald W. 2005. Construction grammars: cognitive, radical, and less so. In Francisco J. Ruiz de Mendoza Ibáñez and M. Sandra Peña Cervel (eds.), *Cognitive linguistics. Internal dynamics and interdisciplinary interaction*. Berlin & New York: Mouton de Gruyter. 101–162.

Langacker, Ronald W. 2008. *Cognitive Grammar: A basic introduction*. Oxford University Press, Oxford.

Langacker, Ronald W. 2009. *Investigations in Cognitive Grammar*. Mouton de Gruyter, Berlin & New York.

Langacker, Ronald W. 2012. Substrate, system, and expression: Aspects of the functional organization of English finite clauses. In Mario Brdar, Ida Raffaelli & Milena Žic Fuchs (eds.), *Cognitive linguistics between universality and variation*. Cambridge Scholars Publishing, Newcastle upon Tyne. 3–52.

Langacker, Ronald W. 2015. Descriptive and discursive organization in Cognitive Grammar. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman and Hubert Cuyckens (eds.), *Change of paradigms – new paradoxes: recontextualizing language and linguistics.* [Applications of Cognitive Linguistics 31.] Mouton de Gruyter, Berlin & Boston. 205–218.

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

Newman, Mark E. J. 2006. "Modularity and community structure in networks". *Proceedings of the National Academy of Sciences of the United States of America:* 103(23):8577–8696.

Osborne, Timothy and Thomas Gross. 2012. Constructions are catenae: construction grammar meets dependency grammar. *Cognitive Linguistics* 23:165–216.

Tesnière, Lucien. 2015. *Elements of structural syntax* (Trans. by T. Osborne, & S. Kahane). John Benjamins, Amsterdam. (Trans. of Éléments de syntaxe structurale. Klincksieck, Paris, 1966, second edition; first edition, 1959).

Verschueren, Jeff. 1999. *Understanding pragmatics*. Arnold, London.

Interpreting and defining connections in dependency structures

Sylvain Kahane
Modyco, Université Paris Nanterre & CNRS
sylvain@kahane.fr

Abstract

This paper highlights the advantages of not interpreting connections in a dependency tree as combinations between words but of interpreting them more broadly as sets of combinations between catenae. One of the most important outcomes is the possibility of associating a connection structure to any set of combinations assuming some well-formedness properties and of providing a new way to define dependency trees and other kinds of dependency structures, which are not trees but "bubble graphs". The status of catenae of dependency trees as syntactic units is discussed.

1 Introduction

The objective of this article is twofold: first, to show that dependencies in a dependency tree or a similar structure should not generally be interpreted only as combinations between words; second, to show that a broader interpretation of connections has various advantages: It makes it easier to define the dependency structure and to better understand the link between different representations of the syntactic structure. We will focus on the possible syntactic combinations (what combines with what), without considering the fact that combinations are generally asymmetric, linking a governor to a dependent. We use the term *connection* to denote a dependency without the governor-dependency hierarchy.

Section 2 presents the notions of combination and connection, which are based on the notion of catena (Osborne et al. 2012). The properties of the set of catenae are studied and used to define an equivalence relation on the set of combinations, which is used to define the connections. Section 3 draws two consequences concerning the interpretation of phrase structure trees and the granularity of dependency structures. Section 4 reverses the process presented in Section 2 and shows how a set of units can be used to define a dependency structure, including structures which are not trees but what we call bubble graphs. We conclude by relating our interpretation of dependency structures to the cognitive process of parsing.

2 How to interpret a dependency tree

After presenting various views on combinations proposed in syntactic theories (section 2.1), we introduce the notion of catenae (Section 2.2), which allows us to introduce a new interpretation of connections (Section 2.3). A formal definition of connections is given in Section 2.4.

2.1 Various views on syntactic combinations

Let us start with an example.

(1) A photo of her room is hanging on the wall.

All syntactic theories agree on the fact that there is a subjectal construction, what Bloomfield (1933) calls an actor-action construction. But theories disagree on what units exactly are involved in this construction. For immediate constituency analysis and phrase structure grammars the combination is between the subject NP/DP *a photo of her room* and the VP *is hanging on the wall*. For dependency grammars, it is often considered that combinations are between words (see especially Mel'čuk (1988), who would consider a combination between *photo* and *is*).[1] Tesnière (1959) considered that connections were between nuclei, a nucleus generally combining a lexical word and function word.[2]

[1] Other dependency approaches make different choices: For Hudson (1984, 1987), who considers the determiner as the head of the "noun phrase", the combination is between *a* and *is*, while for the Universal Dependencies scheme, the combination is between the content words *photo* and *hanging*. We will see in Section 4 that it is possible not to choose between these different views.

[2] In the very beginning of his book, Tesnière (1959[2015]: ch. 1) says: "The sentence is an **organized set**, the constituent elements of which are the **words**. Each word in a sentence is not isolated as it is in the dictionary. The mind perceives **connections** between a word and its neighbors. The totality of these connections forms the scaffold of the sentence." But later in the same book (ch. 22), he revises this and introduces

For our example, the nuclei in question will be *a photo* and *is hanging*. This can also be compared with linguists who consider that combinations are between chunks (Abney 1991, Vergne 2000), following Frazier & Fodor (1978), who argue that utterances are processed in small chunks of about six words that are then connected to each other. Another view is to consider that the governor of a dependency is a word but that the dependent is a constituent.[3] Beauzée (1765) pointed out more than 250 years ago that it was necessary to consider as word complements both words (which he called initial/grammatical complements) and their projections (which he called total/logical complements): "For instance, in the sentence *with the care required in circumstances of this nature*; […] the preposition *of* is the initial *complement* of the appellative noun *circumstances*; and *of this nature* is its total *complement*; *circumstances* is the grammatical *complement* of the preposition *with*; and *circumstances of this nature* is its logical *complement*." (Beauzée 1765:5, cited by Kahane, forthcoming).[4]

As we will see all these views on syntactic combinations are compatible with dependency syntax.

2.2 Dependency trees and catenae

Dependency trees generally link words, which are considered as the minimal units. Figure 1 shows a surface-syntactic dependency tree (Mel'čuk 1988). We focus on the structure strictly speaking and do not introduce relation labels on dependencies.

(2) *Mary looks at a photo of her room.*

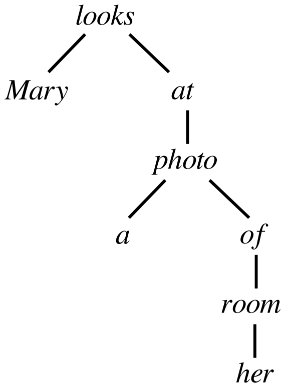


Figure 1. The dependency tree D0 of (2)

It has already been noted by various authors that dependency trees define a large number of units. Osborne et al. (2012) call *catenae* all units which are connected portions of a dependency tree and note that almost all of the units considered in syntax are catenae, starting with the constituents of phrase structure grammars.

The most interesting of these units are the (maximal) projections: To each node *x* of a dependency tree, we can associate its maximal projection, which is the unit formed by the nodes dominated by *x*, including *x* (Lecerf 1960, Beauzée 1765). For instance, the maximal projection of *photo* is *a photo of her room*.[5] An ordered dependency tree is said to be projective iff all its maximal projections are continuous.[6] As showed by Lecerf, an ordered dependency tree is projective iff its dependencies do not cross each other and no dependency covers the root.

Note that some catenae, such as *looks at photo of,* are not very relevant syntactic units. We will see how to avoid to consider such units in Section 4.

2.3 Connections and combinations

The narrow interpretation of a dependency tree is to consider that dependencies correspond to combinations of words, but other interpretations of a dependency can be made. For instance, a connection can be interpreted as a combination between the governor word and the projection of the dependent word (Beauzée 1765). The broadest interpretation that can be made of connections in a dependency tree is to consider that each dependency is potentially

the nucleus as "the set which joins together, in addition to the structural node itself, all the other elements for which the node is the structural support, starting with the semantic elements."

[3] The term *constituent* will be always used in reference to immediate constituent analysis and phrase structure grammars.

[4] Tesnière (1959[2015]: ch. 3) also considered that the dependent can be a projection: "We define a *node* as a set consisting of a governor and all of the subordinates that are directly or indirectly dependent on the governor and that the governor in a sense links together into a bundle."

[5] By removing branches formed by some of the dependents of *x*, we obtain partial projections of *x,* which are also catenae. The node *photo* has two partial connections: *a photo* and *photo of her room*.

[6] We call dependency tree only the tree structure. The dependency is generally combined with a linear order on its nodes. The projectivity is a property of the ordered dependency tree.

the combination of any catena containing the dependent node with any catena containing the governor node. For instance, the dependency between *at* and *photo* in our example can be interpreted as a combination between *at, looks at,* or *Mary looks at* on the one hand and *photo, a photo, a photo of her room* on the other hand (Figure 2). It is the latter interpretation that is the most valuable and that we will formalize now.

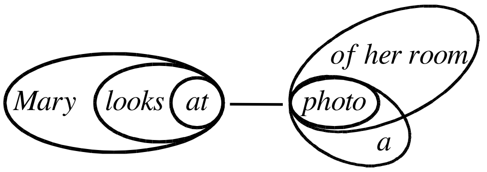


Figure 2. Broader interpretation of a connection

2.4 Formal definition of connections

Let us consider a sentence *s* which is segmented into a set X of (elementary) linguistic units (words, morphemes, etc.). A dependency tree D on X is a tree whose vertices are the elements of X.

The elements of X are linearly ordered by the fact that *s* is a string of elements of X. We call Unit(X) the set of all strings on X which are a subset of *s* (such strings may be a discontinuous subset of *s*). We call Catena(D) the subset of units in Unit(X) that are catenae on D, that is, connected portions of D.

Let us consider a set U of units. For the moment, we are interested in U = Catena(D), but our definitions will be applied to other sets of units in the following sections. A *combination* on U is any pair {A,B} of disjoint units (A ∩ B = ∅) such that A ∪ B is a unit. For instance, { *looks at*, *a photo*} is a combination on Catena(D0).

The set of combinations on U is called Combi(U). A dependency tree D defines a set of combinations, which are all the combinations of adjacent catenae, that is Combi(Catena(D)).

We want to group the combinations of Combi(Catena(D)) that correspond to the same connection of the dependency tree D. Two such combinations are said to be *compatible*. The relation of compatibility is noted ≈. For instance, all the combinations illustrated by Figure 2 should be compatible.

For a dependency tree, the compatibility can be defined by different properties. We propose three of them.

{A,B} ≈ {A',B'} iff {A ∩ A', B ∩ B'} is a combination [P1]
iff {A ∪ A', B ∪ B'} is a combination [P2]
iff A ∩ A' and B ∩ B' are not empty and A ∪ A' and B ∪ B' are disjoint. [P3]

Consequently, if {A,B} ≈ {A',B'}, then {A,B} ≈ {A',B'} ≈ {A ∩ A', B ∩ B'} ≈ {A ∪ A', B ∪ B'}.

The relation of compatibility can be represented by the configuration in Figure 3.

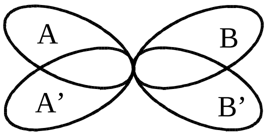


Figure 3. Two compatible combinations {A,B} and {A',B'}

The equivalence between our three characterizations of the compatibility ([P1], [P2], and [P3]) is not completely trivial. The equivalence is due to three more general properties of Catena(D) that we call the Intersection Property, the Acyclicity and the Sticking Property:

- Intersection Property: A set U of units has the Intersection Property iff for every A and B in U such that A ∩ B is non empty, then A ∩ B is in U;
- Acyclicity: A set U of units is acyclic iff for every A, B, and C in U such that A ∩ B, B ∩ C, and C ∩ A are non empty (i.e. A, B, and C form a potential cycle), then A ∩ B ∩ C is non empty;
- Sticking Property: A set U of units has the Sticking Property iff for every A and B in U, if A ∩ B is in U, then A ∪ B is in U.

Let us show the equivalence.

- [P1] → [P3]: If {A ∩ A', B ∩ B'} is a combination, then, by definition, A ∩ A' and B ∩ B' are not empty units. Due to the Acyclicity, A' ∩ B must be empty: if it was not, then A' , B, and A ∪ B' would form a cycle and their intersection, equal to (A ∩ A' ∩ B) ∪ (A' ∩ B ∩ B'), would be non empty, which is contradictory with the fact that A ∩ B and A' ∩ B' are empty. By symmetry, A ∩ B' must also be empty. We deduce that A ∪ A' is disjoint from B and B' and then from B ∪ B'.

- [P3] → [P2]: It is a direct consequence of the Sticking Property: A ∪ A' and B ∪ B' are units, because A ∩ A' and B ∩ B' are non empty, and A ∪ A' ∪ B ∪ B' is a unit, because A ∪ B and A' ∪ B' are non disjoint units.
- [P2] → [P1]: If {A ∪ A', B ∪ B'} is a combination, then A ∪ A' and B ∪ B' are disjoint. As A ∪ A', A ∪ B ∪ B', and A' ∪ B ∪ B' form a potential cycle, their intersection, which is A ∩ A', is non empty (Acyclicity). Similarly, B ∩ B' is non empty. Finally, (A ∩ A') ∪ (B ∩ B'), which is equal to (A ∪ A') ∩ (B ∪ B') in this particular case (because A ∩ B and A' ∩ B' are empty), is a unit (Intersection Property).

The relation ≈ is reflexive and symmetrical on Unit(X), but it is not transitive whenever X has more than 3 elements.[7] Nevertheless, the relation ≈ is transitive on Combi(Catena(D) and is then an equivalence relation. More generally, the transitivity of ≈ on Combi(U) is a consequence of the Acyclicity and the Sticking Property. Suppose that we have {A,B} ≈ {A',B'} and {A',B'} ≈ {A",B"}. This entails that A ∩ A" is non empty, because it is the intersection of A ∪ A' ∪ A", A ∪ B ∪ B' ∪ B", and A" ∪ B ∪ B' ∪ B", which form a potential cycle. Similarly, B ∩ B" is non empty and then {A,B} ≈ {A",B"}.

An equivalence relation on a set E induces a partition of E. The subsets of E forming this partition are called the equivalence classes on E. An equivalence class is a subset of equivalent elements. The set of equivalence classes on E for an equivalence relation R is called the quotient of E by R and is noted E/R. The elements of Combi(Catena(D)/≈ are the *connections*.

For those who are not familiar with quotient sets, it is certainly no doubt to draw a parallel with the set of rational numbers. We know that 1/2 or 2/4 or 50/100 represent the same rational number. In other words, the set of rational numbers is a quotient set obtained from the set of couples of integer numbers and the following equivalence relation: two couples (a,b) and (a',b') represent the same rational number if and only if ab' = a'b. Each couple (a,b), or fraction a/b to take the usual notation, is a representative of the relational number. Similarly, combinations are representatives of connections. For instance, { *at, a photo of her room* } and { *Mary looks at, a photo* } are two representatives of the same connection, represented in Figure 3.

Connections in a dependency tree have a minimal representative, which is a combination between two elementary units, that is, two elements of X. The minimal representative of the connection of Figure 3 is { *at, photo* }. By considering only the minimal combinations, we obtain a graph on X that we call the *connection structure* underlying the dependency tree (see Figure 4).

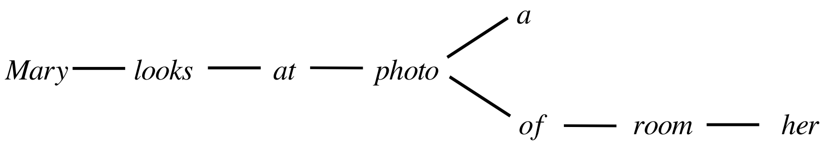


Figure 4. The connection structure associated to D0

The connection structure underlying a dependency tree D and the set of connections on D, Combi(Catena(D)/≈, are equivalent: one can be deduced from the other.

3 First consequences of our interpretation of dependency structures

3.1 Interpretation of phrase structure trees

Let us take a very simple sentence such as *Mary loves Peter*, whose dependency tree D1 contains a subject dependency from *loves* to *Mary* and an object dependency from *loves* to *Peter*. Catena(D1) = {*Mary, loves, Peter, Mary loves, loves Peter, Mary loves Peter*}. With our interpretation of connections we consider that the subject dependency indicates both a combination between *Mary* and *loves* and a combination between *Mary* and *loves Peter*. And symmetrically to this, we also consider that the object dependency indicates both a combination between *loves* and *Peter* and between *Mary loves* and *Peter*. Seen from the point of view of decomposition, this means that we consider that the phrase *Mary loves Peter* can be decomposed into both *Mary* + *loves Peter* and *Mary loves* + *Peter*. This fundamentally distinguishes dependency-based analyses from constituency-based analyses, which require that only one of

[7] Consider for instance a subject combination c0 = {*Mary, is sleeping*} and two incompatible refinements of this combination: c1 = {*Mary, is*} and c2 = {*Mary, sleeping*}. We have c1 ≈ c0 and c0 ≈ c2, but c1 ≉ c2, which shows the non-transitivity.

the two possible decompositions (the combination of the subject and the verb phrase) be retained.[8] Similarly, for a phrase such as *a photo of Mary*, our dependency analysis considers both the decomposition *a* + *photo of Mary* and the decomposition *a photo* + *of Mary*.

A binary-branching constituency tree also defines combinations and connections. Starting from the set U of constituents, we can define Combi(U): {A,B} is a combination if A and B are two constituents that combine to form a constituent. Combinations on a constituency tree are pairwise incompatible (for the relation ≈) and each connection has a unique representative. If we start with the same set X of elementary units, a dependency tree on X and a binary-branching constituency tree on X define the same number of connections. (If X has *n* elements, we need *n*–1 combinations to combine all the elements of X.) But, while a dependency tree propose many ways to combine the elements (with our interpretation of combinations in a dependency tree), a constituency tree only considers one way.

As a consequence, there are many constituency trees that can be derived from a given dependency tree and, conversely, there are many dependency trees which contain the set of combinations of a given constituency tree. More formally, a constituency tree C is said to be compatible with a dependency tree D if each combination defined by C is in Combi(Catenea(D)). As we just said, every constituency tree is compatible with several dependency trees and vice versa. Every constituency tree is obtained by choosing a representative in each connection of a compatible dependency tree.

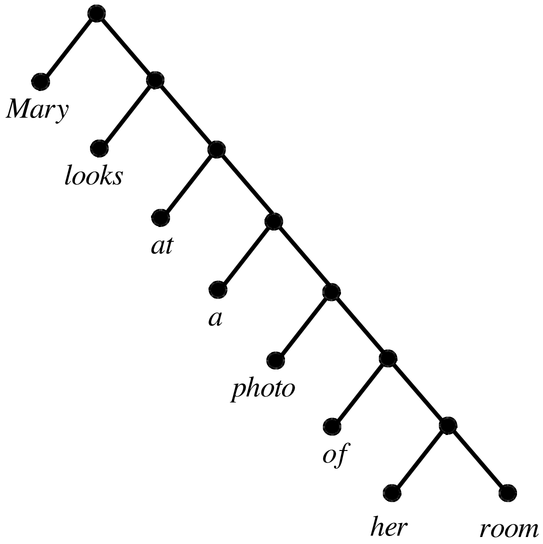


Figure 5. A binary-branching constituency tree C0 for (1)

Let us continue to see a constituency tree from the point of view of dependency. The question that arises is how a particular constituency tree is chosen from a given compatible dependency tree. Let us see how we obtain the constituency tree C0 of Figure 5 from D0. We start from the root *looks* and we choose one of the two dependencies starting from it: the subject dependency between *looks* and *Mary*. Now we choose the maximal representative of this connection {*Mary, looks at a photo of her room*}. We have now two units to analyze and we adopt the same strategy at each step: we consider each unit we have obtained, we take the root, we choose a dependency starting from the root and we choose the maximal representative of this connection inside the unit we are dealing with. In other words, all the constituency trees compatible with a given dependency tree D are obtained by ordering the dependents of every node and applying the same strategy that consists in taking the maximal representative at each step. For each order, we obtain a different binary-branching constituency tree. A dependency tree with such an order on the dependents has been called a stratified dependency tree by Kahane (1997) and is illustrated by Figure 6.

[8] Gerdes & Kahane (2013) give the following citation from Gleason (1955[1961]): “We may, as a first hypothesis, consider that each of [the words of the considered utterance] has some statable relationships to each other word. If we can describe these interrelationships completely, we will have described the syntax of the utterance in its entirety. [...] We might start by marking those pairs of words which are felt to have the closest relationship. […] We will also lay down the rule that each word can be marked as a member of only one such pair.”

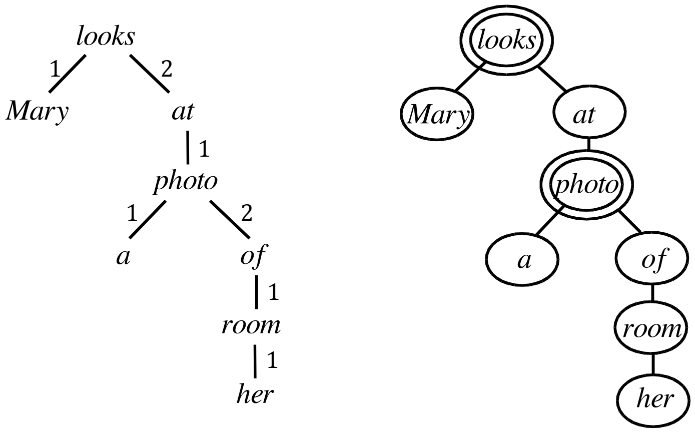


Figure 6. The stratified dependency tree D0+ giving C0
a. With order on dependencies. b. With the conventions of Kahane (1997)

Note that D0 and C0 are not equivalent: C0 induces stratification on D0 which is absent from D0 and conversely D0 induces headedness which is absent from C0. As stated by Kahane & Mazziotta (2015), a dependency tree is a connection structure plus headedness, while a constituency structure is a connection structure plus stratification. But the connection structure induced by a constituency tree is less fine-grained than the connection structure induced by a dependency structure, because it contains only large representatives and finer representatives cannot be deduced without adding additional information (such as headedness for instance).

From the theoretical point of view, the question that arises is whether or not the stratification is useful and need to be encoded in the syntactic structure. From the point of view of dependency syntax, it is an artifact of phrase structure grammars that has no real theoretical foundations.[9]

3.2 Granularity

We have just seen that our interpretation of connections in dependency trees was useful to compare dependency trees and constituency trees. It is also useful to compare dependency representations with one another.

Suppose that we have two structures S1 and S2 respectively defining the sets of combinations Combi1 and Combi 2. We say that S1 is finer than S2 if Combi1 $\supseteq$ Combi2. We have seen that dependency trees are finer than constituency trees in Section 3.1. (We will see in Section 4 that, from some points of view, traditional dependency trees can be considered as too fine.) Now consider a dependency tree D on a set of elementary unit X for a sentence *s* and take another segmentation Y of *s* less fine than X, which means that elements of Y are units on X. The dependency tree on X induces a dependency tree on Y. We can build it geometrically by collapsing connections in order to obtain only units in Y. It can also be defined algebraically by only considering combinations between units on Y, that is, Combi(Catena(D) $\cap$ Unit(Y)). The connections are still the equivalence classes for $\approx$. The new connections we obtain are subsets of the previous connections.

Changing the granularity of the analysis is quite useful. Traditional dependency trees consider words as the basic units, but some authors have considered dependency trees between morphemes (Gross 2011), while others consider that chunks (Abney 1991) are the units that combine together (Vergne 2000). Mel'čuk (1988) considers two levels of syntactic analysis: the Surface-Syntactic Structure (SSS) is a dependency tree between words (even if he considers that words are decomposed in lexemes plus inflectional morphemes), while the Deep-Syntactic Structure (DSS) is a dependency tree between (semantically full) lexical units which can be multi-word expressions. The DSS can be seen as a less granular structure than the SSS, which is interfaced with the semantic structure, while the SSS is interfaced with the phonological structure.

In Machine Translation, various granularities of the structure are involved. Current strategies based on translation memories consist in searching for the largest sub-units of a sentence for which a translation is available and combining their translations. This amounts to instantiating the connections in different ways, the units considered being generally not constituents.

[9] Scope phenomena are often considered as an argument for constituency (that is, stratification from the dependency point of view). But for instance, negation can have various scopes in a sentence such as *Mary did not give the book to Peter* and this is unrelated to the constituency structure.

3.3 Units and connections

The most important consequence of our interpretation of connections concerns the notion of connection itself. The same connection can be apprehended at different levels of granularity: between words, morphemes, chunks, constituents, etc. Whatever the level we consider, it remains more or less the same connection. As seen in Section 1, both dependency grammars and phrase structure grammars consider a subject connection for (1) even if they do not retain the same combinations as representatives. (Even inside dependency grammar, different representatives can be chosen, such as UD that favors combinations between content words, while many others favor functional heads (Osborne & Gerdes 2019).)

Hence, the connection strictly speaking is not subject to a particular level of granularity. The notion of connection is an abstraction on the notion of combination. Whereas the notion of combination is inseparable from the notion of unit, the notion of connection is not attached to a particular type of units.

From the theoretical point of view, it means that the definition of a dependency structure is not subject to a prior definition of the minimal units, and in particular to the controversial notion of word (Haspelmath 2011). As we will see now, we need to consider units to start the definition of the syntactic structure, but the units we consider at the outset are not necessary determining.

4 How to define the syntactic structure

In the first part of this paper, we started from a syntactic structure in order to see how to interpret it. We will now see how our interpretation of the structure can help us to define it.

4.1 From units to connections

We have seen that a syntactic structure such as a dependency tree defines a set of units we called catenae, following Osborne et al. (2012). We will now reverse the process and see how a set of units can define a syntactic structure. This idea was first developed by Gerdes & Kahane (2013). They propose to call fragment any part of an utterance that "is a linguistically acceptable phrase with the same semantic contribution as in the initial utterance". The fragments of a sentence are in some sense the syntactic units that are contained in the sentences, but, contrary to constituency-based analysis, it is not excluded that two syntactic units may overlap.

Gerdes and Kahane introduce the following example:

(3) *Peter wants to read the book.*

For (3), they consider the following set U of fragments: U = {*Peter, wants, to, read, the, book, Peter wants, wants to, to read, the book, Peter wants to, wants to read, read the book, Peter wants to read, to read the book, wants to read the book, Peter wants to read the book*}. We can remark that *read the book, read, the book, the,* and *book* are fragments, but not *read the* or *read book*, which are not acceptable phrases in English. Our purpose here is not to discuss the definition of fragments, but to see how the fragments can be used to define a structure.

Starting from U they build a structure they call the connection structure by first building all the binary-branching constituency trees whose constituents belong to U and then refine any of these trees by a geometric method. What we propose here is to directly build the connection structure.

Our first step is to define Combi(U), the set of combinations on U, as proposed in Section 2.3: A combination on U is any pair {A,B} of disjoint units in U ($A \cap B = \emptyset$) such that $A \cup B$ is still a unit in U.

The second step is to define the connections. Our set U verifies the three properties introduced in Section 2.3: Intersection Property, Acyclicity, and Sticking Property. Consequently, we can define the relation ≈ on Combi(U) by any of the three properties [P1], [P2], or [P3], and the relation of compatibility ≈ is an equivalence relation on Combi(U). The connections are the five equivalence classes of ≈ on Combi(U):

c1 = { {*Peter, wants*}, {*Peter, wants to*}, {*Peter, wants to read*}, {*Peter, wants to read the book*} }
c2 = { {*wants, to*}, {*wants, to read*}, {*wants, to read the book*}, {*Peter wants, to*}, {*Peter wants, to read*}, {*Peter wants, to read the book*} }
c3 = { {*to, read*}, {*to, read the book*}, {*wants to, read*}, {*wants to, read the book*}, {*Peter wants to, read*}, {*Peter wants to, read the book*} }
c4 = { {*read, the book*}, {*to read, the book*}, {*wants to read, the book*}, {*Peter wants to read, the book*} }
c5 = { {*the, book*} }.

We can now associate a connection structure to this set of connections . We will see that it does not correspond to the connection structure of a dependency tree

4.2 From connections to the connection structure

To build a connection structure, we choose in any connection the minimal representative. We obtain the five following combinations: {*Peter, wants*}, {*wants, to*}, {*to, read*}, {*read, the book*}, {*the, book*}. With these combinations we can build a structure we call a bubble graph, because some edges link non-elementary units which are represented by bubbles. (Cf. the notion of bubble tree in Kahane 1997.) It is given in Figure 6.

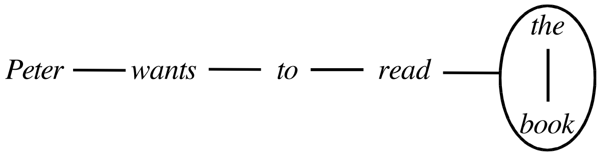


Figure 6. The connection structure associated to the set of fragments of (3)

We can obtain more complex bubble trees. Let us consider the French sentence (4).

(4) *Peter a parlé de Mary*
Peter has talked about Mary

Contrary to English, the subject in French cannot combine with the auxiliary alone (*Mary a* 'Mary has' is not an acceptable sub-phrase of the sentence) and the preposition cannot combine with the verb alone (*parlé de* 'talked about' is not an acceptable phrase). We therefore have the following set of fragments: U = { *Peter, a, parlé, de, Mary, a parlé, de Mary, Peter a parlé, parlé de Mary, a parlé de Mary, Peter a parlé de Mary* }. As before we can calculate Combi(U) and its partition by ≈. The minimal representatives of the connections are: {*Peter, a parlé*}, {*a, parlé*}, {*parlé, de Mary*}, {*de, Mary*}. We obtain the bubble tree of Figure 7.

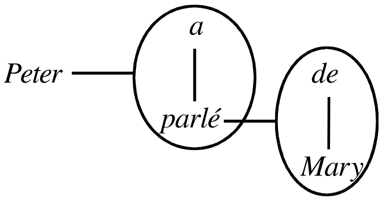


Figure 7. The connection structure associated to the set of fragments of (4)

The same methodology can apply if the elementary units we consider are morphemes. Consider the following sentence:

(5) *Peter reads two books.*

The words *reads* and *books* can be decomposed in *read-s* and *book-s*. The set of possible fragments is: U = { *Peter, read, -s, two, book, -s, read-s, book-s, Peter read-s, two book-s, read two book-s, read book-s, read-s book-s, read-s two book-s, Peter read-s book-s, Peter read-s two book-s* }. We obtain the bubble graph of Figure 8.

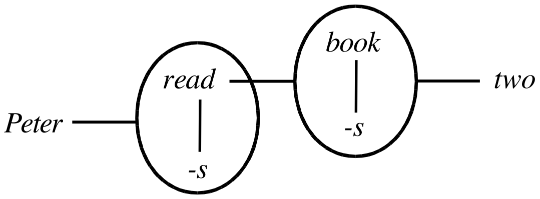


Figure 8. The connection structure associated to the morpheme-level fragments of (8)

4.3 From connection structures to catenae

It is interesting to remark that the set of fragments that gives us the connection structure can be recovered from the connection structure. It can be done by generalizing the notion of catena to bubble graphs. Let G be a bubble graph: Catena(G) contains all the units labeling the vertices of G, that is, all the elementary units as well as all the units corresponding to the content of a bubble. We add to Catena(G) all the units A ∪ B where A and B are the two vertices of an edge in the bubble graph. And then Catena(G) is satured by sticking all units that intersect: if A and B are in Catena(G) and A ∩ B is non empty, then A ∪ B must be added to Catena(G). The sticking operation is iterated until it produces no new unit.

The catenae can also be built "geometrically" by cutting edges in the bubble graph. For instance, if we take our last example, the bubble graph of Figure 8, we can cut the edge between *read* and its inflection, as well as the edge between *Peter* and the bubble containing *read-s*. We obtain a subgraph corresponding to *read two book-s*, which is therefore a catena. We can cut the edge between *two* and *book-s* and obtain *read book-s*. It is possible to separate *read* and *book-s* by cutting the edge linking them, But if we maintain this edge, it is not possible to separate *book* and *-s*, even if we cut the edge linking them. Consequently, it is not possible to obtain units such as *read book* or *read -s (*where *-s* is the plural morpheme on *book*).

4.4 From connection structures to dependency structures

A dependency is a hierarchized connection linking a governor to a dependent. A dependency structure is then a bubble graph where all (or a part) of the edges are directed. This can be done by adding criteria, essentially distributional criteria to define a head in units (see Bloomfield, 1933; Hudson 1984; Mel'čuk, 1988). We will not develop this point. See Gerdes & Kahane (2013) for a description of this step applied to connection structures. As they remark, it is possible to refine some connections when adding new criteria (refining a connection means adding combinations to it). But it must be noted that if we do that, our set of catenae will no longer correspond to the syntactic units that were identified for building the connection structure, that is, the fragments. For instance, distributional criteria applied to the French sentence (4) will show that the subject is linked to the auxiliary, as in English, because it agrees with it and that the preposition is the head of the complement of the verb. In this way, we can obtain a dependency tree for (4), but this tree produces more catenae than the fragments considered for the building of the connection structure. These new catenae are *Peter a* or *parlé de*, as well as larger units. Contrary to the fragments considered at the beginning of the process, any of these units is very legitimate from the syntactic point of view. In other words, saying that A and B are syntactically linked does not necessarily mean that A ∪ B forms an acceptable syntactic unit (even if it is true in many cases). The same problem will arise with our morpheme-level analysis of (5): For instance, if we consider, following Gross (2011) (as well as most analyses in phrase structure grammars), that phrases have functional heads, we will obtain catenae such as *Peter -s* (where *-s* is the inflection morpheme of *reads*) and *read -s* (where *-s* is the plural morpheme on *books*), which are not acceptable units. This problem can lead us to consider richer syntactic structures where the different criteria used to define the structure are not merged and we keep some traces of the role of each of them. An attempt in this direction is proposed in Kahane & Mazziotta (2015).

5 Conclusion

From the mathematical point of view, we have shown that a connection structure can be built from a set of units verifying some properties: Acyclicity, Intersection Property, and Sticking Property. The structures we obtain are what we called bubble graphs. These structures reflect the properties of the set of units they come from and, in particular, they are acyclic and all their edges are binary. It could be interesting to generalize this approach to cases with ternary connections or with cycles.

From the linguistic point of view, we have shown that the connections considered by the dependency structures can be considered as sets of combinations. Most connections have a minimal representative which is a combination of two words, but some connections are not instantiated by a combination of words. First, combinations between morphemes inside a word are not combinations of words of course. Conversely, it is possible that some combinations have a minimal representative that involves units larger than words. For example, considering that, in a sentence such as *the dog slept,* the unit *the dog* has a well identified head and that we can refine the combination {*the dog, slept*} in a unique way is likely to be a bias of the analysis by dependency trees, which requires instantiating each connection by a combination between words or, which ultimately amounts to the same, choosing a head word for each unit. In fact, by their extreme nature, the two modes of representation, dependency and constituency, are biased. While the bias for dependency trees is systematic headedness, the bias for constituency trees is stratification: At each step of the immediate constituent analysis, it is necessary to decide which two units connect, even when there are several connections available. It is not possible with a constituency tree to simply say that, in *Mary loves Peter*, *loves* combines with *Mary* and *Peter*. We must stratify, i.e. treat the connections in a certain order (considering for example that *Mary* combines with *loves Peter,* which is itself the combination of *loves* and *Peter*). As a result, there are as many constituent trees associated with a given connection structure as there are ways to order the connections.

Lastly, a few concluding remarks concerning the cognitive point of view. Even if we think that dependency structures are the best way to encode the syntactic structure, we do not think that connections are instantiated between words. We postulate that, when a hearer analyzes a sentence, connections are instantiated by particular combinations of particular units and that these instantiations may differ from one situation to another and do not correspond a priori to either words or constituents. We believe that prosody (including the silent prosody of the reader) plays an important role and that prosodic units are essential candidates for this instantiation (see Steedman (2014) for a similar approach in Categorial Grammars). As said in Section 2, Frazier & Fodor (1978) claim that main connections are instantiated by combinations of chunks of about six words. The fact that connections are not necessarily instantiated by their minimal representative can have interesting consequences from the NLP point of view: It means that parsing algorithms could take into account the fact that we are not trying to build a particular instantiation of the syntactic structure (a phrase structure tree or a dependency tree in the current state of systems), but to build any instantiation of the connections, even if it means refining them later.

We think that speakers manipulate units of different levels both when producing and analyzing statements. We have shown that our formalization of connections makes no assumptions about the nature of the units involved in the combinations instantiating the connections. We thus argue that we can study syntax without a priori asking the question of units, which is a very delicate question, since it is so difficult to define concepts such as words or sentences.[10] From this point of view, a definition that claims that syntax is the study of the organization of words within the sentence seems to us to have to be totally rejected. For us, syntax is above all the study of (free, regular) combinations between linguistic signs, without prejudging the level of granularity of these signs.

Acknowledgments

I would like to acknowledge Guy Perrier, as well as the three reviewers, for valuable comments.

References

Steven P. Abney. 1991. Parsing by chunks. In R. C. Berwick, S. P. Abney, and C. L. Tenny, *Principle-Based Parsing: Computation and Psycholinguistics*. Springer, Dordrecht, 257-278.

Nicolas Beauzée. 1765. Régime, in Denis Diderot and Jean Le Rond D'Alembert (eds.), *Encyclopédie*, vol. 14, 5-11. On line edition at enccre.academie-sciences.fr.

Leonard Bloomfield. 1933. *Language*, The University of Chicago Press.

Lyn Frazier, Janet D. Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4), 291-325.

H. A. Gleason. 1955. *An Introduction to Descriptive Linguistics.* New York: Holt, Rinehart & Winston, 503 p. Revised edition 1961.

Kim Gerdes, Sylvain Kahane. 2013. Defining dependency (and constituency). In K. Gerdes, E. Hajičová, L. Wanner (eds.), *Computational Dependency Linguistics*, IOS Press.

Thomas Gross. 2011. Catenae in morphology. *Proceedings of the first in ternational conference on Dependency Linguistics (Depling)*, Barcelona, 47-57.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica,* 45(1), 31-80.

Richard A. Hudson. 1984. *Word grammar*, Oxford: Blackwell.

Richard A. Hudson. 1987. Zwicky on heads. *Journal of linguistics*, *23*(1), 109-132.

Sylvain Kahane. 1997. Bubble trees and syntactic representations. *Proceedings of the 5th conference on Mathematics of Language (MoL)*, 70-76.

Sylvain Kahane. Forthcoming. How dependency syntax appear in the French Encyclopedia: from Buffier (1709) to Beauzée (1765). In A. Imrényi and N. Mazziotta, *History of dependency-based approaches to grammatical theory*, Benjamins.

Sylvain Kahane, Nicolas Mazziotta. 2015. Dependency-based analyses for function words – Introducing the polygraphic approach, *Proceedings of the 3rd international conference on Dependency Linguistics (Depling)*, Uppsala.

Yves Lecerf. 1960. Programme des conflits, modèle des conflits. *Bulletin bimestriel de l'ATALA,* 1(4), 11–18.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax,* 15(4), 354-396.

[10] Gerdes & Kahane (2013: note 3) defends a similar position when they say: "It may seem paradoxical that we do not think that fragments are syntactic primitives. In some sense we agree with Tesnière when he says that "the mind perceives connections". A fragment is a witness to a possible connection and it allows us to postulate one connection or another."

Timothy Osborne, 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).

Mark Steedman. 2014. The Surface Compositional Semantics of English Intonation. *Language*, 90, 2-57.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris, [transl. Timothy Osborne and Sylvain Kahane, *Elements of structural syntax*, Benjamins, 2015].

Jacques Vergne. 2000. *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur - Analyse syntaxique automatique non combinatoire*, Habilitation thesis, Université de Caen.

Identifying grammar rules for language education with dependency parsing in German

Eleni Metheniti **Pomi Park** **Kristina Kolesova** **Günter Neumann**

DFKI
Stuhlsatzenhausweg 3
66123, Saarbrücken
firstname.lastname@dfki.de

Abstract

We propose a method of determining the syntactic difficulty of a sentence, using syntactic patterns that identify grammatical rules on dependency parses. We have constructed a novel query language based on *constraint-based dependency grammars* and a grammar of German rules (relevant to primary school education) with patterns in our language. We annotated these rules with a difficulty score and grammatical prerequisites and built a *matching* algorithm that matches the dependency parse of a sentence in CoNLL-U format with its relevant syntactic patterns. We achieved 96% precision and 95% recall on a manually annotated set of sentences, and our best results on using parses from four parsers are 88% and 84% respectively.

1 Introduction

Language teaching on beginner and elementary levels, even for native speakers, brings the challenge of presenting grammatical phenomena which are familiar, unconsciously familiar or unknown to the learner, in a formal and repetitive way so that the learner will be able to understand and remember them. The presentation of these phenomena to the learner should be consistent, to establish correct patterns, repeated, to facilitate learning, and of gradual difficulty and infrequency, to ensure that the easier structures are acquired before the more difficult ones. The iRead project, in which we are scientific partners, aims to create learning applications for children in primary education, in which the user will be able to read and play with language content tailored to their learning needs, e.g. games that require the user to choose the correct morpheme, phoneme or part-of-speech to complete a pattern. Our roles are, first, to provide learning resources for native German primary school learners (ages 6-9) and, second, to provide a syntactic tool for the analysis of sentences and texts (a CoNLL–U multilingual dependency parser (Volokh and Neumann, 2012)) and a formalism that can be used to represent grammatic phenomena and query them from dependency parses. In this paper, we will be focusing on how we created our syntactic pattern formalism, the algorithm to match patterns with sentences, and the language resources that we used alongside our pattern matching tool, in order to find the grammatical rules that are applicable in a sentence.

The reason we decided to create our own query language was the need to be able to create very restrictive patterns that would almost never be found in the text erroneously or overzealously; these patterns express grammatical phenomena taught in school to young learners, and our margin for incorrect matches of a grammar rule with text is very limited. In addition, our language should be very descriptive but also human readable, so that our partners will be able to create grammatical patterns for other languages without extensive knowledge on logical operators and regular expressions. Finally, we opted to create a query language whose search relies on dependency parsing, and not on the surface structure of a clause. We will present our query language and the grammatical rule patterns that we have created for German primary school learners, and we will also present the matching algorithm we built to match these rule patterns to sentences from our corpus of children's texts. Moreover, we will be evaluating our matching algorithm's performance on this corpus with parses from our and other parsers; the reason we are not using more complex text is because our patterns are made to reflect syntactic phenomena appropriate for child learners.

2 Related Work

We are aware that many query languages have been created over the years, in which the researcher can create a pattern to extract one or multiple words with specific syntactic, morphological, orthographic etc. features from text. However, most of them do not support queries from dependency parses, but require annotated text with parts-of-speech, and only a few such as *ANNIS* (Zeldes et al., 2009) allow for patterns to look for relationships between two nodes of a syntactic tree. Other languages require the position of extra words given explicitly relative to the first word (*COSMAS II*; (Bodmer, 1996)), or rely on neighbouring words without capturing any dependencies (*Poliqarp*; (Przepiórkowski et al., 2004)). In addition, these query languages require a certain level of expertise with regular expressions and the syntax of the language; efforts have been made to simplify the syntax of these languages, for example *Coral* (Kuhn and Höfler, 2012) is a controlled natural language that translates natural language queries to the *ANNIS* syntax.

Query languages tailored for use with dependency parses also have existed for a while; for example *PML-TQ* (Pajas and Štěpánek, 2009) contains a very robust query language which is able to search for one, two or multiple constituents of a syntactic tree, either terminal non-terminal nodes. It is versatile and dynamic, and it would allow us to define patterns between words and phrases to cover the simplest rules (e.g. the presence of predicate) to more complex (e.g. constituents of a question clause), but its syntax is very complex for us to use throughout our project. *TüNDRA* (Martens, 2012) is another query language which also supports queries of one or multiple words based on annotation, deep or surface structure, negation, etc. and uses a similar syntax and the TIGERSearch annotation schema (Lezius, 2002). For our intents and purposes, it would be a fairly complete approach for our task of querying grammatical rules; however, we still wanted to attempt an approach that would be inspired by the successes of the predecessors and offer even better readability and adherence to the theory of dependency parsing, instead of also offering a search for serialized words, syntactically meaningless strings etc.

To create the queries for the grammatical rules, as explained in Section 1, we avoided the use of an automatic method to extract syntactic patterns automatically from text. Pattern induction would not be accurate and informative enough to create patterns for the specific grammatic rules that we have declared. For example, a statistical extraction (Ammar, 2016) that created pairs of a *dependent* and *head* word from dependency parses of English sentences would probably not be sufficient in capturing all the constituents of a grammatical rule, and in any case would require human annotation to the corresponding grammatical rule and its difficulty and frequency. A statistic approach close to our needs involves extracting syntactic patterns based on syntax trees from a large English corpus and scoring their difficulty based on a Zipfian distribution (Kauchak et al., 2017). However, as they discuss in their paper and in previous research (Kauchak et al., 2012), frequency is a solid but not determining factor to the difficulty of a pattern, and surface syntactic structure is not sufficient to describe a grammatical phenomenon.

3 Query Language

Our goal is to create syntactic patterns that reflect grammatical phenomena, as taught in primary school education, in a formal language that could be machine-readable, by using the dependencies of words in a sentence, and also adequately user-friendly. Our syntax should be able to map the dependencies among two or more words, use syntactic features (parts-of-speech, dependency labels), morphosyntactic features (lemma, case, number etc.) and orthographic features (one-on-one match with a word, punctuation etc.), and also be position-independent, so that it can find dependencies that span across the sentence.

Our approach is based on the theory of *abstract role values* of *constraint-based dependency grammars* (White, 2000). These grammars possess a set of *lexical categories* of the elements of a phrase, a set of their *roles*, a set of their *labels*, and these sets are governed by a set of *constraints* (Nivre, 2005). In our approach, we create sets of possible syntactic features for each word of the phrase separately (set of part-of-speech tags, set of dependency labels, set of morphosyntactic information) that should match the features of a word in a dependency parse. Then, we pair these sets with the sets of features that the word's head should possess (if a head-dependent connection is needed in the parser), and add more sets of features or tuples of dependent-head features if needed by the pattern. By *head*, we are referring to the head of a two-word phrase, not to the *root* of the sentence; this will allow us to build patterns referring to one-word rules or rules with words that are not directly dependent on the *root*.

We developed an extendable structure to cater to simple and complex structures. The first word that needs to be matched in a pattern is called *comp_word*, after the term *complement* in a head-driven phrase structure. This may have a set of possible parts-of-speech, labels, morphosyntactic features, lemmata, word forms, and morphemes. The second word is the *head_word*, the head of the first word as defined by the dependency parse. This one also has its own set of possible features, and the pattern will only be valid if both words are matched. A pattern template is presented in Figure 1.

comp_word: **POS=**{A,B}&label={c}&
feature={d,e}&**lemma=**{'e'}&
wordform={'f','g'}&
wordform={h-,i-}&
wordform={-j-},
head_word: **POS=**{K}&**label=**{l,m}&
feature={o}&**wordform=**{-p,-q},
tokenID(head_word) = headID(comp_word)

Figure 1: Template for a pattern with a *head-dependent* relation.

Every field may have one or more possible values. The fields **POS**, **label**, **lemma**, and **wordform** will be matched with one of the corresponding features of the word. The field **feature** requires all listed morphosyntactic features of the pattern to match the morphology of the word. Not all possible sets need to be filled, as shown in the *head_word* features; the pattern can include as much relevant information as needed in each grammatical phenomenon. Values should be separated by a comma in every set, and brackets should be used when a word is used, e.g. in **lemma** and **wordform**. Concerning **wordform**, this field can contain either a specific word (preferably inflected), one or multiple prefixes, one or multiple suffixes, or one or multiple infixes. Different types of values should exist in their own **wordform** field, as demonstrated in *comp_word*.

In order to understand better how patterns are created and match words, we will examine a pattern to find a simple noun phrase with a definite article, in Figure 2.

comp_word: **POS=**{DET}&**label=**{det}&
feature={Definite=Def,PronType=Art},
head_word: **POS=**{NOUN},
tokenID(head_word) = headID(comp_word)

Figure 2: Pattern to identify a noun phrase with a definite article.

In order for this pattern to exist in a sentence or phrase, we need to have a word that is a *determiner* as part-of-speech, labeled as *determiner* by the dependency parser, have the features of *definiteness* and being an *article*, and be dependent to a word that is a *noun*. For example, this pattern would be found in the German sentence, *Die Katze schläft.* "The cat sleeps.". According to the dependency tree of the sentence in Figure 3 and the parse in Figure 1, there is a word matching the dependent (*Die*) and its head (*Katze*) matches the *head_word* of the pattern. Therefore, the pattern, and the rule for noun phrase, will be found.

This structure can also support simpler grammatical rules that only require matching one element of the sentence. All the fields that were used above can also be applied here. The word to be matched is tagged as *head_word*, as there is no dependency to create a *head-complement* set, e.g. a one-word pattern grammatical rule that looks for the presence of a definite article (regardless of its dependencies) shown in Figure 4. This pattern would be found in the previous example sentence, because the word *Die* matches all these requirements.

In order to describe more composite grammatical structures, we can use multiple syntactic patters of one or two words, combined. All separate patterns should be matched with the words in the sentence, in order of this compound syntactic pattern to be matched. Since in dependency parsing there is always a pair of *head-complement* no longer than two words, in order to describe phenomena that involve more

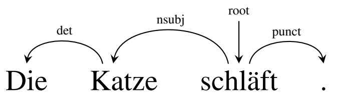


Figure 3: Dependency tree of the sentence *Die Katze schläft.*

Die	**Katze**	**schläft**	**.**
“der"	“Katze"	“schlafen"	.
POS=DET	POS=NOUN	POS=VERB	POS=PUNCT
label=det	label=nsubj	label=root	label=punct
Case=Nom	Case=Nom	Number=Sing	
Definite=Def	Gender=Fem	Person=3	
Gender=Fem	Number=Sing	VerbForm=Fin	
Number=Sing			
PronType=Art			
head=“Katze"	head=“schläft"		head=“schläft"

Table 1: A CDG parse of the sentence *Die Katze schläft.*

head_word: **POS=**{DET}&**label=**{det}&
feature={Definite=Def,PronType=Art}

Figure 4: Pattern to identify a definite determiner.

than two words, first we make patterns of one or two words and connect these patterns by finding their common denominator. This has to be a unique word in the utterance on which all the other words are *dependent*– the *root*. For example, in order to create a pattern for a simple sentence with a mono-transitive verb, e.g. *Er liebt Maria.* “He loves Maria.", our course of action would be to create a pattern that matches a *nominal subject* with a *verb* which is the *root* of the sentence, and a second pattern which matches a *direct object* with a *verb* that is also the *root* of the sentence. In a sentence, only one *root* should exist. Therefore, both patterns have the same *head_word*.

As shown in Figure 6 and Table 2, the compound pattern in Figure 5 will match the sentence ‘Er liebt Maria’, because both patterns in the compound pattern are matched.

comp_word: **label**={nsubj},
head_word: **POS**={VERB}&**label**={root},
tokenID(head_word) = headID(comp_word)
AND
comp_word: **label**={obj},
head_word: **POS**={VERB}&**label**={root},
tokenID(head_word) = headID(comp_word)

Figure 5: Pattern for a simple mono-transitive sentence.

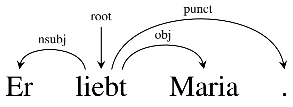


Figure 6: Dependency tree of sentence *Er liebt Maria.*

Er	**liebt**	**Maria**	**.**
“er"	“lieben"	“Maria"	.
POS=PRON	POS=VERB	POS=PROPN	POS=PUNCT
label=nsubj	label=root	label=obj	label=punct
Case=Nom	Number=Sing		
Gender=Masc	Person=3		
Number=Sing	VerbFrom=Fin		
Person=3			
head=“liebt"		head=“liebt"	head=“liebt"

Table 2: CDG parse of the sentence *Er liebt Maria.*

Our previous pattern only used dependency labels and part-of-speech tags for a good reason; in the grammar rule we defined, we are looking for sentences with a nominal subject and a direct object, regardless of their part-of-speech (pronoun, a noun, a proper noun etc.) and their morphosyntactic features. However, this under-defining could prove problematic. Suppose we have a *reflexive sentence*, e.g. *Ich wasche mich.* “I wash myself." (Figure 8 and Table 3). This is a reflexive sentence, because the object of the sentence has the same reference as the subject. Reflexivity is a more complex syntactic structure than a simple sentence with two different entities as subject and object, and we would like to create a special pattern for reflexive sentences. See a pattern for such cases of simple reflexive sentences in Figure 7.

comp_word: **label**={nsubj},
head_word: **POS**={VERB}&**label**={root},
tokenID(head_word) = headID(comp_word)
AND
comp_word: **label**={obj,iobj}& **feature**={PronType=Prs, Reflex=Yes},
head_word: **POS**={VERB}&**label=**{root},
tokenID(head_word) = headID(comp_word)

Figure 7: Pattern for a simple reflexive sentence.

