ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 11, Number 2
december 2025
Special Issue

Bridging Theoretical Linguistics and Automated Language Processing: Emerging Synergies and Advances

Accademia university press

editors in chief

Roberto Basili | Università degli Studi di Roma Tor Vergata (Italy)
Simonetta Montemagni | Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)

advisory board

Giuseppe Attardi | Università degli Studi di Pisa (Italy)
Nicoletta Calzolari | Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Nick Campbell | Trinity College Dublin (Ireland)
Piero Cosi | Istituto di Scienze e Tecnologie della Cognizione - CNR (Italy)
Rodolfo Delmonte | Università degli Studi di Venezia (Italy)
Marcello Federico | Amazon AI (USA)
Giacomo Ferrari | Università degli Studi del Piemonte Orientale (Italy)
Eduard Hovy | Carnegie Mellon University (USA)
Paola Merlo | Université de Genève (Switzerland)
John Nerbonne | University of Groningen (The Netherlands)
Joakim Nivre | Uppsala University (Sweden)
Maria Teresa Pazienza | Università degli Studi di Roma Tor Vergata (Italy)
Roberto Pieraccini | Google, Zürich (Switzerland)
Hinrich Schütze | University of Munich (Germany)
Marc Steedman | University of Edinburgh (United Kingdom)
Oliviero Stock | Fondazione Bruno Kessler, Trento (Italy)
Jun-ichi Tsujii | Artificial Intelligence Research Center, Tokyo (Japan)
Paola Velardi | Università degli Studi di Roma "La Sapienza" (Italy)

editorial board

Pierpaolo Basile | Università degli Studi di Bari (Italy)
Valerio Basile | Università degli Studi di Torino (Italy)
Arianna Bisazza | University of Groningen (The Netherlands)
Cristina Bosco | Università degli Studi di Torino (Italy)
Elena Cabrio | Université Côte d'Azur, Inria, CNRS, I3S (France)
Tommaso Caselli | University of Groningen (The Netherlands)
Emmanuele Chersoni | The Hong Kong Polytechnic University (Hong Kong)
Francesca Chiusaroli | Università degli Studi di Macerata (Italy)
Danilo Croce | Università degli Studi di Roma Tor Vergata (Italy)
Francesco Cutugno | Università degli Studi di Napoli Federico II (Italy)
Felice Dell'Orletta | Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Elisabetta Fersini | Università degli Studi di Milano - Bicocca (Italy)
Elisabetta Jezek | Università degli Studi di Pavia (Italy)
Gianluca Lebani | Università Ca' Foscari Venezia (Italy)
Alessandro Lenci | Università degli Studi di Pisa (Italy)
Bernardo Magnini | Fondazione Bruno Kessler, Trento (Italy)
Johanna Monti | Università degli Studi di Napoli "L'Orientale" (Italy)
Alessandro Moschitti | Amazon Alexa (USA)
Roberto Navigli | Università degli Studi di Roma "La Sapienza" (Italy)
Malvina Nissim | University of Groningen (The Netherlands)
Nicole Novielli | Università degli Studi di Bari (Italy)
Antonio Origlia | Università degli Studi di Napoli Federico II (Italy)
Lucia Passaro | Università degli Studi di Pisa (Italy)
Marco Passarotti | Università Cattolica del Sacro Cuore (Italy)
Viviana Patti | Università degli Studi di Torino (Italy)
Vito Pirrelli | Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Marco Polignano | Università degli Studi di Bari (Italy)
Giorgio Satta | Università degli Studi di Padova (Italy)
Giovanni Semeraro | Università degli Studi di Bari Aldo Moro (Italy)
Carlo Strapparava | Fondazione Bruno Kessler, Trento (Italy)
Fabio Tamburini | Università degli Studi di Bologna (Italy)
Sara Tonelli | Fondazione Bruno Kessler, Trento (Italy)
Giulia Venturi | Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Guido Vetere | Università degli Studi Guglielmo Marconi (Italy)
Fabio Massimo Zanzotto | Università degli Studi di Roma Tor Vergata (Italy)

editorial office

Danilo Croce | Università degli Studi di Roma Tor Vergata (Italy)
Sara Goggi | Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Manuela Speranza | Fondazione Bruno Kessler, Trento (Italy)

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)
© 2025 Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

BY NC ND

direttore responsabile
Michele Arnese

isbn 9791255001522

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_11_2

aAccademia university press

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 11, Number 2
december 2025
Special Issue

Bridging Theoretical Linguistics and Automated Language Processing: Emerging Synergies and Advances

Guest Editors:
Alessandro Lenci, Marco Passarotti, Rachele Sprugnoli, Fabio Tamburini

CONTENTS

Bridging Linguistics and Computational Linguistics: Insights into Synergies and Challenges from a Case Study

Simonetta Montemagni*
Istituto di Linguistica Computazionale
"Antonio Zampolli", CNR

This paper outlines the evolving interplay between Linguistics and Computational Linguistics, aiming to map the current state of their interactions and to identify areas where deeper integration could drive significant advancements in both areas. Since the early days of Computational Linguistics as an autonomous discipline, the synergy has developed in parallel with progress in both computational methods and linguistic theory. Computational modeling of language offers a powerful framework to investigate core questions of linguistics, from how language works and is acquired, to how it changes across time, space, communicative situations, and domains.