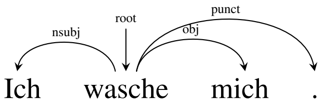


Figure 8: Dependency tree of sentence *Ich wasche mich.*

Ich	**wasche**	**mich**	**.**
“ich"	“waschen"	“ich"	.
POS=PRON	POS=VERB	POS=PRON	POS=PUNCT
label=nsubj	label=root	label=obj	label=punct
Case=Nom	Number=Sing	Case=Acc	
Gender=Masc	Person=3	Gender=Masc	
Number=Sing	VerbFrom=Fin	Number=Sing	
Person=1		Person=1	
PronType=Prs		PronType=Prs	
		Reflex=Yes	
head=“wasche"		head=“wasche"	head=“wasche"

Table 3: CDG parse of the sentence *Ich wasche mich.*

The sentence *Ich wasche mich.* would match the reflexive sentence pattern, but it would also match the aforementioned pattern for simple mono-transitive sentences, because in dependency parsing, reflexive pronouns are dependent on the head of the clause and not on the entity they reference. While this reflexive sentence is a mono-transitive sentence and the mono-transitive sentence pattern correctly matches it, we would like to keep these two structures separate from each other because of their different difficulties. We could add a constraint to the pattern for reflexive sentences that would state that if both the pattern for mono-transitive sentences and reflexive sentences is matched, then the most ‘relevant’ one is reflexive sentences. However, this approach would be difficult as our set of grammar rule patterns grows and we would have to keep track of all pre-existing possible matching patterns. Our second option would be to revise the way we define simple patterns and add *exclude operators* that would prevent more complex cases to be matched with simpler patterns. An exclude operator (tilde and parentheses) is wrapped around a pattern or a simple pattern and can be used in one or more patterns in a compound pattern. If the pattern inside the exclude operator is found, then the pattern is deemed to not be a match. For example, we would revise our simple mono-transitive sentences pattern to exclude the presence of an indirect object (hence, not matching bi-transitive sentences and the presence of reflexivity) in Figure 9.

comp_word: **label**={nsubj},
head_word: **POS**={VERB}&**label**={root},
tokenID(head_word) = headID(comp_word)
AND
comp_word: **label**={obj},
head_word: **POS**={VERB}&**label**={root},
tokenID(head_word) = headID(comp_word)
AND
∼***(comp_word:*** **label**={iobj},
head_word: **POS**={VERB}&**label**={root},
tokenID(head_word) = headID(comp_word))
AND
∼***(comp_word:*** **label**={obj,iobj}& **feature**=
{PronType=Prs,Reflex=Yes},
head_word: **POS**={VERB}&**label=**{root},
tokenID(head_word) = headID(comp_word))

Figure 9: The revised pattern for simple mono-transitive sentences.

While this approach may seem more arduous, since we would have to take into account multiple cases when making a pattern, it enables the definition of very specific patterns that cater to specific grammatical phenomena, like this case of reflexive sentences. It can also help us define differences between patterns that cannot be taken account by using only dependencies. For example, a simple yes-no question in German, e.g. *Hast du Zeit?* “Do you have time?" (Figure 10) would have the same dependency structure as the sentence *Du hast Zeit.* “You have time." (Figure 11). Therefore, it is not possible to discern between these two cases with a pattern, unless we use an exclude operator that excludes the presence of a specific punctuation mark. The reason we are not using the positions of words in a sentence in our patterns for this case or any other pattern so far is because dependencies are meant to capture deep structural relationships, regardless of position. Declaring strict positions for arguments in a pattern could be problematic for languages that allow even small liberties in word order such as German. *Das Buch lese ich!* and *Ich lese das Buch!* both translate to “I read the book!" despite the surface structures being OVS and SVO, respectively. Ultimately, the choices on how patterns will match grammatical rules and sentences belong to the creators of the patterns for each language.

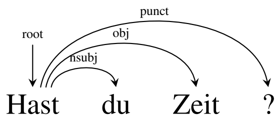


Figure 10: Dependency tree of *Hast du Zeit?*

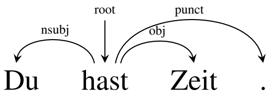


Figure 11: Dependency tree of *Du hast Zeit.*

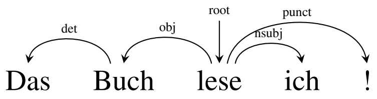


Figure 12: Dependency tree of *Das Buch lese ich!*

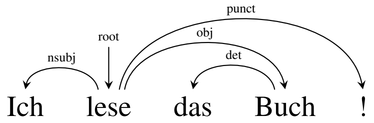


Figure 13: Dependency tree of *Ich lese das Buch!*

4 The matching process

4.1 Building syntactic patterns for German

Now that we have defined our query language, we will present the process of collecting the appropriate grammar rules and creating the patterns to find these rules in a sentence-level. Since our target demographic was primary school children, we had to focus on simpler grammar rules and syntactic structures, and pay close attention to what difficulty level we will assign to them, so that students would be introduced to concepts with a gradual difficulty and based on already acquired rules. It is important to understand which syntactic phenomena are used at each age. While Kauchak et al. (2007) have mentioned that the frequency of a parse tree structure correlates to its difficulty, there are more factors to how difficult a grammar rule is, e.g. young German students are not introduced to complex cases such as passive voice in German until 10/11 years old (*Klasse 5/6*). (Note that Germany does not have a unified school curriculum and syllabus, and every state defines their own standards; we consulted the school curricula of the German states of Saarland and Rheinland-Pfalz to understand which syntactic phenomena are used at each age.) To further study the syntactic phenomena, we consulted linguistics textbooks (Altmann and Hahnemann, 2007).

As was discussed in Section 3, we built the patterns following grammar rules as close as possible, excluding cases where the pattern would be too general. We used the Universal Dependencies 2.3 annotation schema for our patterns (Nivre et al., 2018a). So far, we have created 135 patterns for morphosyntactic and syntactic rules in German with their syntactic categories, a human-readable description, a difficulty score, and their prerequisite rules (a list of what rules need to be already known in order for this rule to be taught. It is used by our partners in the project to automatically curate content according to the user's level.). We present an abridged version of a few of syntactic rules, their difficulty, and the patterns we have created to match them; Table 4 with simple rules, Table 5 with complex rules and Table 6 with compound rules.

ID	Description	Dif.	Pattern
218	Auxiliary verb “sein", present indicative	1	head_word: POS={AUX}&wordform={“bin",“bist",“ist",“sind",“seid",“sein"}&feature={VerbForm=Fin}
222	Auxiliary verb “haben", present indicative	1	head_word: POS={AUX}&wordform={“hab",“habe",“hast",“hat",“haben"}&feature={Mood=Ind,VerbForm=Fin}

Table 4: A few simple syntactic patterns to match one word. ‘Dif’ is the assigned difficulty.

ID	Description	Dif.	Pattern
240	Composed forms: Perfect indicative	1	comp_word: {<222>,<218>}, head_word: POS={VERB}&feature={VerbForm=Part}, tokenID(head_word)=headID(comp_word)
261	Adjective is Predicate to Noun	1	comp_word: POS={NOUN,PROPN,PRON}, head_word: POS={ADJ}&label={root}, tokenID(head_word)=headID(comp_word)
281	Two-part Coordinate conjunctions	2	comp_word: POS={CCONJ}&label={cc}, head_word: POS={CCONJ}&label={cc}, tokenID(head_word)=headID(comp_word)
287	Prepositions with accusative	2	comp_word: POS={ADP}&label={case}, head_word: POS={NOUN,PROPN}&feature={Case=Acc}, tokenID(head_word)=headID(comp_word)

Table 5: A few complex syntactic patterns for one dependent word (261) or a dependent word and its head (240, 281, 287). Note that the complement side of rule 240 is the simple rules 222 or 218 from Table 4.

ID	Description	Dif.	Pattern
288	Simple clause with intransitive verb	1	(comp_word: label={nsubj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: label={obj}) AND ∼(head_word: label={iobj}) AND ∼(head_word:POS={VERB}&label={root}&feature={VerbForm=Part}) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: feature={Mood=Imp}&label={root})
289	Simple clause with intransitive verb, with auxiliary verb	1	(comp_word: label={nsubj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: POS={AUX}&label={aux}, head_word: POS={VERB}&label={root}&feature={VerbForm=Part}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: label={obj}) AND ∼(head_word: label={iobj}) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: feature={Mood=Imp}&label={root})
290	Simple clause with transitive verb	1	(comp_word: label={nsubj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: label={obj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: label={iobj}) AND ∼(head_word:POS={VERB}&label={root}&feature={VerbForm=Part}) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: label={obj,iobj}&feature={Reflex=Yes}) AND ∼(head_word: feature={Mood=Imp}&label={root})
292	Simple clause with bitransitive verb	2	(comp_word: label={nsubj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: label={obj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: label={iobj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: POS={VERB}&label={root}&feature={VerbForm=Part}) AND ∼(head_word: POS={obj,iobj}&feature={Reflex=Yes}) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: feature={Mood=Imp}&label={root})
294	Reflexive sentence with transitive verb	1	(comp_word: label={nsubj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: label={obj,iobj}&feature={Reflex=Yes}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word:POS={VERB}&label={root}&feature={VerbForm=Part}) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: feature={Mood=Imp}&label={root})
296	Simple clause with predicate	1	(comp_word: label={nsubj}, head_word: POS={ADJ}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: POS={VERB,AUX}&label={cop}, head_word: POS={ADJ}&label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: feature={Mood=Imp}&label={root})
298	Simple clause with separable verb	2	(comp_word: POS={ADP}&label={compound:prt}, head_word: POS={VERB}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: POS={PUNCT}&wordform={“?"}) AND ∼(head_word: feature={Mood=Imp}&label={root})
299	Simple w- question (yes-no)	1	(comp_word: label={nsubj}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: POS={PUNCT}&wordform={“?"}, head_word: label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: feature={PronType=Int}) AND ∼(head_word: POS={VERB}&label={root}&feature={VerbForm=Part}) AND ∼(head_word: feature={Mood=Imp}&label={root})
301	Simple w- question where adverb/pronoun is Subject	1	(comp_word: POS={PRON}&label={nsubj}&feature={Case=Nom,PronType=Int}, head_word: POS={VERB}&label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: POS={PUNCT}&wordform={“?"}, head_word: label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: feature={Mood=Imp}&label={root})
302	Simple question, adverb or pronoun is Complementizer	3	(comp_word: label={advmod}&feature={PronType=Int}, head_word: label={root}, tokenID(head_word)=headID(comp_word)) AND (comp_word: POS={PUNCT}&wordform={“?"}, head_word: label={root}, tokenID(head_word)=headID(comp_word)) AND ∼(head_word: feature={Mood=Imp}&label={root})

Table 6: A few of the compound patterns that will match an entire (simple) clause.

4.2 Dictionaries and the case of multi-word expressions

The patterns need to be matched with a parsed sentence. We use the *MUNDERLINE* parser of (Volokh and Neumann, 2012) for dependency parsing in CoNLL-U trees with Universal Dependencies annotation tags. However, we require some further linguistic information for some patterns, which cannot be provided from a parser, e.g. morphemes. As part of the German language resources, we have created a dictionary of 15,000 German words, from our 117K corpus of age-appropriate texts: children's texts from children's magazines and newspapers (*GEOlino*[1], *GEOlino Extra*[2], *Dein SPIEGEL*[3]), children's

[1] https://www.geo.de/geolino

[2] https://www.geo.de/magazine/geolino-extra

[3] https://www.deinspiegel.de/

literature (*Phontasia*[4]) and pedagogical material (works of *Ursula Rickli*[5]). These words are annotated with lemma and stem information, their phonological and morphological features, their morphemes and orthographic syllables. Our dictionary is relatively small, because children's texts tend to be simple, but it should cover the most important words for sentences relevant to German primary school students.

Concerning multi-word expressions (MWEs), there are still open discussions on how they should be handled (Gerdes and Kahane, 2016). Even so, parsers are usually unable to identify them. For example, while 'as well as' is considered a fixed MWE, parsers tend to fail to capture the dependencies between the words of the MWE and the MWE's head. Our approach to identifying multi-word expressions and making use of them in syntactic patterns was to reduce a multi-word expression post-parse to a single phrase retaining the features of only one word of the MWE (Kato et al., 2016). It was only marginally worse than training a parser with MWE awareness (Candito and Constant, 2014). For example, in the sentence "We like John as well as Mary.", the MWE *as well as* is a coordinate conjunction between *John* and *Mary* (Figure 14 and Table 7). The first *as* has the features of the coordinate conjunction, and *well* and *as* are the subsequent words dependent to *as*. We concatenate the MWE as one word, *as well as*, into one word, retaining only the features of the first *as*, since it is the head of the MWE.

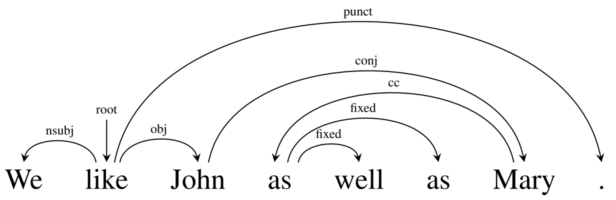


Figure 14: Dependency tree of sentence "We like John as well as Mary."

John	as	well	as	Mary
"John"	"as"	"well"	"as"	"Mary"
POS=PROPN	POS=ADV	POS=ADV	POS=ADP	POS=PROPN
label=obj	label=cc	label=fixed	label=fixed	label=conj
		Degree=Pos		
head='like'	head='Mary'	head='as'	head='as'	head='John'

Table 7: Part of the CDG parse for the sentence "We like John as well as Mary.". Note that the parsers we tested were not able to successfully annotate 'well' and 'as' with 'fixed'– this is a gold parse.

In order to recognize MWEs and which one of their components is syntactically important (whose features we are going to annotate the joined MWE with), we compiled a list of the MWEs we deem relevant for primary school level texts. This allows us to expand if necessary and handle other languages of the project.

4.3 Matching algorithm

By using our syntactic patterns for German and the required resources (dictionary, list of MWEs), we now proceed to build an algorithm to match patterns with dependency parses. In Section 3, we presented three different types of patterns: one-word patterns (Fig. 4), two-word patterns (Fig. 2) and compound patterns (Fig. 5, 7, 9). In order to match these rules to a parse, we first transform the parse from CoNLL-U format to a dictionary of dictionaries, where every word of the sentence has its own dictionary of features (part-of-speech, dependency label, head, morphosyntactic features). We also read the patterns we have made to a dictionary. Our goal is to check how many patterns matched our sentence's words, which patterns matched, and what their dependencies are (if applicable). We have created three algorithms which step-by-step look up every word in the sentence and try to match it (and its head, if applicable) to patterns. First, it tries to match the word's features with features of simple rules (one-word patterns). Then, it tries to match the word with the complement side of a complex rule. If the complement side is a match, it finds the head of the word (if applicable) and tries to match the head word's features with the head word's features in the complex rule. Finally, for rules composed of multiple patterns (compound rules), it tries to match every pattern of the rule, simple or complex, by nesting the algorithms described above.

[4] `http://www.phontasia.de/`

[5] `http://www.ursularickli.ch/`

5 Results and Discussion

From our corpus mentioned in Section 4.2, we created two sets: a development set of 152 sentences, which we used during the process of creating the syntactic patterns to fine-tune them, and a test set of 101 sentences to test the performance of our patterns and matching algorithm. We created their dependency parses manually, as a gold standard, and annotated them with the ideal grammar rules that they should be matched with. First of all, we would like to present a few sentences from the test set with the annotated matches and the matches that the algorithm returned with the use of gold standard parses, in Table 8. Some of the patterns for the rules can be found in Tables 4, 5 and 6. The matcher also returns the position of the *head_word* and the *comp_word* if applicable. As shown, the matches are mostly correct for all three types of patterns, meaning that our query language and our patterns are robust enough to describe and find the syntactic phenomena that we aimed to identify.

Sentence	Gold standard rule		Matched rule
Die Reise hat mehrere Tage gedauert. “The journey took several days."	205	Function words – Definite article	205
	269	Discourse anaphors – NP with definite article	269
	222	Function words – Auxiliary verb “haben", present indicative	222
	217	Function words – Indefinite pronouns	217
	240	Morphosyntax – Composed forms: Perfect indicative	240
	289	Clause structure – Simple clause, intransitive verb, with auxiliary verb	289
Ihre Mutter hat eine gute Idee. “Your mother has a good idea."	208	Function words – Possessive pronoun, Nominative	208
	273	Discourse anaphors – NP with possessive pronoun	273
	206	Function words – Indefinite article	206
	270	Discourse anaphors – NP with indefinite article	270
	260	Adjectives – Attribute to noun	260
	290	Clause structure – Simple clause, transitive verb	290
Jetzt konnte sich die Raupe am Ästchen festhalten. “Now the caterpillar could hold on to the branch."	307	Adverbs	307
	230	Function words – Auxiliary verbs, past	230
	274	Binding – Reflexive pronouns	274
	205	Function words – Definite article	205
	269	Discourse anaphors – NP with definite article	269
	294	Clause structure – Reflexive sentence, transitive verb	294
	233	Function words – Prepositions with dative	
	286	Discourse anaphors – Prepositional phrase in dative	
Du bist aber schick. “But you are chic."	212	Function words – Personal pronouns, nominative	212
	265	Discourse anaphors – Personal pronouns as Subject	265
	238	Function words – Particles	238
	307	Adverbs	307
	261	Adjectives – Predicate	261
	296	Clause structure – Simple clause with predicate	296
Ein Schiff fährt auf dem Meer entlang. “A ship sails along the sea."	206	Function words – Indefinite article	206
	270	Discourse anaphors – NP with indefinite article	270
	235	Function words – Prepositions with accusative and dative	235
	205	Function words – Definite article	205
	269	Discourse anaphors – NP with definite article	269
	284	Discourse anaphors – Separable prefix	284
	286	Discourse anaphors – Prepositional phrase in dative	286
	288	Clause structure – Simple clause, intransitive verb	288
	298	Clause structure – Simple clause, with separable verb	298
Hattest du denn keine Arbeit? “Did you have no work then?"	212	Function words – Personal pronouns, nominative	212
	265	Discourse anaphors – Personal pronouns as Subject	265
	238	Function words – Particles	238
	307	Adverbs	307
	207	Function words – Negative Indefinite article	207
	271	Discourse anaphors – NP with negation	271
	299	Clause structure – Simple question (yes-no)	299
	283	Negation	283
Wer steht da neben deinem Vater? “Who stands there next to your father?"	216	Function words – Interrogative pronouns	216
	275	Discourse anaphors – Interrogative pronoun, determiner, numeral or adverb	275
	307	Adverbs	307
	235	Function words – Prepositions with accusative and dative	235
	286	Discourse anaphors – Prepositional phrase in dative	286
	301	W-clauses – Simple question, adverb or pronoun is Subject	301

Table 8: Seven sentences from our test set, their ideal matches and the matches that the algorithm returned. Rules that were incorrectly matched/not matched are marked in bold.

In addition, we used several dependency parsers to parse the sentences and then evaluated their CoNLL-U trees to our gold standard. We wanted to ensure that our parser would have adequate performance and wouldn't cause mismatches that could be avoided with the use of another parser. The parsers we chose are all either pre-trained or trained on the German GSD Universal Dependencies treebank (McDonald et al., 2013): *UDPipe* (Straka and Straková, 2017)), *jPTDP* (Nguyen and Verspoor, 2018) and *Turku neural parser pipeline* (Kanerva et al., 2018). The results for the development set can be found in Table 9, and for the test set in 10. *Turku* was the most successful of the parsers, creating more accurate CoNLL-U trees with significantly higher recall than *MUNDERLINE* and *UDPipe*. *jPTDP* performed poorly because it does not output morphosyntactic features (FEATS) in its CoNLL-U trees, which is an important input to the matcher. However, in our parse tree evaluations, *jPTDP* recreated UPOS, HEAD, and DEPREL columns at least as well as –if not better than– *Turku*. *jPTDP* can be executed on top of CoNLL-U trees with FEATS, but for the purposes of our experiment, we only considered direct output of our input of a list of sentences for each parser.

	Gold Standard	***MUNDERLINE***	***UDPipe***	***jPTDP***	***Turku***
Total	978	978	911	978	911
TP	922	742	638	249	788
FP	26	92	105	98	89
FN	56	236	273	729	123
Precision	0.9726	0.8897	0.8587	0.7176	**0.8985**
Recall	0.9427	0.7587	0.7003	0.2546	**0.8650**
F1	0.9574	0.8190	0.7715	0.3758	**0.8814**

Table 9: Matcher results on the development set with gold standard parses and parse results from the parsers.

	Gold Standard	***MUNDERLINE***	***UDPipe***	***jPTDP***	***Turku***
Total	776	776	664	776	664
TP	734	601	496	246	587
FP	32	80	89	68	96
FN	42	175	168	530	77
Precision	0.9582	**0.8825**	0.8479	0.7834	0.8594
Recall	0.9459	0.7745	0.7470	0.3170	**0.8840**
F1	0.9520	0.8250	0.7942	0.4514	**0.8716**

Table 10: Matcher results on the test set with gold standard parses and parse results from the parsers.

At first glance, it may seem odd that the matcher was not able to match 100% of the rules corresponding to the gold standard. This may be due to our choices on the way we built the patterns; we did not aim for a complete representation of the German language and all the possible expressions of a grammar rule because our goal was to successfully match sentences with grammatical phenomena that are taught in primary school. Therefore, if a sentence has a grammatical rule that is expressed in a way not covered by our patterns, the pattern will not be matched. For example, the sentence *Jetzt konnte sich die Raupe am Ästchen festhalten.* (Table 8) contains the prepositional phrase *am Ästchen*. Even though there is a syntactic pattern to match prepositional phrases with the dative case, the pattern is not matched because it requires a noun as the head of the phrase, and it does not support substitution, which is a more complex syntactic phenomenon. However, the sentence is annotated with the grammar rule because the rule is present.

Additionally, the way that dependency parsing expresses some structures may cause some matches to not occur. For example, in the sentence, *Alles war grün und gelb.* "Everything was green and yellow.", only *grün* will be matched with the rule for predicate (261, 296), because *grün* is labeled as *adjective* which is the *root* of the sentence, but *gelb* is correctly labeled as *conjunct* to *grün*, because they are connected with a conjunction. Even though they have the same syntactic role, conjunct parts of speech cannot be matched to patterns that require a specific label. In the future, we will consider ways to overcome this problem, for example by adding rules to add enhanced dependencies as described in (Nivre et al., 2018b).

We also noticed that problems occur in prepositional phrases when the preposition and the determiner are contracted to one word. Ideally, in German treebanks they are analysed to preposition and determiner, but parsers overall failed to decompose these contractions. For example, in the sentence *Rotkäppchen musste zum Hause gehen.* "Little Red Riding Hood had to go home.')", *zum Hause* would be decomposed to *zu dem Hause*, and then the rule for prepositional phrases with dative would be found. Since parsers are not analysing the contraction, the pattern will not be matched. Multi-word tokens like *zum* caused differences in parse trees for UDPipe and Turku compared to our gold standard. In those instances, we excluded them from the results, which is why the total number of features matched for the two parsers is lower than the other parsers in Tables 9 and 10.

6 Future work

Our future work will be to extend the matcher to other languages. Our partners are working on creating syntactic patterns for grammar rules on a primary education level for English, Greek and Spanish. We would like to assess the pattern quality and the matcher performance for other languages as we did for German. Since our matcher is language-independent and Universal Dependencies includes annotations that cover the majority of documented languages, we are optimistic that we will have satisfactory results. Additionally, we would like to integrate the matcher to the text difficulty metric that has been developed by our other partners, since our patterns correspond to grammar rules with annotated difficulty.

As part of our ongoing research, we would like to further explore how the matcher and the syntactic patterns could be used in other NLP applications. We would like to solve problems such as conjunctions and contractions, and eventually graduate to more complex patterns. For example, we would like to create patterns for analysis of more complex sentences, something that could easily be achieved since our matcher successfully recognizes simple sentences in a clause (e.g. the matcher will return the rule for a subordinate clause and for a simple sentence, in the case of a conditional sentence).

Acknowledgements

We would like to thank our reviewers for their valuable insight. This work was partially funded by the European Union's Horizon 2020 grant agreement No. 731724 (iREAD).

References

Hans Altmann and Suzan Hahnemann. 2007. *Syntax fürs Examen: Studien-und Arbeitsbuch*, volume 1. Vandenhoeck & Ruprecht.

Waleed Ammar. 2016. *Towards a Universal Analyzer of Natural Languages*. Ph.D. thesis, Google Research.

Franck Bodmer. 1996. Aspekte der Abfragekomponente von COSMAS II. *LDV-INFO*, 8:142–155.

Marie Candito and Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 743–753.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142.

Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Construction of an English Dependency Corpus incorporating Compound Function Words. In *LREC*.

David Kauchak, William Coster, and Gondy Leroy. 2012. A Systematic Grammatical Analysis of Easy and Difficult Medical Text. In *AMIA*.

David Kauchak, Gondy Leroy, and Alan Hogue. 2017. Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.

Tobias Kuhn and Stefan Höfler. 2012. Coral: Corpus access in controlled language. *Corpora*, 7(2):187–206.

Wolfgang Lezius. 2002. TIGERSearch—ein Suchwerkzeug für Baumbanken. In *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache*, volume 6, pages 107–114.

Scott Martens. 2012. Tündra: TIGERSearch-style treebank querying as an XQuery-based web service. In *Proceedings of the joint CLARIN-D/DARIAH Workshop'Serviceoriented Architectures (SOAs) for the Humanities: Solutions and Impacts', Digital Humanities*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Dat Quoc Nguyen and Karin Verspoor. 2018. An Improved Neural Network Model for Joint POS Tagging and Dependency Parsing. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 81–91, Brussels, Belgium, October. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, and Željko Agić et al. 2018a. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018b. Enhancing Universal Dependency Treebanks: A Case Study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107.

Joakim Nivre. 2005. Dependency grammar and dependency parsing. *MSI report*, 5133(1959):1–32.

Petr Pajas and Jan Štěpánek. 2009. System for querying syntactically annotated corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36. Association for Computational Linguistics.

Adam Przepiórkowski, Zygmunt Krynicki, Lukasz Debowski, Marcin Wolinski, Daniel Janus, and Piotr Banski. 2004. A Search Tool for Corpora with Positional Tagsets and Ambiguities. In *LREC*.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. 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.

Alexander Volokh and Günter Neumann. 2012. Transition-based Dependency Parsing with Efficient Feature Extraction. In *35th German Conference on Artificial Intelligence (KI-2012)*, Saarbrücken, Germany, September.

Christopher M White. 2000. Rapid grammar development and parsing: Constraint dependency grammars with abstract role values. *West Lafayette, Indiana, USA: PhD Thesis, Purdue University*.

Amir Zeldes, Anke Lüdeling, Julia Ritz, and Christian Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora.

How to Parse Low-Resource Languages: Cross-Lingual Parsing, Target Language Annotation, or Both?

Ailsa Meechan-Maddon
Uppsala University
Department of Linguistics and Philology
aime5651@student.uu.se

Joakim Nivre
Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se

Abstract

To develop a parser for a language with no syntactically annotated data, we either have to develop a (small) treebank for the target language or rely on cross-lingual learning or projection, or possibly use some combination of these methods. In this paper, we compare the usefulness of cross-lingual model transfer and target language annotation for three different languages, with varying support from closely related high-resource languages. The results show that annotating even a very small amount of data in the target language is superior to any cross-lingual setup and that accuracy can be further improved by adding training data from related languages in a multilingual model.

1 Introduction

Despite significant advances in natural language processing over several decades, even basic technologies like part-of-speech tagging and syntactic parsing are still available only for a tiny fraction of the languages of the world. This observation has led to an increasing interest in techniques for supporting low-resource languages, typically by making use of data from high-resource languages together with methods for cross-lingual learning or transfer. These techniques include annotation projection (Hwa et al., 2002), model transfer (Zeman and Resnik, 2008; McDonald et al., 2011), treebank translation (Tiedemann et al., 2014), and multilingual parsing models (Duong et al., 2015a; Ammar et al., 2016). Despite the undeniable progress in this line of research, the question always looms large whether it is not more effective to simply annotate a small amount of training data in the target language of interest. Daniel Zeman, one of the inventors of delexicalized transfer parsing, maintains that you can get over 50% accuracy for many languages with just 100 annotated sentences, citing as evidence the results of Ramasamy (2014) for some Indian languages. Further support comes from the study of Garcia et al. (2018), who compares cross-lingual parsing to target language annotation in the context of building a treebank for Galician.

In this paper, we approach this question by comparing three ways of training dependency parsers for low-resource languages: *monolingual* models trained on small amounts of *target* language data; *cross-lingual* models trained only on data from related *support* languages; and *multilingual* models trained on both support and target language data. We perform experiments on three target languages with varying support from related high-resource languages: Faroese (supported by Danish, Norwegian, and Swedish), Upper Sorbian (supported by Czech, Polish, and Slovak), and North Saami (supported by Estonian, Finnish, and Hungarian). Our results show that monolingual models consistently outperform cross-lingual models even with very limited amounts of training data. In addition, there is always a multilingual model that outperforms the best monolingual model. Taken together, these results suggest that the most effective strategy for low-resource parser development may well be to annotate as much data as you can afford in the target language and then add training data from related languages if available.

2 Methodology

To be able to compare monolingual, cross-lingual and multilingual models, we adopt the multilingual parsing approach pioneered by Ammar et al. (2016) and deployed on a large scale by Smith et al. (2018a)

Language	**Treebank**	**Train**	**Dev**	**Test**
Faroese	OFT	4.9k	2.5k	2.5k
Danish	DDT	80k		
Norwegian	Nynorsk	245k		
Swedish	Talbanken	67k		
Upper Sorbian	UFAL	5.8k	2.7k	2.7k
Czech	PDT	300k		
Polish	LFG	105k		
	SZ	63k		
Slovak	SNK	81k		
North Saami	Giella	14.3k	2.5k	10k
Estonian	EDT	288k		
Finnish	FTB	128k		
	TDT	163k		
Hungarian	Szeged	20k		

Table 1: Data sets (UD v2.3) with number of tokens.

in the 2018 CoNLL shared task on universal dependency parsing (Zeman et al., 2018). This approach differs from early work on model transfer, which relied on delexicalized models with part-of-speech tags as pivot features (Zeman and Resnik, 2008; McDonald et al., 2011). Although these models initially gave encouraging results, especially for closely related languages, the results were mostly based on experiments with gold part-of-speech tags, severely overestimating the accuracy achievable under more realistic conditions (Tiedemann, 2015). We instead use lexicalized models, which do not presuppose part-of-speech tagging or any other preprocessing except tokenization for the target language, and instead rely on word, character and language embeddings. Besides being more realistic in a low-resource setting, this is justified by the reduced importance of part-of-speech tagging for neural dependency parsers (Dozat et al., 2017; Ma et al., 2018; Smith et al., 2018b).

2.1 Languages and Treebanks

From Universal Dependencies v2.3 (Nivre et al., 2016; Nivre et al., 2018), we select three language clusters with one low-resource language and three related support languages with larger treebanks: a Scandinavian cluster with Faroese supported by Danish, Norwegian (Nynorsk) and Swedish; a West Slavic cluster with Upper Sorbian supported by Czech, Polish and Slovak; and a Uralic cluster with North Saami supported by Estonian, Finnish and Hungarian. It is worth noting that the support languages are much more closely related to the target language in the Scandinavian and West Slavic clusters than in the Uralic cluster. Table 1 lists the treebanks used for each language and the number of tokens in each data set.

2.2 Parser

We use UUParser v2.3 (de Lhoneux et al., 2017a; Smith et al., 2018a), which is an adaptation of the transition-based parser of Kiperwasser and Goldberg (2016) specifically for multilingual models. The original parsing architecture relies on a BiLSTM to learn representations of tokens in context and a multi-layer perceptron to predict transitions and arc labels based on a few BiLSTM vectors. The multilingually motivated extensions in UUParser include an extended transition system for handling non-projective structures (de Lhoneux et al., 2017b) and a richer representation of input tokens. More specifically, each input token w_i in language l is represented by:

$$\mathbf{x} = e(w) \circ \text{BiLSTM}(ch_{1:m}) \circ e(t)$$

Here $\mathbf{x}$ is the concatenation of a word embedding $e(w)$, a character-based vector $\text{BiLSTM}(ch_{1:m})$ obtained by running a BiLSTM over the characters $ch_{1:m}$ of w, and a treebank embedding $e(t)$ representing the

	Scandinavian				West Slavic					Uralic				
	−Target		+Target			−Target		+Target			−Target		+Target	
	UAS	LAS	UAS	LAS		UAS	LAS	UAS	LAS		UAS	LAS	UAS	LAS
Faroese			78.6	71.1	Upper Sorbian			66.0	58.4	N Saami			66.0	58.6
Dan	45.9	30.6	81.5	74.2	Cze	33.1	23.5	71.5	64.2	Est	22.4	8.5	**68.8**	**60.1**
Nor	47.4	35.7	84.0	76.2	Pol	44.9	34.3	71.3	64.2	Fin	22.5	7.5	67.3	59.5
Swe	45.9	24.9	81.1	73.9	Slo	41.2	29.5	68.6	61.9	Hun	19.4	4.9	65.6	57.5
Dan+Nor	48.5	34.5	83.7	76.7	Cze+Pol	51.1	41.4	72.2	63.9	Est+Fin	24.7	9.4	64.5	55.3
Dan+Swe	55.9	35.9	82.7	75.5	Cze+Slo	44.1	32.6	**72.5**	**65.3**	Est+Hun	23.8	8.9	67.0	58.4
Nor+Swe	56.4	39.6	**83.9**	**77.0**	Pol+Slo	47.5	38.8	72.2	64.9	Fin+Hun	20.8	8.0	65.9	57.7
All	57.7	44.4	82.8	75.3	All	52.4	43.3	69.6	62.8	All	27.1	11.6	65.4	56.4

Table 2: Test set accuracy for target languages (UAS, LAS). −Target = cross-lingual models trained without target language data. +Target = models trained on target language data; monolingual (first row) and multilingual.

treebank *t* that the input comes from. The treebank embedding is used to distinguish data from different languages as well as different treebanks from the same language (Stymne et al., 2018; Smith et al., 2018a). We drop the treebank embedding when training models on a single treebank and otherwise train all models with default settings and no pre-trained embeddings. For cross-lingual and multilingual models, word, character and language embeddings are thus learned jointly for all languages.

2.3 Experimental Setup

Within each cluster, we train cross-lingual models on data from every combination of one, two or three support languages (7 models), multilingual models on the same data sets plus target language data (7 models), and a monolingual model only on target language data, for a total of 15 models. For the support languages, we only use the dedicated training sets from Universal Dependencies v2.3 (Nivre et al., 2018). We do not standardize training set sizes, since the parser has been shown to be robust to size differences when training multi-treebank models (Stymne et al., 2018), but we limit the size of the Czech training set (which is about four times bigger than any other) to 300k tokens. For the target languages, we need a training set for the mono- and multilingual models, a development set to tune hyper-parameters, and a test set for the final evaluation. For Faroese and Upper Sorbian, there is only about 10k tokens of data, which we subdivide into 50% training, 25% development, and 25% test. For North Saami, there is more data, so we leave the dedicated test set of 10k tokens intact and extract a development set of 2.5k tokens from the training set, leaving 14.3k words for training.

The development sets for target languages are used for model selection as follows:

- For support languages with two treebanks of roughly equal size, we run preliminary experiments with cross-lingual models to decide whether to use both or only one. The resulting selection can be seen in Table 1.
- To improve compatibility of character-based representations across (support and target) languages, we try mapping characters that exist only in a target language to characters that exist in one or more support language. This is helpful only for Faroese, where we map {ÍÚýúðí} to {IUyudi}.
- For cross-lingual models, the parser does not learn a language embedding for the target language, so we select a support language to use as proxy during parsing based on LAS on the target language development set.
- All models are trained for 30 epochs, and the best epoch is selected according to LAS on the target language development set.

Finally, the development sets are also used in learning curve experiments (see Section 3).

3 Results and Discussion

Table 2 reports results on the test sets for all cross-lingual, multilingual and monolingual models on our three target languages, both labeled attachment score (LAS) and unlabeled attachment score (UAS). The first thing to note is that the monolingual models, trained only on about 5k tokens for Faroese and Upper Sorbian and 14k tokens for North Saami, consistently outperform all cross-lingual models by a wide margin. The difference is especially large for the Uralic cluster, where the target language is in a different branch of the language family from all support languages, and where the best cross-lingual model does not even reach 10% for LAS (25% for UAS). But even for the Scandinavian and West Slavic clusters, where languages are more closely related and the best cross-lingual models get LAS over 40% and UAS over 50%, the monolingual model gives a higher LAS score by at least 15% absolute (12% absolute for UAS). This indicates that annotating a relatively small amount of training data in the target language is generally superior to using cross-lingual model transfer.

The second main trend is that, despite the poor results for cross-lingual models, the best multilingual models consistently outperform the monolingual models. For the Scandinavian and West Slavic clusters, *all* multilingual models outperform the monolingual model and the best model improves by as much as 5.9/6.9 LAS and 5.3/6.5 UAS. But even for the Uralic cluster, where data from the related support languages seem completely useless in the cross-lingual scenario, using the same data in a multilingual model improves on the monolingual model in 3 out of 7 cases for UAS (2 out of 7 for LAS). The relative improvement for the best multilingual model is smaller than in the other two clusters, but it should be kept in mind that the target language training set is almost three times bigger for North Saami than for Faroese and Upper Sorbian. These results suggest that, even if target language annotation is more effective than cross-lingual transfer, adding data from related support languages can nevertheless lead to further improvements.

To understand why multilingual models work so much better than cross-lingual models, it is important to note that the former learn word, character and language embeddings for the target language and that these embeddings are learned together for all languages. The cross-lingual models have no target language specific representations and have to rely on a proxy language embedding and the existence of cognates for matching word and character representations. This works especially poorly for the Uralic cluster, where the distance from the target to the support languages is much larger.

So how much target language data do we need to outperform a cross-lingual model? To answer this question, we run learning curve experiments for the monolingual and best multilingual models, using the development sets for evaluation, and gradually increasing the amount of target language training data from 0 to 50, 100, 500, 1k, 3k, 5k and 10k tokens (Figure 1). For the Scandinavian and Uralic clusters, we only need 1k tokens for the monolingual model to surpass the cross-lingual model with respect to LAS. For the West Slavic cluster, results are slightly erratic for the smallest training sets, but 3k tokens definitely suffice to reach the accuracy of the best cross-lingual model.[1] In all three cases, this is less than 200 sentences,[2] so the results seem to support Daniel Zeman's claim that something like 100 sentences can be sufficient to train a decent parser, although in our study it is only Faroese that reaches a (labeled) accuracy of 50% with only 100 sentences.

4 Related Work

Work on cross-lingual learning for parsing and related tasks has focused on three main approaches: annotation projection (Hwa et al., 2002; Hwa et al., 2005; Tiedemann, 2014), model transfer (Zeman and Resnik, 2008; McDonald et al., 2011), and (to a lesser extent) treebank translation (Tiedemann et al., 2014). Annotation projection and treebank translation presupposes parallel data, so we will focus on model transfer, which is closest to our work. Model transfer was pioneered for closely related languages by Zeman and Resnik (2008), using delexicalized models and relying on a common part-of-speech tagset for the source and target language. The idea was refined and generalized to multi-source transfer by

[1] If we consider UAS instead of LAS, the patterns are very similar, with the curves crossing at around 1k tokens for the Scandinavian and Uralic clusters and just under 3k tokens for the West Slavic cluster, so we omit these figures to save space.

[2] Upper Sorbian has significantly longer sentences than the other two target languages.

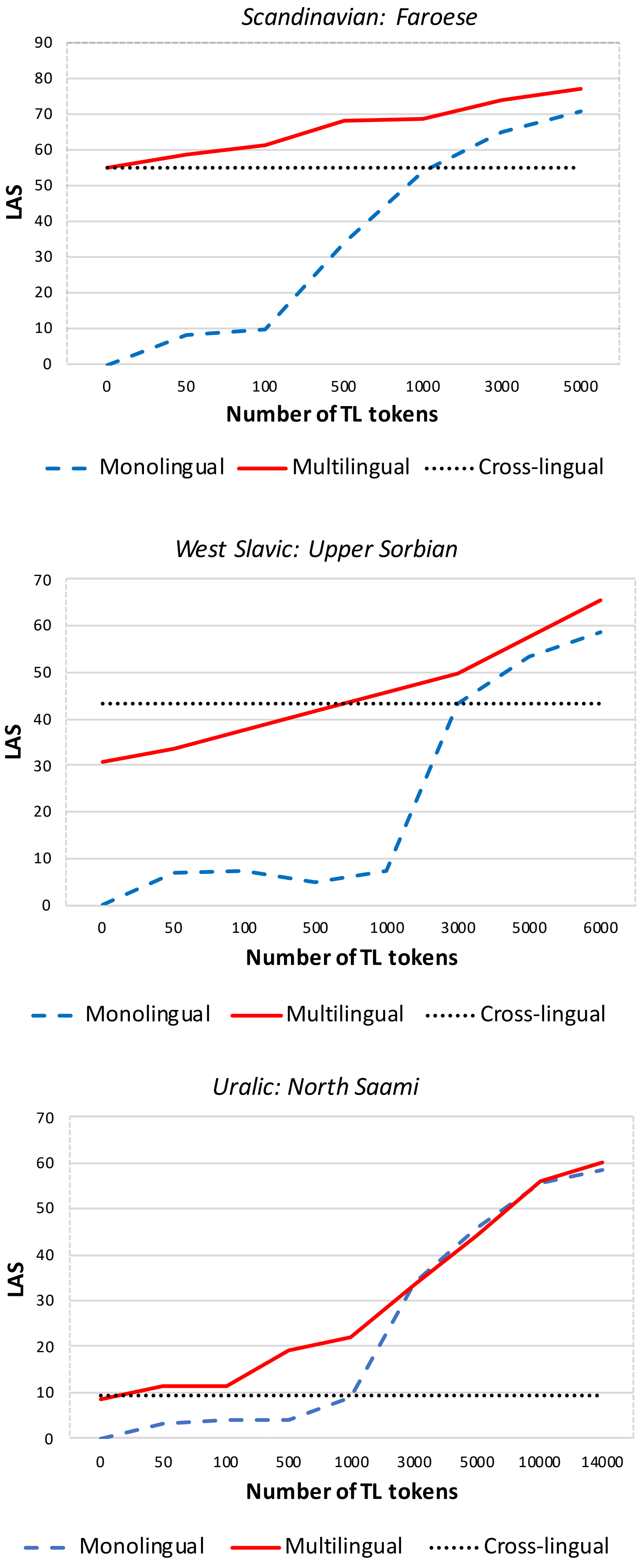


Figure 1: Learning curves (LAS) for the monolingual model (blue dashed line) and the best multilingual model (red solid line), compared to the best cross-lingual model (black dotted line).

McDonald et al. (2011) and gained further momentum with the advent of cross-linguistically consistent syntactic annotation, which facilitated evaluation (McDonald et al., 2013). Other studies concerned methods for selecting optimal source languages (Søgaard and Wulff, 2012; Rosa and Zabokrtsky, 2015). However, most of the early studies of model transfer relied on evaluation with gold part-of-speech tags on the target side, which was later shown to give over-optimistic results (Tiedemann, 2015).

A study of special relevance to our own work is that of Garcia et al. (2018), who specifically study the amount of target language training data needed to outperform cross-lingual model transfer in the context of building a UD treebank for Galician. Drawing on data from 7 other Romance language varieties (Brazilian Portuguese, Catalan, European Portuguese, French, Italian, Romanian and Spanish), they show that a single-source transfer parser achieves LAS corresponding to about 3,000 tokens of target language training data and UAS corresponding to about 7,000 tokens. However, they also show that careful combination and adaptation of source language data from multiple languages can increase these numbers to 16,000 (LAS) and 20,000 (UAS). One difference between their study and ours is that they make use of part-of-speech tags as pivot features, which may explain why especially the adapted multi-source transfer parsers seem to perform better than in our study. In addition, the similarity between Galician and some of the Romance languages is probably greater than in most of our support-target language pairs. Another difference is that Garcia et al. (2018) find that cross-lingual parsers are more competitive with respect to UAS than LAS, whereas we find that about the same number of target language training tokens is needed to reach cross-lingual performance with respect to both metrics. It is possible but by no means obvious that this difference is also related to the presence or absence of part-of-speech tags.

The increasing use of neural networks and distributed representations in syntactic parsing has led to more flexible models for cross-lingual and multilingual learning embeddings that go beyond delexicalized models and their reliance on part-of-speech tags (Duong et al., 2015a; Duong et al., 2015b; Guo et al., 2015a; Guo et al., 2015b). Especially important for our own work is the multilingual model of Ammar et al. (2016) with its use of language embeddings, which were later generalized to treebank embeddings that allow seamless integration of multiple languages as well as heterogeneous treebanks for a single language (de Lhoneux et al., 2017a; Stymne et al., 2018; Smith et al., 2018a). A more recent line of research involves the use of synthetic treebanks (Wang and Eisner, 2016; Wang and Eisner, 2018), an approach recently applied to parser development for one of our target languages, Faroese (Tyers et al., 2018). Finally, it is worth noting that the superiority of annotating target language data over using cross-lingual methods has also been demonstrated for the related part-of-speech tagging problem, in the context of historical text processing, by Schultz and Kuhn (2016) and Schultz and Ketchik (2019).

5 Conclusion

We have compared cross-lingual, multilingual and monolingual parser training for three low-resource languages, supported to different degrees by related languages with more resources. Our main conclusion is that training a monolingual model on target language data gives better performance than any cross-lingual model as soon as we have at least 200 annotated target language sentences. Moreover, adding data from related languages to train a multilingual model can improve performance further by up to 7 LAS points. In conclusion, to develop a parser for a low-resource language, annotate as much data as you can afford and add data from related languages if available.

Acknowledgments

We would like to thank our friends and colleagues, Ali Basirat, Miryam de Lhoneux, Harald Hammarström, Christian Hardmeier, Artur Kulmizev, Paola Merlo and Gongbo Tang, for their invaluable comments on this paper. We also wish to extend our thanks to the anonymous reviewers who provided great insight and suggestions.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017a. From raw text to Universal Dependencies – Look, no tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 207–217.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017b. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 845–850.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. A neural network model for low-resource universal dependency parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 339–348.

Marcos Garcia, Carlos Gómez-Rodríguez, and Miguel A. Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering*, 24:91–122.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015a. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1234–1244.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015b. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2734–2740.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 392–399.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1403–1414.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Kamil Kopacewicz, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 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.

Logathan Ramasamy. 2014. *Parsing Under-Resourced languages: Cross-Lingual Transfer Strategies for Indian Languages*. Ph.D. thesis, Charles University in Prague.

Rudolf Rosa and Zdenek Zabokrtsky. 2015. KLcpos3 – a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–249.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018a. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *Proceedings of the 2018 CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Aaron Smith, Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2018b. An investigation of the interactions between pre-trained word embeddings, character models and pos tags in dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Anders Søgaard and Julie Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1181–1190.

Sara Stymne, Miryam Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 130–140.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1854–1864.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with Universal Dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling)*, pages 340–349.

Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Dingquan Wang and Jason Eisner. 2018. Synthetic data made to order: The case of parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1325–1337.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthtyersast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Word order variation in Mbyá Guaraní

Angelika Kiss
Department of Linguistics
University of Toronto
angelika.kiss@mail.utoronto.ca

Guillaume Thomas
Department of Linguistics
University of Toronto
guillaume.thomas@utoronto.ca

Abstract

This paper presents the preliminary results of a multifactorial analysis of word order in Mbyá Guaraní, a Tupí-Guaraní language spoken in Argentina, Brazil and Paraguay, based on a corpus of written narratives with multiple layers of annotation. Our goals are to assess the validity of previous claims about Mbyá word order (Martins, 2003; Dooley, 1982; Dooley, 2015), and to explore the effects of different types of factors on the position of core arguments relative to their verb. We show that SV and VO are the most frequently attested orders in matrix clauses and that subordinate clauses favour the OV order. Givenness, transitivity and clause type (root vs subordinate) are found to be significant predictors of word order. We identify differences in object position between Mbyá and Paraguayan Guaraní (Tonhauser and Colijn, 2010), and we argue that these differences support Dietrich (2009)'s proposal that Tupí-Guaraní languages are undergoing a change in word order from OV to VO, induced by contact with Spanish and Portuguese.

1 Introduction

This paper presents the preliminary results of a multifactorial analysis of the relative order of subject, object and verb in Mbyá Guaraní, a Tupí-Guaraní language spoken in Argentina, Brazil and Paraguay, which is closely related to Paraguayan Guaraní. To the best of our knowledge, Mbyá word order has only been investigated by Martins (2003), and by Dooley (1982, 2008, 2015). However, these studies do not include detailed reports of word order frequencies, nor do they engage in quantitative modelling of word order variation.

A first goal of the study is to provide statistics that will put the description of word order in Mbyá on a more solid foundation. A second goal is to explore constraints on word order variation in the language through multifactorial techniques. More precisely, we ask what factors affect the position of core arguments relative to their verb, and whether these factors are predominantly syntactic (clause type, grammatical function), discourse-pragmatic (givenness), lexical (animacy, transitivity) or related to processing (argument length). To this end, we annotated a corpus of 1,046 sentences with interlinear glosses, parts of speech tags, syntactic dependency relations and coreference relations, which forms the basis of the present study.

We compare our results to the findings of Tonhauser and Colijn (2010), who investigated subject and object placement in Paraguayan Guaraní. We find notable differences between these two languages, which we interpret in the light of Dietrich (2009)'s analysis of word order change in the Tupí-Guaraní family.

2 Some relevant aspects of Mbyá grammar

Mbyá is a head-marking language. There is no case marking on nouns. Verbs agree in person and number with their core arguments. Intransitive verbs belong to one of two classes, called active and inactive, which use different paradigms of prefixes to cross-reference their subject, as illustrated by the following examples:[1]

(1) a.

Xee	a-	a	ju	ma.
I	A1.SG-	go	again	already

‘*I am already going again.*’ (Dooley 2015)

b.

Xe-	kangy	vaipa.
B1.SG-	feel_weak	very

‘*I feel very weak.*’ (Dooley 2015)

With transitive verbs, the active paradigm is used to cross-reference subjects, and the inactive paradigm is used to cross-reference objects. However, only one argument can be cross-referenced.[2] If both arguments are third person, the subject is cross-referenced. Otherwise, the highest argument on the person hierarchy 1 > 2 > 3 is cross-referenced. In the following example, the verb *xe-r-exa* cross-references its 1^{st} person object. Its implicit subject must be 2^{nd} or 3^{rd} person:

(2)

Xe-	r-	exa.
B1.SG-	R-	see

‘*They/(s)he/you saw me.*’ (Dooley, 2015)

Note that Mbyá is a pro-drop language. All core arguments can be omitted, even if they are not cross-referenced on the verb, as illustrated in example (2) for the subject.

Dooley (1982) reports that SVO is the unmarked order, and that SOV, OSV and OVS orders are also attested. Martins (2003) argues that both SOV and SVO are basic word orders, the latter being more prevalent among younger speakers. However, Martins reports that all six permutations of the subject, verb and object were accepted by native speakers.

(3)

kuee	Maria	o-	jogua	jety	(SVO)
yesterday	Maria	A3	buy	potato	

‘Yesterday Maria bought potatoes’

a. kuee Maria jety o-jogua (SOV)

b. kuee jety o-jogua Maria (OVS)

c. kuee jety Maria o-jogua (OSV)

d. kuee o-jogua jety Maria (VOS)

e. kuee o-jogua Maria jety (VSO) (Martins 2003, p. 154)

Note that Dooley (1982)’s observations are based on his description of Mbyá in the Rio das Cobras community in the Brazilian state of Paraná, while Martins (2003) describes the language spoken in the Morro dos Cavalos and Maciambu communities in the state of Santa Catarina, also in Brazil.

3 Corpus Construction

The corpus used in the present study consists of narratives written between 1976 and 1990 by two Mbyá speakers from the Rio das Cobras community in Paraná, Brazil. These narratives were collected and interlinearized by Robert Dooley. This corpus is available on the Archive of the Indigenous Languages of the America (Dooley, nd).

[1] Glosses: A1.SG: first person singular ‘active’ inflection; B1: first person singular ‘inactive’ inflection; R: linking morpheme.

[2] With the exception of combinations of 1^{st} person subject and 2^{nd} person object, which are cross-referenced with a portmanteau prefix *ro-*.

The 32 narratives used in this study contain 1046 sentences and 1803 tokens. One author, Nelson Florentino, contributed more than 95% of the tokens. The other narratives were written by Darci Pires de Lima.

The corpus was annotated by the authors and research assistants[3]. It contains five layers of annotation: interlinear morphological glosses, parts of speech tags, syntactic dependency relations, coreference annotation and animacy annotation.

Dooley's interlinearization was revised in SIL FieldWorks Language Explorer (Black and Simons, 2008). The interlinearization includes morphological segmentation and glosses, syntactic category annotation using language specific tags, and a free translation into Brazilian Portuguese.

Syntactic annotation was done by the authors in dependency grammar, in the Universal Dependency v2.4 framework (Nivre et al., 2019). Universal POS tags and morphological features were converted automatically from the language specific POS tags and glosses included in the interlinearization layers. Dependency relations were added manually in Arborator (Gerdes, 2013). While the syntactic annotation of Mbyá in Universal Dependencies v2.4 involves a number of non-trivial analytical decisions, the present study only exploits part of the information encoded in the dependency annotation, namely syntactic relations between predicates and their subject and objects, as well as relations of clausal subordination (relative, adverbial and complement clauses). The identification of these relations using Universal Dependency guidelines did not present any particular challenge, and we refer the reader to these guidelines for further information (UD Guidelines, n.d.).

The layer of coreference annotation was created in WebAnno 3 (de Castilho et al., 2016), following Komen (2009)'s annotation guidelines. We understand coreference in a general sense to be a relation between expressions that introduce discourse referents, both referential expressions properly speaking and quantifiers. When a referring expression or a quantifier is used, we call it a *mention* of its discourse referent. Sequences of mentions that have identical or related discourse referents form *referential chains*. Following Bentivoglio (1983), we include implicit mentions of arguments in our referential chains. When an argument is dropped or only expressed in the form of a cross-reference marker on the verb, we consider it an implicit mention and include it in a referential chain. Implicit arguments were annotated by adding null subject and/or object tags on their verb.

A version of the annotated corpus that includes UD dependency annotations is available in a delexicalized form as a part of Universal Dependencies 2.4 (Thomas, 2019). The coreference annotation layer is not yet publicly available at the date of writing of this paper.

4 Data Extraction and Coding Decisions

We exported our corpus to a WebAnno tab-separated file, from which we extracted relevant observations using a Python script. Statistical analysis was performed in R (R Core Team, 2013). Two R data frames were created. In the first one, each observation corresponds to a verb, which is coded for its Transitivity, and for the Word Order of its clause: V (no overt argument), VS, SV, VO, OV, SVO, SOV, OSV, OVS. VOS and VSO orders are unattested in the corpus.

In the second data frame, each observation corresponds to an overt subject or object, which is coded for its position relative to the verb: pre-verbal (XV) or post-verbal (VX). In addition, subjects and objects were coded for several independent variables that have been used in quantitative studies of word order (Prince, 1981; Givón, 1983; Ariel, 1988; Hawkins, 1994; Tonhauser and Colijn, 2010; Heylen, 2005): Animacy (animate/inanimate), Clause Type (root/subordinate), Givenness (new/given), Grammatical Function (subject/object), Length (numeric) and Transitivity of the verb (intransitive/transitive).

We excluded dependent verbs in serial verb constructions, as well as identificational constructions and interrogative clauses. Our counts of subjects and objects only include noun phrases, and excludes clausal arguments.[4] Some coding decisions should be noted:

- *Clause Type*: we coded independent clauses and main clauses of direct reported speech as 'root'. Clausal complements, adverbial clauses and relative clauses were all coded as 'subordinate.'

[3] Gregory Antono, Laurestine Bradford, Vidhyia Elango, Jean-François Juneau, Barbara Peixoto, Darragh Winkelman.

[4] Note that we did analyze word order *within* clausal arguments.

- *Givenness*: mentions that do not have an antecedent in the coreference annotation of our corpus were coded as ‘new’. We coded as ‘given’ all mentions that are related to an antecedent through coreference, bridging anaphora or through a partitive relation.
- *Length*: length was coded as the number of characters making up the relevant mention. Since the orthography used in our corpus makes restricted use of digraphs for simple segments, and the phonology of Mbyá does not contrast long and short vowels, this is a reasonable approximation of the number of phonological segments. In several studies, length is coded as number of words of the mention (Jacennik and Dryer, 1992; Siewierska, 1993; Arnold et al., 2000; Rosenbach, 2005), or number of syllables (Heylen, 2005). There have been proposals for substituting length by different measures of syntactic complexity (e.g. the number of syntactic nodes), but length has been argued to be a good enough predictor of syntactic complexity, at least in English (Wasow, 1997; Szmrecsányi, 2004).

5 Analysis

Table 1 presents counts and proportions of word orders in our data set:

		Clause Type		Transitivity	
		root	sub	vi	vt
Word Order	V	546 (52.3)	497 (47.7)	532 (51.0)	511 (49.0)
	SV	359 (80.0)	90 (20.0)	284 (63.3)	165 (36.7)
	VS	60 (85.7)	10 (14.3)	53 (75.7)	17 (24.3)
	OV	59 (67.8)	28 (32.2)		87 (100.0)
	VO	80 (87.0)	12 (13.0)		92 (100.0)
	SOV	19 (76.0)	6 (24.0)		25 (100.0)
	SVO	33 (94.3)	2 (5.7)		35 (100.0)
	OSV	1 (100.0)	0 (0.0)		1 (100.0)
	OVS	0 (0.0)	1 (100.0)		1 (100.0)
Total		1157	646	869	934

Table 1: Word Order Overview

Out of 1803 clauses, 1043 have no overt subject or object. Subjects are omitted on 58% of verbs, and objects on 74% of transitive verbs. Note that only 62 clauses had both overt subjects and objects, out of 934 transitive clauses. VSO and VOS are unattested in the corpus, and object first orders (OVS/OVS) have only one occurrence each, which shows a tendency for subjects to precede objects.

Table 2 gives an overview of our predictors in the subset of 760 clauses with at least one overt argument, which includes a total of 822 core arguments. The last column reports the p-value of Chi-Square tests for categorical predictors, and of Kruskal-Wallis tests for numeric predictors (Length). Subjects generally precede their verb, while the distribution of objects is more balanced. Animate and given arguments also tend to occur in pre-verbal position. Post-verbal arguments tend to be longer than pre-verbal ones.

Table 3 presents our predictors separately for subject and object positions. We see that animacy and clause type are not significant predictors of subject position, and only clause type and givenness are significant predictors of object position.

In order to explore the combined effects of our predictors on word order, we turn to multifactorial classification models. We fitted conditional inference tree and random forest models to our data set, using the `ctree` function from the `party` package in R (Hothorn, 2019). These models have the advantage of being appropriate for unbalanced designs with multicollinearity (Tagliamonte and Baayen, 2012). We first fit a conditional inference tree to the whole data set, which lets us explore interactions between our predictors. The tree represented in figure 1 includes all splits that are significant at the level of 0.05.

Position		XV (pre-verbal)	VX (post-verbal)	p
Animacy	animate	503 (82.3)	108 (17.7)	<0.001
	inanimate	121 (57.3)	90 (42.7)	
Clause Type	root	491 (73.9)	173 (26.1)	0.007
	sub	133 (84.2)	25 (15.8)	
Givenness	given	533 (83.2)	108 (16.8)	<0.001
	new	91 (50.3)	90 (49.7)	
Grammatical Function	S	510 (87.8)	71 (12.2)	<0.001
	O	114 (47.3)	127 (52.7)	
Length	Mean (SD)	7.5 (3.9)	8.9 (3.6)	<0.001
Transitivity	vi	284 (84.3)	53 (15.7)	<0.001
	vt	340 (70.1)	145 (29.9)	

Table 2: Predictors of Argument Position

		Subjects			Objects		
		XV	VX	p	XV	VX	p
Animacy	animate	470 (88.2)	63 (11.8)	0.326	33 (42.3)	45 (57.7)	0.283
	inanimate	40 (83.3)	8 (16.7)		81 (49.7)	82 (50.3)	
Clause Type	root	412 (87.3)	60 (12.7)	0.452	79 (41.1)	113 (58.9)	<0.001
	sub	98 (89.9)	11 (10.1)		35 (71.4)	14 (28.6)	
Givenness	given	457 (92.0)	40 (8.0)	<0.001	76 (52.8)	68 (47.2)	0.038
	new	53 (63.1)	31 (36.9)		38 (39.2)	59 (60.8)	
Length	Mean (SD)	7.2 (3.6)	8.6 (3.5)	<0.001	9.1 (4.7)	9.0 (3.7)	0.54
Transitivity	vi	284 (84.3)	53 (15.7)	<0.001			
	vt	226 (92.6)	18 (7.4)				

Table 3: Predictors of Argument Position by Grammatical Function

Examination of the conditional inference tree shows that grammatical function is the most important predictor of core argument placement. We also observe a complex interaction between grammatical function, givenness and transitivity. While subjects tend to be preverbal, new subjects of intransitive verbs are more likely to be post-verbal than other subjects. Grammatical function also interacts with clause type, objects being more likely to be pre-verbal in subordinate than in root clauses.

In order to obtain a more robust assessment of the importance of each variable in predicting word order, we fit a random forest model of 1000 trees to our data set, with three variables available for splitting at each node (mtry = 3). Each tree in the forest is built on a random sample of the data set, which serves as a learning-sample for this tree. Some observations, the out-of-bag observations, are held off and used as a built-in test sample for the tree. The prediction accuracy of each tree is calculated on its associated out-of-bag sample (Strobl et al., 2009). The model has an out-of-bag accuracy of 79.8%. Table 4 shows a confusion matrix for the model.

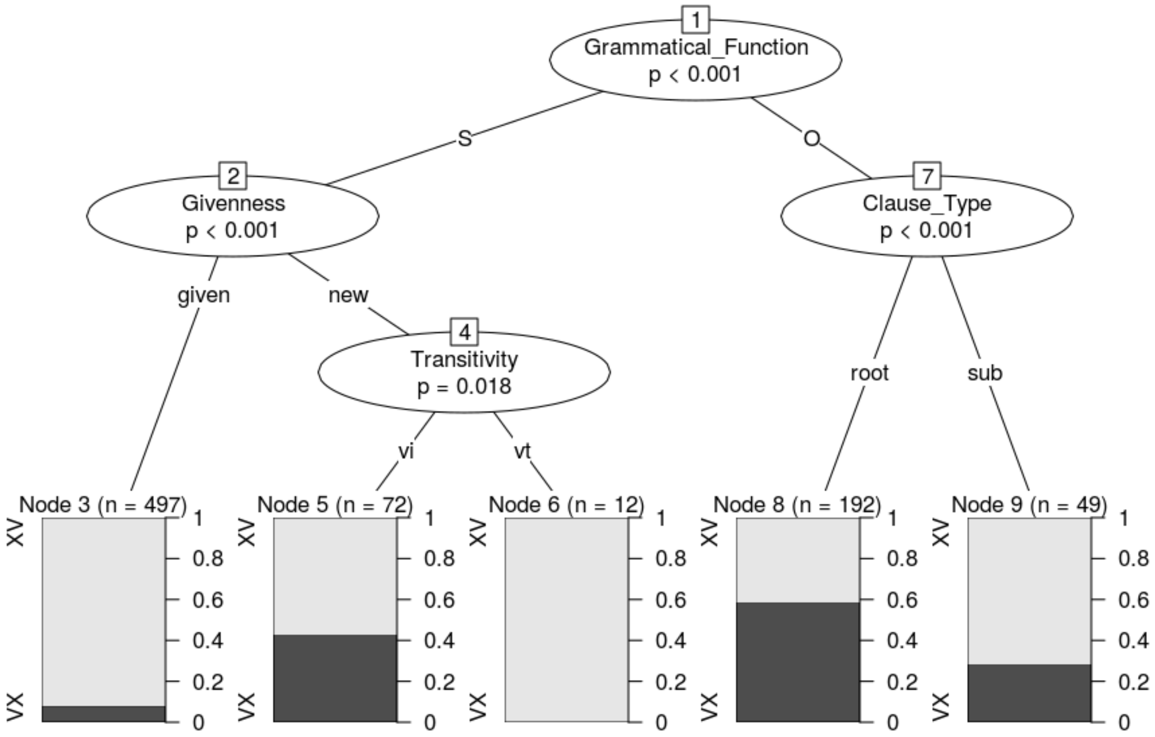


Figure 1: Conditional Inference Tree model of Argument Position

	Predicted: XV	Predicted: VX
Observed: XV	556	68
Observed: VX	98	100

Table 4: Observed values and predictions of the random forest.

Table 5 shows the conditional variable importance (Strobl et al., 2008) for all predictors in our random forest. We see that grammatical function is by far the most important predictor, followed by givenness and clause type. The least important predictors are animacy, length and transitivity. These results are consistent with the conditional inference tree presented in figure 1, where transitivity was only selected to split the class of new subjects.

Transitivity	Length	Animacy	Clause Type	Givenness	Grammatical Function
0.00264	0.00477	0.00703	0.01638	0.02254	0.10165

Table 5: Variable Importance in the Random Forest.

The conclusions drawn from the recursive partitioning models are supported by a logistic regression model, which we report in table 6. Again, we observe that grammatical function is the most important predictor, followed by givenness, clause type and transitivity. Animacy and length are not significant predictors in that model.

6 Discussion

We found that while subjects are mostly preverbal in Mbyá (87.8% of all subjects in the corpus), the position of objects is more variable, with 47.3% of pre-verbal objects and 52.7% of post-verbal objects.

	Intercept	Length	Animacy (inanimate)	Transitivity (transitive)	Cl. Type (sub.)	Givenness (new)	Gram. Funct. (object)
Coef.	-1.86	0.01	-0.37	0.69	-0.81	1.15	2.58
S.E.	0.25	0.02	0.25	0.30	0.27	0.21	0.33
Z	-7.31	0.21	-1.45	-2.34	-3.05	5.44	7.78
p	<0.0001	0.8349	0.1460	0.0190	0.0023	<0.0001	<0.0001

Table 6: Logistic Regression model of Argument Placement (reference level: pre-verbal).

Given arguments are more likely to be pre-verbal, in keeping with proposals that old information tend to precede new information across languages (Clark and Clark, 1977; Siewierska, 1993). In addition, givenness interacts with transitivity in the placement of subjects, new intransitive subjects being more likely to follow the verb than transitive ones. Objects are more likely to be pre-verbal in subordinate than in root clauses.

Our results support Martins (2003)'s observation that both (S)OV and (S)VO orders are frequently attested in Mbyá, when both arguments are expressed. At the same time, we also found support for Dooley (2015)'s claim that the (S)OV order is more frequent in subordinate clauses.

It is interesting to compare constraints on word order in Mbyá with those that Tonhauser and Colijn (2010) observed for Paraguayan Guaraní. Note that Tonhauser and Colijn (2010) only investigated word order in matrix clauses. While 87.3% of subjects are pre-verbal in matrix clauses in our corpus, Tonhauser and Colijn (2010) found that matrix subjects exhibit a greater variability in Paraguayan Guaraní, with only 55% of subjects occurring in pre-verbal position. By contrast, the distribution of objects was found to be less variable in Paraguayan Guaraní, with 95% of direct objects occurring post-verbally compared to Mbyá matrix clauses where 41.1% of the objects are preverbal.

The differences we observed between Mbyá and Paraguayan Guaraní object placement support Dietrich (2009)'s analysis of word order change in Tupí-Guaraní languages. Dietrich argues that Tupí-Guaraní languages are undergoing a change from OV to VO order due in part to contact with Spanish and Portuguese. Of all Tupí-Guaraní languages, Paraguayan Guaraní has had the most sustained contact with Spanish and Portuguese (Melia, 2003), and is also argued to be the language with the most prevalent VO order. Because Mbyá has undergone less contact with Spanish or Portuguese, we expect that OV order will be more frequent overall. Dietrich's hypothesis is also supported by the greater frequency of OV order in subordinate clauses in Mbyá. Since subordinate clauses tend to be more conservative than root clauses (Givón, 1979; Hock, 1986; Bybee, 2002), the lesser frequency of VO order in this environment supports the view that this feature is an innovation in the language.

7 Conclusion

Our study confirmed previous descriptions of word order variation in Mbyá (Martins, 2003; Dooley, 1982; Dooley, 2015). It was found that the position of core arguments relative to the verb is affected by a combination of factors, which are syntactic (clause type, grammatical function), discourse-pragmatic (givenness) and lexical (verb type). The different frequencies of OV order in Mbyá and Paraguayan Guaraní might be explained by an ongoing change from OV to VO in Tupí-Guaraní languages due to contact with Spanish and Portuguese, which has been more intense in the case of Paraguayan Guaraní.

References

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity. *Language*, 76(1):28–55.

Paola Bentivoglio. 1983. Topic Continuity in Spoken Latin-American Spanish. In Talmy Givón, editor, *Topic Continuity in Discourse*, pages 255–312. John Benjamins, Amsterdam/Philadelphia.

Andrew Black and Gary Simons. 2008. The SIL Fieldworks Language Explorer Approach to Morphological Parsing. In *Computational Linguistics for Less Studied Languages: Texas Linguistics Society, 10*, pages 37–55. CSLI Publications.

Joan Bybee. 2002. Main clauses are innovative, subordinate clauses are conservative. Consequences for the nature of constructions Joan L. Bybee and Michael Noonan, editors. *Complex sentences in grammar and discourse: essays in honor of Sandra A. Thompson*, pages 255–312. John Benjamins, Amsterdam/Philadelphia.

H. H. Clark and E. V. Clark. 1977. *Psychology and Language: An Introduction to Psycholinguistics*. Harcourt Brace Jovanovich, New York.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities LT4DH*, pages 11–17, Osaka, Japan. https://webanno.github.io/webanno/.

Wolf Dietrich. 2009. Cambio del orden de palabras en lenguas tupí-guaraníes [Word order change in Tupi-Guarani languages]. *Cadernos de Etnolingüística*, 1:1–11.

Robert A. Dooley. 1982. Options in the pragmatic structuring of Guaraní sentences. *Language*, 58:307–31.

Robert A. Dooley. 2008. Pronouns and topicalization in Guarani texts. Associação Internacional de Lingüística - SIL Brasil, Cuiabá MT.

Robert A. Dooley. 2015. Léxico guarani, dialeto mbyá. Summer Institute of Linguistics.

Robert A. Dooley. n.d. Mbyá Guaraní collection of Robert Dooley. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: text. Access: 100

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97.

Talmy Givón. 1979. *On understanding grammar*. New York: Academic Press.

Talmy Givón (editor). 1983. *Topic Continuity in Discourse. A quantitative cross-language study*. John Benjamins, Amsterdam/Philadelphia.

Talmy Givón. 1988. The pragmatics of word-order: predictability, importance and attention. In Michael Hammond, Edith A. Moravcsik, and Jessica R. Wirth, editors, *Studies in Syntactic Typology*, pages 243–285. John Benjamins, Amsterdam/Philadelphia.

John Hawkins. 1994. *A performance theory of order and consituency*. Cambridge University Press, Cambridge, Massachusetts.

Kris Heylen. 2005. A quantitative corpus study of German word order variation. In St. Kepser and M. Reis, editors, *Linguistic evidence: Empirical, theoretical and computational perspectives*, pages 241–264. Mouton de Gruyter, Berlin & New York.

Hans H. Hock. 1986. *Principles of historical linguistics*. Berlin/New York: Mouton deGruyter.

Torsten Hothorn, Kurt Hornik, Carolin Strobl and Achim Zeileis. 2019. *party: A laboratory for recursive part(y)itioning (R package version 1.3–3).* https://cran.r-project.org/web/packages/party/

Barbara Jacennik and Matthew S. Dryer. 1992. Verb-subject order in Polish. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 209–242. John Benjamins, Amsterdam/Philadelphia.

Erwin R. Komen. 2009. Coreference Annotation Guidelines. http://repository.ubn.ru.nl/bitstream/handle/2066 /78810/78810.pdf.

Marci Fileti Martins. 2003. *Descrição e análise de aspectos de gramática do guarani mbyá [Description and analysis of some grammatical aspects of Guaraní Mbyá].* Ph.D. thesis, State University of Campinas.

Bartomeu Meliá 2003. *La Lengua Guaraní del Paraguay*. Asunción: CEPAG.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Joakim Nivre, Mitchell Abrams, and Agić Željko 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.

Ellen F. Prince. 1981. Toward a taxonomy of given–new information. In Cole, Peter, editors, *Radical Pragmatics*, pages 223–255. New York: Academic Press.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/.

Anette Rosenbach. 2005. Animacy Versus Weight as Determinants of Grammatical Variation in English. *Language*, 81:613–644.

Anna Siewierska. 1993. Syntactic weight vs information structure and word order variation in Polish. *Journal of Linguistics*, 29:233–265.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9: 307 https://doi.org/10.1186/1471-2105-9-307

Carolin Strobl, James Malley and Gerhard Tutz. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, 14(4):323–348.

Benedikt Szmrecsányi. 2004. On Operationalizing Syntactic Complexity. In G. Purnelle, C. Fairon, and A. Dister, editors, *Le poids des mots. 7es Journées internationales d'Analyse statistique des Données Textuelles*, pages 1031–1039, Louvain-la-Neuve. Presses universitaires de Louvain.

Sali A. Tagliamonte and R. Harald Baayen 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice *Language Variation and Change*, 24:135–178.

Guillaume Thomas. 2019. UD Mbya_Guarani_Dooley, Mbyá Guaraní treebank based on narratives collected by Robert Dooley. In Nivre et al. 2019.

Judith Tonhauser and Erika Colijn. 2010. Word order in Paraguayan Guaraní. *International Journal of American Linguistics*, 76:255–288.

Universal Dependencies n.d. Universal Dependencies Guidelines.
https://universaldependencies.org/guidelines.html

Thomas Wasow. 1997. Remarks on grammatical weight. *Language Variation and Change*, 9:81–105.

Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data

Saeko Komori
Chubu University / JAPAN
komori@isc.chubu.ac.jp

Masatoshi Sugiura
Nagoya University / JAPAN
sugiura@nagoya-u.jp

Wenping Li
Dalian Maritime University / CHINA
lwplovely1023@gmail.com

Abstract

The purpose of this study is to examine methods of measuring syntactic complexity by analyzing an original corpus of written Japanese data from native speakers and learners of Japanese. We compared two measures, mean dependency distance (MDD) and mean hierarchical distance (MHD), which have been examined using in English in previous studies. Our research question is to compare the two methods and evaluate them in order to develop an index for measuring Japanese learner's syntactic complexity.

1 Introduction

Ortega (2015) overviewed recent SLA writing and syntactic complexity studies and discussed the reasons for inconclusive results among the studies. She observed that there are some factors that might affect differences in results across studies. One of them is a factor of measurements, and three measurements were discussed: 1) Subordination measures, 2) Length-based measures, and 3) Frequency-based measures. We believe that this factor needs to be studied further, and more precise indexes are necessary to measure syntactic complexity. This paper will examine mean dependency distance (MDD) and mean hierarchical distance (MHD) as good candidates for measuring L2 development of Japanese syntactic complexity.

2 Previous Studies on MDD and MHD

We will first review five studies using MDD and MHD as measures for syntactic complexity. Three of them were studies using native speaker (NS) data, and two used non-native speaker (NNS) data as summarized in Table 1.

Study	MDD/MHD	Language	NS/NNS
Jing and Liu (2015)	MDD and MHD	English and Czech	NS
Jing and Liu (2016)	MHD and other measures	English	NS
Liu et al. (2017)	MDD	20 natural languages	NS
Ouyang and Jiang (2017)	MDD	English	NNS
Komori et al. (2018, 2019)	MDD and MHD	Japanese	NNS

Table 1: Summary of previous studies of the MDD and MHD

First, Jing and Liu (2015) studied both MDD and MHD using English and Czech as the first language. In order to examine the structural complexity of language, they compared two SVO languages: English with rigid word order and Czech with relatively free word order. They reported significant positive correlations between sentence lengths (SL), MDD, and MHD. They also discovered that “for longer sentences, English prefers to increase the MDD, while Czech tends to enhance the MHD” (Jing and Liu 2015, 161).

Second, the purpose of Jing and Liu (2016) was to analyze the hierarchical structure of English sentences, and they examined several different measures, including the MHD using a large English dependency treebank. As a result, they found significant positive correlations between the Vertices number (VN), the Hierarchical number (HN) and the MHD.

Third, Liu et al. (2017) was a cross-language examination of the MDD using 20 natural languages. They posited that dependency distance minimization is probably a universal regularity in human languages (Liu et al. 2017, 176).

Fourth, Ouyang and Jiang (2017) adopted the same calculation method as Liu et al. (2017) in order to examine if the MDD works as a measure of the language proficiency of second language learners. They conducted a study using Chinese EFL learners' compositions in eight grades from the first year of junior high school to the second year of university and reported the MDD increase from 1.845 in the first year of junior high school to 2.466 in the second year of university (Jiang and Ouyang 2017, 210). This results showed that the MDD could indicate the syntactic complexity of the learners' English. Jiang and Ouyang (2017) reported that the MDD measured sentence difficulty and how the MDD changed with the increase of learners' language proficiency across their learning levels.

Lastly, Komori et al. (2018 and 2019) examined the MDD and MHD with Chinese L1 learners of Japanese using Yokohama National University corpus (YNU, Kanazawa, ed., 2014). The learners in the YNU were all advanced learners, and were further divided into three levels: high (H), mid (M), low (L). As a result, there was not a significant difference in the MDD among the three levels of advanced learners. A gradual increase from L to H in the MHD, on the other hand, was found as their levels progressed as shown in Table 2.

Group	MDD	MHD	Words	Number of Sentences
L	2.16	1.75	8,806	1,316
M	2.08	1.84	10,525	1,523
H	2.16	1.98	10,810	1,391
NS	2.07	1.97	9,022	1,209

Table 2: MDD and MHD scores of YNU data

Komori et al. (2018 and 2019) examined advanced learners' syntactic complexity using the MDD and the MHD, but they examined only advanced learners. It is still unclear if the MDD and the MHD can measure language proficiency or language development. Therefore, in this study, we will examine if we can use the MDD and the MHD in order to measure Japanese learners' syntactic complexity using intermediate learners' corpus data. We also see if there are any differences between the two measures of the MDD and the MHD with intermediate learners' data to figure out what kind of differences the MDD and the MHD are measuring.

3 The Current Study

In order to examine the MDD and the MHD as syntactic complexity measures with Japanese learners, we collected our original written data from both learners and native speakers of Japanese. The following will describe the methods and materials of this study.

3.1 Participants

We started the data collection in 2018 with the aim to analyze learners' syntactic development. We collected written data and observed their development over time as their learning progressed. We asked each participant to write an argumentative essay on a manuscript paper of more than 600 characters without referring to any dictionaries. For native speakers, there was a time limit of 30 minutes, but the learners had 50 minutes to write an essay. The university students who participated in this project were the second (C2) and third-year (C3) university students. They were all Chinese native speakers majoring in Japanese in China. We analyzed the data from the intermediate level learners as well as Japanese native speakers (JP). For this particular study, there are 38 C2, 33 C3, and 35 JP compositions for comparison.

3.2 Corpus Data

We manually input each hand-written composition into the computer to compile corpus data. Table 3 shows the outline of the current corpus data. The topic of the composition used for the current study is "Will you decide your plans for life after graduation by yourself or will you consult other people?" which was in Japanese.

Group	Participants	Sentences	Type	Token
C2 (second year university learners)	38	721	1,269	10,296
C3 (third year university learners)	33	605	1,519	11,786
JP (Japanese university students)	35	463	1,462	12,495

Table 3: Outline of the current corpus data

After the data collection, we excluded outlier sentences with less than 4 words and also more than the number of the upper limit, which is upper quartile plus 1.5 interquartile range of the data in each group. As a result, we eliminated 129 (18%), 56(9%), and 34 (7%) of C2, C3, and JP outliers from the data, respectively.

3.3 Analysis

To parse the data, we formatted each composition to one sentence per line. Then, each sentence was parsed syntactically with Cabocha, a Japanese dependency structure analyzer (Kudo and Matsumoto, 2002) and IPADic, and the data was edited by retrieving dependent ID, governor ID and the original word as illustrated in Table 4. After editing, we used the dependent ID and governor ID to calculate the dependency distance (DD), the difference between governor ID and dependent ID. Then, we used the following two formulas (1) and (2) to calculate the MDD of a sentence or text, according to Liu et al. (2017). Finally, we used the dependent ID and governor ID to construct dependency trees and calculated the MHD for each sentence with Python scripts, as shown in Figure 1.

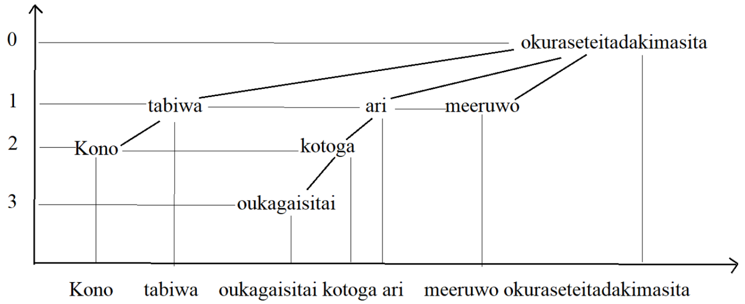
Dependent		Governor	Dependent ID	Governor ID	DD	HD
Kono	=>	tabiwa	0	1	1	2
tabiwa	=>	okuraseteitadakimasita	1	6	5	1
oukagaisitai	=>	kotoga	2	3	1	3
kotoga	=>	ari	3	4	1	2
ari	=>	okuraseteitadakimasita	4	6	2	1
meeruwo	=>	okuraseteitadakimasita	5	6	1	1

Table 4: Method of calculating DD and HD

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

$$MDD(\text{the text}) = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \qquad (2)$$

In formula (1), n is the number of words in the sentence, and DD_i is the DD of the i-th syntactic link of the sentence. In formula (2), n is the total number of words in the text, s is the total number of sentences in the text.



HD = 2 + 1 + 3 + 2 + 1 + 1

MHD = HD / (V - 1)

= 10 / 6

= 1.67

Figure 1: MHD calculation

3.4 Results

Our analysis shows that both the MDD and the MHD increased from C2 to C3 as is shown in Table 5. This means that the increase may reflect their syntactic complexity development as their Japanese learning progressed.

		Median		
	Number of	SL	MDD	MHD
Group	Sentences	(Min, Max)	(Min, Max)	(Min, Max)
C2	592	6	1.91	1.67
		(4, 4)	(1.00, 4.00)	(1.00, 4.00)
C3	547	8	2.00	2.00
		(4, 18)	(1.00, 4.21)	(1.00, 4.64)
JP	429	10	2.00	2.50
		(4, 24)	(1.00, 3.96)	(1.00, 8.17)

Table 5: SL, MDD and MHD comparison of C2, C3 and JP

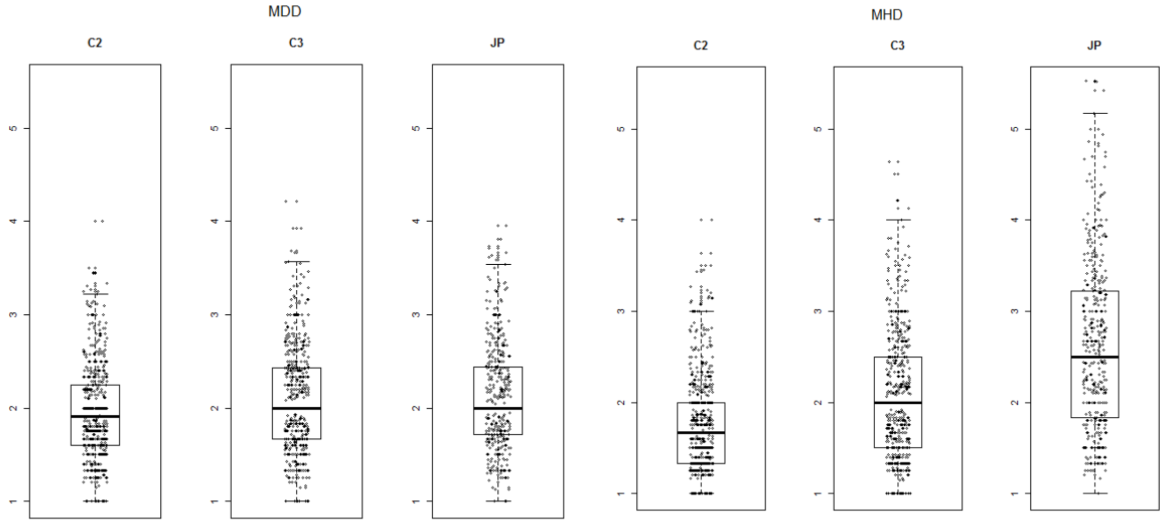


Figure 2: Boxplots with jitter of the MDD and the MHD for C2, C3 and JP

Figure 2 shows the boxplots of the MDD (on the left) and MHD (on the right). It is easy to see a gradual increase of score from C2 to C3 to JP for the MHD. Non-parametric statistical analyses of multiple comparisons were conducted. Table 6 shows Brunner-Munzel (BM) Test results as well as effect sizes (Cliff's delta).

	MDD			MHD		
	BM	*p*	Cliff's delta	BM	*p*	Cliff's delta
C2 v C3	3.88	.0001	.13 (negligible)	7.73	<.0001	.25 (small)
C3 v JP	1.04	.2988	.04 (negligible)	10.26	<.0001	.35 (medium)
C2 v JP	4.86	<.0001	.17 (small)	19.22	<.0001	.56 (large)

Table 6: Brunner-Munzel Test and Cliff's delta of the MDD and MHD

The results of the analyses along with the interpretation of effect sizes indicated that the MHD scores demonstrated significant group differences but the MDD scores did not. There was only a small difference between C2 and JP, but no other significant group differences were observed in the MDD scores. As for the MHD, on the other hand, significant increases can be observed. From our current data, we may conclude that intermediate Japanese learners' syntactic complexity increased in terms of the MHD, but it is difficult to conclude that the MDD showed any increase.

Figure 3 shows correlations between sentence length (SL), MDD, and MHD. They are all significantly correlated (*p* < 0.01). The correlation coefficients between SL and MHD in JP are highest (0.72), and those in C3 and C2 are also moderate (0.67 and 0.62). Correlations between MDD and MHD are not observed in any of the three groups. It can be interpreted that both MDD and MHD are measuring syntactic complexity, but they do not measure the same complexity. Further study is necessary to uncover what the differences are between the syntactic complexities measured by the MDD and the MHD.

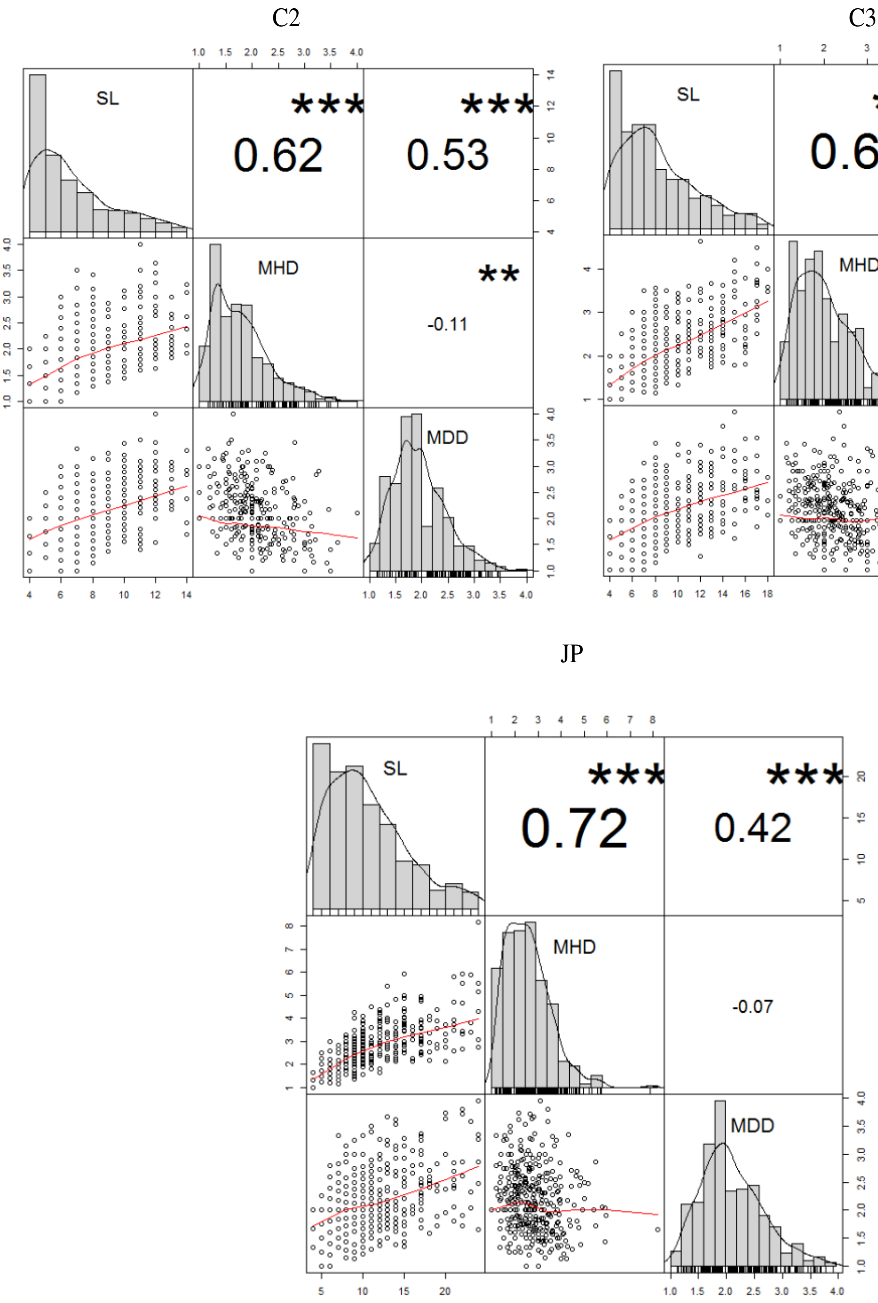


Figure 3: Correlations between SL, MDD and MHD of C2, C3 and JP

4 Discussion

From our data analyses of the intermediate learners and native speakers of Japanese, we showed that Japanese learners' syntactic complexity can be measured with the MHD, but it is not as clear with the MDD. As for the learners' proficiency levels, learners in C2 and C3 of the current study were intermediate learners who studied Japanese for about 13 months

(C2) and 24 months (C3) in China, whereas participants in YNU data in Komori et al. (2018, 2019) were all living in Japan and had studied 20 months to16 years. The MDD from the YNU learners did not show any increase, which may indicate that they might have reached a plateau period. The MDD scores of the intermediate learners in this current study show some increase between groups (C2 and C3), but it is not statistically significant and its effect size is negligible, thus MDD may not denote learners' syntactic development. As for the MHD, the previous study also showed an increase even among advanced learners. In this respect, the MHD might be a better measure to show Japanese learners' syntactic development for both intermediate and advanced learners (Komori, et al., 2019). There may be some linguistic preferences between the MDD and the MHD in Japanese, as is discussed in Jing and Liu (2015) with English and Czech for longer sentences. It may also be argued that some of the characteristics of Japanese syntactic complexity appeared with MHD rather than MDD. As for the composition in terms of genre, the current study used argumentative essays which may contain relatively longer sentences, while the data in YNU consist of 12 different topics and they include short email messages as well (Kanazawa ed. 2014). These two factors (level of learners and genre) may have influenced the results, which we need to control in future studies.

As we have seen above, the MHD may be used to measure learners' syntactic development, but we need to further scrutinize and define the MDD and the MHD as syntactic complexity measures. There are also some problems to be solved in future studies. First of all, the learners' compositions contain errors, and they may cause analytical errors of syntactic complexity. There is also a matter of genre. We only analyzed one topic of compositions in the current study. We are planning to collect compositions with several different topics. Finally, a longitudinal study is necessary to examine the learners' development over time.

Acknowledgements

This study is supported by JSPS KAKENHI Grant Number JP19K00749 and JSPS KAKENHI Grant Number 16H03444.

References

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.

Jinghui Ouyang and Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguisitics,* October 2017, 1-20.

Jingyang Jiang and Jinghui Ouyang. 2017. Dependency distance: A new perspective on the syntactic development in second language acquisition Comment on "Dependency distance: A new perspective on syntactic patterns in natural languages" by Haitao Liu et al. *Physics of Life Reviews* 21, 209-210.

Hiroyuki Kanazawa ed. 2014 *Nihongo kyoiku no tame no tasuku betsu kakikotoba kopasu* (Corpus of task-based writing for Japanese language education), Hitsuji, Tokyo.

Lourdes Ortega. 2015. Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing,* 29, 82–94.

Saeko Komori, Masatoshi Sugiura and Wenping Li. 2018. Examining the applicability of the mean dependency distance (MDD) for SLA:A case study of Chinese learners of Japanese as a second language. *Proceedings of the 4th Asia Pacific Corpus Linguistic Conference (APCLC 2018),* 237-239.

Saeko Komori, Masatoshi Sugiura and Wenping Li. 2019. Evaluating mean dependency distance (MDD) and mean Hierarchical distance (MHD) to measure development of Japanese syntactic complexity. *The 2019 conference of the American Association for Applied Linguistics (AAAL).*

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking*, CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002*, 63-69.

Yingqi Jing and Haitao Liu. 2015. Mean Hierarchical Distance Augmenting Mean Dependency Distance. *Proceedings of the Third International Conference on Dependency Linguistics,* 161-170.

Yingqi Jing and Haitao Liu. 2016. A quantitative analysis of English hierarchical structure. *Journal of Foreign Languages,* 39, 2-11.

Dependency Parsing as Sequence Labeling with Head-Based Encoding and Multi-Task Learning

Ophélie Lacroix
Siteimprove
Sankt Annæ Plads 28
DK-1250 Copenhagen, Denmark
ola@siteimprove.com

Abstract

Dependency parsing as sequence labeling has recently proved to be a relevant alternative to the traditional transition- and graph-based approaches. It offers a good trade-off between parsing accuracy and speed. However, recent work on dependency parsing as sequence labeling ignore the pre-processing time of Part-of-Speech tagging – which is required for this task – in the evaluation of speed while other studies showed that Part-of-Speech tags are not essential to achieve state-of-the-art parsing scores. In this paper, we compare the accuracy and speed of *shared* and *stacked* multi-task learning strategies – as well as a strategy that combines both – to learn Part-of-Speech tagging and dependency parsing in a single sequence labeling pipeline. In addition, we propose an alternative encoding of the dependencies as labels which does not use Part-of-Speech tags and improves dependency parsing accuracy for most of the languages we evaluate.

1 Introduction

Traditional dependency parsers are transition based (Kuhlmann et al., 2011) or graph based (McDonald, 2006). In contrast to previous studies, Strzyz et al. (2019) recently showed that dependency parsing reframed as a sequence labeling problem is also a competitive strategy. The idea is, for a given token in a sentence, to encode into a single tag the information about which token is its parent in the dependency tree (and the label of the incoming dependency). These tags can be predicted in a sequence labeling process and then be decoded in order to rebuild the dependency tree. Strzyz et al. (2019) compare the performance of dependency parsing as sequence labeling using several encodings of the dependencies which have been presented in previous work[1] and show that the best encoding leads to state-of-the-art performance.

One of the main arguments for performing dependency parsing as sequence labeling is to achieve a good speed-accuracy tradeoff (leveraging the efficiency of deep learning frameworks running on GPUs). However, the encoding that is reported as the best one in (Strzyz et al., 2019), requires Part-of-Speech (PoS) tags to encode and decode the dependencies. The method thus involves a pre-processing step of PoS-tagging which is not considered in the evaluation of the parsing speed, whereas previous studies (Ballesteros et al., 2015; de Lhoneux et al., 2017a) showed that PoS-tagging is not a requirement for neural transition-based parsers – using word embeddings as input – in order to achieve state-of-the-art performance. In this work, we set up a single pipeline that performs both PoS-tagging and dependency parsing in order to study the performance of several architectures. We compare the *shared* (Søgaard and Goldberg, 2016) and *stacked* (Hashimoto et al., 2017) multi-task learning strategies to a strategy that combines both, with the aim of identifying a proper trade-off between parsing accuracy and speed.

We also present an alternative encoding that does not use PoS-tags to encode the dependencies. It, however, requires an additional step of *head* tagging which consists of predicting which tokens in a sentence are parents of other tokens (i.e., have dependents in the dependency tree). Hence, the following task of dependency parsing consists of predicting to which of these parents the tokens are attached. We use a similar encoding as in Strzyz et al. (2019). This new encoding aims at reducing the complexity

[1] Such as: the relative positional encoding (Li et al., 2018; Kiperwasser and Ballesteros, 2018), the relative PoS-based encoding (Spoustová and Spousta, 2010) and the bracketing-based encoding (Yli-Jyrä and Gómez-Rodríguez, 2017).

of the attachment step correcting some of the flaws of the original PoS-based encoding. We finally evaluate whether ablating PoS-tagging in the pipeline using the new encoding affects dependency parsing performance.

Contribution We (i) combine two multi-task learning strategies to set up an efficient pipeline for PoS-tagging and dependency parsing as sequence labeling and (ii) propose a new encoding of the dependencies as labels that does not use PoS-tags.

2 Sequence Labeling Pipeline

We propose to perform several sequence labeling tasks, such as PoS-tagging and dependency parsing, in a neural network pipeline architecture which combines *shared* and *stacked* strategies for multi-task learning.

In the *shared* multi-task learning architecture of Søgaard and Goldberg (2016), several tasks are trained simultaneously through the same layers (they share parameters). A single input is given to the network but it feeds different outputs. While Hashimoto et al. (2017) propose a *stacked* multi-task learning architecture, in which each layer is dedicated to the training of one task and layers are stacked on top of each other in a specific order. The calculated output of the final layer dedicated to a given task is concatenated with the input sequence of the network and then feeds the first layer dedicated to the next task.

In our architecture, we combine the two strategies in order to benefit from the strength of both. We define groups of tasks to train sequentially. In a given group, tasks are trained simultaneously using the *shared* multi-task learning strategy (multiple layers can be stacked for one group). They share the same input and feed different outputs. The outputs of the final layer of each group are concatenated with the input sequence to feed the first layer dedicated to the training of the next group of tasks. We name it the *combined* strategy.

In all strategies, each layer is a bi-LSTM (Graves and Schmidhuber, 2005). The input sequence of the network is a concatenation of word embeddings (pre-trained) and character embeddings (trained using an additional bi-LSTM layer). The outputs are calculated through a Softmax layer.

3 Dependency Encodings

Strzyz et al. (2019) observe that the *relative PoS-based* encoding of the dependencies inspired by Spoustová and Spousta (2010) outperforms other encodings. Given a sentence $w_1 \dots w_n$ and its respective sequence of PoS-tags $p_1 \dots p_n$, an incoming dependency to a token w_j, such as w_i is its parent (i.e., $w_i \rightarrow w_j$), is described as a tuple of:

- the PoS-tag p_i of its parent w_i, and
- the relative position n of p_i to w_j with respect to the PoS-tags of the same value p, i.e., p_i is the nth PoS-tag of value p to the right (if $n > 0$) or to the left (if $n < 0$) of w_j.[2]

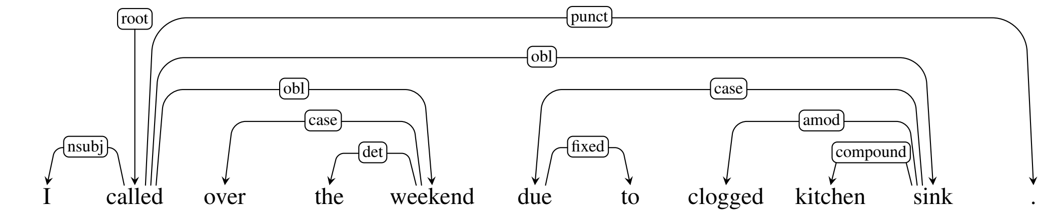
See the RPT tags in Figure 1 as an example of *relative PoS-based* encoding. Note that, in contrast to Strzyz et al. (2019) who predict the relation and the encoding of a dependency as one concatenated tag, in this work, we predict the dependency relations (labels) independently from the dependencies (attachment), as it has been applied to constituent parsing as sequence labelling (vil, 2019). This approach reduces the size of the tagset for each task (label tagging and dependency attachment).

In particular, we identify two flaws with the *PoS-based* encoding:

- the tagset includes many infrequent tags (due to infrequent PoS-tags and long distance dependencies) which are difficult to predict;[3]

[2] The root has no actual parent, thus no relative position.

[3] For instance, on the trainset of the Universal Dependencies (UD) for English (EWT) (Nivre et al., 2018), 90% of the tokens are tagged with the same 15 tags among the 198 encoded tags.



I	called	over	the	weekend	due	to	clogged	kitchen	sink	.	
PRON	VERB	ADP	DET	NOUN	ADP	ADP	VERB	NOUN	NOUN	PUNCT	PoS
VERB^{+1}	ROOT	NOUN^{+1}	NOUN^{+1}	VERB^{-1}	NOUN^{+2}	ADP^{-1}	NOUN^{+2}	NOUN^{+1}	VERB^{-2}	VERB^{-2}	RPT
	X			X	X				X		U.Head
X^{+1}	ROOT	X^{+1}	X^{+1}	X^{-1}	X^{+1}	X^{-1}	X^{+1}	X^{+1}	X^{-3}	X^{-4}	RUH
	VP			NP	X				NP		C.Head
VP^{+1}	ROOT	NP^{+1}	NP^{+1}	VP^{-1}	NP^{+1}	X^{-1}	NP^{+1}	NP^{+1}	VP^{-1}	VP^{-1}	RCH

Figure 1: Dependencies and encoded tags on an English sentence from the EWT treebank. RPT is the encoding based on PoS-tags (PoS). RUH and RCH are the relative encodings based on head tags (respectively: unique head –U.Head– tags and chunk head –C.Head– tags).

- consecutive PoS-tags which have similar roles (such as NOUN and PROPN or VERB and AUX) make the prediction of the relative position less accurate (i.e., biased towards short relative position) due to the difficulty of identifying which token is the head of a subtree (in a group of tokens which constitute a phrase, e.g., the main noun in a noun phrase or verb in a verb phrase).

In order to alleviate the impact of these flaws, we propose a new encoding strategy that we name *relative head-based* encoding. It requires a first step of *head* tagging in which we identify the heads/parents, i.e., the tokens which have children in the dependency tree. We propose two approaches for tagging the heads:

- a first approach (*Unique Head*) is to tag all parents with a unique tag X (and all non-parents with a NONE tag);
- a second approach (*Chunk Head*) is to see parents as heads of syntactic chunks. We define their roles (tags) as such. In this case, the tagset of the head tagging task includes 5 tags: VP (for heads which are VERBs and AUXs), NP (for NOUNs, PROPNs and PRONs), AP (for ADJs and NUMs), X (for the remaining heads) and NONE for the non-parents.[4]

With this approach, disambiguating between PoS-tags with similar roles rests on the head tagging step instead of the actual dependency parsing step which focuses on attaching the children to the correct head.

The *relative head-based* encodings (RUH for *Relative Unique Head* and RCH for *Relative Chunk Head*) are thus deduced from these head tags in the same way as the *relative PoS-based* encoding with the PoS-tags. The encoding of a dependency is defined as a tuple of **the head tag of the parent** and **its relative position** to the child in regards to other head tags with the same role. Hence, the dependency attachment step consists in predicting these encoded tags and then building the dependency tree using in addition the information about the heads from the previous head tagging step.

Using the *relative head-based* encoding reduces the size of the tagset (for the dependency attachment task) by 65% (RUH) and 52% (RCH) on average[5] compared to the *relative PoS-based* encoding. See an example of the *relative head-based* encodings (RUH and RCH) in Figure 1. In this sentence, "*clogged*" and "*kitchen*" have the same tag with both *head-based* encodings because they have the same head while they have different tags with the *PoS-based* encoding.

[4] The NPs and VPs represent respectively 57% and 27% of the heads of the UD trainsets we train on (see section 4).

[5] On the UD trainsets we train on (see section 4).

	PoS-tag based dep. encoding										
	STR19		**Shared**			**Stacked**			**Combined**		
Lang.	UAS	LAS	UAS	LAS	sent/s	UAS	LAS	sent/s	UAS	LAS	sent/s
cs	89.82	87.63	85.36	81.29	$163_{\pm 1}$	**87.50**†	**83.66**†	$100_{\pm 2}$	86.84	82.92	$124_{\pm 2}$
en	82.22	78.96	80.33	76.17	$159_{\pm 2}$	**82.50**	**78.41**	$70_{\pm 2}$	81.88	77.87	$108_{\pm 2}$
fi	80.31	76.39	77.05	71.37	$143_{\pm 2}$	**80.80**†	**75.95**†	$63_{\pm 1}$	79.85	74.85	$97_{\pm 1}$
grc	76.58	71.70	67.98	60.28	$157_{\pm 4}$	68.61	61.29	$81_{\pm 2}$	**68.96**	**61.41**	$105_{\pm 1}$
he	67.23	62.86	72.28	65.52	$89_{\pm 2}$	**77.80**†	**71.56**†	$37_{\pm 1}$	75.53	69.27	$58_{\pm 1}$
kk	32.14	17.03	42.89	18.88	$180_{\pm 2}$	41.27	17.36	$82_{\pm 2}$	**44.08**†	**19.36**†	$127_{\pm 3}$
ta	73.24	66.51	62.89	50.65	$113_{\pm 6}$	63.11	51.37	$48_{\pm 2}$	**63.45**	**52.29**†	$76_{\pm 2}$
zh	61.01	57.28	68.28	61.90	$92_{\pm 1}$	70.91	64.66	$40_{\pm 1}$	**71.00**	**65.00**	$62_{\pm 1}$
avg	70.32	64.79	69.63	60.76	137	**71.56**	**63.03**	65	71.45	62.87	95

Table 1: Dependency parsing scores (+ average sentence per second on CPU) using the PoS-tag based encoding for the different learning strategies (best in bold; † marks statistical significance; T-test with $p<0.05$). **STR19** scores are reported from Strzyz et al. (2019) (besides from Tamil for which they use gold PoS-tags).

4 Experiments

Models We design three types of experiments. In a first set of experiments, we compare the *shared* and *stacked* learning strategies with the *combined* strategy. For each experiment, we train four tasks (simultaneously or sequentially): PoS-tagging, (morphological) feature tagging, label (dependency relation) tagging and dependency attachment. For the *combined* strategy, we define two groups of tasks (trained in the following order): PoS-tagging/feature tagging, followed by label tagging/dependency attachment.

As a second experiment, we compare the performance of the *combined* system using different encodings of the dependencies (*PoS-based* and *head-based*). When using our proposed *head-based* encodings, the groups are (trained in this order): PoS-tagging/feature tagging/head tagging, followed by label tagging/dependency attachment.

Third, we train the pipeline without PoS-tagging and feature tagging (-PoS/feats), using only head tagging as a first group.

Setup We use the pre-trained word embeddings of Grave et al. (2018).[6] For each task or group of tasks, we use 2 hidden layers of dimension 256. Dimension of the hidden layer for training character embeddings is 128.

Data We use the Universal Dependencies 2.2 (Nivre et al., 2018) dataset for training and evaluating. Following de Lhoneux et al. (2017b), we select a subset of the treebanks: Czech-PDT (cs), English-EWT (en), Finnish-TDT (fi), Ancient Greek-PROIEL (grc), Hebrew-HTB (he), Kazakh-KTB (kk), Tamil-TTB (ta) and Chinese-GSD (zh). Universal PoS-tags (UPOS) are used for PoS-tagging and *PoS-based* encoding. Head tags are deduced from the gold universal dependencies and PoS-tags as stated in Section 3.

Evaluation. We average the scores on 5 runs (with different random seeds) for each experiment. We calculate the unlabeled attachment score (UAS) and the labeled attachment score (LAS) following the guideline of the CoNLL 2018 Shared Task (Zeman et al., 2018).[7] We also evaluate precision on heads, i.e., percentage of correctly tagged parents.[8]

5 Results

5.1 Multi-task Learning Strategies

We compare the learning strategies (*shared*, *stacked* and *combined* in Table 1) when using the *relative PoS-based* encoding. The *shared* strategy always leads to the lowest scores, which support the idea

[6] For Ancient Greek, we use the embeddings provided for the CoNLL 2017 Shared Task (Ginter et al., 2017).

[7] Only universal dependency labels are evaluated (ignoring language-specific subtypes); punctuation is included.

[8] For the RPT encoding, PoS-tagging on heads only is evaluated.

	Rel. PoS-Tag based enc.			Rel. Unique Head-based enc.			Relative Chunk Head-based encoding				
										-PoS/feats	
Lang.	**hd prec.**	UAS	LAS	**hd prec.**	UAS	LAS	**hd prec.**	UAS	LAS	UAS	LAS
cs	98.16	**86.84**†	82.92	94.79	86.24	**83.11**	93.28	86.09	82.31	85.96	82.06
en	92.61	81.88	77.87	93.11	81.48	77.34	90.17	**82.70**†	**78.76**†	81.61	77.33
fi	94.64	79.85	74.85	91.37	77.33	72.36	88.25	**79.89**	**75.08**	78.43	72.64
grc	84.90	**68.96**	**61.41**	88.31	67.61	59.72	78.59	68.71	61.39	67.91	60.44
he	91.85	75.53	69.27	94.98	**81.48**†	**74.12**†	88.89	76.93	70.13	77.49	69.97
kk	53.26	44.08	19.36	73.76	**47.61**†	**21.70**†	41.23	40.19	18.95	37.30	17.04
ta	83.13	63.45	52.29	79.75	62.13	50.52	76.67	**65.48**†	**54.32**†	60.70	49.04
zh	92.55	71.00	65.00	92.67	71.85	65.26	88.67	**73.02**†	**66.82**†	71.17	64.34
avg.	86.39	71.45	62.87	88.59	**71.97**	63.02	80.72	71.63	**63.47**	70.07	61.61

Table 2: Dependency parsing scores (+ precision on heads) with the different dependency encodings, using the *combined* learning strategy (best in bold; † marks statistical significance).

that PoS-tagging and dependency parsing must not be trained simultaneously. On average, the *stacked* learning strategy leads to slightly higher performance (+0.11 UAS/+0.16 LAS) than the *combined* strategy. Both strategies lead to highest performance for half of the languages, but the *stacked* strategy significantly outperforms the *combined* strategy on 3 of the 8 languages while this last strategy gives the (significantly) best scores for 2 languages. However, it is worth noting that the parsing speed is much lower with the *stacked* strategy than with the *combined* strategy, which increases the parsing speed by 48% on average. With comparable scores on average, the *combined* strategy is a good trade-off between speed and accuracy.

5.2 Relative Head-Based Encoding

We compare the scores of the *combined* models using the three different encodings (Table 2). We see that the *relative PoS-based* encoding outperforms the other encodings for only one language (cs: +0.6 UAS) while the RUH encoding is significantly better for 2 languages (he: +5.95/4.85 UAS/LAS; kk: +3.53/2.34) and the RCH for 3 languages (en: +0.82/0.89 UAS/LAS; ta: +2.03/2.03; zh: +2.02/1.82).

Overall, the *relative head-based* encoding is a good approach for parsing as sequence labeling. However, from these results, no clear decision can be made on which tagset for head tagging (RUH vs RCH) would be the most adapted to other languages. The intuition behind the RCH encoding is well-suited to languages which are adapted to a structure in syntactic chunks, such as English (which is reflected in the scores). [9] It is worth noting that the variation in the scores between the two head-based encodings is substantial and when one is the best option the other often leads to low scores, which shows that the choice of the tagset for the head tagging is crucial and might require fine-tuning for the different languages. For instance, head tagging performs very well on Hebrew and Kazakh[10] using the *Unique head* tagset leading to high parsing scores.

Furthermore, although the *relative head-based* encoding requires an additional step of head tagging (i.e., one more task in the pipeline), the parsing time is equivalent to the RPT encoding since the head tagging task is performed at the same level as PoS-tagging and feature tagging.

In general, long dependencies are especially difficult to predict correctly. While local dependencies (neighbouring child) achieve more than 80% UAS on average, dependencies of length more than 6 do not overcome 50% UAS. We expect the RCH encoding to alleviate the difficulty of the prediction by artificially reducing the distance between the tokens. We analyse the dependency parsing scores in regards to the length of the dependencies. See the comparison between the encodings in Figure 2. Overall, the RCH encoding outperforms the *PoS-based* encoding for all dependency length but the neighbouring children[11] while the RUH encoding is especially good on local dependencies but performs poorly on long

[9] Dependency parsing as sequence labeling seems to be more adapted to languages with few non-projective dependencies.

[10] Which could be explained by the small training set for Kazakh that makes learning of PoS-tags or chunk head tags more difficult.

[11] The RCH encoding is particularly damaging on Kazakh: -10.3 UAS on dependencies of length 1 while positive on other

dependencies – the dependency attachment tagset for the RUH encoding includes more rare tags with high relative positions which are then more difficult to predict.

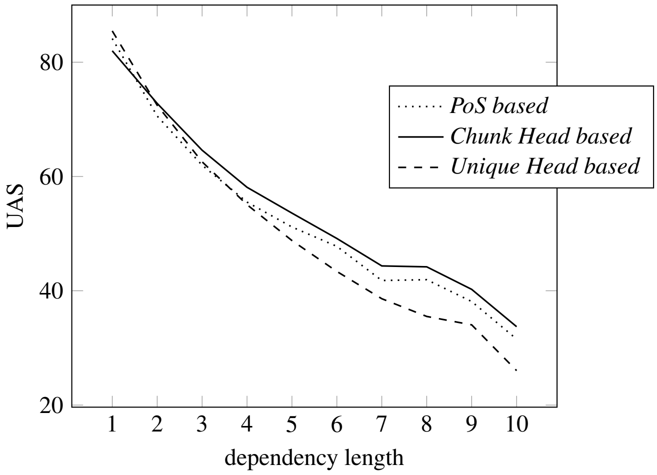


Figure 2: Averaged UAS (on the 8 languages) as a function of the dependency length for the three encodings (using the *combined* learning strategy).

5.3 Ablating PoS-tagging

As previously studied for transition-based parsers (de Lhoneux et al., 2017a; Smith et al., 2018), we want to assess whether dependency parsing as sequence labeling can achieve state-of-the-art performance without PoS-tagging (and feature tagging) as a pre-processing step. We compare the performance of the *combined* strategy using the RCH encoding with and without PoS-tagging (last two columns of Table 2) as part of the first group of tasks to train in the pipeline.

The results are noticeably lower for the ablated model (-1.56 UAS/-1.86 LAS on average) than when using PoS-tagging as an auxiliary task for training head tagging. Determining PoS-tags is essential for most of the languages. Only Hebrew does not suffer from the ablation.

Moreover, it is worth noting that ablating PoS-tagging does not increase the parsing speed since the tasks are performed simultaneously. The *combined* strategy (with PoS-tagging) thus remains a valid trade-off between speed and accuracy.

6 Conclusion

We showed that a combined strategy for multi-task learning using *shared* and *stacked* strategies is on par with a sequential approach while significantly faster at parsing sentences. It provides a good speed-accuracy tradeoff for PoS-tagging and dependency parsing in a single pipeline.

Besides, we proposed a new encoding of the dependencies as labels which does not use PoS-tags. It splits the parsing task in two steps but does not affect negatively the parsing time when performed simultaneously with PoS-tagging. We test two alternatives of this encoding, comparing fine and coarse tagsets for tagging the heads. It shows that the choice of the tagset is crucial: the performance of dependency attachment depends on the performance of head tagging and on how it performs regarding the length of the dependencies. Finally, this suggests that fine-tuning the tagset in regards to properties of the languages could improve overall performance. Globally, the *head-based* models outperform the *PoS-based* model for a majority of the languages.

lengths.

References

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017a. From Raw Text to Universal Dependencies-Look, No Tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017b. Old School vs. New School: Comparing Transition-Based Parsers with and without Neural Network Enhancement. In *The 15th Treebanks and Linguistic Theories Workshop (TLT)*.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, pages 5–6.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled Multi-Task Learning: From Syntax to Translation. *Transactions of the Association of Computational Linguistics*, 6:225–240.

Marco Kuhlmann, Carlos Gómez-Rodríguez, and Giorgio Satta. 2011. Dynamic Programming Algorithms for Transition-Based Dependency Parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq Dependency Parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Ryan McDonald. 2006. Discriminative Training and Spanning Tree Algorithms for Dependency Parsing. *University of Pennsylvania, PhD Thesis*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, 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, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayo Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Aaron Smith, Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2018. An Investigation of the Interactions Between Pre-Trained Word Embeddings, Character Models and POS Tags in Dependency Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Drahomíra Spoustová and Miroslav Spousta. 2010. Dependency Parsing as a Sequence Labeling Task. *The Prague Bulletin of Mathematical Linguistics*.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Viable Dependency Parsing as Sequence Labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*.

2019. Better, Faster, Stronger Sequence Tagging Constituent Parsers, author = vilares, david and abdou, mostafa and søgaard, anders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Anssi Yli-Jyrä and Carlos Gómez-Rodríguez. 2017. Generic Axiomatization of Families of Noncrossing Graphs in Dependency Parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Towards Deep Universal Dependencies

Kira Droganova
Charles University
Faculty of Mathematics and Physics
Praha, Czechia
droganova@ufal.mff.cuni.cz

Daniel Zeman
Charles University
Faculty of Mathematics and Physics
Praha, Czechia
zeman@ufal.mff.cuni.cz

Abstract

Many linguistic theories and annotation frameworks contain a deep-syntactic and/or semantic layer. While many of these frameworks have been applied to more than one language, none of them is anywhere near the number of languages that are covered in Universal Dependencies (UD). In this paper, we present a prototype of Deep Universal Dependencies, a two-speed concept where minimal deep annotation can be derived automatically from surface UD trees, while richer annotation can be added for datasets where appropriate resources are available. We release the Deep UD data in Lindat.

1 Introduction

Universal Dependencies (UD) (Nivre et al., 2016) annotation guidelines have become a de-facto standard for cross-linguistically comparable morphological and syntactic annotation. A significant factor in the popularity of UD is a steadily growing and heavily multilingual collection of corpora: release 2.4 (Nivre et al., 2019) contains 146 treebanks of 83 languages. The UD guidelines have been designed as surface-syntactic, although their emphasis on cross-linguistic parallelism sometimes leads to decisions that are normally associated with deeper, semantics-oriented frameworks (the primacy of content words and the second-class citizenship of function words may serve as an example).

Many theories and annotation frameworks have been proposed that contain a deep-syntactic, tectogrammatical, or semantic dependency layer; to name just a few: Meaning-Text Theory (Žolkovskij and Mel'čuk, 1965), Functional Generative Description (Sgall, 1967), the Proposition Bank (Kingsbury and Palmer, 2002), Sequoia (Candito and Seddah, 2012), or Abstract Meaning Representation (Banarescu et al., 2013). Names vary and so does the extent of 'deep' phenomena that are annotated; the common denominator is that these phenomena are closer to meaning on the meaning-form scale than anything we find in a typical surface-syntactic treebank. By definition, deep representation is more useful for natural language understanding (but it is also more difficult to obtain).

Many of the deep frameworks have been applied to more than one language, sometimes just to demonstrate that it is possible; but none of them is anywhere near the number of languages covered by UD.

UD itself contains a diffident attempt to provide deeper annotations, dubbed the Enhanced Universal Dependencies (Schuster and Manning, 2016). While it is a step in the right direction, it is just the first step: we argue that it should be possible to go deeper. Moreover, Enhanced UD is an optional extension, which is only available in a handful of treebanks (Table 1). Enhanced UD faces the same threat as the other deep frameworks mentioned above: more complex annotation requires more annotation effort, and semantic annotations are often coupled with huge lexical resources such as verb frame dictionaries. Therefore, it is less likely that sufficient manpower will be available to annotate data in a new language. Our principal question is thus the following: is it possible to create a multilingual data collection (and annotation guidelines) that will be as popular and widely used as UD, but deeper?

In our view, the key is to identify a subset of deep annotations that can be derived semi-automatically from surface UD trees, in acceptable quality. These annotations will not be as precise as if they were carefully checked by humans, but they will be available for (almost) all UD languages. More importantly, it will be possible to generate them for new UD languages and the deep extension will thus keep up with the growth of UD. For languages that have better resources available, one could convert them to the deep UD format and provide them instead of the corresponding semi-automatic annotation. Note that there are two dimensions along which a resource can be ‘better’. It can provide the same type of annotation as the light, semi-automatic version, just verified by human annotators. But it may also provide additional types of annotations that cannot be obtained automatically. The Deep UD guidelines should thus cover a broad selection of phenomena that are annotated in popular semantic dependency frameworks.

The present paper reports on work in progress. We have prepared the first prototype of the semi-automatic Deep Universal Dependencies, based on UD release 2.4. The resource is available in the LINDAT/CLARIN repository (`http://hdl.handle.net/11234/1-3022`) under the same set of licenses as the underlying UD treebanks. In the following sections we describe what types of annotation this first version contains and how the annotation is derived from the surface trees; we also offer an outlook on possible future development.

2 Related Work

Manual semantic annotation is a highly time-consuming process, therefore a number of authors experimented with (semi-)automatic approaches to semantic annotation. Padó (2007) proposed a method that uses parallel corpora to project annotation to transfer semantic roles from English to resource-poorer languages. The experiment was conducted on an English-German corpus. Van der Plas et al. (2011) experimented with joint syntactic-semantic learning aiming at improving the quality of semantic annotations from automatic cross-lingual transfer. An alternative approach was proposed by Exner et al. (2016). Instead of utilizing parallel corpora, they use loosely parallel corpora where sentences are not required to be exact translations of each other. Semantic annotations are transferred from one language to another using sentences aligned by entities. The experiment was conducted using the English, Swedish, and French editions of Wikipedia. Akbik et al. (2015) described a two-stage approach to cross-lingual semantic role labeling (SRL) that was used to generate Proposition Banks for 7 languages. First, they applied a filtered annotation projection to parallel corpora, which was intended to achieve higher precision for a target corpus, even if containing fewer labels. Then they bootstrapped and retrained the SRL to iteratively improve recall without reducing precision. This approach was also applied to 7 treebanks from UD release 1.4.[1] However, the project seems to be stalled.

Mille et al. (2018) proposed the deep datasets that were used in the Shallow and Deep Tracks of the Multilingual Surface Realisation Shared Task (SR’18, SR’19). The Shallow Track datasets consist of unordered syntactic trees with all the word forms replaced with their lemmas; part-of-speech tags and the morphological information are preserved (available for 10 languages). The Deep Track datasets consist of trees that contain only content words linked by predicate-argument edges in the PropBank fashion (available for English, French and Spanish). The datasets were automatically derived from UD trees v.2.0. Gotham and Haug (2018) proposed an approach to deriving semantic representations from UD structures that is based on techniques developed for Glue semantics for LFG. The important feature of this approach is that it relies on language-specific resources as little as possible.

3 Enhanced Universal Dependencies

The Enhanced UD (Schuster and Manning, 2016)[2] represents a natural point of departure for us. UD v2 guidelines define five types of enhancements that can appear in treebanks released as part of UD. All the enhancements are optional and it is possible for a treebank to annotate one enhancement while ignoring the others. The enhanced representation is a directed graph but not necessarily a tree. It may contain ‘null’ nodes, multiple incoming edges and even cycles. The following enhancements are defined:

[1] `https://github.com/System-T/UniversalPropositions`

[2] While Schuster and Manning (2016) remains the most suitable reference for Enhanced UD to date, its publication predates the v2 UD guidelines and the proposals it contains are only partially compliant with the guidelines. See `https://universaldependencies.org/u-overview/enhanced-syntax.html` for the current version.

Null nodes for elided predicates. In certain types of ellipsis (*gapping* and *stripping*), multiple copies of a predicate are understood, each with its own set of arguments and adjuncts, but only one copy is present on the surface. Example: *Mary flies to Berlin and Jeremy [flies] to Paris.* The enhanced graph contains an extra node for each copy of the predicate that is missing on the surface. Note that the guidelines do not license null nodes for other instances of ellipsis, such as dropped subject pronouns in pro-drop languages.

Propagation of conjuncts. Coordination groups several constituents that together play one role in the superordinate structure. They are all equal, despite the fact that the first conjunct is formally treated as the head in the basic UD tree. For example, several coordinate nominals may act as subjects of a verb, but only the first nominal is actually connected with the verb via an `nsubj` relation. In the enhanced graph, this relation is propagated to the other conjuncts, i.e., each coordinate nominal is directly connected to the verb (in addition to the `conj` relation that connects it to the first conjunct). Likewise, there may be shared dependents that are attached to the first conjunct in the basic tree, but in fact they modify the entire coordination. Their attachment will be propagated to the other conjuncts, too. (Note that not all dependents of the first conjunct must be shared. Some of them may modify only the first conjunct, especially if the other conjuncts have similar dependents of their own.)

External subjects. Certain types of non-finite, 'open' clausal complements inherit their subject from the subject or the object of the matrix clause. Example: *Susan wants to buy a book.* In the basic tree, *Susan* will be attached as `nsubj` of *wants*, while there will be no subject dependent of *buy*. In contrast, the enhanced graph will have an additional `nsubj` relation between *buy* and *Susan*.

Relative clauses. The noun modified by a relative clause plays a semantic role in the frame of the subordinate predicate. In the basic UD tree, it is represented by a relative pronoun; however, in the enhanced graph it is linked from the subordinate predicate *instead* of the pronoun. (The pronoun is detached from the predicate and attached to the noun it represents, via a special relation `ref`.) This is the reason why enhanced graphs may contain cycles: in *The boy who lived*, there is an `acl:relcl` relation from *boy* to *lived*, and an `nsubj` relation from *lived* to *boy*.

Case information. The labels of certain dependency relations are augmented with case information, which may be an adposition, a morphological feature, or both. For example, the German prepositional phrase *auf dem Boden* "on the ground" may be attached as an oblique dependent (`obl`) of a verb in the basic tree. The enhanced label will be `obl:auf:dat`, reflecting that the phrase is in the dative case with the preposition *auf*. This information is potentially useful for semantic role disambiguation, and putting it to the label is supposed to make it more visible; nevertheless, its acquisition from the basic tree is completely deterministic, and there is no attempt to translate the labels to a language-independent description of meaning.

Several extensions of the enhanced representation have been proposed. The *enhanced++* graphs proposed by Schuster and Manning (2016) extend the set of ellipsis-in-coordination types where null nodes are added; they also suppress quantifying expressions in sentences like *a bunch of people are coming*.

Candito et al. (2017) define the *enhanced-alt* graphs, which neutralize syntactic alternations, that is, passives, medio-passives, impersonal constructions and causatives. They also suggest to annotate external arguments of other non-finite verb forms than just open infinitival complements and relative clauses: most notably, for participles, even if they are used attributively. Hence in *ceux embauchés en 2007* "those hired in 2007", *embauchés* heads a non-relative adnominal clause (`acl`) that modifies the nominal *ceux*, but at the same time *ceux* is attached as a passive subject (`nsubj:pass`) of *embauchés*.

4 Pre-existing Enhancing Tools

Enhanced UD contains information that cannot be derived automatically from the basic UD tree; additional human input is needed in order to fully disambiguate all situations. Nevertheless, it is believed that automatic 'enhancers' can get us relatively far. Schuster and Manning (2016) described and evaluated the Stanford Enhancer,[3] which is available as a part of the Stanford CoreNLP suite.

Nyblom et al. (2013) reported on the Turku Enhancer, a hybrid approach (consisting of rule-based heuristics and machine-learning components) to enhancing Stanford Dependencies of Finnish. The enhancements tackled were conjunct propagation, external subjects, and syntactic functions of relativizers; the first two are thus relevant also in Enhanced UD. Their system achieved F_1 score of 93.1; note however that labeled training data is needed for the approach to work.

Nivre et al. (2018) compares the Stanford Enhancer with an adapted version of the Turku Enhancer. They trained it on the Finnish labeled data, but in a delexicalized fashion (only non-lexical features were considered). The Turku Enhancer does not predict null nodes, and for external subjects it only considers subject control (or raising), but not object control. On the other hand, Stanford Enhancer only predicts core arguments as controllers while in some languages non-core dependents can control subjects too. Nevertheless, both enhancers are found usable for other languages, as shown on Swedish and Italian. The paper also evaluates an Italian-specific rule-based enhancer, which does not predict null nodes.

Candito et al. (2017) took a rule-based approach to produce their *enhanced-alt* graphs for French: they developed two sets of rules, using two different graph rewriting systems. However, they only focus on two of the five enhancements (external subjects and conjunct propagation), and they only do it for French. Some of their heuristics are very French-specific and they assume that information needed for disambiguation is available in the source annotation (which is the case of the Sequoia French treebank).

Several other UD treebanks come from sources where some enhanced annotation is available and can be converted to Enhanced UD. Bouma (2018) demonstrates how original annotations from the Alpino treebank can help enhance the Dutch UD treebanks. Patejuk and Przepiórkowski (2018) discuss conversion from an LFG treebank of Polish and note that not only there is more information than in basic UD, some information cannot be captured even by Enhanced UD. Another example is the distinction between private and shared dependents in coordination: for treebanks converted from Prague-style annotation (Arabic, Czech, Lithuanian, Slovak, Tamil), this distinction is readily available.

5 Data Preparation

The first version of Deep UD is based on UD release 2.4 (Nivre et al., 2019) but we intend to generate updates after each future UD release. While we foresee improved semantic annotation for some languages (based on additional lexical resources, for example), the current version is derived just from the annotation available in UD itself (though we use heuristics that may be language- or treebank-specific). UD 2.4 contains 146 treebanks of 83 languages. We exclude 6 treebanks that are distributed, for copyright reasons, as hollow annotations without the underlying text. We further exclude 19 treebanks with incomplete or non-existent lemmatization.[4] Consequently, our resource contains 121 treebanks of 73 languages.

We take enhanced UD graphs (Section 3) as the point of departure for deep UD. However, only a small fraction of the UD treebanks have some enhanced annotation, and if they do, then they often omit one or more of the five types of enhancements defined in the guidelines. There are 24 treebanks of 16 languages that have enhanced graphs (Table 1). We will refer to these enhanced graphs as *trusted enhanced annotations*. Some of them were converted from non-UD manual annotations, some were probably generated with the help of automatic enhancers, but at least they were overseen by the teams responsible for the given language.

We use the Stanford Enhancer[5] to generate enhanced graphs for corpora that lack them. For the six treebanks in Table 1 that contain trusted annotation of all five enhancement types, we take the trusted annotation. For the other 18 treebanks in the table, ideally we should merge the trusted annotation with the output of the enhancer so that all enhancement types are present. However, merging may not be trivial

[3] The Stanford UD Enhancer was adapted from an older tool that was designed to work with the Stanford Dependencies, a predecessor of UD.

[4] Note that we do not exclude some other treebanks where lemmas exist but have been assigned by a stochastic model instead of human annotators.

[5] The README file of the released data provides details on what version we used and how we ran it.

Language	Treebank	Gapping	Coord	XSubj	RelCl	CaseDeprel
Arabic	PADT		yes			
Bulgarian	BTB		yes	yes	yes	yes
Czech	CAC		yes			
Czech	FicTree		yes			
Czech	PDT		yes			
Dutch	Alpino	yes	yes	yes	yes	yes
Dutch	LassySmall	yes	yes	yes	yes	yes
English	EWT	yes	yes	yes	yes	yes
English	PUD	yes	yes	yes	yes	yes
Estonian	EWT	yes				
Finnish	PUD	yes	yes			
Finnish	TDT	yes	yes	yes		
Italian	ISDT		yes	yes	yes	yes
Latvian	LVTB	yes	yes	yes		yes
Lithuanian	ALKSNIS		yes			
Polish	LFG		yes	yes		yes
Polish	PDB		yes			
Polish	PUD		yes			
Russian	SynTagRus	yes				
Slovak	SNK		yes			
Swedish	PUD	yes	yes	yes	yes	yes
Swedish	Talbanken	yes	yes	yes	yes	yes
Tamil	TTB		yes			
Ukrainian	IU	yes	yes	yes	yes	

Table 1: Overview of enhanced annotations in UD 2.4 treebanks. Gapping: there are empty nodes representing elided predicates. Coord: dependencies (both incoming and outgoing) are propagated to all conjuncts. XSubj: higher argument is linked as the subject of a controlled verb. RelCl: nominal modified by a relative clause is linked as argument or adjunct in that clause. CaseDeprel: case markers are added to the dependency labels of adverbial and oblique dependents.

in sentences where multiple enhancement types interact, and we leave it for future work. In the current version, the enhanced graphs in these 18 treebanks are replaced by the output of the Stanford Enhancer.

Note that using the Stanford Enhancer does not guarantee that the resulting annotation identifies all five types of enhancements—even if the phenomenon exists in the language and the treebank is large enough to provide examples. Identification of relative clauses relies on a language-specific list of relative pronouns and on the optional dependency label `acl:relcl`, but some treebanks use `acl` instead. Gapping, besides being relatively rare, is not annotated properly in the basic representation of some UD languages. Consequently, only 58 enhanced treebanks have some null nodes (gapping) and only 54 treebanks have edges specific to relative-clause enhancements. Most treebanks have the other three types; a remarkable exception is Japanese where the three treebanks have only one enhancement type, namely the case-augmented dependency relations. 37 treebanks feature all five types. We plan to expand the relative clause annotation to other treebanks in the future; listing relative pronouns (a closed class) is quite feasible, and we can utilize the morphological feature `PronType=Rel` where available.

6 Delving Deeper

There are numerous phenomena that various semantic frameworks strive to capture. Without precluding any of them from future versions of Deep UD, we believe that the core of sentence understanding is its predicate-argument structure. We start with verbal predicates and identify their arguments, if present in the same sentence. We number the arguments roughly reflecting their decreasing salience and making

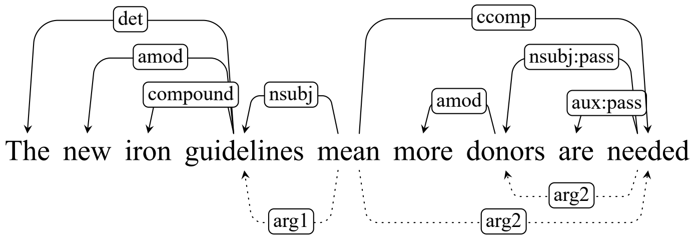


Figure 1: An example of a deep graph for the English sentence *The new iron guidelines mean more donors are needed.*

sure that for the same predicate (sense), the argument with a particular semantic role will always get the same label/number, regardless the syntactic environment. That means that we have to neutralize valency-changing operations such as passivization; here we are very close to the *enhanced-alt* representation proposed by Candito et al. (2017). For example, in *George killed the dragon* as well as in *The dragon was killed by George*, *George* will be `arg1` and *the dragon* will be `arg2`. We do not label the actual semantic roles (i.e., agent / actor / killer for *George* and patient / killed for *the dragon*) directly in the text. Instead, the predicate instance can be linked to a frame dictionary (if available) where the corresponding frame will provide interpretation of the numbered arguments. Linking of frame instances to dictionary frames will not be trivial and the concrete approach will depend on the language and on the nature of the target lexical resource. Valency frame dictionaries often contain information on morphological and syntactic properties of the arguments. A verbal lemma will typically correspond to several (sometimes dozens of) different frames. Sometimes the forms of the arguments (their morphological case, preposition etc.) will narrow down the search; but full disambiguation may not be possible without a statistical model or a human annotator. Once we have the correct frame, identification of individual arguments is (again) just matching their properties against those specified by the frame.

We follow the CoNLL-U Plus file format[6] with two new columns: DEEP:PRED and DEEP:ARGS. These columns contain annotation we add on top of Enhanced UD; without them, the file is still a valid CoNLL-U file. The value in DEEP:PRED identifies the predicate. It can be a reference to a particular sense (frame) in a dictionary but we currently use just the lemma of the verb, possibly augmented with other lemmas if it is a compound verb (e.g. Germanic phrasal verbs such as *come up*). The value in DEEP:ARGS points to the head nodes of subtrees that represent the arguments. For example, `arg1:33|arg2:12,27` means that the most salient argument (possibly the agent) is headed by node 33, while the second most salient argument (possibly the patient) is coordination and the conjuncts are headed by nodes 12 and 27, respectively. See Figure 1 for an example of a deep graph.

Thanks to Enhanced UD, the annotation resolves some instances of grammatical coreference (Zikánová et al., 2015), i.e., situations where one node serves as an argument of multiple verbs, and it can be inferred from the grammatical rules of the language. On the other hand, the current version does not attempt to address textual coreference, e.g., a personal pronoun that is coreferential with a noun. Arguably, textual coreference cannot be resolved without a human annotator or a trained model.

Some arguments are not accessible through Enhanced UD; similar to Candito et al. (2017), we are experimenting with heuristics that yield additional enhanced dependencies for non-finite verbs:

Infinitives that are not `xcomp`. They can be ordinary clausal complements (`ccomp`) and then we cannot identify their subject, as in Dutch: *Zijlaard adviseerde te gokken op de sprint* (lit. *Zijlaard advised to bet on the sprint*) “Zijlaard advised betting on the sprint”. But they can be also adverbial clauses (`advcl`), or adnominal clauses (`acl`), if the main clause’s predicate is a light verb with a noun, as in Dutch: *had moeite om zich te concentreren* (lit. *had trouble so himself to concentrate*) “struggled to concentrate”. The infinitive *concentreren* “to concentrate” in this case works similarly to an `xcomp`, that is, it should inherit the subject from the matrix clause.

[6] `https://universaldependencies.org/ext-format.html`

Participles. An attributively used participle modifies a noun. If it were a relative clause, the enhanced graph would identify the noun as the "subject" argument of the participle; but it is an `amod` rather than a clause, and no external subject relation is present. A Dutch example: *de afgelopen week* (lit. *the expired week*) "last week". We add a heuristic that participles attached as `amod` shall take the modified noun as their argument; note that we need to distinguish active and passive participles in order to find out whether the noun is argument 1 or 2. Currently we only look for the morphological feature `Voice=Pass` but it is not always available, and some verb forms can be used both in active and passive clauses. Consider English: *the shares reflected on your statement*; *reflected* is used as a passive participle but `Voice=Pass` is not present, it is just a "past participle" without any voice feature. We may need to estimate whether a verb is transitive, and if it is, the participle will be considered passive, otherwise it will be considered active. Nevertheless, no such heuristic was applied to the current version of the data.

Converbs (gerunds). English: *X did Y..., killing several people.* The syntactic annotation does not tell us that X is the argument 1 of *killing.* We work with the hypothesis that a gerund or converb attached as `advcl` inherits the subject of the matrix clause. This is a rule at least in some languages but we have yet to evaluate to what extent the rule may be universal.

Language-specific heuristics. A number of heuristics will be needed that are language- or even treebank-specific. For example, passivization of English ditransitive clauses promotes the indirect object rather than the direct object *(what I was asked).*[7] Therefore, if there is a direct object in a passive clause, the subject should be considered argument 3 and not 2.

7 Conclusion and Outlook

We presented a prototype of Deep Universal Dependencies, a deep-syntactic annotation layer that can be derived semi-automatically from surface UD graphs. Our plan is to accommodate rich semantic annotations in languages where necessary resources are available, and automatically generate the core part for other languages after each UD release. Our contribution at the current stage is threefold: 1. While UD releases still contain Enhanced UD only for a few treebanks, we make sure that enhanced graphs are available everywhere; 2. to find more arguments, we do additional enhancements (infinitives, gerunds, participles) internally but we do not show them in the enhanced graphs so that the graphs stay within the current guidelines; 3. we normalize diathesis and show the numbered arguments (canonical subject and object in the terms of Candito et al. (2017)).

The list of possible future directions is much longer than we can accommodate in a short paper; for instance, we want to take advantage of oblique argument marking in treebanks where it is available, improve recognition of passives and other diathesis alternations, or implement other enhancements from Schuster and Manning (2016)'s *enhanced*++. Nevertheless, the most important next step is to evaluate the quality of the generated annotation (both the output of the Stanford Enhancer and the additional heuristics we applied to the enhanced graphs). Since there is no gold-standard labeled data suitable for such evaluation, we will have to manually inspect random samples of the output, or compare the predicate-argument patterns with existing valency dictionaries (in languages where they exist).

Acknowledgements

This work is partially supported by the GA UK grant 794417, the SVV project number 260 453, and the grant no. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.

We also wish to thank the three anonymous reviewers for their valuable comments.

References

Alan Akbik, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, Huaiyu Zhu, et al. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In Proceedings of the 53rd Annual Meeting of

[7] Of course, one could then question whether *I* is an indirect object in the active clause if it can be promoted by passivization; here we follow the actual approach of the English UD treebanks.

the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 397–407.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186.

Gosse Bouma. 2018. Comparing two methods for adding enhanced dependencies to UD treebanks. In Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), pages 17–30, Oslo, Norway. Linköping Electronic Conference Proceedings.

Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles.

Marie Candito, Bruno Guillaume, Guy Perrier, and Djamé Seddah. 2017. Enhanced UD dependencies with neutralized diathesis alternation. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 42–53, Pisa, Italy.

Peter Exner, Marcus Klang, and Pierre Nugues. 2016. Multilingual supervision of semantic annotation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1007–1017.

Matthew Gotham and Dag Trygve Truslew Haug. 2018. Glue semantics for Universal Dependencies. In Miriam Butt and Tracy Holloway King, editors, Proceedings of the LFG'18 Conference, pages 208–226, Wien, Austria. CSLI Publications.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In LREC, pages 1989–1993.

Simon Mille, Anja Belz, Bernd Bohnet, and Leo Wanner. 2018. Underspecified universal dependency structures as inputs for multilingual surface realisation. In Proceedings of the 11th International Conference on Natural Language Generation, pages 199–209.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Paris, France. European Language Resources Association.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing universal dependency treebanks: A case study. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 102–107, Bruxelles, Belgium. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Sarah McGuinness, Abigail Walsh, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, 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.

Jenna Nyblom, Samuel Kohonen, Katri Haverinen, Tapio Salakoski, and Filip Ginter. 2013. Predicting conjunct propagation and other extended Stanford Dependencies. In Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 252–261, Praha, Czechia. Matfyzpress.

Sebastian Padó. 2007. Cross-lingual annotation projection models for role-semantic information. Saarland University.

Agnieszka Patejuk and Adam Przepiórkowski. 2018. From Lexical Functional Grammar to Enhanced Universal Dependencies. Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa, Poland.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association.

Petr Sgall. 1967. Functional sentence perspective in a generative description. Prague Studies in Mathematical Linguistics, 2:203–225.

Lonneke Van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pages 299–304. Association for Computational Linguistics.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. 2015. Discourse and Coherence. From the Sentence Structure to Relations in Text. Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia.

Aleksandr K. Žolkovskij and Igor A. Mel'čuk. 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza (on a possible method and instruments for semantic synthesis). Naučno-texničeskaja informacija, 5:23–28.

Delimiting Adverbial Meanings. A corpus-based comparative study on Czech spatial prepositions and their English equivalents

Marie Mikulová
Faculty of Mathematics and Physics
Charles University, Prague
mikulova@ufal.mff.cuni.cz

Veronika Kolářová
Faculty of Mathematics and Physics
Charles University, Prague
kolarova@ufal.mff.cuni.cz

Jarmila Panevová
Faculty of Mathematics and Physics
Charles University, Prague
panevova@ufal.mff.cuni.cz

Eva Hajičová
Faculty of Mathematics and Physics
Charles University, Prague
hajicova@ufal.mff.cuni.cz

Abstract

The data of the Prague Czech-English Dependency Treebank (a member of the family of Prague Dependency Treebanks) have served as a basis for the comparative study of delimiting adverbial meanings of the local relation “within the given place”. The Czech prepositional groups containing the prepositions *v, na*, and *u* with the above meaning are compared with their English equivalents, using a more subtle differentiation into three semantic subgroups of "inside", "on the surface" and "at the given place". Our analysis confirms that though every language structures the reality in a different way, certain tendencies may be observed in the relation of the forms and their functions that eventually result in a more detailed classification. The contribution presents results of an ongoing work.

1 Introduction

The description of adverbial meanings has a long tradition in linguistics, varying in its attention to detail (e.g. Quirk et al., 1985; for Czech: Šmilauer, 1947). However, it is well known that the traditional classification of adverbials is not fine-grained enough, either in theoretical description or for NLP tasks.

In the multi-layered scenario of Prague Dependency Treebanks (Sect. 2), linguistic meaning is captured by the deep syntactic layer, where the syntactic relations are represented by the so-called functors. However, the functors capture relatively general categories. E.g., all the following adverbials *na stole* ‘on the table’, *pod stolem* ‘under the table’, *za stolem* ‘behind the table’, *poblíž stolu* ‘near the table’, etc. are represented by a single functor with a static meaning “where” (functor LOC). It is obvious that a differentiation among the partial meanings (“on the given place”, “under the given place”, “behind the given place”, etc.) is needed for a more precise representation of the sentence meaning and for its translation to another language. In order to describe these fine-grained distinctions, a set of so-called subfunctors has been considered (Mikulová et al., 2017).

The area of spatial meanings is wide. It includes the general meanings of "where", “which way”, “to where” and “from where” (which we capture by functors), and also their subtle meanings (“inside”, “on the surface”, “next to”, “under”, etc.), for which we propose subfunctors.[1] In the paper, we analyze only a narrow, highly problematic set of meanings within the LOC functor ("where"). We focus on the specification of spatial adverbial meanings expressed by prepositional groups (Sect. 3). Our Czech-English parallel data (Sect. 2) make it possible to compare corresponding expressions in the two languages and to explore differences in forms and meanings, in particular those expressing localization “within the given place“ (Sect. 4). We believe that such an analysis will help us to evaluate the universality and language specificity of the suggested subset of adverbial meanings and thus to make the description of this subset for Czech more precise.

[1] For the delimitation of the functors, the lexical meaning of the verb and its valency properties may be a useful clue, whereas subfunctors are primarily expressed within prepositions.

2 Theoretical Background and Data Resources

2.1 Functional Generative Description

We base our investigation on the theoretical framework of the Functional Generative Description (Sgall et al., 1986), a language-oriented rather than ontology-oriented dependency syntax theory. As for the relationship between language meaning and ontological content, the FGD works with language meaning in the sense of structural linguistics, treating meaning as a linguistically structured phenomenon. When describing attributes necessary for the layer of language meaning, we inevitably tackle the boundary between meaning and content, for example by differentiating homonymy (properties of a form in relation to meaning) and vagueness (properties of meaning in relation to content). We search for testable criteria to be able to account for these distinctions and also to specify synonymy (Sect. 3).

Compared with other descriptions of spatial relations,[2] our approach is characterized especially by the following aspects:

- An exclusive focus on the way how the given language in its structure reflects the reality
- Dependency syntax approach
- A detailed corpus-based research.

2.2 Prague Czech-English Dependency Treebank

The ideas of the Functional Generative Description were applied in the annotation scenario of the Prague Dependency Treebanks (Hajič et al., 2017).[3] The Prague treebanks are complex linguistically motivated corpora with interlinked hierarchical layers of standoff annotation (on morphological, surface and deep syntactic layer). The pilot Prague Dependency Treebank (Hajič et al., 2018) was built in 1996 through 2018. A slightly modified scenario was then used for the annotation of the other treebanks.

The Prague Czech-English Dependency Treebank (PCEDT; Hajič et al., 2012), which is used for our comparative study, is an annotated Czech-English corpus. The English part consists of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). Czech part was translated from the English source sentence by sentence.

3 Methodology of Delimiting Adverbial Meanings

Our analysis of adverbial meanings is based on the assumption that there is no one-to-one relation between the underlying syntactic function represented by the functor-subfunctor combination and its formal expression, in this case the preposition(al group). One syntactic function can be expressed by several different forms whereas one form can be used to express different syntactic functions.

Analyzing fine-grained adverbial meanings, we apply the following principles 3.1, 3.2, and 3.3.

3.1 Substitutability of forms

When deciding which forms are synonymous and thus can be described by the same subfunctor we test whether the forms are substitutable in different contexts and how their meaning is influenced by the substitution. The forms may be semantically the same, they can only partially correspond to each other, or they are not substitutable at all.

E.g., the two prepositions *v* 'in' and *na* 'in/on' are substitutable in cases when the semantic distinctions between them are obscured due to the fact that they form a prepositional group with words denoting objects that do not distinguish „inside“ and „surface“, cf. (1), or this distinction is excluded by a broader sentential context, cf. (2). However, when it comes to localizations beyond the meaning "inside", the preposition *na* 'in/on' cannot be substituted by *v* 'in' (cf. (3), where the greenhouse is supposed to be placed in the garden).

(1) Umíte se dobře zorientovat **v mapě**? /→na mapě
'Are you able to read the map (lit. to orient yourself in a map)? /→on a map'

[2] There is a brief list of various analysis of the spatial prepositions: Bennett, 1975; Herskovits, 1986; Aurnague, 1995; Garrod et al., 1999; Lindstromberg, 2010; Talmy, 2006; Vandeloise, 1991; etc.

[3] https://ufal.mff.cuni.cz/pdt-c

(2) Přespali jsme tam **na** té **chatě**. /→v chatě /≠u chaty
'We slept at the cottage. /→in the cottage /≠by the cottage.'
(3) Mám **na chatě** skleník. /?v chatě
'I have got a greenhouse at my cottage. /?in the cottage'
(4) Má **na hlavě** čepici. /*v hlavě
*'He has a cap on his head. /*in his head'*

The preposition *na* 'in/on' can be substituted by the preposition *v* 'in' only in case the real localization is „inside", which is however impossible to determine without the knowledge of the situation or without a clue from the context. We conclude from this that the semantic feature „inside" (subfunctor *inside*) does not appertain to the preposition *na* 'in/on'. The preposition *na* 'in/on' introduces an object as a whole, covering several different localizations within the given place (subfunctor *at the given place*). Similarly, forms *v* 'in' and *na* 'in/on' are not substitutable if the given place is a 3D object[4] and the real localization is "surface", cf. (4); thus the meaning "on the surface" (subfunctor *surface*) only appertains to the preposition *na* 'in/on'.

3.2 Partial synonymy

When delimiting subfunctors, we differentiate forms that are typical for the given function from those that are untypical for this (e.g. secondary prepositions). The untypical forms are always associated with certain connotations which do not arise with the typical ones. The test of substitutability is thus directed from the untypical forms to typical ones, e.g. *uvnitř* 'inside'→ *v* 'in', cf. (5). A substitution in the opposite direction thus does not work in general, cf. (6).

(5) **Uvnitř těchto zemí** jsme navštívili hlavní a známá města. /= v těchto zemích
*'**Inside these countries** we visited the capitals and some famous cities. /= in these countries'*
(6) Byli jste někdy **v zahraničí?** /*uvnitř zahraničí
*'Have you ever been abroad (**in a foreign country**)? /*inside a foreign country'*

3.3 Disjunction of forms

One function (subfunctor) can be expressed by two or more forms that are not substitutable (in which case their meaning has to be inferable from the context). However, more forms do not imply more subfunctors. This is the case of forms *u* 'by' and *na* 'in/on' used for localization "within the given place"; cf. (2) and (7). The form *u* 'by' can only be used in this meaning in contexts restricted to a certain group of lexical items, bearing some animate and institutional features.

(7) Přespal jsem **u kamaráda**. /*na kamarádovi/*v kamarádovi
*'I slept over **at (my) friend's (place)** / *on (my) friend / *in (my) friend.'*

4 Comparative Study: Czech Spatial Prepositions and Their English Equivalents

In any language, prepositions for expressing localization are few in number but allow for a wide range of uses; this discrepancy presents a challenge for semantic analysis of spatial prepositions in a cross-linguistic perspective (Levinson - Wilkins, 2006).[5] Based on the material of the PCEDT corpus (Sect. 2.2) we compare formal realizations of the corresponding deep syntactic units, focusing on the most frequently analyzed area of spatial meanings, namely localization "within the given place". Applying

[4] Whenever we refer here to a 2D or 3D object, we have not in mind the real dimension, but we refer rather to the speaker's actual conception of the given place.

[5] Studies exploring the way how Czech and English structure spatial relationships focus especially on equivalents of particular prepositions in a parallel corpus data, either from the Czech-English perspective (Novotná, 2010; the preposition *na* 'on/in'), or from the English-Czech perspective (Kirschner, 1974; the preposition *in*). Investigating English equivalents of the most frequent Czech prepositions (i.e. *na* 'on/in', *v* 'in' and *s/se* 'with'), a semantic analysis is also carried out by Klégr et al. (2012), who classify spatial meanings 'where' into (i) location on the surface, (ii) a point in the space and (iii) a point inside the space. A systematic contrastive analysis of the meaning of English forms vis-à-vis their Czech counterparts is given by Strnadová in Dušková et al. (2006); she observes that English spatial prepositions can express more specific features of reality than the Czech ones. This corresponds to the observations of Hruška (1976), who states that English spatial adverbials display ability to differentiate more precisely various notions of place by means of a wider choice of prepositions (cf. *between* and *among*). The relations between Czech and English forms are also described in Czech textbooks of English (cf. e.g. Vít, 2019).

the principles described in Sect. 3, we have subcategorized this localization into a set of three subfunctors associated with the corresponding Czech forms, as illustrated in Table 1.

In the Czech part of the PCEDT corpus we have searched for adverbials with the LOC functor (depending on a verb) expressed by prepositional groups containing the prepositions *v* 'in', *na* 'in/on' or *u* 'by',[6] and then looked for their most frequent equivalents in the English part. The Czech-English pairs of sentences were then sorted out according to the form of the English equivalent. Finally, we have manually assigned the subfunctor of the local specification to the respective adverbials in each Czech sentence (see Table 2).[7]

Subfunctor	**Form**	**Example**
inside	*v* 'in'	(1), (6)
	uvnitř 'inside'	(5)
surface	*na* 'in/on'	(4)
at the given place	*na* 'in/on'	(2), (3)
	u 'by'	(7)

Table 1: Subfunctors for localizations "within the given place" (of LOC functor).

Czech form	**Subfunctor**	**English form**	**Number of pairs**	**Example**
na 331	*surface*	**on**	4	*na stole – on the desk*
		at	1	*na moři – at the sea*
	at the given place	**on**	147	*na trávníku – on the lawn*
		in	93	*na světě – in the world*
		at	86	*na škole – at a school*
v 3061	*inside*	**in**	2913[8]	*ve věži – in the tower*
		at	88	*v továrně – at a factory*
		on	60	*v televizi – on television*
u 18	*at the given place*	**at**	8	*u agentury – at the agency*
		in	6	*u soudu – in the court*
		with	3	*u příbuzných – with relatives*
		on	1	*u soudu – on the court*

Table 2: Czech prepositions for localization "within the given place" and their English equivalents.

In spite of the fact that the collected material is not large, certain tendencies can be followed:

(A) The equivalent for ***v*** with the subfunctor ***inside*** is mostly the form ***in*** (e.g. **inside a 3D object:** *ve vozidlech – in cars, v garáži – in a garage, v košíku – in a bask□t;* **inside a 2D area:** *v regionu – in the district, v zemi – in the country, v Číně – in □hina;* **in a piece of art**: *v knize – in a book, ve filmu – in a film, v dopisech – in th□l□tt□s;* **in a domain:** *v průmyslu – in the industry, v technologii – in technology*).

(B) The equivalent for ***na*** with the subfunctor ***surface*** is mostly the form ***on*** (e.g. **on the surface of a 3D object:** *na stole – on the table, na čepicích – on caps, na kopci – on a hill*).[9]

[6] We have not examined here the secondary preposition *uvnitř* 'inside'.

[7] We exclude cases where the equivalent in the English sentence is not a prepositional group. Since the texts in the corpus are mostly mono-thematic (economic and political texts from journals), the lexically identical pairs are counted as a single case (e.g. in Table 2, 128 occurrences of *na trhu - in the market* are counted as a single case of the equivalence). We also exclude cases of annotation mistakes and we do not work with idiomatic and fixed phraseological expressions.

[8] For the most frequent occurrence of *v-in* (2913 pairs) the first 200 different pairs have been analyzed, other figures in the Table 2 are the total numbers of the given pairs in the material analyzed.

[9] There are only few examples in our data, but the observation is confirmed by the conclusions in Klégr et al. (2012).

Other English equivalents for the subfunctors *inside* and *surface* are rather rare (cf. Table 2) and concern an oscillation described below. Only two rather conspicuous subgroups expressing localization ***inside*** can be distinguished, both with the English form *on* corresponding to the Czech form *v*, i.e. **means of communication** (e.g. *ve vysílání – on a broadcast, v rádiu – on the radio, v televizi – on television*) and **transport** (e.g. *ve vlaku – on the train*).

(C) The equivalents for ***na*** with the subfunctor ***at the given pla*****ce** are almost evenly distributed among the forms ***at, in, on.*** The prevailing tendencies are as follows:

(C-i) The form *na* with the subfunctor *at the given place* is equivalent to ***on*** first of all with the localization **on a 2D area** (e.g. *na pozemku – on the property, na podlaží – on the floor, na trávníku – on the lawn*) and **on a "line"** (e.g. *na cestě – on a path, na silnicích – on roads, na skluzavce – on the slide*).

(C-ii) The form *na* with the subfunctor *at the given place* is equivalent to ***at*** in case the localization is understood as **a special-purpose place** (such as an institution or an event*: na škole – at the college, na Institutu – at the Institute, na večírku – at a party, na konferenci – at the conference*) and in case the location is understood as **a point** (e.g. *na zastávce – at the station, v centru – at the Center*).

(C-iii) The form *na* with the subfunctor *at the given place* is equivalent to ***in*** first of all in case of the localization **inside a 2D area** (e.g. *na dvorku – in the yard, na hřbitově – in the cemetery, na severozápadě – in the Northwest*).

(D) The equivalent forms for ***u*** with the subfunctor ***at the given place*** are the prepositions ***at*, *in*** and ***with.*** If the given location is **an institution**, all the above three forms may occur (e.g. *u agentury – at the agency, u soudu – in the court, u firmy – with the firm*). If the given location is **a person**, the equivalent is primarily the preposition *with* (e.g. *u příbuzných – with relatives, u ředitele – with the director*).

4.1 Discussion

The tendencies (A) and (B), i.e. a clear equivalence of the forms ***v*** – ***in*** and ***na*** – ***on***, are very strong and support the differentiation of the opposite locations *inside – surface*. Originally, we have delimited the subfunctor *surface* as an opposition to the meaning of *inside* just with 3D objects (cf. Sect. 3). However, the tendency in (C-i) indicates a possibility to expand the scope of this subfunctor to localization "on the surface" of both 2D areas and 3D objects.

The tendency (C) confirms the vague character of the preposition *na* in Czech; it is evident that the subfunctor *at the given place* covers several meanings, which are not fixed in Czech, in contrast to English. The preposition *at* makes it possible to differentiate further semantic nuances in English, described here in a simplified way as localization at a special-purpose place or at a point, cf. (C-ii) and parallel Czech-English examples (8) and (9). In Czech, for the localization perceived as "at a special-purpose place" the preposition *u* (primarily expressing the localization „beside") is used; however, its coverage is narrower than with the English form *at*, cf. (D).

(8) Až dosud se inzeráty společnosti objevovaly téměř výlučně **v novinách** a časopisech.
Until now, the corporate ads have appeared almost exclusively **in newspapers** and magazines.
(9) Podle podmínek smlouvy, která byla uzavřena **v novinách** Toronto Star, se 500 zaměstnanců…
Under the terms of the contract reached **at** Torstar **newspaper**, the 500 workers…

The analysis of our material has also demonstrated that both languages provide a high degree of contextual substitutability of two or even more forms expressing localization with a very slight difference in meaning (cf. the three English equivalents of the only Czech expression *na trhu* 'in the market' (10)-(12)). A localization can be perceived and structured in language in the different ways with different (language) meanings. Our material reflects a specially high degree of oscillation between the expression of the meanings **inside a 2D area** versus **on the surface of a 2D area** (e.g. *v ulicích/na ulicích – on the streets/in the streets, ve světě/na světě – in th□wo□ld; na ost□ově – on the island/in the island*) and **inside**

a 3D object versus **at a special-purpose place** (e.g. *ve škole/na škole – in the school/at the school, ve/u společnosti – in the company/at the company*).

(10) There is finally some sort of sense **in the market**.
(11) It had to buy sugar **on the** world **market** to meet export commitments.
(12) They graze **at the** Farmers **Market**, a combination gourmet food court and grocery store.

There is also an appreciable established correlation between the given lexical unit and a certain preposition, which is especially frequent with toponyms but occurs also in other cases and which affects the validity of general tendencies (e.g. ***na*** *Havaji* (***v** Havaji) – ***in*** *Hawaii*).

5 Conclusion

Our analysis has confirmed that every language may structure the reality in a different way and that there may be an „*overwhelming diversity, and apparently endless mismatches between any two languages in both the formal coding of distinctions, and semantical basis for them*“ (Levinson – Wilkins, 2006, 550). Our analysis has also supported the conclusions of previously published studies that English spatial prepositions can express more specific features of reality than the Czech ones. The deep syntactic representation of the Prague Dependency Treebanks decreases the “distance” between languages, yet there does not exist a universal set of subfunctors. Cross-language studies help to explore the differences in structuring the reality and their description is useful for teaching and translation applications.

Acknowledgements

The research has been supported by the Czech Science Foundation under the project GA17-12624S and by the LINDAT/CLARIAH-CZ project of Ministry of Education, Youth and Sports of the Czech Republic (LM2018101). This research also has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

Reference

Michel Aurnague. 1995. Orientation in French spatial expressions: formal representations and inferences. *Journal of Semantics*, 12(3), 239-268.

David C. Bennett. 1975. *Spatial and Temporal Use of English Prepositions: An Essay in Stratificational Semantics*. Longman, London.

Libuše Dušková et al. 2006. *Mluvnice současné angličtiny na pozadí češtiny*. Academia, Prague.

Simon Garrod, Gillian Ferrier and Siobhan Campbell. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72, 167-189.

Jan Hajič, Eva Hajičová, Marie Mikulová and Jiří Mírovský. 2017. Prague Dependency Treebank. In: *Handbook on Linguistic Annotation*. Springer, Verlag, Berlin, 555-594.

Jan Hajič, Eduard Bejček, Allevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anja Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Maga Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, Zdeněk Žabokrtský. 2018. *Prague Dependency Treebank 3.5*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID: http://hdl.handle.net/11234/1-2621.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Jiří Toman, Zdeňka Urešová and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association, Istanbul, 3153-3160.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English.* Studies in Natural Language Processing, Cambridge: Cambridge University Press.

Jiří Hruška. 1976. An Attempt at Linguistic Characterology of Prepositions in Present Day English in Comparison with Czech. *Brno Studies in English*, 12, 125–144.

Zdeněk Kirschner (published anonymously). 1974. Some Problems of the Automatic Analysis of English Prepositional Constructions. In: *Automatické zpracování textů* (Natural Language Processing). SNTL, Prague, 86-156.

Aleš Klégr, Markéta Malá and Pavlína Šedová. 2012. *Anglické ekvivalenty nejfrekventovanějších českých předložek.* Karolinum, Prague.

Seth Lindstromberg. 2010. *English Prepositons Explained*. John Benjamins, Amsterdam/Philadelphia.

Stephen C. Levinson and David P. Wilkins. 2006. *Grammar of Space. Explorations in Cognitive Diversity*. Cambridge University Press, Cambridge.

Mitchell Marcus, Beatrice Santorini and Mary A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank.

Marie Mikulová, Eduard Bejček, Veronika Kolářová and Jarmila Panevová. 2017. Subcategorization of adverbial meanings based on corpus data. *Jazykovedný časopis*, 68(2), 268–277.

Renata Novotná. 2010. The Czech preposition *na* and its English Equivalents. In: F. Čermák, K. Kučera, V. Petkevič (eds). *Intercorp: Exploring a Multilingual Corpus*. Lidové noviny, Prague, 138–145.

Randolph Quirk, Sidney Greenbaum, Geoffrey N. Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Petr Sgall, Eva Hajičová and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Vladimír Šmilauer. 1947. *Novočeská skladba*. Ing. Mikuta, Prague.

Leonard Talmy. 2006. The fundamental system of spatial schemas in language. In: B. Hampe (ed.). *From Perception to Meaning: Image Schemas in Cognitive Linguistics*. Mouton de Gruyter, Berlin, 37–47.

Claude Vandeloise. 1991. *Spatial Prepostions: A Case Study from French*. The University of Chicago Press, Chicago/London.

Marek Vít. 2019. *Help for English*. URL: https://www.helpforenglish.cz/article/2006060502-mistni-predlozky-in-on-at.

A Spanish e-dictionary of collocations

María Auxiliadora Barrios
Universidad Complutense de Madrid
auxibarrios@filol.ucm.es

Igor Boguslavsky
Universidad Politécnica de Madrid /
IITP, Russian Academy of Sciences
iboguslavsky@fi.upm.es

Abstract

We present a new e-dictionary of Spanish (in progress) called *Diretes* (DIccionario RETicular de ESpañol). It contains descriptions of collocations by means of Lexical Functions (LFs), both standard and non-standard, in the sense of the Meaning – Text Theory by Igor Mel'čuk. At present, *Diretes* contains about 50,000 collocations. This paper concentrates on the collocations in which the collocate is an adjectival or an adverbial phrase. These collocations are mostly extracted from the *Práctico* combinatorial dictionary of modern Spanish. We explain the structure of the e-dictionary, the types of information it contains and the way it is presented. We also show how the LF-interpreted collocations can be used in NLP applications. We demonstrate it with the SemETAP semantic analyzer, in which LFs are used to normalize semantic structures and make inferences.

1 Introduction

This paper presents a Spanish e-dictionary called *Diretes*. It has several sources. The first of them is the BADELE.3000 database (Barrios and Bernardos, 2007; Barrios, 2010), which contains 25,000 collocations described by means of Lexical Functions (LFs) of the Meaning-Text Theory (MTT) (Mel'čuk, 1996, 2014). Recently, we built a new database and doubled the number of collocations, so that now *Diretes* totals about 50,000 items. An important source of data is the EsTenTen corpus (SketchEngine, https://www.sketchengine.eu/estenten-spanish-corpus). Our next step consists in incorporating the data of *Práctico* – a well-known dictionary of Spanish collocations. We aim at interpreting the *Práctico* collocations in terms of LFs, as we did in previous portions of *Diretes.* Lexical Functions have been proposed in MTT as a tool for the formalization of lexical relations and classifying collocations and some paradigmatic relations (such as synonymy, antonymy and semantic derivatives). However, standard MTT LFs cannot cover the whole material of *Práctico.* A significant part of uncovered material is presented by expressions containing adjectives and adverbs, which are our primary concern in this paper. To bring some order into this group of collocations, we widely use non-standard LFs and a set of semantic features.

This paper is structured as follows. In section 2, we summarize some relevant characteristics of two Spanish combinatorial dictionaries particularly useful for our task. In section 3, we present the Spanish e-dictionary we are building, *Diretes*. In section 4, we relate lexical resources such as *Diretes*, that store LFs, to NLP applications, that make use of LFs. Drawing on the example of the SemETAP semantic analyzer we show that LFs can be effectively used for the normalization of semantic structures and for drawing inferences. Finally, we present our conclusions and outline future work.

2 *Redes* and *Práctico*, two Spanish combinatorial dictionaries

Redes and *Práctico* (Bosque, 2004, 2006) are two Spanish combinatorial dictionaries containing a carefully selected set of collocations, which constitute the collocational knowledge of educated native speakers of Spanish. *Redes* contains 7,115 entries and *Práctico* 14,000 (Barrios, 2007). Both of them are relevant sources of semantic information, particularly *Redes*, a dictionary that offers a detailed lexical and semantic analysis. The combinatorial data are presented in this dictionary by means of lexical classes, each one described by semantic features. For instance, the entry of the adjective *férreo* '(referring to) iron' reflects first of all the primary meaning of the adjective ('made of iron') and then its figurative meanings, in which it modifies action nouns such as *control férreo* 'iron grip', nouns of physical objects used in figurative sense, as *mano férrea* 'iron fist', phrases as *regla férrea* 'iron rule', etc. For each collocation there is a real example taken from a corpus of more than 250 million words. *Redes* is a dictionary mostly intended for research purposes.

Práctico is conceived as a dictionary for practical purposes. It includes all the collocations from *Redes* and many more. It is useful, first of all, for native speakers interested in perfecting their mastery of language, for authors, translators and language learners. It gives fewer examples than *Redes* and does not use the explicit semantic classification of *Redes*, but it preserves its semantic structure. In both dictionaries, each entry contains a large number of collocations: for instance, the *Práctico* entry of the adjective *aromático* 'aromatic' shows thirteen nouns (such as the Spanish equivalents of *plant, herb, drink, wine, oil*, etc.) but not *flor* 'flower' nor *rosa* 'rose', even though in the real world flowers in general and roses in particular are aromatic often enough to expect the existence of these collocations.

Redes and *Práctico* are valuable sources of combinatorial information. As opposed to collocational material extracted from large collections of texts automatically, which often contains a lot of rubbish, materials offered by *Redes* and *Práctico* are a result of thorough individual research and exhibit the highest standard of quality. What they lack is some degree of formalization, which could render them more useful for applications. This is what we are trying to achieve in the *Diretes* project.

3 *Diretes*: A Spanish e-dictionary supplied with Lexical Functions

Electronic dictionaries are structured sets of lexicographic data in numerical form accessible in different ways and having multiple functionalities (De Schryver, 2003). Some of them are targeted at humans and some are machine-readable, which means that they are useful not only for humans but also for computers that can read their contents (Dziemianko, 2017). The problem that arises here is that even if an e-dictionary is machine-readable, its contents are designed for human consumption: in many cases, text understanding requires inferences of diverse kinds, which is still unfeasible for the machines; some of these inferences need to be based on dictionary definitions, some others are not linguistic but pragmatic or cultural (Barrios, in press).

On the other hand, NLP tools reuse different linguistic resources, such as dictionaries. In recent years, many NLP researchers are actively developing practices oriented to sharing data on the web, which are called *linked data* (Bizer et al, 2011). Different models to represent linguistic linked data have been proposed, some of them focused on lexical resources, and some others on ontologies, catalogues of linguistic data or even corpora models (Bosque-Gil et al, 2016, 2018).

Many of the new electronic dictionaries are human-oriented: collocations and meanings of lexical units are explained in a natural language rather than in a formalism suitable for machines. What we propose in *Diretes* is to create contents accessible to machines, in a way similar to some other dictionaries within the Meaning – Text approach, such as the French dictionaries Dicouèbe[1] and DiCoInfo (L'Homme, 2008), the English and Russian ETAP-4 dictionaries[2] and the Spanish dictionary of emotions DICE[3] and DiCoEnviro (Ortego Antón, 2011). At present, *Diretes* contains about 50,000 collocations. Among them, there are 551 adjectival and adverbial collocations extracted from *Práctico* and *Redes* (beginning with the letter *a*). In this paper, we concentrate on these collocations.

[1] http://olst.ling.umontreal.ca/dicouebe/index.php

[2] http://cl.iitp.ru

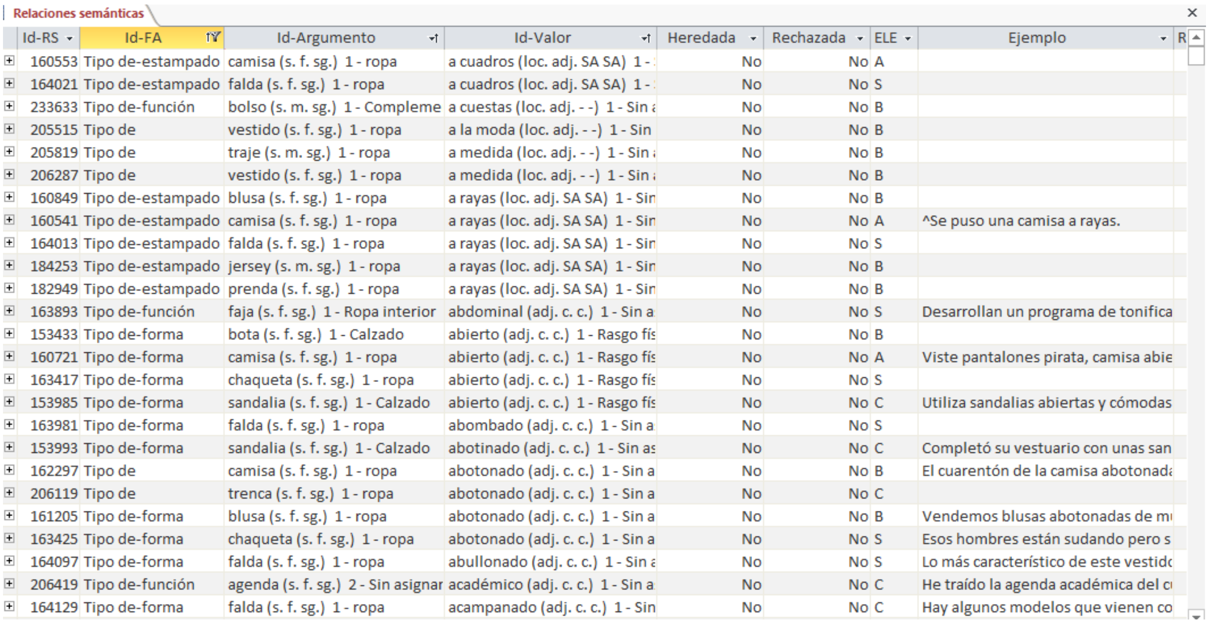
[3] http://www.dicesp.com/paginas. The Spanish dictionary DICE provides LFs for the semantic field of emotions, but emotions is only one of the 664 semantic fields of Diretes. DICE contains 200 entries, and all of them belong to the field of emotions; Diretes contains 372 entries for emotions, which provide 7737 collocations.

Diretes assigns a large amount of LF information to words. First of all, one should distinguish between standard and non-standard LFs (Mel'čuk, 2014: 173-174). As for standard LFs, adjectives and adverbs can act as values of several of them, including semantic derivatives A_i and Adv_i, plus a number of syntagmatic LFs, such as Magn (meaning 'very, to a high degree', such as *infinite* in *infinite patience*), Ver (meaning 'such as should be', e.g. *legitimate* in *legitimate demand*), Bon (meaning 'good', such as *fruitful* in *fruitful analysis*), Pos (meaning 'positive evaluation', e.g. *favourable* in *favourable opinion*), Epit (meaning 'redundant clichéd modifier', such as *sweet* in *sweet dream*). All of them are useful when formalizing not only adjective collocations but also adverbial ones. All of these LFs can combine with the LF Anti (meaning 'opposite'): if the expression *to pay an arm and a leg* is covered by Magn, then *to pay a mere trifle* and *to cost peanuts* are covered by AntiMagn.

In *Diretes,* we widely use the conceptual relation TypeOf, which denotes hypernymy (similar to LF Gener of MTT). To make the description more precise, we introduced several semantic variants of the TypeOf relation: TypeOf–form (Sp. Tipo de–forma), TypeOf–function (Sp. Tipo de–función), TypeOf–print (Sp. Tipo de–estampado) and some others.

In Fig. 1, one can see a fragment of *Diretes* illustrating some collocations of this class. Here the first column shows the identification number of each lexical relation; the second one, the name of this relation; the third, the argument of the lexical relation, its grammatical features and its semantic label (which is the name of the hypernym); the fourth shows the value of the lexical relation, its grammatical features and its semantic label; the fifth signals if this lexical relation was automatically inherited from the relation between the hypernym and the value; the sixth one is filled in manually if the automatically inherited lexical relation is incorrect (such as *ponerse el bolso* 'to put on one's bag'); the seventh suggests of the level at which this lexical relation could be learnt by students of Spanish as a second language; and the last one shows a real example of use taken from *SketchEngine.*

Finally, there is a large portion of collocations formalized by means of non-standard LFs. We classified them using some of the most productive semantic features shown in Table 1.



Relaciones semánticas

Id-RS	Id-FA	Id-Argumento	Id-Valor	Heredada	Rechazada	ELE	Ejemplo	R
160553	Tipo de-estampado	camisa (s. f. sg.) 1 - ropa	a cuadros (loc. adj. SA SA) 1 -	No	No	A		
164021	Tipo de-estampado	falda (s. f. sg.) 1 - ropa	a cuadros (loc. adj. SA SA) 1 -	No	No	S		
233633	Tipo de-función	bolso (s. m. sg.) 1 - Compleme	a cuestas (loc. adj. - -) 1 - Sin a	No	No	B		
205515	Tipo de	vestido (s. f. sg.) 1 - ropa	a la moda (loc. adj. - -) 1 - Sin	No	No	B		
205819	Tipo de	traje (s. m. sg.) 1 - ropa	a medida (loc. adj. - -) 1 - Sin a	No	No	B		
206287	Tipo de	vestido (s. f. sg.) 1 - ropa	a medida (loc. adj. - -) 1 - Sin a	No	No	B		
160849	Tipo de-estampado	blusa (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	B		
160541	Tipo de-estampado	camisa (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	A	^Se puso una camisa a rayas.	
164013	Tipo de-estampado	falda (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	S		
184253	Tipo de-estampado	jersey (s. m. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	B		
182949	Tipo de-estampado	prenda (s. f. sg.) 1 - ropa	a rayas (loc. adj. SA SA) 1 - Sin	No	No	B		
163893	Tipo de-función	faja (s. f. sg.) 1 - Ropa interior	abdominal (adj. c. c.) 1 - Sin a	No	No	S	Desarrollan un programa de tonifica	
153433	Tipo de-forma	bota (s. f. sg.) 1 - Calzado	abierto (adj. c. c.) 1 - Rasgo fís	No	No	B		
160721	Tipo de-forma	camisa (s. f. sg.) 1 - ropa	abierto (adj. c. c.) 1 - Rasgo fís	No	No	A	Viste pantalones pirata, camisa abie	
163417	Tipo de-forma	chaqueta (s. f. sg.) 1 - ropa	abierto (adj. c. c.) 1 - Rasgo fís	No	No	S		
153985	Tipo de-forma	sandalia (s. f. sg.) 1 - Calzado	abierto (adj. c. c.) 1 - Rasgo fís	No	No	C	Utiliza sandalias abiertas y cómodas	
163981	Tipo de-forma	falda (s. f. sg.) 1 - ropa	abombado (adj. c. c.) 1 - Sin a	No	No	S		
153993	Tipo de-forma	sandalia (s. f. sg.) 1 - Calzado	abotinado (adj. c. c.) 1 - Sin as	No	No	C	Completó su vestuario con unas san	
162297	Tipo de	camisa (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	B	El cuarentón de la camisa abotonada	
206119	Tipo de	trenca (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	C		
161205	Tipo de-forma	blusa (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	B	Vendemos blusas abotonadas de mu	
163425	Tipo de-forma	chaqueta (s. f. sg.) 1 - ropa	abotonado (adj. c. c.) 1 - Sin a	No	No	S	Esos hombres están sudando pero s	
164097	Tipo de-forma	falda (s. f. sg.) 1 - ropa	abullonado (adj. c. c.) 1 - Sin a	No	No	S	Lo más característico de este vestido	
206419	Tipo de-función	agenda (s. f. sg.) 2 - Sin asignar	académico (adj. c. c.) 1 - Sin a	No	No	C	He traído la agenda académica del cu	
164129	Tipo de-forma	falda (s. f. sg.) 1 - ropa	acampanado (adj. c. c.) 1 - Sin	No	No	C	Hay algunos modelos que vienen co	

Fig.1. Some adjectival collocations formalized by means of TypeOf relations in *Diretes*

Semantic Feature	Adjective/adverbial expression	Example of use
Material	*abonado* 'fertilised'	*tierra abonada* 'potting soil'
Appearance	*abierto* 'open'	*mente abierta* 'open mind'

Place	*aéreo* 'aerial'	*tráfico aéreo* 'air traffic'
Manner	*a boca jarro* 'point-blank'	*decir algo a boca jarro* 'to say something bluntly'
Cause	*abrasador* 'burning'	*sol abrasador* 'blazing sun'
Able to	*accesible* 'accessible'	*lugar accessible* 'accessible place'
Quantity	*a partes iguales* 'in equal parts'	*dividir a partes iguales* 'divide in equal parts'
Time	*anual* 'annual'	*convocatoria anual* 'annual call'
Recurrence	*asiduo* 'regular'	*orador asiduo* 'regular guest speaker'
Speed	*a toda máquina* 'at full speed'	*trabajar a toda máquina* 'to work at full speed'

Table 1. Semantic features used to characterize non-standard LFs.

In *Diretes* the words are organized as a net, not as a hypernym/hyponym hierarchy. In the latter case, a "mother" may have several "children", while a "child" may have only one mother. In *Diretes* this is not so: a "child" may have several "mothers". For instance, a word such as *reloj* 'watch' is labeled as belonging both to the class 'artifact' and 'accessory'. One of the salient features of *Diretes* is that the database was designed to implement the LF Domain Principle (which is similar but not identical to the lexical inheritance principle of (Mel'čuk & Wanner, 1996: 229)). According to this principle, most words sharing a hypernym usually develop similar collocations (Barrios, 2009; Barrios, 2010, Barrios, Bernardos 2007). Below, we present the structure of the database and then we illustrate the LF Domain Principle.

In *Diretes* the data are organized in several tables. The four most important tables are: a) lemmas; b) the hierarchy of semantic labels; c) semantic predictions; and d) semantic and lexical relations.

In the first table, the lemmas of the dictionary are tagged by semantic labels (i.e. hypernyms); for instance, *camisa* (shirt) is labeled as 'piece of clothing' and *calcetín* (sock) as 'underwear'.

In the second table, the semantic labels are structured in a hierarchy of nine levels; for instance, '*ropa y accesorios'* ('clothing and accessories') is the "mother" of *'ropa', 'zapatos'* and *accesorios'* ('clothing', 'shoes' and 'accessories'; and '*ropa*' ('clothing') is in its turn the "mother" of '*ropa interior'* ('underwear').

In the third table we predict some relations that can be inherited from "mothers" to "children"; for instance, '*ropa y accesorios'* ('clothing and accessories') is related to four verbs and its Lexical Functions are: *llevar (puesto)* 'to wear' (Real$_1$), *ponerse* 'to put sth on' (IncepReal$_1$), *quitarse* 'to take sth off' (FinReal$_1$), *estropearse* 'to get damaged' (Degrad). The semantic label '*ropa*' ('clothing') inherits these verbs and then we add manually thirteen new verbs or verbal phrases, such as *sentar bien* 'look good' (BonFact$_1$), *sentar mal* 'look bad' (AntiBonFact$_1$), *arreglar* 'to fix' (CausPredPlusVer), etc. Some of them are inherited by the "grandchild" label '*ropa interior'* 'underwear', and we add some new particular verbs such as *quedar apretado* be tight' (AntiBonFact$_1$). To sum up, in this particular case, the table of semantic predictions contains forty-six collocations represented by lexical functions to be inherited by nouns such as *camisa* 'shirt', *calcetín* 'sock', *anillo* 'ring' *or reloj* 'watch'.

In the fourth table, we collect all the collocations and semantic relations attached to each corresponding Lexical Function. To conclude with the 'clothing and accessory' example, all the nouns labeled as 'clothing', 'shoes', 'accessories', 'complements', and 'underwear' inherit all the verbs from the third table; then, in the fourth table there are 4989 collocations for nouns such as *camisa* 'shirt', labeled as 'clothing' and *calcetín* 'sock', labeled as 'underwear' (2567 were automatically inherited and 2422 were manually added); 909 collocations for nouns such as *botas* 'boots' labeled as 'shoes' (539 were automatically inherited); 1060 collocations for nouns such as *anillo* 'ring' labeled as 'accessory' (626 were automatically inherited). There are also 987 collocations for nouns such as *billetera* 'wallet' labeled as 'complements' (151 were automatically inherited).

The LF Domain Principle says that some collocations can be predicted on the basis of the LF domain, i.e. the list of words likely to be keywords of this LF. For example, we can predict that all words denoting fruits, vegetables and objects made from organic materials can be keywords of Degrad (which means 'to become permanently worse or bad'); and all words denoting artifacts can be keywords of CausFunc$_0$;

the domain of CausFunc$_0$ is the set of nouns denoting things that can be created. Semantic labels for each domain allow us to predict groups of collocations, such as *to build* for 'building' (*temple, tower, concert hall*, *castle,* etc.) and 'housing' (*apartment, flat, duplex*, etc.); *to compose* for 'text' (*poem, novel, essay*, etc.) and 'music' (*symphony, melody, sing*, etc.); *to make* for 'clothes' (*shirt, trousers, coat*, etc.) and 'food' (*cake, paella, soup*, etc.). This is applicable to many other LFs, such as LiquFunc$_0$ ('to cause something to not exist anymore'), IncepFunc0 ('to start existing'), FinFunc0 ('to finish existing'), CausFact0 ('to cause something to start to work'), LiquFact0 ('to cause something to finish working'), IncepPredPlus ('to increase'), FinPredPlus ('to decrease'), Son ('to emit a characteristic sound') (Barrios and Goddard, 2013).

The implementation of the LF domain principle, as well as the lexical inheritance principle, allows us to generate automatically thousands of collocations, and consequently it is possible to complete the lexicographic task in less time. Once both principles are applied, we can analyze the meaning in a deeper way, by means of dimensions of meaning, as Mel'čuk and Wanner propose, or even by means of primes and molecules, as Barrios and Goddard proposes for the LF Degrad after analysing some English and Spanish collocations related to this LFs: "The intuition behind the Degrad function is that there is a common semantic core to all the verbs (…) First, all the Lexico-Syntactic Frames include the following pair of components: something bad happens to something for some time; because of this, after this, this something is not like it was before. Second, with one partial exception, the explications all share the following component in the Process section: when it happens, it happens slowly, people can't see it. These three components are (arguably) enough to capture a serviceable core or 'prototype' for the intuition behind the 'Degrad' notion" (Barrios and Goddard, 2013: 239).

4 Adjectival and adverbial Lexical Functions in semantic analysis

In this section, we show that LFs stored in e-dictionaries such as *Diretes* can be effectively used in NLP applications. Let us recall that the interest that LFs aroused in the community from their very inception was largely motivated by the fact that they can be useful for different tasks, both lexicographic and related to computational linguistics. To name but a few publications, early attempts of using LFs in NLP are described in (Arsentjeva et al., 1969; Streiter, 1996; Wanner, 1996; Polguère, 1998; Mel′čuk, Wanner, 2001). Apresjan et al. (2007) explains how LFs can be used in language learning. Apresjan et al. (2002) presents LFs included in the electronic combinatorial dictionaries of Russian and English. In these dictionaries, about 50,000 Russian and 25,000 English words are supplied with LFs. It is shown that LFs can improve lexical and syntactic disambiguation during parsing, idiomatic translation in machine translation and synonymous paraphrasing. The latter task is described in detail in (Apresjan, Tsinman, 2002). In (Lambrey, Lareau, 2015) LFs are used in language generation. Formalization of LFs carried out in (Jousse, 2010; Fonseca et al., 2016) can be used for the development and efficient consulting of lexical databases.

Here, we present yet another application in which LFs can be put to good use. It is semantic analysis as implemented in the SemETAP system (Boguslavsky, 2011). The task of the semantic analyzer is to represent the meaning of the text in an explicit and unambiguous way. Two levels of semantic structure are distinguished: Basic Semantic Structure (BSemS) interprets the text in terms of ontological concepts; Enhanced Semantic Structure (EnSemS) extends BSemS by means of a series of inferences. LFs are used in SemETAP at two stages: in constructing and normalizing BSemS and in drawing inferences thereof. Below, we will illustrate both of these types.

We will call a syntactic derivative of word L such a word, or phrase, L′ that has the same (or very close) meaning as L, but belongs to a different syntactic category and hence displays a different behavior. Actantial syntactic derivatives (S_i, A_i, Adv_0, Adv_i) are in some way oriented towards one of the actants of the keyword. Such derivatives are supplied with a numerical index, which corresponds to the number of this actant.

In BSemS all predicates should be brought to the normalized form, which means that syntactic derivatives should be replaced by their keyword. In case of actantial derivatives, normalization also requires that the i-th actant of the keyword be explicitly established. Here are some examples of actantial derivatives and normalizing operations they trigger: A_1(*fear*) = *fearful, frightened* (≈ 'such that fears something'), A_2(*fear*) = *fearsome* (≈ 'such that is feared'); Adv_1(*hurry*) = *hastily* (≈ 'hurrying'), Adv_2(*permit*) = *with the permission* (≈ 'being permitted').

A_1: The *child was fearful <frightened>* ==> 'the child feared something'

A_2: *The consequences were fearsome* ==> 'one could fear the consequences'

Adv_1: *He said good bye hastily* ==> 'he said good bye; while saying it he was hurrying'

Adv_2: *The evidence was examined by the experts with the permission of the court* ==> 'the evidence was examined by the experts; the court permitted the experts to examine the evidence'.

Some examples of other types of LFs that also trigger inferences:

$Real_1$(*promise*) = *fulfil: He fulfilled his promise to help me.* Inference: 'he helped me'.

$CausFunc_0$(*crisis*): *bring about (a crisis).* Inference: 'the crisis takes place'.

$LiquFunc_0$(*beard*): *shave off (one's beard).* Inference: 'the beard exists no longer'.

5 Future work

We presented a project which aims at compiling a new e-dictionary of Spanish supplied with Lexical Functions and other information. In its current state, it contains about 50,000 lexical relations: 20,000 cover the most frequent collocations of Peninsular Spanish; that is, the collocations that any student of B2 level should master (based on frequency corpus data, Barrios, 2010). 30,000 other collocations correspond to the domain of the body and body parts, emotions, clothing and accessories. We are working now on the domain of the house, and in the next months we will work on artifacts, food and evaluation domains. Our goal is to obtain a database of 75,000 collocations described in terms of Lexical Functions by the end of 2020. Another immediate task is to significantly enlarge the set of adjectival and adverbial non-standard LFs. We have a large number of collocations in our database that are still lacking adequate description in terms of LFs. We are also planning to bring our adjectival semantic classification closer to WordNet standards.

Acknowledgements

This work is developed within the project DiRetEs (Diccionario Reticular Español: Diccionario analógico y relacional con acceso en red desde el sentido y desde la forma) and was partially supported by the Ministerio de Ciencia, Innovación y Universidades (FFI2017-83293) and the RSF grant No. 16-18-10422-P, which is gratefully acknowledged.

References

Apresjan, Ju., I. Boguslavsky, L. Iomdin, and L. Tsinman. 2002. Lexical Functions in NLP: Possible Uses. *Computational Linguistics for the New Millennium: Divergence or Synergy? Festschrift in Honour of Peter Hellwig on the occasion of his 60th Birthday*, edited by Manfred Klenner and Henriette Visser, 55–72. Frankfurt/M., Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Apresjan, Ju., L. Tsinman. 2002. Formal'naja model' perifrazirovanija predlozhenij dlja system pererabotki tekstov na estestvennyx jazykax, *Russkij jazyk v nauchnom osveschenii* 2 (4):102-146. ("A formal model of paraphrasing for NLP sysems"; In Russian)

Apresjan, Ju., P. Diachenko, A. Lazurskiy, L. Tsinman. 2007. O kompjuternom uchebnike leksiki russkogo jazyka. *Russkij jazyk v nauchnom osveschenii* 2 (14):48-112. ("Computer manual of the Russian vocabulary"; In Russian)

Arsentjeva, N., N.Balandina, A.Krasovskaja. 1969. O mashinnoj realizatsii sistemy perifrazirovanija. Institut prikladnoj matematiki AN SSSR, preprints 25, 26, 27. Moscow. ("Computer implementation of the paraphrasing system"; In Russian)

Barrios Rodríguez, Mª Auxiliadora. 2007. Diccionarios combinatorios del español: diferencias y semejanzas entre Redes y Práctico. *RedELE. Red electrónica de Didáctica del Español como Lengua Extranjera*, 11, 1 - 14.

Barrios Rodríguez, María Auxiliadora. 2009. Domain, domain features of lexical functions and generation of values by analogy according to the MTT approach. Proceedings of Fourth International Conference on Meaning-Text Theory. http://olst.ling.umontreal.ca/pdf/ProceedingsMTT09.pdf.

Barrios Rodríguez, María Auxiliadora. 2010. *El dominio de las funciones léxicas en el marco de la Teoría Sentido-Texto. Estudios de Lingüística del español* (ELiEs), vol 30. http://elies.rediris.es/elies30/index30.html.

Barrios Rodríguez, M.A. (in press) ¿Aún queda alguien para quien no exista un diccionario? *Diretes*, un diccionario electrónico apto para máquinas. In M.C. Cazorla, M.A. García Aranda, & P. Nuño (Eds.), *Homenaje a Manuel Alvar*. Lugo: Axac.

Barrios, Mª Auxiliadora; Bernardos, Socorro. 2007. BaDELE3000: An implementation of the lexical inheritance principle. In Reuther, Tillman; Wanner, Leo. (eds.) Wienner Slawistischer Almanach. Sonderband 69. 68 - 77.

Barrios, Mª Auxiliadora; Goddard, Cliff. 2013. 'Degrad verbs' in Spanish and English Collocations. Lexical Functions and contrastive NSM semantic analysis. Functions of Language, 20:2, 219-249.

Bizer, Christian, Tom Health and Tim Berners-Lee. 2011. In Amit P. Sheth (ed.), *Linked Data: The Story so Far. Semantic Services, Interoperability and Web Apllications: Emerging Concepts*. Ringgold Inc, Portland, 1–23.

Boguslavsky I. 2011. Semantic analysis based on linguistic and ontological resources. *Proceedings of the 5th International Conference on Meaning-Text Theory (MTT'2011)*. Barcelona, September 8–9, 2011. pp. 25–36.

Bosque I. 2004. REDES. Diccionario combinatorio del español contemporáneo. Las palabras en su contexto. Ediciones SM, Madrid.

Bosque I. 2006. Diccionario combinatorio PRÁCTICO del español contemporáneo. Las palabras en su contexto. Ediciones SM, Madrid.

Bosque-Gil, Julia, Jorge Gracia, Elena Montiel-Ponsoda y Guadalupe Aguado-de-Cea. 2016. Modelling multilingual lexicographic resources for the web of data: the K dictionaries case. In Kenerman, Ilan, Iztok Kosem, Simon Krek y Lars Trap-Jensen (eds.), *Proceedings GLOBALEX Lexicographic Resources for Human Language Technology* http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf

Bosque-Gil, Julia, Jorge Gracia, Elena Montiel-Ponsoda y Asunción Gómez-Pérez. 2018. Models to represent linguistic linked data, *Natural Language Engineering*, 24(6), pp. 811-859. doi:10.1017/S1351324918000347

De Schryver, G.-M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography,* 16(2):143–199.

Dziemianko, Anna. 2018. Electronic dictionaries. In Pedro A. Fuertes Oliveras (ed.), *The Routledge handbook of Lexicography*. Routledge, Oxon.

Fonseca, A., Sadat, F., and Lareau, F. 2016. A lexical ontology to represent lexical functions. In Proceedings of LangOnto'16 at LREC, Portorozˇ.

Jousse, A.-L. 2010. Modèle de structuration des relations lexicales basée sur le formalisme des fonctions lexicales. Ph.D. thesis, Université de Montréal/Université Paris 7.

L'Homme , M.C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. Traduire, 217, 78-103.

Mel′čuk, I. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in Lexicon. *Lexical Functions in Lexicography and Natural Language Processing*, edited by L. Wanner, 37–102. Amsterdam: John Benjamins Publishing Company.

Mel'čuk I. 2014. *Semantics. From meaning to text.* Volume 3. John Benjamins Publishing Company. Amsterdam/Philadelphia.

Mel′čuk I., and L. Wanner. 1996. Lexical Fucntions and Lexical Inheritance. *Lexical Functions in Lexicography and Natural Language Processing*, edited by L. Wanner, 209-278. Amsterdam: John Benjamins Publishing Company.

Mel′čuk I., and L. Wanner. 2001. Towards a Lexicographic Approach to Lexical Transfer in Machine Translation (Illustrated by the German-Russian Language Pair). *Machine Translation* 16:21–87.

Ortego Antón M.T. 2011. La compilación de DiCoEnviro en español. *Las TIC: Presente y futuro en el análisis de corpus.* Universitat Politècnica de Vaència

Polguère, A. 1998. Pour un model stratifié de la lexicalisation en génération de texte. *Traitement Automatique des Langues* 39 (2):57–76.

Streiter, O. 1996. Linguistic Modeling for Multilingual Machine Translation. Aachen: Shaker Verlag.

Wanner, L. 1996. *Lexical Functions in Lexicography and Natural Language Processing*, edited by L. Wanner. Amsterdam: John Benjamins Publishing Company.

Exceptive constructions: A Dependency-based Analysis

Mohamed Galal
Faculty of Arts
Sohag University
82524 Sohag, Egypt
mohamed_mostafa1@art.sohag.edu.eg

Sylvain Kahane
MoDyCo-UMR 7114
Paris Nanterre University
92001 Nanterre Cedex, France
sylvain.kahane@parisnanterre.fr

Yomna Safwat
Faculty of Languages
Ain Shams University
11566 Cairo, Egypt
yomna_safwat@alsun.asu.edu.eg

Abstract

The goal of this paper is to provide a description of the syntax of exceptive constructions within a dependency framework. These constructions are introduced in English by the markers *except, but, except for, apart from, other than,* etc. Examining their syntactic properties across a variety of languages shows that they imply two main types of constructions: The *paradigmatic*-EC and the *hypotactic*-EC. The first type shares many properties with coordination, and it can be integrated into the paradigmatic lists/piles phenomena in which two segments of the utterance pile up on the same syntactic position and whose most famous case is coordination.

1 Introduction

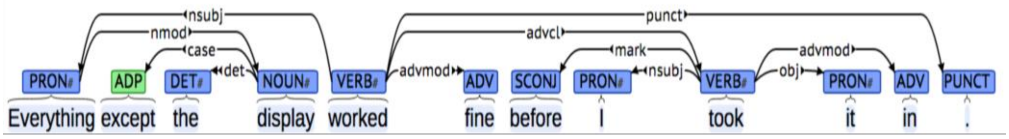
The paper aims to discuss the syntax of the exceptive constructions (henceforth ECs) within a dependency framework and across a variety of languages. These constructions are introduced in English by the markers *except, but, except for, apart from*, *other than*, etc., as exemplified by (1):

(1) a. I want to clear all variables except one. (mathworks.com)
b. We talked about everything but mock trial. (nytimes.com)
c. Netflix operates pretty much everywhere in the world except for China. (shanghai.ist)
d. No one apart from the man making the threats had been injured. (thelocal.se)

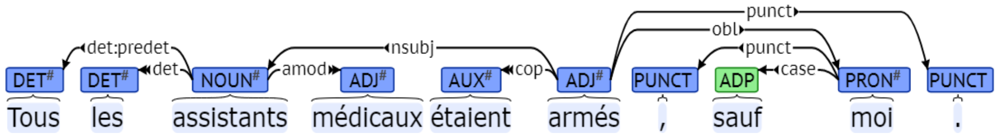
The exception is an understudied phenomenon in syntax. Many studies have been conducted on formal semantics, especially on the theory of Generalized Quantifiers (von Fintel, 1993; Gajewski, 2008; García Álvarez, 2008; Hoeksema, 1987; 1995; Lappin, 1996; Moltmann, 1992; 1995), but quite few on syntax (see Pérez-Jimenéz & Mareno-Quibén, 2012, for Spanish; Soltan, 2016, for Egyptian Arabic; and Piot, 2005; Galal & Kahane, 2018 for French).

In many languages, exceptive markers are traditionally analyzed as a preposition in dictionaries and grammars. This is the case of *but* /*except* in English (Eastwood 1994/2002) and *sauf* /*excepté* in French (Grevisse & Goosse 2008). It is also the analysis that is used in the multilingual treebanks annotated corpus Universal Dependencies (hereafter UD): *except* (2a) and *sauf* (2b) are ADP and linked by the relation *case*[1].

(2) a.



b.



Indeed, the authors consider these analyses problematic. These markers, in their exceptive use, do not have the properties of prepositions but rather those of coordinating conjunctions, since they can be followed, in addition to NPs, by PPs (3a) or AdvPs (3b). Moreover, they commute with a coordinating conjunction like *but* (3c) or a paradigmatizing adverb (see Nølke, 1983) like *even* (3d).

[1] Available at: universaldependencies.org. (2a) is from UD_English_GSD 2.4 and (2b) from UD_French_PUD 2.4.

(3) a. These snakes are found everywhere in Florida except in the Keys. (news-press.com)
b. I'm here every day except when it's a holiday and they're closed. (katc.com)
c. These snakes are found everywhere in Florida but not in the Keys.
d. These snakes are found everywhere in Florida even in the Keys.

Based on a corpus of authentic examples collected from several sources (treebanks, corpora, web, etc.), the authors suggest a binary classification of exceptive constructions. While the first construction is called the *paradigmatic*-ECs[2], which are syntactically related to coordination, the second is called the *hypotactic*-ECs, which are contrarily related to subordination. The authors tackle the exceptive markers in the paradigmatic use and analyze them as a particular case of paradigmatic lists/piles (Blanche-Benveniste 1990) in which two segments of the utterance pile up on the same syntactic position and whose most famous case is coordination.

The paper is structured as follows: Section 2 presents the two common classes of ECs observed in English and in other languages, including French, Arabic, and Spanish, and exposes the criteria adopted for the classification of the data. Section 3 is dedicated to the analysis of the paradigmatic-ECs as a particular case of paradigmatic lists/piles; a generic notion that can subsume exception and coordination.

2 The two common types of exceptive constructions in English and other languages

In this section, the two common classes of ECs observed in English and other languages are presented. In the literature on exception, a binary classification of ECs in English has been identified since Hoeksema (1987, 1995): *connected exceptives* and *free exceptives*. This classification has been adopted in other languages, e.g. Spanish (Perez-Jimenéz & Mareno-Quibén, 2012) and Egyptian Arabic (Soltan, 2016). The two types are canonically illustrated in English in (4).

(4) a. Every day but/except Sunday it was raining.
b. Except for Sunday, it was raining every day. (Hoeksema, 1987, p. 100)

The study adopts the term exceptive phrase (hereafter EP) to refer to the group consisting of an exceptive marker and a following XP such as *except Sunday* in the example (4a). Furthermore, the NP that an exception relates to *every day* is called antecedent, while the XP argument of the exceptive marker *Sunday* is be referred to as the excepted element.

Many authors, in the literature on exception, postulate that, on the one hand, the EP in the connected exceptives is associated with an NP that must contain a universal quantifier and that, on the other hand, the free exceptives are compatibles with non-universal quantifiers such as *most*, *many* and *few*, quasi-universal like *the majority*, as well as generic sentences. Note that, on the basis of attested data, this characterization is rejected in English (cf. García Álvarez, 2008)[3], in French (cf. Galal & Kahane, 2018)[4] and in Arabic (cf. Galal, 2019)[5]. The universal quantifiers are not the only ones possible in the connected exceptives. Quantifiers such as *most*, *many* and *few* are also possible[6].

Furthermore, the authors of this paper prefer to use the terms *paradigmatic*-ECs and *hypotactic*-ECs to *connected exceptives* and *free exceptive* because the term *connected* belong to English *but*-phrase that can only occur in contiguous position relative to the antecedent (García Álvarez, 2008, p. 113). On the contrary, the EP introduced by *except* in English, *sauf*/*excepté* in French and *ʾillā* in Arabic can occur in noncontiguous positions, as shown below.

[2] *Paradigmatic* vs *syntagmatic* and *hypotactic* vs *paratactic* are generally opposed. As Blanche-Benveniste (1990) has pointed out, paradigmatic constructions are also syntagmatic, since the conjuncts maintain both a paradigmatic relationship (possibility of commuting with each other) and a syntagmatic relationship (they can combine with each other). Moreover, it is not a paratactic construction, since the construction has a clearly identifiable marker (*except*, *sauf*, *ʾillā*, etc.).

[3] (i) a. Kate is an actress who has played many roles except that of a real woman.
b. Karadzic is a moderate man in most things but politics. (García Álvarez, 2008, p.13, 114)

[4] (ii) Le temps […] sera ensoleillé sur la plupart des régions française, sauf le Sud-Ouest […] (rtl.fr)

Le	temps	sera	ensoleillé	in	la plupart	régions	française	sauf	le Sud-Ouest
The	weather	will be	sunny	in	most	regions	French	except	the South-West

'The weather will be sunny in most French regions, except the South-West [...]'

[5] (iii) (الإرهاب ضرب معظم الدول إلا بريطانيا) (albawabhnews.com)

al-ʾirhāb	ḍaraba	muʿẓam	ad-duwal	ʾillā	brīṭānyā
DEF-terrorism	hit.PRES.3SG	most	DEF-country.PL	except	Great Britain

'Terrorism has hit most countries except Great Britain'

[6] However, this constraint is confirmed in Spanish. According to Pérez-Jimenéz & Mareno-Quibén (2012, p. 585), the connected constructions whose main clause does not include an expression of universal quantifiers is ungrammatical.

The classification is based on strictly syntactic criteria: (i) The linear position of the exceptive phrase, (ii) the syntactic category of the excepted element and (iii) the possibility or not to coordinate the EP.

The paradigmatic-ECs are introduced in English by the items *but* and *except*; while the hypotactic-ECs are introduced by the lexical units *except for, apart from*, *other than*, etc.

2.1 The linear position of the exceptive phrase

The EP in the paradigmatic-ECs allows only two positions. While the first position is adjacent to the antecedent (5a), the second is at the right periphery either adjacent (5b) or nonadjacent to the antecedent (5c).

(5) a. All children, except one, grow up. (goodreads.com)
b. The discount applies to everything except fuel [...]. (moneytalksnews.com)
c. Everything was great, except the weather. (tripadvisor.com)

The EP in the paradigmatic-ECs is not allow to be before the antecedent and particularly in the fronted position (6a). It does not also accept to be noncontiguous without being at the right periphery (6b).

(6) a. *Except the weather, everything was great.
b. *Everything was, except the weather, great.

The hypotactic-ECs behave differently. These constructions allow the abovementioned two syntactical positions. They can be adjacent to the antecedent (7a), postposed in a position either contiguous (7b) or noncontiguous (7c). They also, unlike the paradigmatic-ECs, allow the fronted position (7d) and the insertion in the VP (7e).

(7) a. All data except for Head Start data are from the U.S. Department of Labor [...]. (ed.gov)
b. Extreme right is gaining ground in all of Europe, except for Wallonia. (brusselstimes.com)
c. Everything is right except for the Price. (seekingalpha.com)
d. Except for killings, all crimes drop in Duterte's 1st year. (rappler.com)
e. No one was, except for the man who played him, Marion Morrison. An actor and man with true grit. (manchesterinklink.com)

2.2 The syntactic categories of the excepted element

In this section, the possibilities of the connection between the markers and the different syntactic categories of the excepted element are presented. The examination of naturally occurring data shows that the exceptive markers in the paradigmatic-EC can be combined with constituents of different parts of the speech. They can be combined with NPs, as shown in the example below, but more interestingly is that they can be followed by a PP (8a) or an AdvP (8b).

(8) a. The prison has closed-circuit cameras in every corner except in her cell. (The New York Times)
b. Lorraine Bower is just a regular graduate student, except when she's in her Army uniform. (The Daily Orange)[7]

On the contrary, the exceptive markers in the hypotactic-ECs can only be combined with an NP (9a vs b).

(9) a. I agree with everybody except with John.
b. *I agree with everybody except for with John.

2.3 The possibility or not to coordinate the sequence introduced by the markers

In the corpus, the authors have not found occurrences introduced by *but/except* in English, by *sauf/excepté* in French and by *'illā* in Arabic where the EP presents the possibility to coordinate, like in the constructed example (10a). On the contrary, the exceptive markers in the hypotactic-ECs allow the repetition before each coordinated phrase (10b).

(10) a. *I will be there every day but/except Monday and but/except Tuesday.
b. The incidence of cancer (except for cervical cancer, and except for the north-eastern state of Mizoram) is

[7] Note that the EC, in this example, is realized without the explicit presence of the antecedent. The example can, therefore, be interpreted, as follows: *Lorraine Bower is just a regular graduate student {on all times}, except when she's in her Army uniform*. This case is discussed in more detail in the following section.

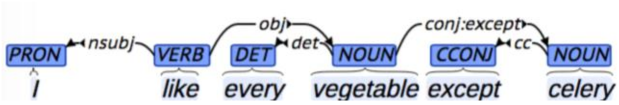
much lower than that in countries that can be said to be in a similar epidemiological transition as India […] (thelancet.com).

3 The analysis of the paradigmatic-ECs as a case of paradigmatic lists/piles constructions

The syntactic behavior of the paradigmatic-ECs testifies that these constructions behave very similarly as coordination. The fact that exception is not coordination necessarily leads the authors to introduce a notion that subsumes exception and coordination. This notion is the paradigmatic lists/piles, constructions in which two segments of the utterance pile up on the same syntactic position and whose most famous case is coordination along with other phenomena like reformulation (Blanche-Benveniste, 1990; Gerdes & Kahane, 2009; Nølke, 1983). Exception can be, therefore, analyzed in the same way as coordination.

In the UD annotation scheme, the coordination is encoded by the relation *conj* between the two conjuncts and a relation *cc* from the second conjunct to the coordinating conjunction (CCONJ). The study uses the relation *conj* for all paradigmatic relations and indicates that it is a coordination or an exception by an extension to the label: *conj:coord* for coordination and *conj:except* for exception[8].

(11)



In this construction, the EP forms a phrase with its antecedent because the EP must always be after the antecedent, but it is not necessarily contiguous to it. It must be noted that this also arises with coordination, such as the French example (12) in which the second conjunct is placed in a postponed position of the statement, without being adjacent to the first conjunct, even if it is much more common and grammaticalized with paradigmatic-ECs. This is a special case of extraposed complement (Botalla, 2019).

(12) Cela vient de l'école, ici, on est puni si on coupe la parole à un camarade. Et d'une tradition rurale encore très forte. (Est Républicain journal)

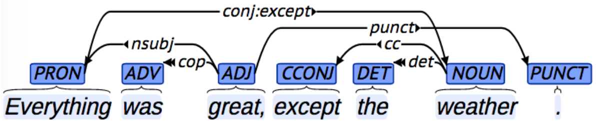
Cela	vient	de	l'école,	ici,	on	est puni	si	on	coupe	la	parole
this	comes	from	school	here	we	are punished	if	we	cut	the	word

à	un	camarade.	Et	d'une	tradition	rurale	encore	très	forte
to	a	comrade	And	from a	tradition	rural	still	very	strong

'This comes from school, here, we are punished if we cut a comrade word. And from a rural tradition still very strong'

When a coordination phrase is discontinuous, the second conjunct is systematically rejected at the right periphery, forming a new illocutionary unit. In other words, the discontinuity of ECs, as illustrated by (13), does not invalidate their analysis as paradigmatic constructions.

(13)



The ECs can occur without the explicit presence of the antecedent, especially in the case where the excepted element fulfills the function of an adverbial clause, such as in example (8) above. This property does not distinguish between

[8] There are other types of paradigmatic relations that should be considered as particular cases of *conj*. This concerns apposition such as *John, one of my friend*, which is annotated with the relation *appos* in UD, but could be annotated better *conj:appos*. For reformulation, UD proposes the relation *reparandum,* which goes from the second to the first conjunct. In some sense, paradigmatic relations are not as directed as pure dependency relations between a governor and subordinated element. In the case of a reparation, the second conjunct replaces the first and it can make sense to allocate the relation coming from the governor. Another solution would be to use a sub-type of *conj*. As shown by Blanche-Benveniste (1990), there are many cases where a reformulation is not a reparation and cannot easily be differentiated from a coordination (*she is a good linguist, a computational linguist*).

paradigmatic-EC and coordination. The absence of a first conjunct also occurs with coordination. The coordination with *and* and *but* also may have no antecedent (Gerdes & Kahane, 2009):

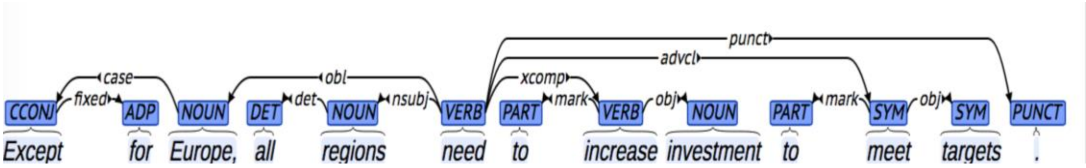
(14) a. He speaks French and well.
b. He speaks English, but badly.

In these examples, there is a coordination with two illocutionary units (Gerdes & Kahane, 2015, p. 109). In (14a), the speaker makes two assertions: *'he speaks French*' and *'he speaks French well*'.

Note that the analysis of the exceptive markers as coordinating conjunctions in this paper is supported by argumentations similar to the ones made in English by Harris (1982), Reinhart (1991), and García Álvarez (2008), in Spanish by Perez-Jimenéz & Mareno-Quibén (2012), and in Egyptian Arabic by Soltan (2016). Nevertheless, none of them introduce the concept of paradigmatic construction and properly explain the link between exception and coordination.

For the hypotactic-ECs, the EP has a much freer order and is not necessarily contiguous to the antecedent. Thus, it is no longer possible to consider that it forms a phrase with its antecedent. We consider that the EP is a PP modifying the main verb. The marker *except for* is analyzed as an idiomatic adposition (marked with the link *fixed* in UD, *except* and *for* keep their POS, the POS of the idiom does not appear, but the relation *case* indicates that it is analyzed as an adposition)[9].

(15)



3.1 The third type of exceptive constructions in Arabic: Paratactic-ECs

In Modern Standard Arabic, there is a problem concerning the analysis of the EC introduced by *'illā* + ACC as a paradigmatic list construction. According to the grammatical system of Arabic, the NP that follows *'illā* in affirmative ECs systematically takes the accusative case whatever the case of its antecedent. In negative ECs, either it takes the accusative case, or it takes the same case as the one assigned to its antecedent. This accusative case goes against the analysis of this construction as a paradigmatic construction and of *'illā* as a coordinating conjunction, since in a coordinating construction the two conjuncts usually carry the same grammatical case. It also goes against the analysis of *'illā* as a preposition because, in Arabic, prepositions are always followed by the genitive, while the accusative is used for direct objects of verbs.

In fact, the identification of the governor of this accusative case in the NP followed by *'illā* in the affirmative construction has been the subject of vivid debates between Arabic grammarians since the eighth century. Eight different analyses have been suggested by the ancient Arab grammarians. One of them is proposed by the grammarians of the Koufa School in the ninth century (Al-Anbary, XII^e [1961, p. 261]) considering that the particle *'illā* itself which imposes the accusative case on its complement. According to this analysis, *'illā* replaces an ellipsed verb meaning *'astaṯnī* (أستثني 'I except/I make the exception') (16). This analysis, therefore, considers the EC as a binary construction formed of two juxtaposed clauses.

(16) a. (حضر الوزراء إلا وزيرَ البترول)
ḥaḍara al-wūzarāʾ ʾillā wazīr-a al-bitrūl
come.PAST.3SG DEF-minister.PL except minister-ACC DEF-petroleum
'The Ministers came except the Minister of Petroleum'

b. (حضر الوزراء، أستثني وزيرَ البترول)
ḥaḍara al-wūzarāʾ ʾastaṯnī wazīr-a al-bitrūl
come.PAST.3SG DEF-minister.PL except.PRES.1SG minister-ACC DEF-petroleum
'[The Ministers came], [I except the Minister of Petroleum]'

[9] It could be possible to introduce a sub-relation *case:except* in order to have a common feature *except* for every exceptive constructions.

The authors argue that the construction *'illā* + ACC is a paratactic construction that is common in Arabic, where two clauses are juxtaposed and form a unique illocutionary unit (17).

(17) (رأي علىٌّ الأولادَ يلعبون)

raʾa	ʿaliyy-u-n	al-ʾawlād-a	yalʿab-ūna
see.PAST.3SG	Ali-NOM-INDEF	DEF.children.PL-ACC	play.PRES.3PL

Lit. Ali saw the children they play
'Ali saw the children playing '

In the *'illā* + ACC construction, the EP must be at the right periphery (which is the canonical position of paratactic clause). It does not allow either the fronted position (18) or the position contiguous to its antecedent but in fronted position relative to the verb (19a vs b).

(18) *(إلا واحداً هذه قطتي مات جميع أولادها)

*ʾillā	wāḥid-a-n	haḏihi	qiṭa=tī	māta	ǧamīʿ	awlāda=hā
except	one-ACC-INDEF	DEM	cat=PRO	die.PAST.3SG	all	children.PL=PRO

Lit. that is my cat, have been dead, except one, all his children'

(19) a. (كل أبطال المشهد رحلوا إلا واحداً) (elwatannews.com)

kull	ʾabṭāl	al-mašhad	raḥalū	ʾillā	wāḥid-a-n
all	star.PL	DEF-scene	die.PAST.3PL	except	one-ACC-INDEF

'All the stars of the scene are dead, except one'

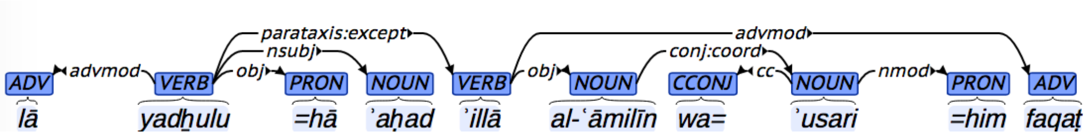
b. *(كل أبطال المشهد إلا واحداً رحلوا)

*kull	ʾabṭāl	al-mašhad	ʾillā	wāḥid-a-n	raḥalū
all	star.PL	DEF-scene	except	one-ACC-INDEF	die.PAST.3PL

Lit. All the stars of the scene, except one, are dead.

The authors agree with the traditional Arabic grammar considering that *'illā* in this construction has a verbal form. According to this analysis, *'illā* + ACC is a binary construction formed of two juxtaposed clauses. In the UD analysis (20), *'illā* will be categorized as a verb and will be linked with the relation *parataxis:except* for the paratactic-EC.

(20)



Nobody enters, except company staff and their families only'

4 Conclusion

In this paper, a syntactic description of exceptive constructions (ECs) within a dependency framework was proposed. Based on the distributional properties of the exceptive phrase, on the combinatorial possibilities of the exceptive markers with different parts of speech and on their (in)ability to coordinate, the authors suggested a binary classification of exceptive constructions observed in a many languages: the *paradigmatic*-ECs and the *hypotactic*-ECs (eventually a tripartite classification in Arabic with *paratactic*-ECs). The study considers, moreover, that the markers in the paradigmatic-ECs are coordinating conjunctions and can be integrated into the paradigmatic lists/piles constructions, a generic notion that can subsume both coordination and exception, and in which two segments of the utterance pile up on the same syntactic position.

References

Al-Anbary (XII^{e}). *ʾasrāru al-ʿarabiyyah* (أسرار العربية). [Habboud, B. Y (ed.). 2010. Beyrouth: Dar Al-Arqam].

Blanche-Benveniste, C. et al. 1990. *Le français parlé : études grammaticales*. Paris: CNRS.

Botalla, M.-A. 2019. *Modélisation de la production des énoncés averbaux: le cas des compléments différés*. PhD Dissertation. Sorbonne Nouvelle.

Eastwood, J. 1994/2002. *Oxford guide to English grammar*. Oxford: Oxford University Press.

Gajewski, J. 2008. NPI *any* and connected exceptive phrases, *Language Semantics*, 16(1). 69-110.

Galal, M., S. Kahane. 2018. Les constructions exceptives vues comme des listes paradigmatiques : à propos de la syntaxe de sauf, excepté, hormis… en français. *Proceedings of the 6th World Congress of French Linguistics* (CMLF), Mons, 1-21.

Galal, M. 2019. *Les constructions exceptives du français et de l'arabe: syntaxe et interface sémantique-syntaxe*. Ph.D. Dissertation, Paris Nanterre University & Sohag University.

García Álvarez, I. 2008. *Generality and exception. A study in the semantics of exceptives*. Ph.D. Dissertation, Stanford University.

Gerdes K., S. Kahane. 2009. Speaking in piles: Paradigmatic annotation of French spoken corpus, *Proceedings of the fifth Corpus Linguistics Conference*, Liverpool.

Gerdes K., S. Kahane. 2015. Non-constituent coordination and other coordinative constructions as dependency graphs, *Proceedings of the third international conference on Dependency Linguistics* (*Depling*), Uppsala, 10 p.

Grevisse, M., A. Goosse. 2008. *Le bon usage. Grammaire française*. Bruxelles: De Boeck-Duculot.

Harris, Z. 1982. *A Grammar of English on Mathematical Principles*. New York: John Wiley & Sons.

Hoeksema, J. 1987. The logic of exception. *Proceedings of the Fourth Eastern States Conference on Linguistics*, Columbus, OH, 100-113.

Hoeksema, J. 1995. The semantics of exception phrases, *Quantifiers, logic, and language*, 145 -177.

Kahane S. (with the participation of K. Gerdes, P. Pietrandrea, C. Benzitoun, R. Bawden). 2013. Protocole of micro-syntactic annotation, Guidelines of Rhapsodie treebank for spoken French, www.projet-rhapsodie.fr, 64 p.

Lappin, S. 1996. Generalized quantifiers, exception phrases, and logicality, *Journal of Semantics*, 13, 197-220.

Moltmann, F. 1992. *Coordination and comparatives*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Moltmann, F. 1995. Exception sentences and polyadic quantification. *Linguistics and Philosophy*, 18, 223–280.

Nølke, H. 1983. *Les adverbes paradigmatisants: fonction et analyse*. Copenhague: Akademisk Forlag.

Pérez-Jimenéz I., N. Mareno-Quibén. 2012. On the Syntax of Exceptions. Evidence from Spanish, *Lingua*, 122, 582–607.

Piot, M. 2005. Sur la nature des fausses prépositions *sauf* et *excepté*, *Journal of French Language Studies*, 5, 297-314.

Reinhart, T. 1991. Elliptic conjunctions – non-quantificational QR. In Kasher, A. (ed.), *The Chomskian Turn*. Blackwell, Cambridge, MA, 360-384.

Soltan, U. 2016. On the syntax of exceptive constructions in Egyptian Arabic, *Perspectives on Arabic linguistics*, XXVII, 35-57.

von Fintel, K. 1993. Exceptive constructions, *Natural Language Semantics*, 1, 360-384.

Quantitative Analysis on verb valency evolution of Chinese[1]

Bingli Liu

Institute of Chinese and Culture Education Studies, Huaqiao University, Xiamen, 361021, China

bingli_liu@yahoo.com

Chunshan Xu

School of Foreign Studies, Anhui Jianzhu University, Anhui, 230601, China

adinxu@126.com

Abstract

The paper aims at studying the evolution of syntactic valency of Chinese verbs. We construct three corpora of ancient classical Chinese， ancient vernacular Chinese and modern vernacular Chinese. From these corpora, ten main verbs are selected to probe into the evolution of their valency, namely, their complements and adjuncts. The paper reveals that the syntactic structures has a trend toward complex. The ancient classical Chinese and the ancient vernacular Chinese are similar in sentence structure. With the transformation from the ancient vernacular to the modern vernacular, syntactic complexity increases dramatically, indicating drastic changes in sentence structure.

1. Introduction

Valency is a property of words (Tesnière, 1959). It refers to the ability of words to syntactically or semantically to combine with other words (Liu, 2007). It is determined by the meaning of the word itself. The valency relations realized in sentences are dependence relationships between words (Liu, 2009). Quantitative investigations into valency may reveal some syntactic and semantic features of human language. Based on the German valency dictionary, Köhler studies some quantitative characteristics of the German verbs valency (Köhler, 2000; 2005; 2007). Čech et al. (2010) studied Czech valency framework distribution and verified the hypothesis about the relationship between the number of valency frames and the word length. And they proposed the concept of "full valency" without distinguishing between complements and declaratives. Liu(2011) conducted quantitative studies into English verb valency and concludes that the number of meanings of English verbs obeys the positive negative binomial distributions -- the more meanings a word has, the bigger the valency is.

However, most of these studies focus on the synchronic description of the word valency. but ignore the diachronic changes of word valency. Meanwhile, the description and study of the valencys in a corpora are not balanced in every historical period and there is a lack of

[1] Huaqiao University’s Academic Project Support by the Fundamental Research Funds for the Central Universities. No.19SKGC-QG04

uniform descriptive principles and methods. Our studies try to answer the question that how the main verbs valency evolve during the long period from ancient classical Chinese to modern vernacular Chinese.

2. Methods and Materials

This study statistically analyzes 2,817 examples from ancient classical Chinese, ancient vernacular Chinese and modern vernacular Chinese. Being different from the existing researches on valency, this paper is concerning with the macroscopic and the diachronic picture of Chinese verb valency. The study explores the evolution of the "syntactic value" of verbs in real corpus. Not only will the present study help us to learn more about the context of language development, but also indicate the possible role of the valency in natural language

processing.

The diachronic researches involve comparison and contrast of different diachronic stages of a language. Historically, Chinese can be roughly divided into ancient classical Chinese, ancient vernacular Chinese and modern vernacular Chinese. The ancient classical Chinese dates back from 1600 BC to 618 AD; the ancient vernacular Chinese dates back from 618 to 1911; 1912 is often taken as the year dividing the modern vernacular Chinese and the ancient vernacular Chinese.

In order to reflect the overall linguistic properties of each period, we try to cover as many genres as possible when constructing the corpora. The corpus of ancient classical Chinese language include *Zuozhuan* (narrative chronicle), *Lv Shi Chun Qiu* (a book on political theory), *Liu Tao* (a book on military strategies), *Shangshu* (government archives), *Mencius* (quotations from a sage), *Xunzi* (a book on philosophical treatise), *Zhanguoce, Shiji, Han Shu, Sanguozhi, Houhanshu* (5 books on history), *Guoxiaoshuogoucheng, Shishuoxinyu* (novels); the corpus of ancient vernacular Chinese include samples from *Dunhuangbianwenji, Qingpingshantanghuaben, Xixiangji, Sanguoyanyi, Chukepaianjinqi, Erkepaianjinqi, Shuihuzhuan* and *Xiyouji* (novels or playbooks); the corpus of modern vernacular Chinese mainly include samples from novels.

Ten verbs are selected because of their diachronic lexeme stability: 走(walk)，听(listen), 到(arrive), 爱(love),有(have), 为(be), 能(can), 来(come), 使(let) , 愿(wish).

In total, we annotated 3,128 sentences, of which 1,383 are from ancient classical Chinese, 813 from ancient vernacular Chinese, and 932 from modern vernacular Chinese. Table 1 shows the frequencies of sentence containing the above 10 verbs.

	ancient classical Chinese	ancient vernacular Chinese	modern vernacular Chinese
到(arrive)	166	52	102
来(come)	91	97	78
爱(love)	98	51	70
能(can)	211	100	100
使(let)	210	62	97

听(listen)	71	76	72
为(be)	201	100	100
有(have)	207	116	104
愿(wish)	53	108	151
走(walk)	75	51	58
Total	1383	813	932

Table 1. Number of sentence containing the 10 verbs

To study the valency of verbs, we need sentences where the verbs appear. We take the following criteria to select sentences:

(1) The verb is used in the active voice

(2) The verb has similar semantic meaning across different periods

Then we begin to annotate the sentences which include the verbs chosen according to the dependency grammar. Basically, syntactic dependency can be roughly divided into two types: complement and adjunct(Liu, 2011). Complement relationships involve arguments like subjects, objects or complements. The adjunct relationships often involve adverbials and attributives.

3. Results and Discussion

3.1 Increasing complexity of subjects and objects

According to their syntactic complexity, subjects and objects of these verbs can be divided into two categories: simple subject and complex subject. Simple subjects are single words, such as nouns and pronouns, while complex subjects include phrases, such as numerical-classifier phrase, noun phrases and verb phrases. Figure 1 and Figure 2 show the ratio of complex subjects and complex objects in three forms of Chinese.

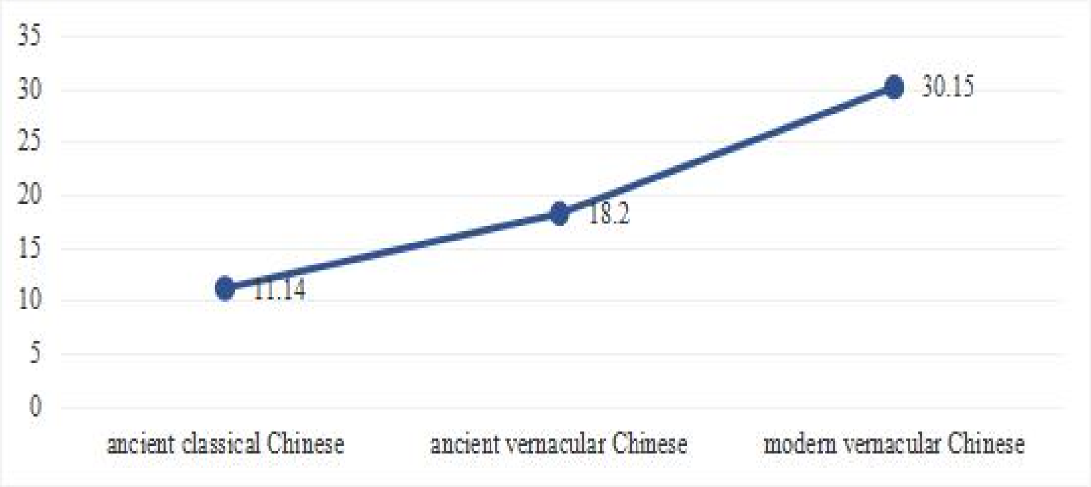


Figure 1. Average ratio of the complex constituent in the subject argument（%）

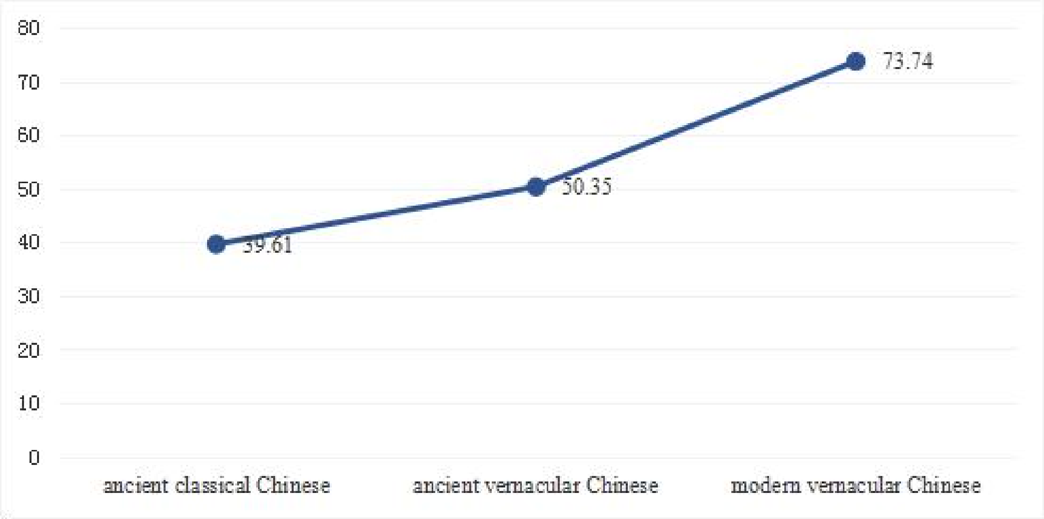


Figure 2. Average ratio of the complex constituent in the object argument（%）

Generally, the complexity of the subjects increases in ancient classical language, ancient vernacular Chinese and modern vernacular Chinese, as indicated by the increasing ratio of complex subjects. The complexity of the objects also increases diachronically, as indicated by the increasing ratio of complex objects in Figure 2. This suggests a tendency in Chinese to evolve into more complexity.

H0 is that all the complex sign has nothing to do with the time. We use SPSS to do the chi-square test. Result for Figure 1 and Figure 2 are *p*<0.001. It means the difference in Figure 1 and Figure 2 are both highly significant. The reason for this significant difference is complex since language is a complex adaptive system. Maybe because writing is more and more convenient with the society development or the human thinking becomes more complex.

3.2 Increasing use of Complements and Adjuncts

Among the ten verbs, seven verbs may take complements: 到(arrive), 听(listen), 走(walk), 来(come), 为(be), 有(have), 爱(love). Figure 3 shows the average ratio of other constituent in the complement.

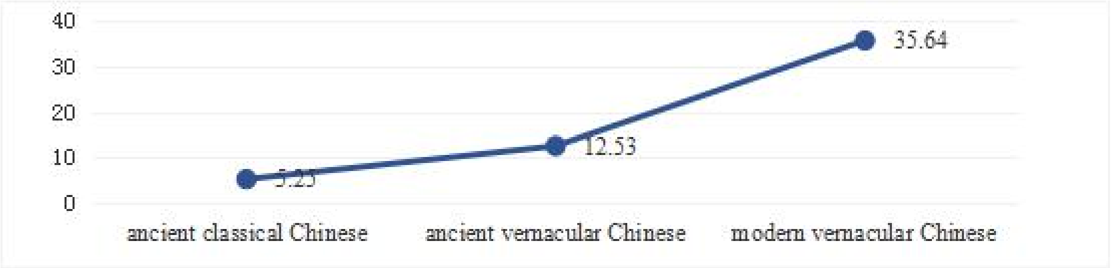


Figure 3. Average Ratio of other constituent in the complement（%）

Diachronically, these verbs show a growing tendency to take complements. From ancient classical Chinese to ancient vernacular Chinese, the proportion of complements increases by 7.28%；from ancient vernacular Chinese to modern vernacular Chinese, the proportion increases by 23.11%. At the same time, the types of complements increase from 5 to 9. In short, the diversity and the frequency of complements both increase.

Chi-square test result for Figure 3 is that *p*<0.001. It means the difference is also highly significant. Besides the reasons mentioned above, maybe it is also relevant to the factors inside the language such as parts of speech function.

Not only have the complements been used more frequently, but also the adjuncts. In the present study, we are mainly concerning with two types of adjunct valency: the adverbial and the topic

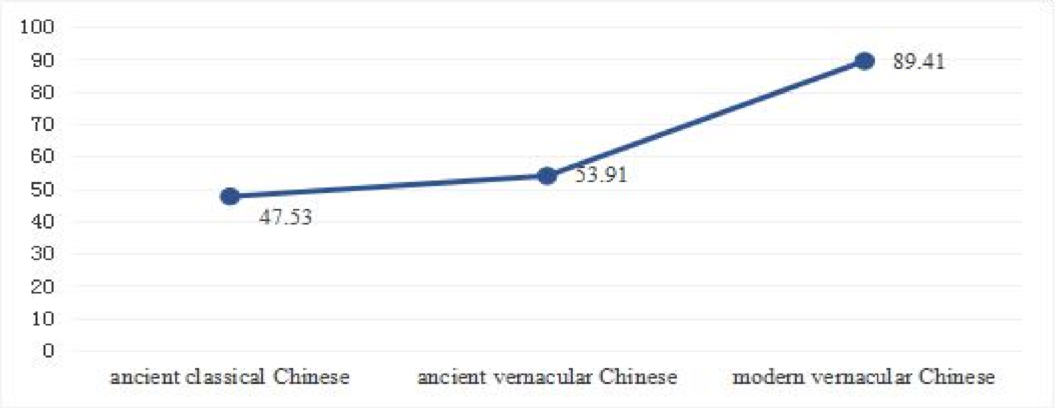


Figure 4. Average Ratio of the adjunct in the sentence (%)

The statistical data from figure 4 shows that the frequencies of these two types of adjunct valency increases diachronically. From the ancient classical Chinese to the ancient vernacular Chinese, the proportion increases by 6.38%, while from the ancient vernacular to the modern vernacular, the percentage increases, drastically, by 35.5%. These results strongly suggest that the tendency toward more complexity is not merely found in nominal constructions, or, in the valency patterns of nouns, but also in verbal constructions, or, in the valency patterns of verbs.

Chi-square test result for Figure 4 is *p*<0.001 It means the difference is also highly significant(*p*<0.001). And it means syntactic structures has a trend toward complex.

The findings presented in the above tables are diagrammed in Figure 5.

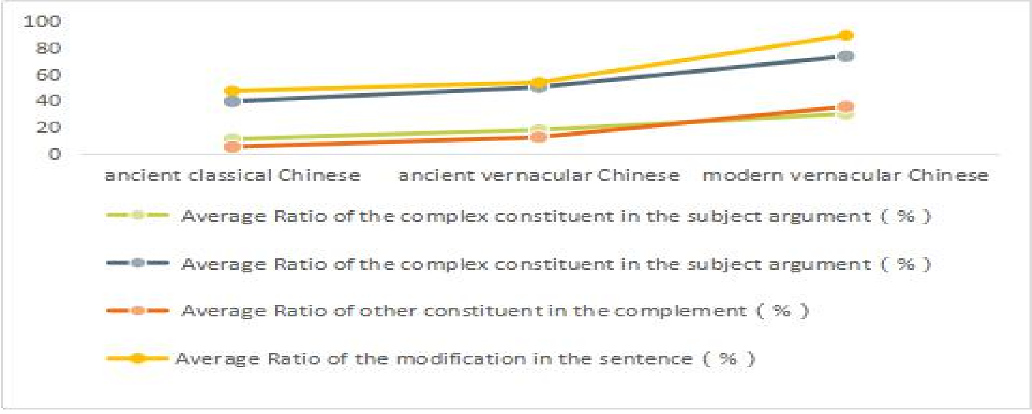


Figure 5. The valency evolution of the main verbs in the three forms of Chinese (%)

4. Conclusion

This quantitative study suggests that Chinese syntax changes gradually with time. As the Chinese language passed through the three stages, that is, the ancient classical Chinese, the ancient vernacular Chinese and the modern vernacular Chinese. The syntactic structure indicates a tendency toward increasing complexity. In other words, the valency patterns of both nouns and verbs have evolved into growing complexity. Moreover, the ancient classical Chinese and the ancient vernacular Chinese are more similar in valency patterns. The transition from the ancient vernacular Chinese to the modern vernacular Chinese seems to be drastically increased in the syntactic complexity.

And the main corpus of this paper is written language, which does not reflect the whole picture of Chinese. The next step will be to expand the scope of the study and adopt more representative oral corpus to carry out statistical analysis so as to explore the evolvement of Chinese valency laydown as far as possible.

References

Liu Haitao. 2007. Building and using a Chinese dependency treebank. *GrKG/Humankybernetik*, 48(1):3 -14.

Liu Haitao. 2009. Probalility Distribution of Dependencies Based on a Chinese Dependency Treebank, *Journal of Quantitative Linguistics*, volume. 16:256-273.

Liu Haitao. 2011. Quantitative Properties of English Verb Valency. *Journal of Quantitative Linguistics*, volume 18:207-233.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Radek Čech, Peter Pajas and Ján Mačutek. 2010. Full Valency. Verb Valency without Distinguishing Complements and Adjuncts. *Journal of Quantitative Linguistics*, Volume 17: 291-302.

Reinhard Köhler. 2005. Quantitative Studies of Valency of German Verbs. *Glottmetrics*, Volume 9:13-20.

Reinhard Köhler and Gabriel Altmann. 2000. Probability Distributions of Syntactic Units and Properties. *Journal of Quantitative Linguistics,* Volume 7:189-200.

Reinhard Köhler. 2007. Quantitative Analysis of Syntactic Structures in the Framework of Synergetic Linguistics. *Studies in Fuzziness and Soft Computing*, Volume 209:191-209.

Association metrics in neural transition-based dependency parsing

Patricia Fischer **Sebastian Pütz** **Daniël de Kok**
SFB 833
University of Tübingen, Germany
{patricia.fischer, sebastian.puetz, daniel.de-kok}@uni-tuebingen.de

Abstract

Lexical preferences encoded as association metrics have been shown to improve performance on structural ambiguities that are still challenging for modern parsers. This paper introduces a mechanism to include lexical preferences into a neural transition-based dependency parser for German. We compare pointwise mutual information (PMI) and embedding-based scores. Both the PMI-based model and the embedding-based model outperform the baseline significantly. The best model is PMI-based and increases overall performance by 0.26 LAS points over the baseline.

1 Introduction

Structural ambiguities that cannot be solved purely on the basis of structural preferences still pose a major challenge to syntactic parsing. Prepositional phrase (PP) attachment and subject-object inversion are two examples of such ambiguities. Table 1 gives an overview of the most frequent parser errors in a German newspaper corpus of 20K sentences and 350K tokens, parsed by the De Kok and Hinrichs (2016) parser with 92.01 labeled attachment score. It shows that more than one third of all errors involves prepositions, subjects and accusative objects.

Relation	**Error count**	**Percent of all errors**
Prepositional phrase/object	6,861	25.62
Adverbial	3,106	11.60
Conjunction	2,391	8.92
Accusative object	1,608	6.00
Subject	1,577	5.81
Total error count	26,775	100.00

Table 1: Five most frequent parser errors by dependency label of the parser by De Kok and Hinrichs (2016) for a German newspaper corpus. More than one third of all errors involves prepositions, subjects and accusative objects.

Resolving such ambiguities often requires context information or world knowledge. In Example 1, the direct object *Problem* ‘problem’ is fronted. The parser, however, learns from training data a preference for the unmarked word order with sentence-initial subject. *Problem* would therefore be misclassified as subject. Additionally, both *Problem* and *Post* ‘post’ are ambiguous between nominative and accusative case. Information on the sentence level thus does not suffice to decide on the correct attachment. Contextual knowledge reveals that *Problem* typically attaches to *lösen* ‘to solve’ as direct object.

Semantic preferences can provide further disambiguation cues. The verb *lösen* prefers an animate subject and an inanimate direct object. In Example 1, both *Problem* and *Post* are inanimate. World knowledge is necessary to interpret *Post* as the (animate) group of postal employees. Such knowledge can be learned from large corpora. Semantic preferences then yield the correct analysis of *Post* as animate subject and *Problem* as inanimate direct object of *lösen*.

(1) *Dieses Problem muß auch die Post noch lösen .*
This problem has-to also the post still solve .

'*The German Federal Post Office still has to solve this problem.*'

Pointwise mutual information (PMI, Fano (1961)) has been used to measure selectional preferences (Church and Hanks, 1990). PMI indicates how much two words occur together more often than chance. In the example above, a high PMI of *lösen* and *Problem* in $verb \rightarrow direct\ object$ relations would already provide enough information to solve the subject-object ambiguity. As PMIs are ideally calculated from large corpora, they provide additional context information.

In more traditional analyses of dependency distributions, it has been shown that PMI is very beneficial to solve structural ambiguities such as PP attachment (Hindle and Rooth, 1993; Ratnaparkhi, 1998; Volk, 2002). In parsing, bilexical preferences have been used by Van Noord (2007) to improve syntactic ambiguity resolution in a Maximum-Entropy parser for Dutch. Kiperwasser and Goldberg (2015) extended bilexical preferences to contextual association scores based on PMI and dependency embeddings (Levy and Goldberg, 2014a) in a graph-based parser. Mirroshandel and Nasr (2016) integrated selectional preferences into a graph-based dependency parser.

Recent approaches to neural dependency parsing (Chen and Manning, 2014; Kiperwasser and Goldberg, 2016; Dozat and Manning, 2017) implicitly encode information about co-occurrences through vector representations of the token input (Mikolov et al., 2013). However, De Kok et al. (2017) have shown for PP attachment that neural models can still benefit from information provided by PMI scores.

This paper argues that bilexical preferences are also useful in neural transition-based dependency parsing. The two main contributions are 1) a methodology to apply bilexical preferences to neural transition-based dependency parsing, and 2) an evaluation of two types of association metrics in a neural dependency parser. Results confirm that association metrics benefit neural dependency parsing. The best association score models outperform the baseline by 0.26 LAS points and improve performance on two ambiguity solving tasks by up to 2.33 points.

2 Bilexical Preferences in Neural Dependency Parsing

2.1 Approach

Transition-based dependency parsing is the task of establishing dependency relations between tokens (Kübler et al., 2009). Typically, unprocessed tokens are put on a buffer β, and a stack σ keeps track of the partially processed tokens. In the transition system used in this work, sometimes called the stack-projective system, attachments are made between the token on top of the stack and the second token on the stack (Nivre, 2004). A LEFTARC transition attaches the second token on the stack as a dependent of the token on top of the stack with relation $r \in R$, and vice versa for a RIGHTARC transition:

$$\begin{array}{ll} \textsc{LeftArc} & (\sigma|i|j,\beta,R) \rightarrow (\sigma|j,\beta,R \cup \{j,r,i\}) \\ \textsc{RightArc} & (\sigma|i|j,\beta,R) \rightarrow (\sigma|i,\beta,R \cup \{i,r,j\}) \\ \textsc{Shift} & (\sigma,i|\beta,R) \rightarrow (\sigma|i,\beta,R) \end{array}$$

Association scores can inform a parser about whether an attachment with a particular dependency relation should be made between two attachment sites. For each parser state, two attachments are possible with any of the dependency relations that are available in that system.[1] Association scores for all possible attachments provide disambiguation cues at each state. They are added to the feature vector that is used as input to the transition classifier. Association score vectors enhance existing vector representations of words, part-of-speech tags, characters, dependency relations and morphological features.

2.2 Parser Integration

For each parser state, association scores are retrieved for LEFTARC and RIGHTARC transitions, and for all possible dependency relations. Equation 1 defines the association score vector for a stack-projective

[1] A third option is to apply a SHIFT transition which does not introduce an attachment.

transition system with transitions between the token on top of the stack s_0 and the second token on the stack s_1:

$$v_{assoc} = [assoc(s_0, s_1, r),\ assoc(s_1, s_0, r) \mid \forall r \in R] \tag{1}$$

Example 2 provides the resulting association score vector in a stack-projective system with a dependency relation set that contains the *subject*, *object* and *preposition* relation.

(2) $$v_{assoc} = [assoc(s_0, s_1, subject),\ assoc(s_0, s_1, object),\ assoc(s_0, s_1, preposition),\\ assoc(s_1, s_0, subject),\ assoc(s_1, s_0, object),\ assoc(s_1, s_0, preposition)]\\ \text{with } R = \{subject, object, preposition\}$$

If no association score is available for a dependency triple, a default is assigned. An optional binary indicator $b \in \{0, 1\}$ specifies whether the dependency triple was known. This makes it possible for the model to distinguish between the default value and association strengths that overlap with the default value. The binary indicators are added to the association score vector. The association score vectors are concatenated with the remaining input feature vectors to represent a parser configuration.[2]

2.3 Association Metric Variants

Pointwise mutual information. Traditionally, PMI has been a means to capture bilexical preferences. Normalized (NPMI, Bouma (2009)) and positive normalized PMI (PNPMI, Van de Cruys (2011)) with add-1 Laplace smoothing have been tested in the parsing model. Given the dependency triple $h \xrightarrow{r} d$, consisting of the head h, dependent d and dependency relation r, PMI is defined as:

$$PMI(h \xrightarrow{r} d) = log \frac{p(h \xrightarrow{r} d)}{p(h \xrightarrow{r})\ p(\xrightarrow{r} d)} \tag{2}$$

The probability of h and d as heads and dependents with relation r is represented as $p(h \xrightarrow{r})$ and $p(\xrightarrow{r} d)$, the dependency triple probability as $p(h \xrightarrow{r} d)$. Normalized PMI

$$PMI_{norm}(h \xrightarrow{r} d) = \frac{PMI(h \xrightarrow{r} d)}{-log\ p(h \xrightarrow{r} d)} \tag{3}$$

is a more easily interpretable variant of PMI, limiting the range of PMIs to lie between -1 and 1. Positive PMI

$$PMI_{pos}(h \xrightarrow{r} d) = max(PMI(h \xrightarrow{r} d), 0) \tag{4}$$

rounds negative PMIs to 0.

Dependency embedding scores. PMI is likely to suffer from sparseness of dependency triples in the training data. Previous attempts have used back-off models (Collins and Brooks, 1995) to counteract this problem. The dependency embedding model by Levy and Goldberg (2014a) estimates probabilities for unseen triples $h \xrightarrow{r} d$ from word embeddings. The model predicts the probability $p(1|h \xrightarrow{r} d)$ of a dependency triple. Words are represented as embeddings that are trained jointly with the classifier $p(1|h \xrightarrow{r} d)$.

An embedding-based association score for the head word embedding W_h and the context embedding $C_{d,r}$ of the dependent d that is related to a head h via the dependency relation r can be formulated as:

$$assoc_{dep}(h \xrightarrow{r} d) = p(1|h \xrightarrow{r} d) = \sigma(W_h \cdot C_{d,r}) \tag{5}$$

where $C \in \mathbb{R}^{|V| \times r \times d}$ and $W \in \mathbb{R}^{|V| \times d}$. In the current model, the maximum entropy probability of 0.5 is assigned as a default when no embedding for h, d or both is available and no score can be calculated. Further model variations also include a binary indicator to distinguish the default score from a calculated embedding-based score. In a more finegrained binary indicator model, the indicator informs the parser

[2] A complete list of parser input features can be found in Appendix A.

for which of the two tokens no embedding was available.

Levy and Goldberg (2014b) have shown that the skip-gram model is an implicit factorization of the shifted PMI matrix of word co-occurrences. Dependency embeddings (Levy and Goldberg, 2014a) therefore implicitly factorize the shifted PMI matrix of head-dependent co-occurrences. Hence, association scores based on dependency embeddings (Kiperwasser and Goldberg, 2015) can be seen as correlated with PMIs.

3 Experiments

3.1 Experimental Setup

The neural transition-based dependency parser of De Kok and Hinrichs (2016) serves as the baseline for the experiments. Words, part-of-speech tags and characters are represented as vectors that were trained with structured skip-gram (Ling et al., 2015). Topological fields are used as additional input features. The parser does pseudo-projective parsing (Nivre and Nilsson, 2005) and was trained on the shuffled TüBa-D/Z (Telljohann et al., 2017) that contains 105K sentences and 1.9M tokens of manually labeled data from the Berliner Tageszeitung (taz). Non-gold part-of-speech tags were trained via 10-fold jackknifing on the TüBa-D/Z.[3] The data was split in a 7:1:2 ratio for respectively training, development and testing. Association scores are retrieved for lowercased word forms to increase lexical coverage. Common and proper nouns are typically capitalized in German and were therefore not lowercased.

Results are presented as labeled (LAS) and unlabeled attachment scores (UAS) including punctuation. Accuracies for inversion and prepositions indicate performance on resolving ambiguities. Inversion accuracy reports correct labeling of subjects and objects in clauses with fronted object. Preposition accuracy comprises all correct heads and labels of prepositional phrases and objects. The test set contains 1,887 cases of inversion (5.82 percent of all clauses) and 31,687 prepositional phrases and objects.

3.2 PMIs in Neural Dependency Parsing

A table of PMIs was generated for dependency triples $h \xrightarrow{r} d$ from the German newspaper taz (393.7M tokens, 22.8M sentences) and a dump of the German Wikipedia from January 2018 (803.5M tokens, 39.9M sentences), two subcorpora of the TüBa-D/DP treebank (De Kok and Pütz, 2019) parsed by the De Kok and Hinrichs (2016) parser without association scores. All dependency triples not contained in the table are mapped to the most neutral value of 0. The PMI table is generated once in linear time. The same holds for the dependency embeddings described in Section 3.3. Each association score retrieval is then done in constant time so that the linear time property of parsing remains unchanged.

Model	**LAS**	**UAS**	**Inversion accuracy**	**Preposition accuracy**
De Kok and Hinrichs (2016)	92.01	93.88	81.03	77.80
+ NPMI, minfreq 5	**92.27**	**94.01**	81.93	78.60
+ NPMI, minfreq 50	92.14	93.92	82.25	78.29
+ NPMI, minfreq 100	92.16	93.92	80.72	78.56
+ NPMI, minfreq 5, binary	92.18	93.94	**82.57**	**78.78**
+ NPMI, minfreq 50, binary	92.16	93.93	81.93	78.35
+ NPMI, minfreq 100, binary	92.18	93.96	81.67	78.29
+ PNPMI, minfreq 5	92.21	93.99	82.09	78.44
+ PNPMI, minfreq 50	92.19	93.95	81.46	78.66
+ PNPMI, minfreq 100	92.17	93.94	82.25	78.57

Table 2: Parser accuracy (overall, inversion, preposition attachment) for neural dependency parsing with PMI-based association scores. The NPMI model with minimum frequency 5 achieves the best overall performance.

[3] Using the *sticker* software package: https://github.com/danieldk/sticker.

PMI models with minimum dependency triple frequencies of 5, 50 and 100 have been trained with both NPMI and PNPMI scores. NPMI models have been tested with and without binary indicator. Results for the PMI models are given in Table 2.

The best PMI model uses normalized PMI with a minimum frequency of 5. The model outperforms the baseline by 0.26 LAS points which is significant in the Wilcoxon test (Dror et al., 2018) with $p < 5.24 \times 10^{-10}$. It also improves the LAS by 0.03 points over the best embedding-based model but the improvement is not statistically significant. Larger improvements can be seen for both sorts of ambiguity. The best model increases inversion LAS by 1.54 points and preposition LAS by 0.98 points over the baseline.

3.3 Dependency Embedding Scores in Neural Dependency Parsing

For the embedding-based model, dependency embeddings with 300 dimensions were trained with the algorithm from Levy and Goldberg (2014a).[4] Different embeddings have been trained on pseudo-projectivized and non-projective versions of taz, Wikipedia, and the German europarl (1.25B tokens and 42.1M sentences in total). The number of dependency relations varies from 38 non-projective to 212 pseudo-projective relations.

All embedding variants have been trained on regular head-dependent and inverse dependent-head relations. A fully typed model was trained on context that includes the token typed per dependency relation. A second semi-typed model includes the token without dependency relations as context. For both models, variants with and without binary indicator have been evaluated. The binary model uses a simple binary indicator which labels association scores as default or as being calculated from dependency embeddings. A more finegrained triple-binary model for fully typed embeddings evaluates the following three conditions to true or false: 1) the head word embedding could be retrieved from the focus matrix, 2) the dependent word embedding, i.e. the combination of the context token and the dependency relation, could be retrieved from the context matrix, 3) an embedding for the context token could be retrieved from the focus word matrix, indicating whether there exists a word embedding for the token at all. The double-binary model for semi-typed embeddings indicates whether an embedding has been found for the focus and the context token. As the context token is not typed for dependencies in the semi-typed model, the context matrix contains entries for tokens without the different dependency relations they occur with.

Model	**LAS**	**UAS**	**Inversion accuracy**	**Preposition accuracy**
De Kok and Hinrichs (2016)	92.01	93.88	81.03	77.80
+ projective, fully typed	92.23	**93.97**	82.57	78.55
+ projective, fully typed, binary	**92.24**	**93.97**	**83.36**	78.47
+ projective, fully typed, triple-binary	92.16	93.88	**83.36**	**78.62**
+ projective, semi-typed	92.11	93.94	80.66	77.99
+ projective, semi-typed, binary	91.98	93.89	80.61	77.71
+ projective, semi-typed, double-binary	92.07	93.93	81.93	77.98
+ non-projective, fully typed	92.17	93.93	81.46	78.17
+ non-projective, fully typed, binary	92.22	**93.97**	82.20	78.45
+ non-projective, fully typed, triple-binary	92.08	93.86	82.99	78.26

Table 3: Parser accuracy (overall, inversion, preposition attachment) for neural dependency parsing with embedding-based association scores. The overall best model uses projectivized, fully typed dependency embeddings with a binary indicator.

Results for parsing with association scores based on dependency embeddings are shown in Table 3. The overall best embedding-based model uses projectivized, fully typed embeddings with a binary indicator. The model outperforms the baseline parser by 0.23 LAS points, significant in the Wilcoxon

[4] Using the *finalfrontier* software package: https://finalfusion.github.io/finalfrontier.

test ($p < 1.94 \times 10^{-7}$), and remains 0.03 points below the best PMI model. Embedding-based models are only superior to PMI models when it comes to inversion LAS. There, the best embedding-based model improves by 2.33 points over the baseline, compared to 1.54 points improvement of the best PMI model.

4 Evaluation

Both the PMI-based and embedding-based models perform better than the baseline. Overall performance will improve by more correctly solved ambiguous attachments. Lexical associations between more than two tokens may be necessary to further improve ambiguity resolution. For PP attachment, the compatibility between the preposition, its modifier noun and the verbal or nominal head candidate of the PP have to be modeled. De Kok et al. (2017) have shown that trilexical preferences help to better capture attachment preferences of the preposition.

It can also be beneficial to make competing attachment sites available to the parser. Currently, association scores are only computed for the two attachment candidates for any given parser state. With beam search, several attachment candidates can compete in different analyses. The best candidate can then be chosen from all or the *n* best candidates (Zhang and Clark, 2008; Andor, 2016).

5 Ambiguity Resolution with Association Metrics

Most parser errors still involve a limited number of dependency relations, as shown in Table 1. Errors in PP attachment, subjects and objects often can be traced back to problems with resolving ambiguities. An evaluation of association scores for particular word pairs can show if such scores can be useful in parsing ambiguous sentences. Table 4 lists PMI- and embedding-based scores for selected word pairs and dependency relations. Random pairs that are common in everyday language are distinguished from pairs that occur in subject-object inversion and have been incorrectly attached by the (best-performing embedding-based) parser. PMIs have been retrieved from the positive normalized PMI table with minimum frequency 5. Embedding-based scores were calculated from projectivized, fully typed dependency embeddings.

	PNPMI		**Embedding-based**		**Example**
Relation	*Subject*	*Object$_{acc}$*	*Subject*	*Object$_{acc}$*	
Random pairs					
isst, sie	–	0.0617	0.9778	0.9863	Sie isst Spaghetti.
isst, Spaghetti	–	–	0.1375	0.9996	*'She eats Spaghetti.'*
trinkt, Mann	0.1341	–	0.9883	0.8776	Der Mann trinkt Milch.
trinkt , Milch	–	0.3627	0.9509	0.9997	*'The man drinks milk.'*
weiß, Computer	–	–	0.9280	0.1397	Ein Computer weiß alles.
weiß, alles	–	0.1995	0.9847	0.9948	*'A computer knows everything.'*
Incorrectly attached inversion pairs					
erstatteten, Angeklagten	–	–	0.9917	0.9545	Strafanzeigen erstatteten die Angeklagten
erstatteten, Strafanzeige	0.4645	0.5604	0.9566	0.9996	*'The defendants pressed criminal charges'*
wollte, niemand	0.1906	0.1750	0.9940	0.9366	Nur wollte den Krempel niemand.
wollte, Krempel	–	–	0.5458	0.0008	*'But nobody wanted that junk.'*
tragen, Studierenden	0.0794	–	0.9645	0.7761	Das Risiko tragen die Studierenden.
tragen, Risiko	0.0825	0.2540	0.9269	0.9972	*'The students take the risk.'*

Table 4: PMI and embedding-based scores for random and incorrectly attached dependency triples.

Problems of data sparsity can indeed be solved by using embedding-based rather than PMI-based scores, as Table 4 shows. In spite of a low frequency threshold of 5, the PMI table is very sparse compared to the embedding-based scores. However, when a PMI is available scores indicate the correct tendency in the majority of the cases. Considering that all unknown values are equal to the default PMI of 0.0, the tendencies are correct for e.g. *trinkt* 'drinks' which prefers to attach *Mann* 'man' as the subject and *Milch* 'milk' as the direct object. The tendencies of embedding-based scores are mostly correct, such as the preference of *Spaghetti* 'spaghetti' to attach to *isst* 'eats' as a direct object. Wider lexical coverage of embedding-based models may not lead to any gains over PMI-based models partially due to the architecture of the neural dependency parser which already encodes information about co-occurrences in the distributional representations of the input tokens.

6 Conclusion

This paper presented a technique to include association metrics into a neural transition-based dependency parser for German. PMI and embedding-based association scores have been tested. Both PMI-based and embedding-based models significantly outperform the baseline. In spite of the wider lexical coverage of embedding-based models, PMI models achieve accuracies on a par with embedding-based models.

A qualitative analysis revealed that association scores in parts provide useful disambiguation cues to the parser. Follow-up experiments in other languages with relatively free word order and moderately complex morphology will further investigate the effect of association metrics on neural transition-based dependency parsing. Due to its similarity to German, Dutch will be the first language to be examined. Trilexical rather than bilexical preferences could further improve results. Keeping more competing attachment candidates through beam search is another promising direction for future work. As an alternative to association scores, a compatibility model that is directly integrated into the parser could be considered.

Acknowledgments

Financial support for the research reported in this paper was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center "The Construction of Meaning" (SFB 833), project A3.

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally Normalized Transition-Based Neural Networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*:2442–2452. Berlin, Germany.

Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*:31–40. Tübingen, Germany.

Danqi Chen and Christopher D. Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*:740–750. Doha, Qatar.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Michael Collins and James Brooks. 1995. Prepositional Phrase Attachment through a Backed-Off Model. *Proceedings of the Third Workshop on Very Large Corpora*:27–38. Cambridge, MA.

Daniël de Kok and Erhard Hinrichs. 2016. Transition-Based Dependency Parsing with Topological Fields. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers*:1–7. Berlin, Germany.

Daniël de Kok, Jianqiang Ma, Corina Dima, and Erhard Hinrichs. 2017. PP Attachment: Where do We Stand? *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*:311–317. Valencia, Spain.

Daniël de Kok and Sebastian Pütz. 2019. Tüba-D/DP Stylebook. *https://github.com/sfb833-a3/tueba-ddp/blob/master/stylebook/stylebook-r4.pdf* [last visited on 05/23/2019].

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. *International Conference on Learning Representations*. Toulon, France.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*:1383–1392. Melbourne, Australia.

Robert Mario Fano. 1961. *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, MA.

Donald Hindle and Mats Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1):103–120.

Eliyahu Kiperwasser and Yoav Goldberg. 2015. Semi-Supervised Dependency Parsing Using Bilexical Contextual Features from Auto-Parsed Data. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*:1348–1353. Lisbon, Portugal.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing.* Morgan and Claypool, San Rafael, CA.

Omer Levy and Yoav Goldberg. 2014a. Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting for the Association for Computational Linguistics, Volume 2: Short Papers*:302–308. Baltimore, MD.

Omer Levy and Yoav Goldberg. 2014b. Neural Word Embeddings Implicit Matrix Factorization. *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2014*:2177–2185.

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*:1299–1304. Denver, CO.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at the International Conference on Learning Representations 2013*:1–12. Scottsdale, AZ.

Seyed Abolghasem Mirroshandel and Alexis Nasr. 2016. Integrating Selectional Constraints and Subcategorization Frames in a Dependency Parser. *Computational Linguistics*, 42(1):55–90.

Joakim Nivre. 2004. Incrementality in Deterministic Dependency Parsing. *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*:50–57. Barcelona, Spain.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-Projective Dependency Parsing. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*:99–106. Ann Arbor, MI.

Adwait Ratnaparkhi. 1998. Statistical Models for Unsupervised Prepositional Phrase Attachment. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*:1079–1085. Montréal, Canada.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2017. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). *http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1707.pdf* [last visited on 05/08/2019].

Tim van de Cruys. 2011. Two Multivariate Generalizations of Pointwise Mutual Information. *Proceedings of the Workshop on Distributional Semantics and Compositionality*:16–20. Portland, OR.

Gertjan van Noord. 2007. Using Self-Trained Bilexical Preferences to Improve Disambiguation Accuracy. *Proceedings of the 10th Conference on Parsing Technologies*:1–10. Prague, Czech Republic.

Martin Volk. 2002. Combining Unsupervised and Supervised Methods for PP Attachment Disambiguation. *Proceedings of the 19th International Conference on Computational Linguistics, Volume 1*:1–7. Taipei, Taiwan.

Yue Zhang and Stephen Clark. 2008. A Tale of Two Parsers: Investigating and Combining Graph-Based and Transition-Based Dependency Parsing Using Beam-Search. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing: Volume 2, Short Papers*:562–571. Honolulu, HI.

Appendix A. Parser Inputs

The parser uses vector representations of the word (`TOKEN`), its part-of-speech tag (`TAG`), dependency relation (`DEPREL`), characters (`CHAR`) and topological field (`TF`) as inputs. Different positions on the stack and buffer are addressed for each feature. The full list of features is provided below. [`BUFFER 0`], for example, refers to the first token on the buffer, [`STACK 0, LDEP 0`] addresses the leftmost dependent of the token on top of the stack. Character representations are included for the word prefix and suffix each of length 4.

```
[STACK 0] TOKEN
[STACK 1] TOKEN
[STACK 2] TOKEN
[STACK 3] TOKEN
[BUFFER 0] TOKEN
[BUFFER 1] TOKEN
[BUFFER 2] TOKEN
[STACK 0, LDEP 0] TOKEN
[STACK 1, LDEP 0] TOKEN
[STACK 0, RDEP 0] TOKEN
[STACK 1, RDEP 0] TOKEN

[STACK 0] TAG
[STACK 1] TAG
[STACK 2] TAG
[STACK 3] TAG
[BUFFER 0] TAG
[BUFFER 1] TAG
[BUFFER 2] TAG
[STACK 0, LDEP 0] TAG
[STACK 1, LDEP 0] TAG
[STACK 0, RDEP 0] TAG
[STACK 1, RDEP 0] TAG

[STACK 0] DEPREL
[STACK 0, LDEP 0] DEPREL
[STACK 1, LDEP 0] DEPREL
[STACK 0, RDEP 0] DEPREL
[STACK 1, RDEP 0] DEPREL

[STACK 0] CHAR 4 4
[STACK 1] CHAR 4 4
[BUFFER 0] CHAR 4 4

[STACK 0] TF
[STACK 1] TF
[STACK 2] TF
[STACK 3] TF
[BUFFER 0] TF
[BUFFER 1] TF
[BUFFER 2] TF
```

Presenting TWITTIRÒ-UD: An Italian Twitter Treebank in Universal Dependencies

Alessandra Teresa Cignarella
Dipartimento di Informatica, Università degli Studi di Torino
PRHLT Research Center, Universitat Politècnica de València
cigna@di.unito.it

Cristina Bosco
Dipartimento di Informatica
Università degli Studi di Torino
bosco@di.unito.it

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
prosso@dsic.upv.es

Abstract

In this paper we describe the early stage application of the Universal Dependencies to an Italian corpus from social media developed for shared tasks related to irony and stance detection. The development of this novel resource (TWITTIRÒ-UD) serves a twofold goal: it enriches the scenario of treebanks for social media and for Italian, and it paves the way for a more reliable extraction of a larger variety of morphological and syntactic features to be used by sentiment analysis tools. On the one hand, social media texts are especially hard to parse and the limited amount of resources for training and testing NLP tools further damages the situation. On the other hand, we thought that adding the Universal Dependencies format to the fine-grained annotation for irony, that was previously applied on TWITTIRÒ, might meaningfully help in the investigation of possible relationships between syntax and semantics of the uses of figurative language, irony in particular.

1 Introduction

In the last decade, the interest towards social networking sites has grown considerably and the NLP community has been relying more and more on data extracted from social media and micro-blogs. In particular, thanks to the APIs provided by the platform, and the fact there is a variety of expressions of people's sentiments and opinions, Twitter has become one of the most exploited sources for the retrieval of data, especially in the fields of Sentiment Analysis (SA) and Opinion Mining. Nevertheless, although humans can understand each other while they exchange social media contents, which are featured by non-standard word-forms, misspelled words, dialectal word-forms, emojis and elongated words, dealing with them still proves to be a very hard challenge for automatic analyses, especially concerning syntax and morphology.

In this paper we introduce a novel Twitter treebank for Italian, i.e. TWITTIRÒ-UD. The data come from a resource originally developed for training and testing irony detection systems, also exploited as a benchmark for the Italian irony detection task held in EVALITA 2018[1] (Cignarella et al., 2018b). In order to pave the way towards collecting evidences about the relationships between syntax and semantic knowledge involved in SA tasks we are developing this project of annotation which encompasses in TWITTIRÒ-UD both the fine-grained annotation for irony applied in a multilingual setting in Karoui et al. (2017) and that morphological and syntactic provided by Universal Dependencies (UD). An alike resource will allow us to extract morphological and syntactic features to be used to improve the performance in irony and stance detection tasks (Duric and Song, 2012; Sidorov et al., 2014). The UD resources available for Italian and social media meaningfully helped us in the morphological and syntactic analysis of the dataset (Bosco et al., 2014; Sanguinetti et al., 2017; Sanguinetti et al., 2018).

This paper is organized as follows. The next section briefly surveys the literature about Italian social media UD resources. Section 3 introduces the dataset used for our project and describes the various

[1] http://www.evalita.it/2018

annotation steps. In Sections 4 we discuss the creation of the gold standard set, and we highlight the findings of a quantitative analysis. Finally, in Section 5 we draw some considerations on the current state of the project and give some insights on future work.

2 Related Work

In recent years UD have become the standard for syntactic annotation (De Marneffe et al., 2014; Nivre et al., 2016) and the repository of UD projects enlarges by the day, also including data for under-resourced languages and less studied varieties, see e.g. Wang et al. (2017). As far as Italian is concerned, the main UD resources, that we exploited as reference, are two: namely, the UD-Italian treebank (Simi et al., 2014) and PoSTWITA-UD (Sanguinetti et al., 2017; Sanguinetti et al., 2018). The former entails standard texts drawn from newspapers, legal codes and Wikipedia, the latter texts from social media.

The genre of social media texts can be a bottleneck for morphological and syntactic analysis, but some experiments are reported in literature about parsing this type of data, see e.g. (Foster et al., 2011) and (Kong et al., 2014), who introduce the dependency parser TWEEBOPARSER and TWEEBANK, a Twitter treebank later extended in TWEEBANK V2 (Liu et al., 2018). In Albogamy and Ramsay (2017) an Arabic dependency treebank of tweets is converted in the UD format, while in (Blodgett et al., 2018) a treebank of tweets in African-American English is created, and in Bhat et al. (2018) a UD treebank of Hindi-English is created focusing on syntactic aspects of code-switching.

Finally, addressing the morphological analysis of social media, the task organized in the 2016's edition of EVALITA[2] can be cited. In this edition of the evaluation campaign for NLP and speech tools for Italian, a task about PoS-tagging of social media texts has been organized (Bosco et al., 2016) which was centered on the POSTWITA corpus, i.e. that later enriched with UD annotation for creating PoSTWITA-UD. This kind of experience encourages the community to adapt NLP tools to this different type of text domain, which is noisy and difficult to deal with automatically.

3 Data and Annotation

The data of TWITTIRÒ-UD are drawn from TWITTIRÒ (Cignarella et al., 2018a; Cignarella et al., 2019), a gold standard Italian corpus for irony detection. It has been firstly annotated according to the fine-grained schema for irony proposed in Karoui et al. (2017). Later it has been extended with the annotation for sarcasm exploited in the EVALITA 2018 task on irony detection in Italian tweets (IronITA[3]) (Cignarella et al., 2018b). The corpus includes 1,424 tweets annotated as follows.

```
# sent_id = 507111702744162304
# twittiro = EXPLICIT HYPERBOLE
# sarcasm = 0
# text = se sento ancora la parola merito vomito #labuonascuola #chenonèquelladirenzi
```

In the tweet[4] two features are marked for irony, i.e. the fact that all the elements necessary for interpreting the irony are lexically represented in the post (EXPLICIT), and that a particular device (HYPERBOLE) triggers irony, while a binary annotation has been applied for marking the (absence of) sarcasm. In TWITTIRÒ-UD, this annotation manually provided and revised in the original resource is enhanced by that for morphology and syntax according to UD (see examples in Sec. 3.1).

In order to create TWITTIRÒ-UD, we applied the full pipeline of tokenization, lemmatization, PoS-tagging and dependency parsing provided by *UDPipe*[5] (Straka and Straková, 2017). For this purpose, we trained UDPipe on two different gold benchmarks, namely PoSTWITA-UD (Sanguinetti et al., 2018) (6,712 tokens) and UD_Italian (Simi et al., 2014) (14,167 tokens). Considering the typology of text and the features of ironic messages, we followed the PoSTWITA-UD tenets, in particular for what concerns segmentation, which is at tweet level rather than at sentence level.

[2] http://www.evalita.it/2016/tasks/postwita.

[3] http://www.di.unito.it/~tutreeb/ironita-evalita18/index.html

[4] Translation: if I hear again the word merit I will throw up #labuonascuola #thatisnotthatofrenzi.

[5] In the UDPipe pipeline, the parsing is performed using Parsito (http://ufal.mff.cuni.cz/parsito).

3.1 Issues in Manual Correction

In this paper we focus on a subset of the original corpus, which includes 897 tweets only, while we plan a second release in the UD repository including the full corpus for November 2019. From the manual correction of this dataset[6] we have already learned some interesting lessons.

Tokenization

Several tokenization errors depend on misspelled words (i.e. not correctly separated by spaces) or punctuation irregularly used, like in the following example.

```
# sent_id = 516493351034826752
# twittiro = EXPLICIT RHETORICAL QUESTION
# sarcasm = 0
# text = @User #labuonascuola deve riconoscere il merito di chi ha superato il concorso...solo in Italia chi
vince perde?#dalleparoleaifatti

1 @User @User SYM SYM _ 4 vocative:mention _ _
2 #labuonascuola #labuonascuola SYM SYM _ 4 nsubj _ _
3 deve dovere AUX VM Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 4 aux _ _
4 riconoscere riconoscere VERB V VerbForm=Inf 0 root _ _
5 il il DET RD Definite=Def|Gender=Masc|Number=Sing|PronType=Art 6 det _ _
6 merito merito NOUN S Gender=Masc|Number=Sing 4 obj _ _
7 di di ADP E _ 8 case _ _
8 chi chi PRON PR PronType=Rel 6 nmod _ _
9 ha avere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 10 aux _ _
10 superato superare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part 8 acl:relcl _ _
11 il il DET RD Definite=Def|Gender=Masc|Number=Sing|PronType=Art 12 det _ _
12 concorso...solo concorso...solo NOUN S Gender=Masc|Number=Sing 10 obj _ _
13 in in ADP E _ 14 case _ _
14 Italia Italia PROPN SP _ 12 nmod _ _
15 chi chi PRON PR PronType=Rel 17 nsubj _ _
16 vince vincere VERB V Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 15 acl:relcl _ _
17 perde?#dalleparoleaifatti perde?#dalleparoleaifatti CCONJ CC _ 4 cc _ SpacesAfter=\n
```

In line 12 and line 17 of the tweet[7] we find "concorso...solo" and "perde?#dalleparoleaifatti", which should be split in three different tokens each. In order to avoid that the failures in tokenization propagate in the other annotation levels, before tokenization we applied an automatic data cleaning which consists in always adding a white space between words and punctuation signs (with the exception of the apostrophe which left attached to the preceding token). We only manually corrected the remaining cases of misspelled tokens, that is not separated by the necessary white space. The result of the correction of the example above can be seen below (where we also corrected the PoS tags).

```
# sent_id = 516493351034826752
# twittiro = EXPLICIT RHETORICAL QUESTION
# sarcasm = 0
# text = @User #labuonascuola deve riconoscere il merito di chi ha superato il concorso...solo in Italia chi
vince perde?#dalleparoleaifatti

1 @User @User SYM SYM _ 4 vocative:mention _ _
2 #labuonascuola #labuonascuola SYM SYM _ 4 nsubj _ _
3 deve dovere AUX VM Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 4 aux _ _
4 riconoscere riconoscere VERB V VerbForm=Inf 0 root _ _
5 il il DET RD Definite=Def|Gender=Masc|Number=Sing|PronType=Art 6 det _ _
6 merito merito NOUN S Gender=Masc|Number=Sing 4 obj _ _
7 di di ADP E _ 8 case _ _
8 chi chi PRON PR PronType=Rel 6 nmod _ _
9 ha avere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 10 aux _ _
10 superato superare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part 8 acl:relcl _ _
11 il il DET RD Definite=Def|Gender=Masc|Number=Sing|PronType=Art 12 det _ _
12 concorso concorso NOUN S Gender=Masc|Number=Sing 10 obj _ SpaceAfter=No
13 ... ... PUNCT FS _ 10 punct _ SpaceAfter=No
14 solo solo ADV B _ 16 advmod _ _
```

[6] We exploited the *Dependency Grammar Annotator*: `http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/`.

[7] Translation: @User #labuonascuola needs to acknowledge the merit of whose who passed the competition...only in Italy who wins also loses? #fromwordstofacts.

```
15 in in ADP E _ 16 case _ _
16 Italia Italia PROPN SP _ 4 obl _ _
17 chi chi PRON PR PronType=Rel 19 nsubj _ _
18 vince vincere VERB V Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 17 acl:relcl _ _
19 perde perdere VERB V Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 4 acl:relcl _
SpaceAfter=No
20 ? ? PUNCT FS _ 19 punct _ SpaceAfter=No
21 #dalleparoleaifatti #dalleparoleaifatti SYM SYM _ 4 parataxis:hashtag _ SpaceAfter=\n
```

Lemmatization and PoS-tagging

Misspelled forms often occurring in social media contents cannot be recognized by lemmatizers and their analysis may result in a failure. Here, as it was done in the annotation of PoSTWITA-UD, we associated the non-standard forms with the lemmas of their normalized versions, thus allowing a correct PoS-tagging. For instance, the typo *anema* is paired with the lemma *anima* (soul), the abbreviation *ke* with *che* (that), the elongated *nooo* with *no* (no), and the abbreviations *X* and *h* respectively with *per* (for) and *ora* (hour). Emoticons, emojis, URLs, email addresses, and Twitter marks (hashtags and mentions) have been instead labelled with the tag `SYM`.

Dependency Relations Attachment

As said above, following the strategy applied in POSTWITA-UD, we did not perform any sentence splitting in the novel dataset. Each syntax tree of TWITTIRÒ-UD corresponds to a tweet in its entirety, and may consist of multiple sentences too. At the same time, provided that the UD scheme poses a single-root constraint, the internal connections between different sentences occurring in a tweet have to be annotated and labeled by the dependency relation `parataxis`. This relation is quite hard to be provided by the parser, which often fails in recognizing this kind of structure. See for instance, Figure 1 where we display a tweet[8] containing more paratactic structures.

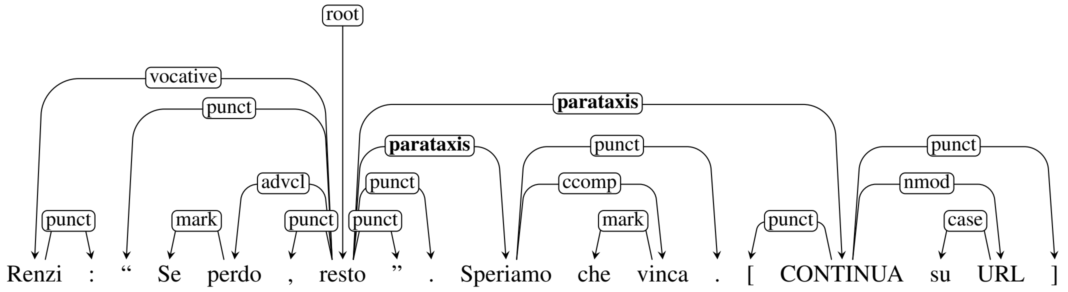


Figure 1: Example of tweet containing multiple sentences.

Another issue is related to the wide presence of Twitter marks. The current limited amount of adequate training data prevents the parser from dealing with them successfully. Within the manual correction phase, we resort to the label `vocative:mention` for Twitter mentions, the label `discourse:emo` for emojis, and `dep` for URLs.

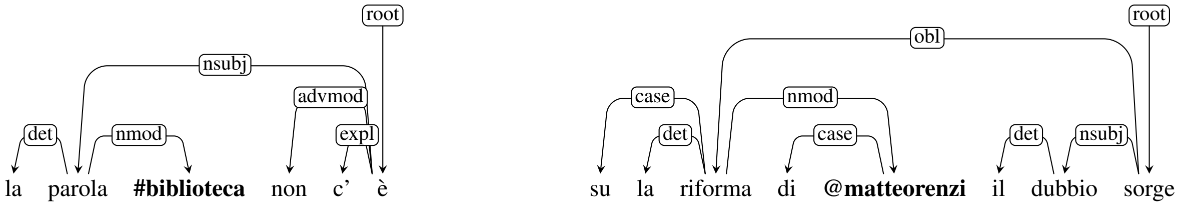

root
nsubj
advmod
det
nmod
expl
la parola #biblioteca non c’ è
obl
root
case
nmod
det
case
det
nsubj
su la riforma di @matteorenzi il dubbio sorge

Figure 2: Examples of tweets containing a hashtag and a mention with syntactic function.

Moreover, hashtags and mentions could be either used at the end of the tweet, to create more emphasis, or with a full syntactic function. In the first case, we resort to the relation (`parataxis:hashtags` and

[8] Translation: Renzi: “If I lose, I stay”. Let’s hope he wins then. [Follows on URL].

`vocative:mention`), while in the second we annotate accordingly to the syntactic role, see for example in Fig. 2 the hashtag and the mention[9] labelled as `nmod`.

		PoSTWITA-UD	TWITTIRÒ-UD
hashtags	`parataxis:hashtag`	40.89%	54.79%
	`nmod`	19.64%	11.55%
	`nsubj`	13.48%	8.59%
	other	25.99%	25.07%
mentions	`vocative:mention`	92.37%	87.41%
	other	7.63%	12.59%

Table 1: Distribution of deprel labels for hashtags and mentions.

Table 1 shows the distribution of the dependency relations (*deprels*), and confirms that there is a syntactic correlate of the peculiar semantic role that hashtags and mentions play in tweets. The labels that are mostly exploited for linking the hashtags to the sentence structure PoSTWITA-UD and in TWITTIRÒ-UD are mostly two: `nmod` and `nsubj`.

4 Analysis and Discussion

Table 2 shows the distribution of deprels in UD_Italian, PoSTWITA-UD and TWITTIRÒ-UD.

	UD_Ita	PoSTW	TWIT		UD_Ita	PoSTW	TWIT
`acl`	0.99	0.48	0.65	`flat`	0.19	0.35	0.10
`acl:relcl`	1.06	0.68	0.71	`flat:foreign`	0.05	0.28	0.05
`advcl`	1.26	1.00	0.90	`flat:name`	1.17	2.18	0.85
`advmod`	3.53	4.85	4.21	`goeswith`	0.00	0.03	-
`amod`	5.59	2.75	3.49	`iobj`	0.23	0.75	0.52
`appos`	0.31	0.43	0.16	`list`	-	0.22	-
`aux`	2.02	1.67	1.80	`mark`	2.11	2.23	2.10
`aux:pass`	0.75	0.12	0.18	`nmod`	8.01	6.84	5.68
`case`	14.03	9.42	10.23	`nsubj`	4.30	4.50	4.40
`cc`	2.73	2.26	1.80	`nsubj:pass`	0.77	0.16	0.26
`ccomp`	0.49	0.80	0.67	`nummod`	1.20	0.88	0.93
`compound`	0.25	0.17	0.27	`obj`	3.43	4.10	4.64
`conj`	3.39	2.95	1.72	`obl`	5.77	4.03	4.80
`cop`	1.15	1.75	1.54	`obl:agent`	0.38	0.12	0.13
`csubj`	0.11	0.17	0.07	`orphan`	0.01	0.05	-
`csubj:pass`	0.00	-	-	`parataxis`	0.14	4.02	4.62
`dep`	0.00	2.34	0.89	`parataxis:appos`	-	0.10	0.01
`det`	15.54	10.97	10.98	`parataxis:discourse`	-	0.02	0.01
`det:poss`	0.63	0.48	0.31	`parataxis:hashtag`	-	1.81	2.15
`det:predet`	0.14	0.12	0.11	`parataxis:insert`	-	0.03	-
`discourse`	0.02	1.18	0.75	`parataxis:nsubj`	-	0.03	-
`discourse:emo`	-	0.59	0.13	`parataxis:obj`	-	0.07	-
`dislocated`	0.01	0.11	0.01	`punct`	11.36	12.08	17.24
`expl`	0.73	0.85	0.96	`root`	4.75	5.39	4.77
`expl:impers`	0.14	0.15	0.13	`vocative`	0.03	0.38	0.09
`expl:pass`	0.13	0.05	0.04	`vocative:mention`	-	2.06	2.89
`fixed`	0.32	0.19	0.30	`xcomp`	0.76	0.76	0.78

Table 2: Dependency relations' distribution across the three main Italian treebanks. The values are expressed in percentage %.

We can observe, despite the sparseness of relations, how their frequency and distribution characterizes the language exploited in the social media data collected in TWITTIRÒ-UD and PoSTWITA-UD with respect to the standard language collected in UD_Italian. As expected, meaningful differences emerge for parataxis and punctuation. Punctuation is indeed exploited more extensively in the two social media datasets (12.08% and 17.24%) than in UD_Italian (11.36%), and the frequency of the `parataxis` deprel is 4.02% and 4.62% in PoSTWITA and TWITTIRÒ-UD, while it is only 0.14% in UD_Italian, marking a

[9] Translation: about the reform of @matteorenzi a doubt rises.

significant difference. The distributions of the relations `vocative:mention` and `parataxis:hashtag` especially features the two social media treebanks. The mentions' deprel is 2.06% in PoSTWITA-UD and 2.89% in TWITTIRÒ-UD, while the hashtags are respectively 1.81% and 2.15%.Furthermore, it is interesting to notice how the use of passive voices (`aux:pass`) is 0.75% in the UD_Italian treebank while only 0.12% in PoSTWITA-UD and only 0.18% in TWITTIRÒ-UD, indicating a preference for the exploitation of active voices in the language used in social media, as it happens in spoken language.

4.1 A Parsing Experiment

In order to preliminary evaluate the similarities between the three datasets, we performed an evaluation of UDPipe using the TWITTIRÒ-UD gold corpus as a test set. The following three settings were exploited.

1) training UDPipe using only UD_Italian (UD_It),
2) training UDPipe using only PoSTWITA-UD (PoSTW),
3) and training UDPipe using both resources (UD_It+PoSTW).

For evaluation we used the script made available for the CoNLL 2018 Shared Task 5[10] with the default setting parameters. Table 3 surveys the resulting scores for precision (P), recall (R) and averaged F1-score (F1).

	UD_It			**PoSTW**			**UD_It+PoSTW**		
	P	R	F1	P	R	F1	P	R	F1
Tokens	66.85	67.28	67.07	66.50	65.15	65.82	67.62	67.63	67.62
Sentences	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18	66.18
Words	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
UPOS	57.10	57.44	57.27	62.71	61.44	62.07	65.75	65.77	65.76
XPOS	56.30	56.63	56.47	62.23	60.97	61.59	65.59	65.61	65.60
Feats	59.35	59.70	59.52	62.17	60.91	61.53	65.64	65.66	65.65
AllTags	55.11	55.43	55.27	60.59	59.36	59.97	65.04	65.06	65.05
Lemmas	60.88	61.23	61.05	62.17	60.91	61.53	65.48	65.50	65.49
UAS	66.73	67.12	66.92	66.36	65.01	65.67	67.54	67.56	67.55
LAS	50.12	50.42	50.27	54.07	52.97	53.51	56.84	56.85	56.85

Table 3: Evaluation of UDPipe.

First of all, it is interesting to notice the variation of the Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS). For what concerns UAS, the first setup, where only the data from UD_Italian have been used for training, allowed a better result than the second one, where PoSTWITA-UD is the training dataset. But the opposite can be seen for LAS. We can hypothesize that the larger amount of data in UD_Italian allowed to build a more representative statistical model. Nevertheless, training on a resource which includes the same typology of data may be crucial for collecting an adequate knowledge about the specific relations exploited. This motivates the best scores for LAS an UAS, which were obtained in the third setup benefiting of both the resources for training. This encourages us to develop more and better gold standard treebanks also for social media to be used for training.

5 Conclusion and Future Work

In this paper we presented an ongoing project for the development of a novel Italian treebank from Twitter in the UD format: TWITTIRÒ-UD. Focusing on the 897 tweets currently annotated for the first release, we discuss the annotation of this resource which encompasses a fine-grained representation of irony and the UD morpho-syntactic analysis.

The preliminary analysis we applied shows some difference in the distribution of dependency relations in standard Italian and social media language, e.g. in the use of verbal active/passive voices, confirming that the language used in social media presents a strong preference for the exploitation of active voices. Furthermore, a simple parsing experiment and a comparison among the novel resource, UD_Italian (Simi et al., 2014) and PoSTWITA-UD (Sanguinetti et al., 2018) are provided, in order to shed light on the

[10] `http://universaldependencies.org/conll17/evaluation.html`

syntactic features of social media texts. Also considering the perspective of the future release of the complete resource (1,424 tweets) to be accomplished before the next UD release in November 2019, the work serves a twofold goal: it enriches the scenario of available resources for a text genre which is especially hard to parse (social media text), and helps in the investigation of possible relationships between syntax and semantics of the uses of figurative language (irony in particular). The availability of a resource whose annotation encompasses both UD relations and a fine-grained description of irony may indeed pave the way for the investigation of whether syntactic knowledge might help in SA and other related tasks.

Acknowledgments

The work of Cristina Bosco was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618L2BOSC01). The work of Paolo Rosso was partially funded by the the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

Fahad Albogamy and Allan Ramsay. 2017. Universal Dependencies for Arabic Tweets. In *RANLP*, pages 46–51.

Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava, and Dipti Misra Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. *arXiv preprint arXiv:1804.05868*.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. In *EVALITA 2014 Evaluation of NLP and Speech Tools for Italian*, pages 1–8.

Cristina Bosco, Tamburini Fabio, Bolioli Andrea, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 part of speech on Twitter for Italian task. In *CEUR Workshop Proceedings*, volume 1749, pages 1–7.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. 2018a. Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018)*.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018b. Overview of the EVALITA 2018 task on Irony Detection in Italian Tweets (IronITA). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. 2019. TWITTIRÒ: an Italian Twitter Corpus with a Multi-layered Annotation for Irony. *IJCoL - Italian Journal of Computational Linguistics*, 4(2):25–44.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–92.

Adnan Duric and Fei Song. 2012. Feature selection for sentiment analysis based on content and syntax models. *Decision support systems*, 53(4):704–711.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. #hardtoparse: POS Tagging and Parsing the Twitterverse. In *Workshops at the 25th AAAI Conference on Artificial Intelligence*.

Jihen Karoui, Benamara Farah, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A Smith. 2018. Parsing Tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975.

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 *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*.

Manuela Sanguinetti, Cristina Bosco, Alessandro Mazzei, Alberto Lavelli, and Fabio Tamburini. 2017. Annotating Italian Social Media Texts in Universal Dependencies. In *Proceedings of the 4th International Conference on Dependency Linguistics (DepLing 2017)*, pages 229–239.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018)*.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as Machine Learning Features for Natural Language Processing. *Expert Systems with Applications*, 41(3):853–860.

Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of LREC 2014*, page 83–90.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies Parsing for Colloquial Singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744.

Pāṇinian Syntactico-Semantic Relation Labels

Amba Kulkarni
Department of Sanskrit Studies,
University of Hyderabad
India
apksh@uohyd.ernet.in

Dipti Misra Sharma
Language Technologies Research Center,
IIIT-Hyderabad
India
dipti@iiit.ac.in

Abstract

We present in this paper a list of dependency relations based on Pāṇini's grammar for Sanskrit. The important feature of this list is that most of the relations represent well defined semantics that can be extracted from the surface string without any extra-linguistic information.

1 Introduction

In the last two decades the researchers in the Natural Language Processing (NLP) community have recognised the importance of dependency parsing. For English, several parsers producing dependency style output were developed. In the initial stages, there was no consensus among the dependency parser developers on the number of dependency relations and their names. The link parser (Sleator and Temperley, 1993) used 106 relations, while Minipar (Lin, 1998) which was based on Chomsky's minimalism and produced dependency parse used only 59 dependency relations. de Marneffe et al. (2006) modified the dependency relations proposed by Carroll et al. (1999) and King et al. (2003). These relations, known as Stanford Dependencies, were originally developed for English. They proposed a universal taxonomy with a total of 42 relations, which are supported across many languages. This set of relations then was adapted for several other languages. With the development of parsers for several languages, a need was felt to arrive at a single coherent standard, and this led to the development of universal dependencies that can be used for developing cross-linguistically consistent treebanks, that can facilitate multilingual parser development (Nivre, 2015; Nivre et al., 2016). All these various lists of relations mentioned above are syntactic in nature. Several NLP tasks such as database query, robot instructions, information extraction, etc. need semantic representations of sentences. Two major efforts viz. Framenet (Fillmore and Baker, 2000) and Propbank(Kingsbury and Palmer, 2002; Kingsbury and Palmer, 2003) concentrated on the development of semantically tagged lexicon and corpus respectively. The first automatic semantic role labelling system was developed by Gildea and Jurafsky (2002). The major problem with semantic roles is the difficulty involved in coming up with a standard set of roles and formal definitions of thematic roles. As a consequence, PropBank uses verb specific semantic roles as well as generalised semantic roles. Framenet uses semantic roles that are specific to a frame. There are also efforts to transform the syntactic dependency analysis to Logical Form (Reddy et al., 2016) for semantic parsing. There are also efforts to use Abstract Meaning Representation extending the existing relations in Propbank for the development of Semantic databanks (Banarescu et al., 2013).

Given this background, now we highlight some of the salient features of a dependency tagset based on the Pāṇinian grammar framework. Bharati et al. (1991) proposed a computational grammar for processing Indian languages based on the Pāṇinian framework. A dependency tagset based on the Pāṇini's grammar is being used for the development of treebanks for Indian languages (Bharati and Sangal, 1990; Bharati et al., 2002; Rafiya et al., 2008; Chaudhry et al., 2013; Chaudhry and Sharma, 2011). These tagsets are also used for the development of dependency parsers for Indian languages (Tandon and Sharma, 2017). Ramakrishnamacharyulu

(2009) compiled a list of relations used in Indian Grammatical Tradition. A rule-based parser for Sanskrit has been developed using these dependency relations (Kulkarni, 2013; Kulkarni et al., 2010; Kulkarni and Ramakrishnamacharyulu, 2013; Kulkarni, 2019b). There are efforts to analyse English through the Pāṇinian framework. Bhatt (1993) and Bharati et al. (1997) extend the notion of case suffixes (*vibhakti pratyaya*) to account for the notions of subject and object which have fixed positions in a sentence. Bharati and Kulkarni (2011) argues further that the concept of subject in English is the same as the concept of *abhihita* (expressed), and how by assigning a fixed position for Subject, and thereby doing away with the accusative marker English gains in the economy. Sukhada and Sharma (2016) and Bharati et al. (2015) compare the dependencies based on the concepts from Pāṇinian grammar (PG) with other dependency relations such as Stanford Dependencies, Link grammar parser dependencies etc. and offers an automatic mapping of dependency relations of these parsers to a PG based syntactico-semantic scheme.

In the next section, we provide a brief introduction to the Pāṇinian grammar. In the third section, we provide the salient features of the Pāṇinian dependency relations. In the fourth section, we describe the semantic content of the *kāraka* roles (to be defined in the next section) and conclude that this set of relations encodes the semantic relations of predicate arguments that can be extracted without appealing to the world knowledge.

2 Pāṇinian theory of *kāraka* in brief

Sanskrit assumes a unique status when it comes to the field of linguistic analysis with its more than 2500 year long and still extant grammatical tradition. Sanskrit grammar enjoys a similar status in India as mathematics in the West. Pāṇini's grammar is an important milestone in the Indian grammatical tradition. It is the first grammar for any language which is almost complete and together with the theories of verbal understanding (*śābdabodha*), it provides a complete system for language analysis as well as generation for Sanskrit in particular. Pāṇini's grammar known as *Aṣṭādhyāyī* is in the form of aphorisms (*sūtra*s)[1], arranged in 8 chapters with four sections each. According to Kiparsky (2009), the grammar analyses sentences at a hierarchy of 4 levels of description, which are traversed by 3 mappings in the direction from semantics to phonology.

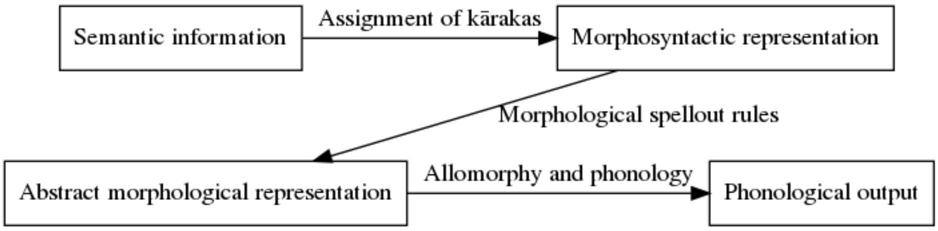

Semantic information
Assignment of kārakas
Morphosyntactic representation
Morphological spellout rules
Abstract morphological representation
Allomorphy and phonology
Phonological output

Figure 1: Levels in the generation process in Pāṇini

The generation starts from the abstract meaning representation and maps it to the surface form incrementally building up from one level to the other. To give an example, the initial semantic representation for the sentence

Skt: Rāmaḥ vanaṁ gacchati
Gloss: Rama {nom.} forest {acc.} go {pr tense, 3p, sg.}
Eng: Ram goes to the forest

may be described as follows:

[1] An aphorism is like a concise formula, the characteristics of which are: minimum number of words, devoid of ambiguity, contains an essence of topic, is universal in nature, without un-meaningful words and without any faults

- there is an activity taking place in the present time,
- there are two participants participating in this activity viz. the doer and the goal.

In the next step, Pāṇini's grammar assigns semantic labels to these various participants. Then the morphological spell out rules assign case suffixes to the participants depending upon their semantic labels, and finally, the phonological rules produce the sentence.

Our main focus of the discussion is on the semantic labels assigned to various participants of the activity. These labels indicate the role (relation) of the participant in the activity, such as *kartṛ*, *karman*, etc. These labels follow directly from the speaker's intention that determines the semantics that would be expressed through the language string. A generic term for such labels is *kāraka*, which literally means "a thing that brings about an action". Pāṇini classifies all these participants into only six categories viz. *kartṛ*, *karman*, *karaṇa*, *sampradāna*, *apādāna* and *adhikaraṇa*. He provides the semantic definitions for them. The definitions go like this.

- The participant which is the most independent to perform the activity is termed as *kartṛ*.[2] (doer of the activity)
- The participant which is the most desired by the *kartṛ* is termed as *karman*.[3] (roughly theme)
- The thing which is most instrumental in bringing the action to accomplishment is called a *karaṇa* (instrument).[4]
- The participant which the agent wishes to reach through the object is termed *sampradāna* (beneficiary).[5]
- The participant which is fixed when there is a movement away is termed as an *apādāna* (source).[6]
- The participant which serves as a locus of an activity is called an *adhikaraṇa* (locus).[7]

These are the general definitions of predicate-argument relations (*kāraka*). Each of these definitions is followed by a list of exceptional cases through which Pāṇini extends the scope of the semantic definitions of the predicate arguments. The extensions are of two types:

- where the associated semantics is totally different from normal expectations and is due to the frozen usages. For example, the verb *sthā* (to stand) takes locus as one argument. But, when this verb is prefixed with *adhi* (the meaning of the verb *adhi-sthā* also has a shade of meaning as 'to govern', in addition 'to stand over', 'to inhabit', etc. by a special rule[8]) the locus gets a *karman* label, as in *saḥ grāmam adhitiṣṭhati* (He inhabits / governs the village). Thus *grāma* (village), here, is not a locus but a theme. Pāṇini lists this rule especially because one may fail to notice this shift in the role when the verbal root has a prefix.
- where the extension to the semantics is not obvious to a layman. In such situations, he lists down special cases making the extension clear and obvious. Such an extension is semantic in nature and is not an idiosyncrasy of Sanskrit. For example, Pāṇini defines the source (*apādānam*) as the participant which is fixed when there is a movement away from it. Thus in *vṛkṣāt parṇam patati* 'The leaf falls from the tree', the tree (*vṛkṣa*) is assigned a role of source (*apādāna*). In the case of a sentence 'The boy fell down from a running horse', the horse is considered to be a source for the action of 'falling down', since the horse, though is running, is stationary relative to the action of falling. He, then, extends this definition to the cases which deal with mental separation and includes verbs such as *bhī* (afraid of) under the purview of this definition. With this, in the sentence, *John is afraid of a lion*, the *lion* gets the source (*apādāna*) role, since John, being afraid of a lion, experiences a mental

[2] *svatantraḥ kartā* (1.4.54) The number in the brackets refer to the chapter.section.*sūtra*

[3] *kartturīpsītatamaṁ karma* (1.4.49)

[4] *sādhakatamaṁ karaṇaṁ* (1.4.42)

[5] *karmaṇā yamabhipraiti sa sampradānaṁ* (1.4.32)

[6] *dhruvam apāye apādānam* (1.4.24)

[7] *ādhāro'dhikaraṇaṁ* (1.4.45)

[8] *adhiśīṅsthāsām karma (1.4.46)* 'in the case of verbal roots *sthā*, *śīṅ* and *as* when prefixed by *adhi* the locus gets a *karman* label

separation from it even when he just thinks of it. Since this extension may not be obvious, Pāṇini provides special aphorisms listing this and all such extensions.

3 Pāṇinian dependency relations for automatic processing

Apart from the predicate-argument relations, Pāṇini also mentions other relations between words such as cause (*hetu*), purpose (*prayojana*), precedence (*pūrvakāla*), etc. without providing any formal definitions for them, and thus implying they carry the same semantics as per their normal language usage. Works, in ancient Indian literature, dealing with grammar (*Vyākaraṇa*), logic (*Nyāya*), and discourse analysis (*Mīmāṁsā*), and especially the texts dealing with the theories of verbal cognition provide a fine-grain classification of such relations.

3.1 Granularity

A list of such relations for Sanskrit was compiled by Ramakrishnamacharyulu (2009). The consortium working on Sanskrit-Hindi Machine Translation adapted a subset of relations from this list for the computational analysis of Sanskrit.[9] It was also noticed that the granularity involved in this collection was too fine for mechanical processing (Kulkarni and Ramakrishnamacharyulu, 2013), and accordingly, a suitable subset was selected that could provide analysis with high accuracy (see Appendix A). The core dependency relations for different modern Indian languages and Sanskrit is common. However, there are a few language specific variations.

3.2 Salient features

Pāṇinian dependency relations have the following features.

- The relations are binary.
- All relations are between words denoting concepts.
- Underspecified relations are provided to handle the complexity in processing.
- Most of the relation names are the same as found in the Pāṇinian tradition. A few new relations, which were not found in Pāṇinian grammar, are added. These correspond to certain accompanying terms (*upapada*) that govern the case markers of the accompanying word. Pāṇini does not discuss the semantics of such relations. Kulkarni (2019a) provides the semantics associated with such relations and thereby elevating the status of such relations from morpho-syntactic to semantic level.
- These dependency relations are found to be suitable for automatic parsing with high accuracy (Kulkarni, 2013).
- The labels are also comprehensible by non-grammarians.
- These relations are also found to be appropriate for both parsing as well as generation (Kulkarni, 2019a).

4 Semantic content

Based on the semantic content, the Pāṇinian dependency relations may be classified into two categories: purely syntactic and purely semantic. We discuss each of them below.

- Purely syntactic
  These tags do not assign any semantic notion to the relation. There are only four such tags.
  - The first one is due to the duplication of a word. There are several meanings associated with the duplication such as pervading, several, successive order, series, distributiveness, repetition, and so on. A Sanskrit word *vīpsā* covers all these meanings. Since in order to decide the exact meaning one needs an access to the extra-linguistic information, we, without analysing this relation further, mark it as *vīpsā*.
  - Another syntactic relation is due to the genitive case marker. The semantic relations associated with this case marker are possession, part and whole relation, kinship rela-

[9] http://sanskrit.uohyd.ac.in/scl/GOLD_DATA/Tagging_Guidelines/tag_proposal_July2019.pdf

  tions, and so on. Here also, we do not sub-classify them providing the semantic labels, but collectively classify all of them under the syntactic label genitive (*śaṣṭhī*).

  - The pair of arguments arg1 (*anuyogin*) and arg2 (*pratiyogin*) correspond to the two arguments of a binary relation. They do not carry any specific meaning. These relations are used to specify the inter-sentential relations with sentential connectors such as if-then (*yadi-tarhi*), where the then-clause is the first argument and the if-clause is the second argument with the terms if and then being co-indexed.

- Purely semantic
  Barring the above relations, all other relations are purely semantic in nature. The relations between action and its participants referred to as *kāraka*, and other relations such as purpose (*prayojana*), cause (*hetu*), precedence (*pūrvakāla*) are some examples. The semantics associated with the predicate-argument relations, however, deserves some explanation. Due to the limitation of space, we discuss the semantics associated with only one relation viz. *kartṛ*, and its practical significance from computational point of view.

4.1 *Kartṛ* is not a subject

Consider the analysis of the following two sentences, one in active, and the other in passive represented in Figures 2 and 3 below.
(1) Skt: Rāmaḥ pāṭhaṁ paṭhati
Gloss: Rama{nom.} lesson {acc.} read {pr tense 3p sg}
Eng: Rama reads a lesson.
(2) Skt: Rāmeṇa pāṭhaḥ paṭhyate
Gloss: Rama{ins.} lesson {nom.} read {passive pr tense 3p sg}
Eng: The lesson is read by Rama.

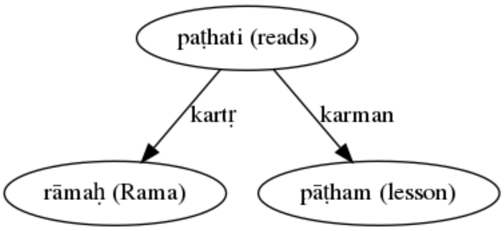


Figure 2: analysis of an active sentence

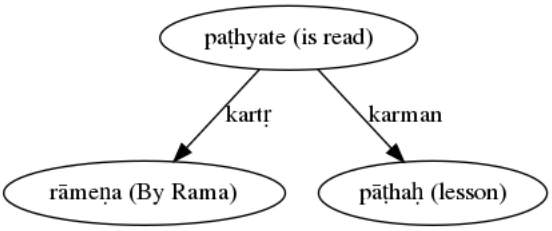


Figure 3: analysis of a passive sentence

We notice that *Rama* which is in the nominative case in the first and in instrument case in the second is marked as *kartṛ* in both the sentences. Special feature of the Pāṇini's grammar is that it does not give two different rules for active and passive, instead handles both by a single rule (Kiparsky, 2009). In other words, there is no transformation rule involved. This brings in uniformity in the analysis of a sentence in the active and passive voice. Now the natural question is, then, is *kartṛ* an agent? And again the answer is No.

4.2 Kartṛ is not an agent

Look at the following three sentences.
1) Skt:*rāmaḥ kuñcikayā tālam udghāṭayati.*
Gloss: Rama{nom.} key{ins.} lock{acc} open{pr tense 3p sg}.
Eng: Rama opens the lock with a key.
In this sentence, *Rama* is a *kartṛ* and an agent, the *key* is an instrument, and the *lock* is the goal. Now consider a situation where somebody is trying to open the lock. He tries with several keys, and finally, with one black key, he could open the lock. In such a situation, he utters,

2) Skt:*śyāmā kuñcikā tālam udghāṭayati.*
Gloss: Black{nom.} key{nom} lock{acc.} open{pr tense 3p sg}.
Eng: The black key opens the lock.
Though thematically, the *key* is still an instrument, according to Pāṇini's grammar, in this sentence it is a *kartṛ*. As a final example, let us consider a situation where somebody is trying to open a lock, and even before inserting the key, the lock gets opened on its own. In such a situation, one may utter 'And then he touches the lock and the lock opens'.
3) Skt:*tālaḥ udghāṭyate.*
Gloss: Lock{nom.} open{pr tense 3p sg}.
Eng: The lock opens.
Here, thematically the lock is a theme. However, according to Pāṇinian analysis, in this sentence, the *lock* is a *kartṛ*. Thus we notice that *kartṛ* in the first sentence is an agent, in the second sentence an instrument and in the third it is the theme. *Kartṛ*, therefore, can be roughly translated as 'doer' which need not be animate.

4.3 What is the semantics associated with the *kartṛ*?

Pāṇini defines *kartṛ*[10] as 'the independent participant in the activity'. An activity typically involves more than one participants. The underlying verb expresses the complex activity which consists of subactivities of each of the participants involved. For example, in the case of opening of a lock, three subactivities are very clearly involved (Bharati et al., 1995) , viz.

1. the insertion of a key by an agent,
2. pressing of the levers of the lock by an instrument (key), and
3. moving of the latch and opening of the lock.

Though in practice, to a large extent all the three subactivities 1 through 3 together constitute the activity 'opening a lock', sometimes the subactivities 2 and 3 together are also referred to as 'opening a lock', as noticed above in the second example, and the activity 3 alone is also referred to as 'opening a lock', as we see in the third sentence. Let us call them open$_1$, open$_2$ and open$_3$, respectively.

Pāṇini draws our attention to the following.

1. The verbal roots are finite in numbers while the conceptual space they cover is infinite. In spite of this, the ambiguity resulting due to the overloading can be resolved from the substantive playing the role of *kartṛ*. Such disambiguation is important in rule-based or knowledge-based Machine Translation systems when the source language and target language map the conceptual space differently. For example, in Hindi open$_1$ and open$_2$ correspond to the verbal root 'khola', while open$_3$ corresponds to the verbal root 'khula'.
2. In order to assign the thematic relations, one has to appeal to the extra-linguistic information.

The greatness of the Pāṇini lies in **"identifying exactly how much information is coded and then giving it a semantic interpretation"** (*sūtras* 1.4.23 - 1.4.55). This level of semantics is the one which is achievable/reachable through the grammar rules and the language string alone. This puts an upper bound on the analysis, making it very clear what is guaranteed by rule-based or knowledge-based analysis and what is not. We can extract only that which is available in a language string 'without any requirement of additional knowledge'.

5 Sanskrit Parser using Pāṇinian Dependencies

A rule-based parser for Sanskrit based on the Indian theories of verbal cognition using the dependency labels provided in Appendix A has been developed which can handle both the prose as well as the verse.[11] For the following verse from the Bhagavadgītā the parser produces the

[10] *svatantraḥ kartā* (1.4.54)

[11] Due to the space constraint, we do not discuss here the performance of this parser. One may refer to (Kulkarni, 2019b) which discusses its performance on the prose.

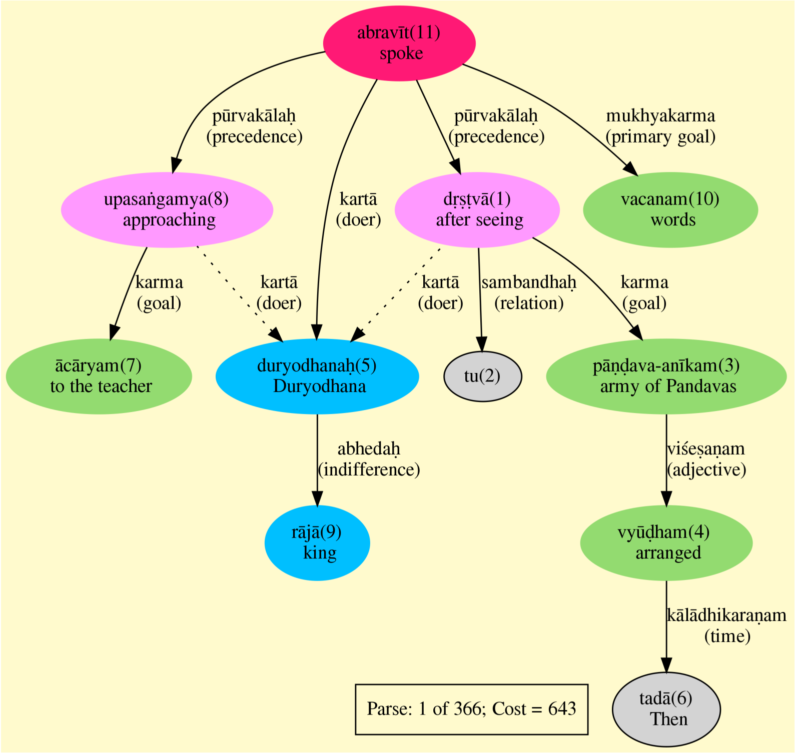


Figure 4: Parsed output of the BhG 1.2 verse

parsed structure as shown in Figure 4. The numbers in the parenthesis indicate the index of the word in the verse. The dotted lines show the shared relations.
Skt: *dṛṣṭvā tu pāṇḍavānīkam vyūḍham duryodhanaḥ tadā |*
ācāryam upasaṅgamya rājā vacanam abravīt || (BhG 1.2)
Gloss: After_seeing[12] the_army_of_the_Pāṇḍavas arranged_in_military_phalanx Duryodhana at_that_time, teacher approached King words spoke
Eng: At that time, after seeing the army of the Pāṇḍavas arranged in military phalanx, King Duryodhana approached (his) teacher and spoke (these) words.

The parser has produced total 366 parses. The first parse is shown here. We note that the parser has gone wrong only in one relation. The sixth word *tadā* (then) should have been connected to the final verb *abravaīt* (spoke). The multiple parses are due to the fact that the parser does not yet have a mechanism to check the mutual compatibility between the word meanings before establishing a relation between them. The current implementation uses this condition only to handle the adjectival relations, where Pāṇini's grammar provides a semantico-syntactic criterian for adjectives, which are otherwise indistinguishable from the substantives morphologically. There are several other cases of ambiguities as well where more than one relation use the same case marker, and the clue is only in the semantics of the word involved. While minimum semantic information such as the classification of the words following the Vaiśeṣka[13] ontology promises better results, the deep learning would complement it further for better results.

6 Conclusion

There are two advantages of using Pāṇinian dependencies. It provides a well-defined semantics that can be extracted purely from the language string. And the same set of relations can be used for both analysis as well as generation. The clear separation of what can be extracted from a language string alone and what can not be helps us plan eclectic use of rule-based and machine

[12] *tu* here is just a filler for metrical purpose
[13] One of the schools of Indian philosophy

learning approaches for developing better parsers.

Acknowledgement

Prof. Vineet Chaitanya deserves special acknowledgement, series of discussions with whom have helped in shaping this manuscript.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Association Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bullgaria. Association for Computational Linguistics.

Akshar Bharati and Amba Kulkarni. 2010. Information coding in a language: Some insights from Paninian grammar. *Dhīmahi, Journal of Chinmaya International Foundation Shodha Sansthan*, I(1):77–91.

Akshar Bharati and Amba Kulkarni. 2011. 'subject' in English is abhihita. In Ashok Aklujkar George Cardona and Hideyo Ogawa, editors, *Studies in Sanskrit Grammars (Proceedings of the Vyakarana Section of the 14th World Sanskrit Conference)*. D.K. Printworld.

Akshar Bharati and Rajeev Sangal. 1990. A karaka based approach to parsing of Indian languages. In *Proceedings of International Conference on Computational Linguistics (Vol. 3)*, Helsinki, Association for Computational Linguistics NY.

Akshar Bharati and Rajeev Sangal. 1993. Parsing free word orderlanguages in the Paninian framework. In *Proceedings of Annual Meetingof Association for Computational Linguistics, Association forComputational Linguistics, New Jersey*, pages 105–111.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1991. A computational grammar for Indian languages processing. *Indian Linguistics*, 52, nos 1–4:91–103.

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall New Delhi.

Akshar Bharati, Medhavi Bhatia, Vineet Chaitanya, and Rajeev Sangal. 1997. Paninian grammar framework applied to English. *South Asian Langauge Review.*

Akshar Bharati, Rajeev Sangal, Vineet Chaitanya, Amba P Kulkarni, Dipti M Sharma, and KV Ramakrishnamacharyulu. 2002. Anncorra: Building tree-banks in Indian languages. In *Proceedings of Workshop on Asian Language Resources, COLING-2002, Taipei.*

Akshar Bharati, Samar Husain, Dipti M Sharma, and Rajeev Sangal. 2009. Two stage constraint based hybrid approach to free word order language dependency parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT). Paris.*

Akshar Bharati, Sukhada, Dipti M. Sharma, and Soma Paul. 2015. Anusaaraka dependency schema from Paninian perspective. In *Sanskrit and Computational Linguistics, Proceedings of 16th World Sanskrit Conference, Bangkok*. D. K. Publishers.

Rajesh Bhatt. 1993. *Paninian Theory for English.* Ph.D. thesis, Department of CSE, IIT KAnpur. B.Tech. thesis.

J. Bronkhorst. 1979. The role of meanings in pāṇini's grammar. *Indian Linguistics*, 40:146–157.

George Cardona. 2007. On the structure of pāṇṇini's system. *Sanskrit Computational Linguistic*, 1&2:1–31.

John Carroll, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL.*

Himani Chaudhry and Dipti M Sharma. 2011. Annotation and issues in building an English dependency treebank. In *Proceedings of 9th international Conference on Natural Language Processing.* Macmillan Publishers, India.

Himani Chaudhry, Himanshu Sharma, and Dipti Misra Sharma. 2013. Divergences in English-Hindi parallel dependency treebanks. In *Procedings of the second international conference on Dependency Linguistics.*

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC.*

Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Samuel R. Bowman, Timothy Dozat, and Christopher D. Manning. 2013. More constructions, more genres: extending Stanford dependencies.

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

Charles J. Fillmore and Collin F. Baker. 2000. Framenet: Frame semantics meets thecorpus. In *Poster presentation, 74th Annual Meeting of the Linguistic Society of America.*

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

S.D. Joshi and J.A.F. Roodbergen. 1980. Patañjali's vyākaraṅa-mahābhāṣya, kārakāhnika.

S. D. Joshi. 2009. Background of the aṣṭādhyāayī. pages 1–5.

Tracy H. King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald Kaplan. 2003. The parc 700 dependency bank. In *4th International Workshop on Linguistically Interpreted Corpora.*

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Third International Conference on Language Resources and Evaluation.*

Paul Kingsbury and Martha Palmer. 2003. Propbank: The next level of treebank. In *The Second Workshop on Treebanks and Linguistic Theories.*

Paul Kiparsky. 2009. On the architecture of pāṇini's grammar. pages 32–94.

Amba Kulkarni and K. V. Ramakrishnamacharyulu. 2013. Parsing Sanskrit texts: Some relation specific issues. In Malhar Kulkarni, editor, *Proceedings of the 5th International Sanskrit Computational Linguistics Symposium.* D. K. Printworld(P) Ltd.

Amba Kulkarni, Sheetal Pokar, and Devanand Shukl. 2010. Designing a Constraint Based Parser for Sanskrit. In G N Jha, editor, *Fourth International Sanskrit Computational Linguistics Symposium*, pages 70–90. Springer-Verlag, LNAI 6465.

Amba Kulkarni. 2013. A deterministic dependency parser with dynamic programming for Sanskrit. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, Prague, Czech Republic, August. Charles University in Prague.

Amba Kulkarni. 2019a. Appropriate dependency tagset for Sanskrit analysis and generation. In *Proceedings of Sanskrit in China International Conference 2019: Sanskrit on Paths.* forthcoming.

Amba Kulkarni. 2019b. *Sanskrit Parsing based on the theories of śābdabodha.* D K Print World and IIAS Shimla.

Dekang Lin. 1998. Dependency-based evaluation of MINIPAR. In *Workshop on the evaluation of Parsing Systems.* Granada, Spain.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC.*

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing.*

Sanjeev Panchal and Amba Kulkarni.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Daniel Jurafsky. 2005. Semantic role labeling using different syntactic views. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 581–588.

Begum Rafiya, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of The Third International Joint Conference on NaturalLanguage Processing (IJCNLP). Hyderabad, India.*

K V Ramakrishnamacharyulu. 2009. Annotating Sanskrit texts based on Śābdabodha systems. In Amba Kulkarni and Gérard Huet, editors, *Proceedings Third International Sanskrit Computational Linguistics Symposium*, pages 26–39, Hyderabad, India. Springer-Verlag LNAI 5406.

Siva Reddy, Oscar Täckstroṁ, Micheal Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. 2016. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140.

Daniel D Sleator and Davy Temperley. 1993. Parsing English with a link grammar. In *Third international workshop on Parsing Technologies.*

Sukhada and Dipti M Sharma. 2016. Analyzing English phrases from Paninian perspective. In *17th International Conference on Intelligent Text Processing and Computational Linguistics.*

Juhi Tandon and Dipti Misra Sharma. 2017. Unity in diversity: A unified parsing strategy for major indian languages. In *Proceedings of the fourth International Conference on Dependency Linguistics, Pisa, Italy*, pages 255–265.

Appendices

A Tagset of Dependency Relations

- sambandhaḥ (relation)
  - **Kāraka sambandhāḥ** (Predicate-argument relations)
  - kartā (roughly agent)
    - prayojaka-kartā (causer)
    - prayojya-kartā (causee)
  - karma (goal)
    - mukhya-karma (primary goal)
    - gauṇa-karma (secondary goal)
    - vākya-karma (sentential argument)
  - karaṇaṁ (instrument)
  - sampradānaṁ (beneficiary)
  - apādānaṁ (source)
  - adhikaraṇaṁ (locus)
    - kāla-adhikaraṇaṁ (time)
    - deśa-adhikaraṇaṁ (space)
    - viṣaya-adhikaraṇaṁ (subject as a locus)
  - **Kāraketara sambandhāḥ** (Relations other than arguments of predicate)
    - **kriyā-kriyā-sambandhāḥ** (verb-verb relations)
      - pūrva-kālaḥ (precedence)
      - Vartamāna-samāna-kālaḥ (present participial)
      - Bhaviṣyat-samāna-kālaḥ (future participial)
      - Bhāvalakṣaṇa-pūrva-kālaḥ (simulateity in past)
      - Bhāvalakṣaṇa-vartamāna-samāna-kālaḥ (simulateity in present)
      - Bhāvalakṣaṇa-anantara-kālaḥ (simulateity in future)
      - Sahāyaka-kriyā (auxiliary-verb)
    - **kriyā-sambandhāḥ**
      - sambodhyaḥ (vocative)
      - hetuḥ (cause)
      - prayojanam (purpose)

    - kartṛ-samānādhikaraṇam (predicative adjective)
    - karma-samānādhikaraṇam
    - kriyāviśeṣaṇam (manner adverb)
    - pratiṣedhaḥ (negation)
  - **Nāma-Nāma-sambandhāḥ**
    - śaṣṭhī-sambandhaḥ (genitive)
    - aṅgavikāraḥ (deformity)
    - vīpsā (reduplication)
    - viśeṣaṇam (adjective)
    - sambodhana-sūcakam (vocative marker)
    - vibhaktam
    - avadhiḥ (interval)
    - abhedhaḥ (indifference)
    - nirdhāraṇam
    - atyanta-saṁyogaḥ
    - apavarga-sambandhaḥ
- **Upapada-sambandhāḥ**
  - svāmī (possessor)
  - saha-arthaḥ (association)
  - vinā-arthaḥ (dis-association)
  - point of reference
  - point of comparison
- **Inter-sentential relations**
  - anuyogī (arg1)
  - pratiyogī (arg2)
  - nitya-sambandhaḥ (co-reference)
- **Conjuncts and disjuncts**
  - samuccitaḥ (conjunction)
  - samuccaya-dyotakaḥ (conjunctive marker)
  - anyataraḥ (disjunction)
  - anyatara-dyotakaḥ (disjunctive marker)

Note: The bold entries are the headings and do not indicate relation labels

Experiments on human incremental parsing

Leonid Mityushin
Institute for Information
Transmission Problems
Russian Academy of Sciences, Russia
lmityushin@gmail.com

Leonid Iomdin
Institute for Information
Transmission Problems
Russian Academy of Sciences, Russia
iomdin@gmail.com

Abstract

Experiments have been conducted in which the subjects incrementally constructed dependency trees of Russian sentences. The subject was successively presented with growing initial segments of a sentence, and had to draw syntactic links between the last word of the segment and the previous words. The subject was also shown a limited right context – a fixed number of words following the last word of the segment. The results of the experiments show that the right context of 1 or 2 words is sufficient for confident incremental parsing of Russian narrative sentences.

1 Introduction

The concept of incremental text comprehension implies that at any moment the reader/listener has a complete or almost complete linguistic and pragmatic interpretation of the part of the text perceived up to that moment, and that this interpretation, as a rule, does not change after new parts of the text have been perceived. Usually, this concept is used with regard to language learning (especially reading learning), literary studies, nontrivial semantic and pragmatic comprehension, and logical inference, which requires full understanding of subtle context; see e.g. a recent paper by E. Fischer et al. (2019). The aim of this work is to evaluate whether this is true for human comprehension of the syntactic structure of a text (as a matter of fact, of an individual sentence).

We have conducted experiments on incremental construction of dependency trees for Russian sentences. The subjects in the experiments were linguists with considerable experience of syntactic annotation. In a single experiment, the subject was successively presented with growing initial segments of a certain sentence, and had to draw syntactic links between the last word of the segment (**the active word**) and the previous words (**the left context**); the syntactic links created up to a certain moment form a partial syntactic structure of the sentence. At each step, the subject was also shown a limited **right context** – a fixed number of words following the active word. Three series of experiments have been conducted for the lengths of the right context 0, 1 and 2, with 100 sentences processed in each series.

2 ETAP syntactic model

We use the representation of syntactic structures of sentences in the formalism of dependency trees adopted in the ETAP multilingual multifunctional linguistic processor (Iomdin et al., 2012) and originally introduced by I. Mel'čuk (1974, 1988). The nodes of a dependency tree are the words of the sentence; punctuation marks are not included and constitute a kind of additional data – unlike, for example, the practice of the Universal Dependencies approach (https://universaldependencies.org). The nodes are connected by directed arcs called syntactic links, which are labelled with names of syntactic relations. The lists of syntactic relations for Russian and English include about 70 and 60 relations respectively.

Based on the ETAP syntactic formalism, a treebank named SynTagRus has been created which at present contains about 1.1 million words of Russian text (Dyachenko et al., 2015; Inshakova et al., 2019). Due to the complexity of the ETAP syntactic model, the developers of SynTagRus have always paid special attention to the reduction of the number of human errors. As a rule, each new sentence in SynTagRus is processed twice, by two different people: the annotator, who creates the complete syntactic structure of the sentence (using the raw results produced by the ETAP linguistic processor), and the editor, whose role is to check the structures created by the annotator.

The syntactic link $a \rightarrow b$, where a and b are words of the sentence, is called projective if all the words between its head node a and dependent node b are directly or indirectly dominated by the word a, and non-projective otherwise. About 8% of syntactic links in SynTagRus are non-projective.

In SynTagRus, dependency trees for sentences with ellipsis contain additional "phantom" nodes that represent omitted words. Although ellipsis is not very frequent in Russian texts, it appears quite regularly; the proportion of elliptical sentences in SynTagRus is about 2%.

3 Modifications to the syntactic model

To facilitate the incremental construction of Russian dependency trees, certain modifications were made to the representation of subtrees containing prepositions and conjunctions; we will describe these changes using similar English examples. In the ETAP syntax, prepositions/conjunctions dominate the noun/verb groups that follow them. For example, the sentences

(1) *He arrived at work* and

(2) *He arrived at noon*

have the following dependency trees:

```
(1)  He       <-.      predic           (2)  He       <-.      predic
     arrived  --' --.  ---                   arrived  --' --.  ---
     at       --. <-'  2-compl               at       --. <-'  adverb
     work     <-'      prepos                noon     <-'      prepos
```

Figure 1. Dependency trees for the sentences beginning with *He arrived at ...*

Here for syntactic links entering the words, the abbreviated names of the assigned syntactic relations are shown; for full names and descriptions of English syntactic relations see (Apresjan et al. 1989). Being presented with the initial segment *He arrived at ...* , the subject cannot confidently decide which type of link connects *arrived* and *at*. The sentences

(3) *He saw Mary and Kate* and

(4) *He saw Mary and smiled*

have the following dependency trees:

```
(3)  He    <-.          predic          (4)  He      <-.          predic
     saw   --' --.      ---                  saw     --' --. --.  ---
     Mary      <-' --.  1-compl              Mary        <-'   |  1-compl
     and   --.     <-'  coord                and     --.     <-'  coord
     Kate  <-'          coord-conj           smiled  <-'          coord-conj
```

Figure 2. Dependency trees for the sentences beginning with *He saw Mary and ...*

Being presented with the initial segment *He saw Mary and ...* , the subject cannot decide which word is the head of the coordinating link: *saw* or *Mary*.

To avoid these difficulties, it was decided to invert the direction of the left-to-right links "preposition → X" and "conjunction → X" so that the links are directed from the word X to the function word; the names of the links remain unchanged. The links that entered a preposition or conjunction will now enter the word X, which in the new situation dominates the preposition/conjunction; again the names of the links remain unchanged. These modifications are purely technical and allow automatic transformation from the old form to the new and vice versa. It is worth noting that the new form of these constructions agrees with the principles of the Universal Dependencies approach; see the discussion in (Osborne and Gerdes, 2019).

The transformation described is not used for prepositions homonymous with adverbs, such as *naprotiv* ('opposite'), *poperek* ('across'), *posle* ('after'), *szadi* ('behind'), *vnutri* ('inside'), *vozle* ('near') etc. Instead, such words are always considered as adverbs, and the dependent of the preposition formally becomes the dependent of the adverb (with the 1st completive syntactic relation instead of prepositive).

4 Tentative links

As shown by garden-path sentences (such as *The horse raced past the barn fell)*, which were first discussed by H.W. Fowler (1926) who actually introduced the incremental approach to syntax, a 100 percent confident incremental parsing of a sentence is impossible. There inevitably arise situations where it is necessary to revise decisions made earlier. We distinguish two types of such situations: those where the necessity of revision is surprising to the subject, and those where the possibility of revision was planned in advance. This "conscious uncertainty" is realized in the experiments in the form of tentative links.

Consider the sentence

(5) *I met her sister yesterday*,

and suppose the subject is given the first three words: *I met her ...* . The subject understands that in this segment the syntactic link *met → her* (1-compl) is possible and quite probable, and at the same time understands that the dependency tree of the complete sentence need not contain this link (as is indeed the case in this example, where the correct links are *met → sister* (1-compl) and *sister → her* (determ)). In this situation the subject inserts the link into the syntactic structure but marks it as "tentative" (the other links are called "final"). It is also allowed to create tentative links and keep them in reserve, without immediate insertion into the structure, – for example, when there is an alternative which seems more probable. While processing a sentence, the subject has the right to freely insert existing tentative links into the structure or remove them from the structure, on the condition that at any moment the syntactic structure should remain a well-formed directed tree or a union of disjoint well-formed trees.

Normally, the process of incremental construction of the syntactic structure consists in augmenting the structure by new syntactic links (final or tentative) that connect the active word and the words of the left context. It is also allowed to make "corrections", that is to insert into the structure or remove from it final links whose both ends belong to the left context. We always presume that processing a given sentence results in producing its correct complete dependency tree. The subject's performance on a sentence is measured with two indicators: the number of corrections and the number of created tentative links. In an ideal situation, both these numbers are equal to zero; in reality the subjects are instructed to avoid making corrections as much as possible and to keep the number of tentative links to a minimum. Accordingly, tentative links should only be created when the use of final links is associated with a significant risk of error.

In principle, the experiments might be conducted in a more straightforward way without an additional type of link. At each moment the subject would create a syntactic structure which is plausible enough for the known part of the sentence, for example, would include the link *met → her* (1-compl) in the structure for the segment *I met her...* If at a later stage certain links turn out to be incorrect, they are simply removed from the structure; similarly, missing links are added to the structure. In this case we have only one indicator of performance: the number of corrections. However, with this metric we cannot distinguish between changing the structure in situations of genuine ambiguity and correcting ordinary human errors such as those caused by carelessness.

5 Setup of the experiment

The experiment is conducted as a dialogue supported by a special program. The program takes as input a sentence in the form of a string of characters and splits it into words. The dialogue consists of N–1 steps numbered 2, 3, ... , N, where N is the number of words in the sentence. At step K the subject is presented with a text file which shows the first K words of the sentence (with the adjacent punctuation) plus the right context, that is, a fixed number of words following the word K. The syntactic links are also shown that were created at previous steps between the words of the left context (1, ... , K–1). When the last word of the sentence is shown, it is accompanied by the message [end of sentence]; until this message appears, the subject has no information about the length of the sentence.

The task of the subject is to create, if needed, new syntactic links between the active word K and the words of the left context. To create a link, the subject writes the name of syntactic relation and, if necessary, the number of the head (and in some cases dependent) in the appropriate field of the file.

Consider, for example, the English sentence

(6) *London Orbital is a 117 mile long motorway, encircling almost all of Greater London,*

and let the size of the right context be set to 1. If everything goes correctly, at step 8 the subject will be presented with the text file shown in Figure 3.

```
 London Orbital is a 117 mile long motorway,
 encircling ......

  1   London     <-.      compos
  2   Orbital    --' <-.  predic
  3   is             --'  ---
  4   a                   ---
  5   117        <-.      quantit
  6   mile       --' <-.  restr
  7   long           --'  ---
  8   motorway,
      encircling

-----------------------------------------------------------
|  * -->  8        |
|  8 -->  3  is    |
|  8 -->  4  a     |
|  8 -->  7  long  |
-----------------------------------------------------------

   TENTATIVE LINKS
-----------------------------------------------------------
|  create and insert into the tree  |    -->
|  create                           |    -->
|  insert into the tree             |    -->
|  remove from the tree             |    -->
-----------------------------------------------------------

   CORRECTION OF FINAL LINKS
-----------------------------------------------------------
|  insert into the tree  |    -->
|  remove from the tree  |    -->
-----------------------------------------------------------
```

Figure 3. The dialogue file at step 8.

The subject should create links between the active word 8 *motorway* and the previous words. In this case tentative links are not needed, and the subject only deals with final links, writing information about them in the first of the three frames (Figure 4).

```
-----------------------------------------------------------
|  * -->  8        |  3 copulat
|  8 -->  3  is    |
|  8 -->  4  a     |  determ
|  8 -->  7  long  |  modif
-----------------------------------------------------------
```

Figure 4. Creating new links between the active word 8 and the left context.

At step 9 these links are inserted into the structure, and word 9 becomes active (Figure 5). At this point the subject creates the link 8 → 9 (modif), and so on. The program keeps a complete record of the subject's actions at all steps of sentence processing.

```
 London Orbital is a 117 mile long motorway,
 encircling almost ......

  1   London     <-.              compos
  2   Orbital    --' <-.          predic
  3   is             --'     --.  ---
  4   a                  <-.   |  determ
  5   117        <-.       |   |  quantit
  6   mile       --' <-.   |   |  restr
  7   long       <-. --'   |   |  modif
  8   motorway,  --'     --' <-'  copulat
  9   encircling
      almost

------------------------------------------------------------
|  * -->  9      |
|  9 -->  3  is  |
------------------------------------------------------------

   TENTATIVE LINKS
   ......

   CORRECTION OF FINAL LINKS
   ......
```

Figure 5. The dialogue file at step 9.

6 Experimental dataset

The sentences for the experiments were taken from the two sets of sentences *dev.csv* and *train.csv* offered as training material for the competition "Automatic Gapping Resolution for Russian" held in association with the conference Dialogue 2019 (Dialogue Evaluation / AGRR-2019). These sets contain over 20,000 sentences of various genres, about one third of which are marked as elliptical. For our experiments, non-elliptical sentences were selected that satisfied the following additional requirements:

(1) the number of words does not exceed 30;
(2) the first alphanumeric character is a Russian capital letter;
(3) the last character is a small Russian letter or full stop;
(4) the proportion of small Russian letters among all alphanumeric characters is at least 90%.

The aim of these requirements was to restrict experimental material to "ordinary narrative Russian sentences of average length". As a result, a set of about 7,700 sentences was formed; the sentences for the experiments were taken from it without replacement using pseudorandom numbers. The distribution of sentence length in this set is rather "flat" on the segment from 7 to 30, with a mean of 17.4 and a standard deviation of 6.4. Hence, in a random sample of 100 sentences the average length has the same mean and a standard deviation of 0.64.

Size of the right context	Total number of links in the trees	Number of tentative links in the trees	Total number of created tentative links	Number of corrections
0	1627	34 (2.23%)	75	3
1	1741	21 (1.21%)	34	0
2	1607	8 (0.50%)	13	0

Table 1. The results of the experiments.

7 Results and future work

Three series of experiments were conducted for the sizes of the right context 0, 1 and 2, with 100 sentences processed in each series. The role of the subjects was played by the authors of this paper. They have considerable practical experience of developing the SynTagRus treebank, each having tagged not less than ten thousand sentences. The results of the experiments are given in Table 1.The figures in the table show that the right context of 1–2 words is sufficient for error-free and confident incremental parsing of Russian narrative sentences.

In the future, we plan to conduct experiments on incremental parsing of Russian elliptical sentences. Processing of a sentence is supposed to be similar to the procedure described in Section 5, but in addition to creating syntactic links, the subject will be able to create new nodes of the syntactic structure representing omitted lexical items.

Another possible area of future work is incremental parsing of English sentences. Generally, the results for English are expected to be more modest than for Russian, partly because the English inflectional system is not as rich as the Russian. However, preliminary experiments did not show a great difference in performance.

8 Conclusion

We believe that the experiments described in this paper characterize certain general features of human text comprehension. It could be argued that in fact we studied a much narrower phenomenon: text comprehension in people who are experts in linguistics. In our opinion, however, text comprehension is a highly automatic subconscious process which, in the case of native speakers, is not influenced by special linguistic training. But linguists, in contrast to ordinary speakers, have the tools which enable them to externalize their understanding of the text – for example, they can assign morphological features to wordforms or identify syntactic dependencies between words – and this is exactly what is required of the subjects in our experiments.

The results of the experiments, namely almost complete absence of errors (i.e. corrections of final links) and a small number of tentative links created, may be regarded as arguments in favour of the following general model of text comprehension. Suppose that while processing a sentence, only final links have been used. This means that the syntactic structure of the sentence was built in a strictly incremental way: links were added to the structure but never removed from it. In this case we can imagine the comprehension process to develop like this: for each new word, the reader/listener adds to the structure the links containing this word that satisfy the syntactic and semantic requirements, and later never returns to them. It may be assumed that this strategy of immediately adding plausible links to the structure is used universally, while relatively infrequent collisions (incompatibility of new potential links with those already in the structure) are successfully resolved on the basis of information available at the moment of collision. For this strategy to be efficient, natural language texts should be specially adapted to it. We assume that this adaptation is provided by their authors, who are interested in successful communication.

Acknowledgements

The authors express their gratitude to the Russian Foundation for Basic Research for their partial support of this work (grant No 19-07-00842).

References

Juri Apresjan, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Nikolai Pertsov, Vladimir Sannikov, and Leonid Tsinman. 1989. *Lingvisticheskoe Obespechenie Sistemy ETAP-2* [The linguistics of the ETAP-2 system]. Nauka, Moscow. (in Russian)

Dialogue Evaluation / AGRR-2019. https://github.com/dialogue-evaluation/AGRR-2019

Pavel Dyachenko, Leonid Iomdin, Alexander Lazursky, Leonid Mityushin, Olga Podlesskaya, Viktor Sizov, Tatiana Frolova, and Leonid Tsinman. 2015. Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (SynTagRus) [A deeply annotated corpus of Russian texts (SynTagRus):

contemporary state of affairs]. *Natsional'nyi korpus russkogo yazyka: 10 let proektu. Trudy Instituta russkogo yazyka im. V.V. Vinogradova. Vyp. 6* [The Russian National Corpus: 10 Years of the Project. Proc. of the V.V. Vinogradov Russian Language Institute. Issue 6]. Moscow. 272–299. (in Russian)

Eugen Fischer, Paul E. Engelhardt, Joachim Horvath, and Hiroshi Ohtani. 2019. Experimental ordinary language philosophy: a cross-linguistic study of defeasible default inferences. Preprint of paper forthcoming in Synthese. https://philarchive.org/archive/PHISEOL2.

Henry Watson Fowler. 1926. *A Dictionary of Modern English Usage.* Oxford University Press.

Evgeniya Inshakova, Leonid Iomdin, Leonid Mityushin, Viktor Sizov, Tatiana Frolova, and Leonid Tsinman. 2019. SynTagRus segodnya [SynTagRus today]. *Trudy Instituta russkogo yazyka im. V.V. Vinogradova* [Proc. of the V.V. Vinogradov Russian Language Institute]. Moscow. (in Russian) (to appear)

Leonid Iomdin, Vadim Petrochenkov, Viktor Sizov, and Leonid Tsinman. 2012. ETAP parser: state of the art. *Computational Linguistics and Intellectual Technologies. International Conference (Dialog'2012).* RGGU Publishers, 2012. Issue 11(18). Moscow. 830–843.

Igor Mel'čuk. 1974. *Opyt Teorii Lingvisticheskikh Modelei "Smysl ⇔ Tekst"* [Towards a theory of Meaning – Text linguistic models]. Nauka, Moscow. (in Russian)

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

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.

Character-level Annotation
for Chinese Surface-Syntactic Universal Dependencies

Chuanming Dong
Institut National des
Langues et Civilisations
Orientales
dongchuanming@yahoo
.com

Yixuan Li
Sorbonne Nouvelle
Lattice (CNRS)
yixuan.li@sorbonne
-nouvelle.fr

Kim Gerdes
Sorbonne Nouvelle
Almanach (Inria)
LPP (CNRS)
kim.gerdes@sorbonne
-nouvelle.fr

Abstract

This paper presents a new schema to annotate Chinese Treebanks on the character level. The original Universal Dependencies (UD) and Surface-Syntactic Universal Dependencies (SUD) projects provide token-level resources with rich morphosyntactic language details. However, without any commonly accepted word definition for Chinese, the dependency parsing always faces the dilemma of word segmentation. Therefore we present a character-level annotation schema integrated into the existing Universal Dependencies schema as an extension.

1 Introduction

With its writing system being a *Scriptua Continua*, Chinese is a language without explicit word delimiters and thus the "wordhood" is a particularly unclear notion. Yet, the vast majority of downstream NLP tasks for any language are based on "tokens", which mostly boils down to some kind of spelling-based tokenizer. Yet, in the case of Chinese, this step requires a preprocessing step called "word segmentation", whose performance has an non-neglectable influence on the final results. While the F-score of the segmentation task of general texts is in the high nineties since more than 10 years (Emerson 2005) and results have even been slightly improved by recent neural models (Chen et al. 2015, Cai & Zhao 2016), these numbers drop to below 10% for Out-of-Vocabulary terms, i.e. where the system has to take educated guesses on where the word borders are. This leads to catastrophic results for domain specialized texts that use a great number of neologisms unknown to the system, such as patent texts (Li & Gerdes 2019).

Since (Zhao 2009) proposed the first method for character-level dependencies parsing on the Chinese Penn Treebank, a series of research involving the character-based annotation (Li & Zhou 2012; Zhang & al. 2014; Li & al. 2018) have already shown the usefulness of the word-internal structures in Chinese syntactic parsing by obtaining limited but real improvements by means of extra character-level information (character POS, head character position and word internal dependency relation). (Zhao 2009) and (Zhang & al. 2013) have annotated a large-scale word list on Penn Treebank (PTB) and constituent Chinese Treebank (CTB) on the morphological level. Other character-based parsing attempts are generally based on these two annotated corpora.

In this work, we report on the integration of character-level annotations into the Chinese UD treebanks with the goal to find a joint segmentation-parsing method, which enables a multi-granularity analysis on Chinese sentences. Besides the final goal to improve the performance of the dependency parser with character-level information, in particular on out-of-domain texts, this work can also be regarded as a new Chinese word segmentation method: As we distinguish the morphological and syntactic relations between characters by a different set of dependency relation labels, we can ultimately fuse the character parsing results into a simple word segmentation, which can be compared to the original UD word segmentation. The character-level parse tree can thus also be projected onto a dependency tree on the words, which allows us to compare our parsing results with a simple token-based model.

In Section 2 we will briefly introduce various internal structures of Chinese words before presenting our annotation scheme for character-level POS and word internal dependency structures. The experiments and the results obtained are shown in Section 3, followed by the conclusion in Section 4.

2 Internal Dependency Structure of Chinese Words

Chinese words can be largely divided into two categories according to the number of morphemes contained:

1. simple words that contain only a single morpheme (monosyllabic (e.g.花, hua, '*flower*') or polysyllabic (e.g. 巧克力, qiao-ke-li, '*chocolate*'))
2. complex words that contains two or more morphemes.

Polysyllabic simple words are often words that have been directly transliterated form foreign languages and in which all characters have a semantically and syntactically equal status in the word formation. On the other hand, polysyllabic complex words, presenting the overwhelming majority of Chinese words, have more complex relations at the character-level and can also be divided into different subcategories. In the most widely accepted Chinese morphological theory (Feng 1997; Zhang 2003; Pan & al. 2004; Dong 2011), complex words are derivative words or compound words. The latter group includes five types: modifier-head type, coordinative type, predicate-object type, predicate-complement type, and subject-predicate type. In this work, without intention to give a theoretical definition of Chinese word, we aim to analyse the inner structure of already segmented words in UD treebanks.

In order to obtain these inter-character relations, we need to establish and apply syntactic tests that allow us to establish the head of a word based on distributional criteria. In this perspective, it is important to fit the new inter-character relations into a dependency tree that has been established based on similar distributional criteria. That is why our work is based on the Surface-Syntactic Universal Dependencies (SUD) variant of UD (Gerdes & al. 2018), which is an near-isomorphic but more surface syntactic alternative schema to UD with a more classical word distribution-based dependency structure that favors functional heads.

In this section, after an introduction of the different types of complex words in Chinese and their character-level dependency structure with examples (Section 2.1), we describe the three levels of our annotation scheme: determination of the head-daughter relations (the dependency structure), the type of the dependency relation, and the words' POS (Section 2.2).

2.1 Dependency Structure of Complex Words in SUD

In order to keep a clear distinction between word-based and character-based dependency relations, we use a set of specific labels starting with *m:* (standing for morphology) for the character-based relations. In the annotation schema, all under-word level structure in Chinese have an internal relation belonging to one of the four following extended morphological syntactic relations in SUD, which largely correspond to its original SUD syntactic relation types:

1. **m:mod** label given to head-modifier relations
   such as 中<**m:mod** 国 for 中国 zhong guo *center country* 'China'
2. **m:con**j label given to coordinative relations
   such as 自>**m:conj**己 for 自己 zi ji *self self* 'self'
3. **m:arg** label given to subject-predicate, e.g. 脸红 lian hong *face red* blush, predicate-object, e.g. 惊人 jing ren *suprise person* 'superising' and predicate-complement relations in which the complement is usually the result of the predicate, e.g. 减少 jian shao *minor less* 'reduce'
   such as 毕>**m:arg**业 for 毕业 bi ye *accomplish study* 'graduate'
4. **m:flat** label given to unheaded word constructions and to unknown kinds of relations, usually transliterated directely form foreign languages
   such as 巴>**m:flat**黎 for 巴黎 ba li *expect dawn* 'Paris'

For the position of the head in a word, we encounter three different categories of head directions (Zhang & al., 2013): left-headed, right-headed, and coordination (arbitrarily left-right, as in UD/SUD).

Another large category of complex words is made up of derivative words, i.e. usually consisting of the combination of a stem and an affix or the duplication of words. This category of words are analyzed by means of two different dependency relations (**m:mod** or **m:arg**) according to our annotation guidelines.

In this case, it can be hard to determine which character acts as head in the word. For this reason, we apply a series of syntactic tests to find the head: in (1a), it is obvious that the plural affix 们men does not change the syntactic distribution of the whole word and 我wo "me" should be considered as the head; in contrast, in (1b) the verbalizing affix 化hua this time changed the distribution from a nominal compound to a verbal compound. Thus we annotate (1a) with 我wo>**m:mod**们men and (1b) with 现代xiandai<**m:arg**化hua.[1] And here we categorise head-modifier and modifier-head relations in a single group as in UD treebanks the modifier can precede or postcede the head.

(1) a. 我 们
wo men
I,me plural
'we, us'

b. 现代 化
xiandai hua
modern -ize
'modernize'

In order to obtain a systematic and reproducible word-internal dependency analysis, our annotation guide uses a detailed decision tree, that cannot be reproduced here for lack of space. For example, for establishing consistent head-daughter relations, we apply the following tests: (1) Does the added character change the entire distribution? (2) Does the individual characters have the same POS as the whole word? (3) For a given character, can we find a complete paradigm of other words or characters that can occupy the character's position? (4) Is it possible to insert the character 的/地(de, genitive marker) into the word (for testing the modifier-head relation)? (5) Is it grammatically possible to inverse the characters in a word (for testing the coordinative relation)?

We finally annotated the 500 most frequent words in the Chinese SUD corpus, among which we count in total 71 left-headed words, 221 right-headed words and 198 coordinative words. For internal relations, we annotated 222 **m:mod**, 198 **m:conj**, 64 **m:arg**, and 16 **m:flat** relations. The degree of inter-annotator agreement over 100 words reached 88%.

For the remaining words of our corpus we provide an automatic character-based analysis by annotating them with the default left-right relation.

2.2 Statistics-based Character POS Annotation

In order to train a joint tagger-parser, we also need to have character-level POS annotation. To tag the part-of-speech of each character in a Chinese word, we make a list of all the multi-character words (except the polysyllabics which are often tagged as PROPN) in the SUD corpus sorted by frequency. Then, using a character POS dictionary, we insert into the list the character level POS for each word. In order to compare the word level POS and the character level POS, we also insert into the list the most frequent POS of each word. To construct the character level POS dictionary, we combine all the Chinese treebanks in the SUD project, forming a corpus of 299 895 words in total, and we apply the following strategy : If the character has appeared in this corpus as a single-character word, we simply select the most frequent POS of this character alone in the treebanks; on the other hand if the character appears only in multi-character words, we will select the most frequent POS of all the words that contain this character. However, since one character can have multiple POS in different words, the dictionary created by this method can cause plenty errors during the tagging. Therefore we manually

[1] The word 现代xiandai 'modern' is itself a compound word that can be analyzed as 现(xian, 'present') >**m:mod** 代(dai, 'era, generation'), giving the complete analysis (现xian>**m:mod**代dai)<**m:arg**化hua

corrected the character POS of the 1000 most frequent multi-character words in the dictionary. Here are some examples of what we obtain in our dictionary in **Table 1**.

FORM	POS:char1	POS:char2	POS:char3	POS:char4	...	POS:word	Frequency
电影 dian-ying '*film*'	NOUN	NOUN	-	-	...	(NOUN)	96
发展 fa-zhan '*developmen*t'	VERB	VERB	-	-	...	(VERB)	95
平方公里 ping-fang-gong-li '*square kilometer*'	NOUN	NOUN	NOUN	NOUN	...	(NOUN)	90

Table 1 Character POS Dictionary

To train the character level POS tagger, we divide the SUD Chinese corpus into 3 sets: a training set of 151 954 words, a developing set of 4 469 words, and a testing set of 4 232 words. We then convert these 3 sets of treebanks from word level to a character level by splitting all the complex words. And by using the dictionary that we obtain from the last step, we insert into these treebanks the character level POS, and we can thus train a POS tagger on the characters with these 3 sets using a proper deep learning algorithm such as LSTM. This approche give us a 91% accuracy of the characters POS tagging when we used the tagger of the Dozat parser (Dozat 2016) to train our character level tagger.

3 Experiments

We have worked on the four Chinese UD treebanks converted into SUD format and simplified characters when necessary: The Traditional Chinese Universal Dependencies Treebank annotated by Google (GSD), the Parallel Universal Dependencies treebanks created for the CoNLL 2017 shared task on Multilingual Parsing (PUD), the Traditional Chinese treebank of film subtitles and of legislative proceedings of Hong Kong (HK), and the essays written by learners of Mandarin Chinese as a foreign language (CFL), also proposed by the City University of Hong Kong.

To train the character-based POS tagger and SUD parser, we choose the Graph-based Neural Dependency Parser developed by Timothy Dozat at Stanford University for its character-based LSTM word representation. This parser contains a tagger training network and a dependency parser training network, but unfortunately these two training processes are separated, meaning that to obtain a corpus tagged and parsed, first we have to train a tagger, use it to tag our corpus, then train a parser and use it to parse our tagged corpus. Before the training process, we have also prepared a character vector file which is trained by BERT, a word embedding model developed by Google with a pre-trained character based Chinese model.

Our experiments consist of using Dozat Parser to train the word-based (WB) tagger and parser, as well as the character-based (CB) tagger and parser. Then by applying them to tag and parse our test corpus we can obtain two versions of our treebank: a word-based and a character-based treebank (see **Annex 2**), so that we can perform systematic tests of comparison on the combined Chinese SUD treebanks and evaluate the performance of our character-based tagger and parser. To sum up, we need to go through at least four training processes: WB tagger training, WB parser training, CB tagger training and CB parser training. Therefore, we have prepared our training data as following: for WB tagger and parser training, we extract the last 10% of the four former mentioned Chinese SUD treebanks to serve as the testing set and the developing set, and we combine the rest 90% to serve as a training set; for CB tagger and parser training, we carry out the exact same arrangement, except this time all the treebanks are converted from word level to character level.

Concerning the tagger, we compare the F-score of the tagger trained on WB and on CB. The **Table 2** display the direct result of the CB tagging.

Category	**Precision**	**Recall**	**F-score**
ADJ	89.37%	87.98%	88.67%
ADP	88.55%	81.38%	84.81%
ADV	89.33%	90.17%	89.75%
AUX	75.46%	89.96%	82.07%
CCONJ	95.92%	63.51%	76.42%
DET	89.36%	77.78%	83.17%
INTJ	66.67%	66.67%	66.67%
NOUN	93.20%	94.10%	93.65%
NUM	93.53%	100.00%	96.65%
PART	96.43%	96.72%	96.57%
PRON	96.06%	97.99%	97.01%
PROPN	73.37%	82.12%	77.50%
PUNCT	100.00%	100.00%	100.00%
SCONJ	0.00%	0.00%	0.00%
SYM	100.00%	100.00%	100.00%
VERB	92.22%	89.89%	91.04%
TOTAL	**91.99%**	**91.87%**	**91.93%**

Table 2 F-score of character level POS for our character-based tagg

As we can see, the Dozat parser achieved a rather high score on CB tagging. Some POS, because of it's absence on a character level, doesn't have a remarkable score, like SCONJ, but regardless of that we believe this tagger can satisfy our basic need in CB tagging. However, we can not compare directly this result with the result of WB tagging, since the words in the treebank for CB tagging has been split into characters, and thus we don't have the exact same number of POS in the WB and CB tagged treebanks. Therefore a recombination of the CB treebank after the tagging is necessary. To facilitate the recombination, we use the XPOS column in our treebank (under Conll-U format) to record the word level POS. When we split a word into characters during the preparation of treebanks for tagger training, we insert the character level POS into the UPOS column, and copy the word's original POS to the XPOS column of each character. And since in Dozat Parser the prediction of XPOS is dependent on the prediction of UPOS, we can thus train a tagger that can tag WB POS based on the CB POS. The following are the results of WB tagging (**Table 3**) and CB tagging after the recombination (**Table 4**)

Category	**Precision**	**Recall**	**F-score**
ADJ	65.69%	50.00%	56.78%
ADP	63.48%	69.75%	66.47%
ADV	80.08%	76.40%	78.20%
AUX	59.84%	81.56%	69.03%
CCONJ	92.68%	58.46%	71.70%
DET	96.81%	68.94%	80.53%
INTJ	100.00%	0.00%	0.00%
NOUN	88.17%	82.27%	85.12%
NUM	63.92%	98.41%	77.50%
PART	84.03%	91.74%	87.72%
PRON	94.06%	93.14%	93.60%
PROPN	38.17%	89.29%	53.48%
PUNCT	99.84%	99.84%	99.84%
SCONJ	100.00%	0.00%	0.00%
SYM	100.00%	0.00%	0.00%
VERB	76.29%	77.56%	76.92%
TOTAL	**81.85%**	**81.62%**	**81.74%**

Table 3 F-score of word level POS (UPOS) for our word-based tagger

Category	**Precision**	**Recall**	**F-score**
ADJ	65.52%	42.54%	51.58%
ADP	60.11%	87.90%	71.40%
ADV	75.00%	70.80%	72.84%
AUX	64.71%	86.03%	73.86%
CCONJ	92.68%	58.46%	71.70%
DET	91.22%	86.45%	88.77%
INTJ	100.00%	20.00%	33.33%
NOUN	77.87%	85.56%	81.54%
NUM	65.14%	93.65%	76.84%
PART	91.56%	94.50%	93.00%
PRON	92.47%	88.24%	90.30%
PROPN	54.05%	71.43%	61.54%
PUNCT	99.84%	100.00%	99.92%
SCONJ	20.00%	4.35%	7.14%
SYM	100.00%	100.00%	100.00%
VERB	83.31%	76.41%	79.71%
TOTAL	**88.85%**	**88.70%**	**88.78%**

Table 4 F-score of word level POS (XPOS) for our character-based tagger after the recombination

As we can see from these two tables above, the training on a character base has greatly improved the performance of the tagger. However for some most common POS, like ADJ and NOUN, there's an obvious decline of f-score. One of the possible reasons is that there's an inconsistency between the word level POS and character level POS in Chinese. For example, 活动 (NOUN, '*activity*') is composed by two verbal character "活" (VERB, '*living*') and "动" (VERB, '*moving*'). But by reviewing our data, we noticed that there's also an inconsistency on the POS annotation of the same words between different treebanks, even if in a similar context. This problem may have a bigger influence on both tagger and parser.

Concerning the parser, we have the usual UAS and LAS, but in addition the Orthogonal Label Unattached Score (OLS) that simply measures whether the word is connected to its governor with the right relation, independently whether the governor is correct (**Table 5**)

	WB	**CB**
UAS	78.96%	81.72%
OLS	81.29%	85.93%
LAS	66.65%	72.99%

Table 5 Comparison between the results of WB and CB parser

By comparing the UAS, OLS and LAS between the WB and CB parser, we can see that although the CB parser can correctly recognise more heads and dependency relations, the score is still relatively low, especially for the recognition of the dependency tree (LAS)

This is due to several possible reasons, including the incomplete character POS annotation and word structure annotation. Since we haven't totally finished the pretreatment process, there's a problem of inconsistency in our data, with the same word in the same context but having different POS or different internal structure annotated.

We can also separately measure the performance on the syntactic and morphological dependencies (**Table 6**). This method has a special function, that is the performance of the segmentation can be evaluated by concerning only about the two main groups of dependency relations: Morphe (relations annotated with m: at the beginning) and Deprel (the original dependency relations in SUD).

	Morph (Gold)	**Deprel (Gold)**	**TOTAL**
Morphe	2099	2	2101
Deprel	0	3128	3128
Wrong Head	4	1092	1096
TOTAL	2103	4222	6325

Table 6 Binary Confusion Matrix for Relations at Word/Character-level

The parsing error analysis has shown that the comparatively inferior recall scores for almost all types of relations are largely caused by the great quantity of false annotation of head-dependent arcs, while the morphe relations is the only one with a high recall (above 99%). Some relations with especially high head-dependency arc errors include clf, conj, dep, flat and punct. In contrast, the precision scores of most of dependency relations have passed 80% or close to it, with the exception of obl (62%, confusing with various types) and parataxis (47%, confusing frequentitly with comp) type relations. See **Annex 3** and **Annex 4** for more details about our evaluation data.

One possible reason behind these errors is the annotation error at previous tagger step, which also involve the dismatch of word POS annotations between different original Chinese Treebanks (e.g. the ordinal numbers are annotated as ADJ in certain corpus and as NUM in others). This the lack of equivalence may later lead the neural parser to some incorrect intuitions from statistics.

The f-score of the morphe relation is about 99.85% (**Table 6**) . The low annotation error (around 0.15%) shows an outstanding capability of the parser to distinguish character-level and word-level relations, and thus has the potential to serve as a decent word segmenter.

4 Conclusion

In this paper, we have presented a character-level annotation schema for modern Chinese and evaluated the state-of-the-art parser trained to annotate character level POS and dependency relations based on this schema. By comparing it with the word-based tagger and parser, we have witnessed a progress in the accuracy of this annotation system. However, after the evaluation we found out that the score for dependency tree annotation are not so satisfying. According to our error analysis, we conclude that there are mainly three reasons: incomplete and incorrect character level POS annotation, incomplete word structure annotation and discorrespondance in annotation between treebanks, all of them causing the irregularity of our data and thus confuse the algorithm to find the pattern. The solution is clear, by normalizing the data we can make further progress at improving the accuracy of our parser. Thus our next step is to establish formal annotation guidelines for this annotation schema in order to refine SUD treebanks so that them can be better adapted to our training system. Also, there's still room for improvement in our character POS annotation and word structure annotation, for example instead of using the most frequent POS for a single character and manually correct the faults, we can use deep learning algorithm to assign the most probable POS to a character judging by its context. And by accomplishing these two tasks we can provide our parser with a more powerful morphological support to achieve a more thorough syntactic analysis.

In spite of a less favorable score, these preliminary results show that it is actually possible to skip the word segmentation task and perform a joint segmentation and parsing. This has been shown to work on the existing Chinese UD dependency treebanks. We expect this to be useful for parsing texts with high rates of neologisms such as technological texts, but we will have to show that the joint parsing performance will not be too negatively affected itself by the unknown words. Yet, intuitively, it seems likely that the new words also show a systematic internal behavior and that many of the head-daughter relations can be correctly predicted because the individual characters have appeared elsewhere in the training corpus even if the combined word is new to the parser. Work is in progress to test this claim on Chinese patent texts.

We consider this work to be a step out of the hen-and-egg problem of tokenization and syntactic analysis: A parser needs tokens and a tokenizer needs syntactic information. Yet, a parser is an optimized tool to predict structure depending on the context. There is no reason that word-internal relations cannot be predicted in the same way as syntactic relations among words, even more so as many of these relations, in particular for compound words, actually correspond and behave very similarly to syntactic relations. This is an interesting result, not only for a scriptua continua on an isolating language such as Chinese but for other languages, too, where a morphological decomposition could be a successful basis for dependency parsing as long as the decomposition is linguistically well-grounded.

References

Cai D., Zhao H. 2016. Neural Word Segmentation Learning for Chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 409-420).

Chang S. 2003. On the Study of Compounds : A Contrastive Analysis of Chinese, English and Japanese. *In Proceedings of the 7th World Symposium On Chinese Language Teaching*. 張淑敏，〈漢英日複合詞的對

比分析：分類、結構與衍生〉，《第七屆世界華語文教學研討會論文集》，世界華語文教育學會，2003年12月。

Chen X., Qiu X., Zhu C., Liu P., Huang X. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1197-1206.

Dong X. 2011. *Lexicalization: The Origin and Evolution of Chinese Disyllabic Words*. Sichuan: Sichuan Minorities Press.

Dozat T., Manning C. D. 2016. Deep Biaffine Attention for Neural Dependency Parsing. arXiv preprint arXiv:1611.01734.

Emerson T. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Feng S. 1997. *Interactions between Prosody, Morphology and Syntax in Chinese*. Beijing: Peking University Press.

Gerdes K., Guillaume B., Kahane S., Perrier G. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *In Proceedings of the Universal Dependencies Workshop* (UDW), EMNLP, Bruxelles.

Li H., Zhang Z., Ju Y., Zhao H. 2018. Neural character-level dependency parsing for Chinese. In The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).

Li Y., Gerdes K. 2019. In *Proceedings of the 13th TOTh International Conference* (TOTh 2019).

Li Z., Zhou G. 2012. Unified dependency parsing of Chinese morphological and syntactic structures. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1445–1454.

Li Z. 2011. Parsing the internal structure of words: a new paradigm for Chinese word segmentation. In *Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, 1405–1414.

Zhang M., Zhang Y., Che W., Liu T. 2013. Chinese Parsing Exploiting Characters. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

Zhang M., Zhang Y., Che W., Liu T. 2014. Character-Level Chinese Dependency Parsing. *In Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL 2014).*

Packard J.L. 2000. *The morphology of Chinese. A linguistic and cognitive approach*. Cambridge University Press, Cambridge.

Pan W., Ye B., Han Y. 2004. *The research on word formation in Chinese*. Shanghai: Huadong Shifan Daxue Chubanshe. 潘文国，叶步青,《汉语的构词法研究》，上海：华东师范大学出版社，2004。

Zhao H., Kit C., Song, Y. 2009. Character dependency tree based lexical and syntactic all-in-one parsing for chinese. In *The 10th Chinese National Conference on Computational Linguistics (CNCCL-2009)*, 82–88.

Zhao H. 2009. Character-level dependencies in Chinese: Usefulness and learning. *In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, 879–887.

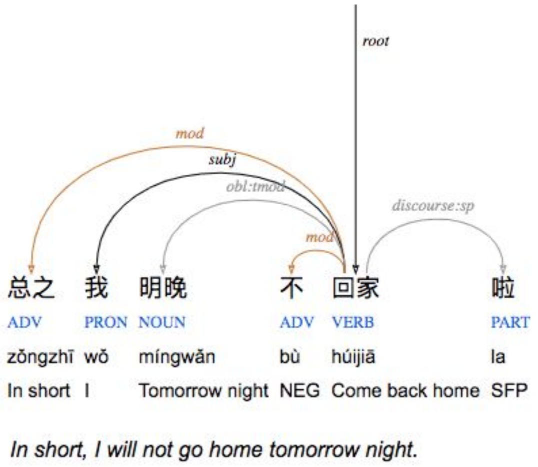
Annex 1

The annotation of the head-dependency relation follows the CoNLL-U Format for UD and SUD (https://universaldependencies.org/format.html), in which every line for a single token including its annotation in 10 fields (ID, FORM, LEMMA, UPOS, XPOS, FEATS, HEAD, DEPREL, DEPS, MISC) separated by single tab characters. In our retokenized Chinese sentences, each line is devoted to a single character. Based on the dictionary of all Chinese words in the SUD corpus annotated with its head position and internal dependency relation type, we automatically integrate these character-level information into the converted CoNLL file with a Python script.

In the actual annotation process, we only indicate the index of the head character in the field of HEAD, as it is done for the syntactic dependencies.

Annex 2

And here is a comparison between the word-based (WB) treebank (Figure 1) and the character-based (CB) treebank (Figure 2) of the same sentence in Chinese.


root
mod
subj
obl:tmod
discourse:sp
mod
总之 我 明晚 不 回家 啦
ADV PRON NOUN ADV VERB PART
zǒngzhī wǒ míngwǎn bù húijiā la
In short I Tomorrow night NEG Come back home SFP
In short, I will not go home tomorrow night.

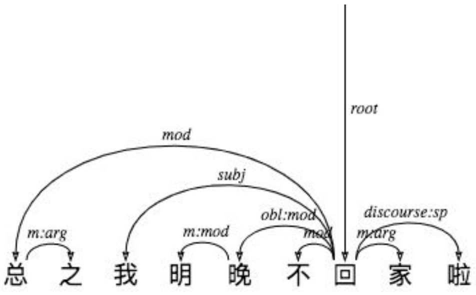


Figure 1 word-based treebank

Figure 2 character--based treebank

Annex 3

Confusion matrix of dependency relations annotated by our character-based parser

	appos (Golden)	case	cc	clf	comp	compound	conj	dep	det	discourse	dislocated	flat	mark	mod	morphe*	obj	obl	parataxis	punct	reparandum	root	subj	vocative
appos	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
case	0	23	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cc	0	0	35	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
clf	0	0	0	23	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
comp	0	0	0	0	791	4	7	3	0	2	0	0	0	23	0	0	1	1	0	0	0	0	0
compound	0	0	0	0	5	136	1	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
conj	0	0	0	0	1	0	35	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
dep	0	1	0	0	1	6	1	265	1	0	0	0	1	6	0	0	1	0	0	0	0	0	0
det	0	0	0	0	0	0	0	3	106	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discourse	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0
dislocated	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
flat	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
mark	0	0	0	0	0	5	0	0	0	1	0	0	55	1	0	0	0	0	0	0	0	1	0
mod	0	0	0	0	1	0	4	1	0	2	0	0	0	449	0	0	0	0	0	0	0	0	0
morphe*	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2099	0	0	0	0	0	0	0	0
obj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
obl	0	0	0	0	0	0	0	2	0	0	0	0	0	4	0	1	23	0	0	0	0	4	3
parataxis	0	0	0	0	11	0	2	1	0	0	0	0	0	4	0	0	0	8	0	0	0	0	0
punct	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	413	0	0	1	0
reparandum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
root	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	266	0	0
subj	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	291	2
vocative	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	8
Wrong Head	4	3	11	39	215	33	63	138	29	8	2	35	15	143	4	6	12	7	219	1	41	60	6

Annex 4

Comparison between the parsing result of our word-based parser and character-based parser on several most frequent relations.

Category	Precision	Recall	F-score
case	89.66%	96.30%	92.86%
cc	70.31%	95.74%	81.08%
clf	89.71%	90.39%	90.04%
comp	80.82%	84.96%	82.83%
compound	66.67%	77.42%	71.64%
conj	56.04%	44.74%	49.76%
det	96.21%	93.38%	94.78%
discourse	93.62%	84.62%	88.89%
mark	76.71%	78.87%	77.78%
mod	90.71%	78.86%	84.37%
obl	45.10%	62.16%	52.27%
parataxis	5.13%	11.11%	7.02%
punct	99.53%	100.00%	99.76%
root	85.34%	85.34%	85.34%
subj	79.27%	84.12%	81.62%
vocative	100.00%	0.00%	0.00%
TOTAL	**81.41%**	**75.49%**	**78.33%**

Table 7 F-score of the most frequent dependency relations of the word-based parser

Category	Precision	Recall	F-score
case	85.19%	85.19%	85.19%
cc	73.77%	95.74%	83.33%
clf	91.94%	91.94%	91.94%
comp	78.74%	85.19%	81.84%
compound	62.93%	78.49%	69.86%
conj	62.32%	37.72%	46.99%
det	96.27%	94.85%	95.56%
discourse	97.78%	84.62%	90.72%
mark	71.43%	84.51%	77.42%
mod	90.94%	78.93%	84.51%
obl	62.00%	70.27%	65.88%
parataxis	47.02%	44.44%	45.69%
punct	99.68%	100.00%	99.84%
root	86.64%	86.64%	86.64%
subj	79.08%	86.35%	82.56%
vocative	81.82%	47.37%	60.00%
TOTAL	**83.67%**	**78.81%**	**81.17%**

Table 8 F-score of the most frequent dependency relations of the character-based parser after the recombination of characters