Despite this potential, the capabilities of state-of-the-art computational methods remain only partially exploited within linguistic research, leaving a gap between advances in Natural Language Processing and the needs of linguistics. This paper seeks to examine the current landscape of this synergy, its scientific and practical implications, and the challenges that must be addressed to fully harness its potential. A pilot study is presented to illustrate how linguistic resources and computational modeling can provide answers to long-standing research questions and, at the same time, open up new avenues for investigating open issues in language typology.

1. Introduction

The interaction between Computational Linguistics (CL) and Linguistics has a long and articulated history, dating back to the 1960s. As noted by Kučera (1982), "computational linguistics provides important potential tools for testing theoretical linguistic constructs and their power to predict actual language use". More than two decades later, Martin Kay, in his ACL Lifetime Achievement Award speech (Kay 2005), echoed this view: "Computational linguistics is trying to do what linguists do in a computational manner". Yet, despite these programmatic statements, the relationship between CL and Linguistics has often been less productive than initially envisioned. Over time, it has undergone multiple shifts, shaped by evolving theoretical paradigms, methodological trends, and technological capabilities.

Today, the dialogue between CL and linguistics reflects a growing recognition of the mutual benefits of integrating technological advances with theoretical insights - a convergence that is reshaping both fields in profound ways. This interaction unfolds along two complementary axes: linguistics for CL and CL for linguistics. These dimensions are intertwined in a virtuous circle. Computational models can refine linguistic theory by deepening our understanding of language structure, acquisition, and its relation

* E-mail: simonetta.montemagni@cnr.it

© 2025 Associazione Italiana di Linguistica Computazionale

to other cognitive capacities. Conversely, advances in understanding how language functions enable the development of more robust and accurate Natural Language Processing (NLP) components. This reciprocity exemplifies a broader scientific principle: deep theoretical knowledge often underpins technological progress, while practical challenges can inspire theoretical innovation.

This paper focuses on the CL for linguistics dimension. Between 2011 and 2014, Mark Liberman repeatedly described that moment as a potential "golden age of speech and language science"[1]. From a linguistic perspective, the emergence of vast digital archives of text and speech, combined with advanced analytical tools and affordable computing power, represented a leap in research capabilities comparable to the invention of the telescope or microscope in the 17th century. These technologies have made it possible to study linguistic phenomena across time, space, domains, and cultural contexts with unprecedented scale and precision, offering empirical insights far beyond earlier possibilities.

More than a decade later, the full potential of CL technologies is still unfolding, yet their contribution to linguistic inquiry is already substantial. Building on the unprecedented opportunities offered by large-scale digital resources and advanced computational tools, new avenues of empirical research have emerged across multiple areas of linguistics. To explore this contribution, the paper illustrates a case study aimed at showing how the integration of computational tools not only enhances the ability to process and analyze language data, but also contributes to generating fine-grained, reliable linguistic knowledge that feeds back into theoretical models of language variation and change. In particular, it presents results from a line of research aimed at identifying patterns of language dynamics - both interlinguistically and intralinguistically - that exemplify how linguistic resources and NLP methods can address current needs of linguistic research.

The case study focuses on linguistic typology, a subfield of linguistics whose main goals, as summarized by Croft (2003), are (i) the classification of languages and (ii) the identification and explanation of cross-linguistic generalizations. In recent years, linguistic typology has increasingly attracted the attention of NLP research, as the rise of machine learning has brought language independence to the forefront of research objectives. However, as Bender (2009) points out, the development of truly language-independent NLP systems crucially depends on linguistic knowledge, particularly on the generalizations about structural variation across languages provided by linguistic typology. Typology, therefore, constitutes a natural meeting ground between computational linguistics and theoretical linguistics.

The paper is organized as follows. Section 2 reviews the evolution of the relationship between CL and linguistics over time, since the CL origins to the present. Section 3 outlines the theoretical background of the case study from the twofold perspective of linguistics and CL/NLP, while sections 3 and 4 focus respectively on the contribution of annotated corpora and on the added value of NLP. Finally, Section 5 discusses the results in the broader context of language typology and highlights key issues for strengthening the synergy between CL and linguistics.

1 See e.g. the invited lecture given by Mark Liberman (University of Pennsylvania) in the framework of the Annual Meeting of the Linguistics Association of Great Britain at the University of Manchester in 2011, entitled *Towards the golden age of speech and language science*.

2. The Evolution of the Relationship Between Computational Linguistics and Linguistics

2.1 Historical Sketch

The first wave of collaboration in the 1960s was driven by efforts to implement transformational grammars for automatic natural language parsing. Although these early systems were constrained by the computational limitations of the time, they laid the foundations for a closer integration of formal linguistic theory and computational modeling. Building on this groundwork, subsequent decades witnessed the development of increasingly sophisticated grammar formalisms, including Lexical Functional Grammar (LFG) (Kaplan and Bresnan 1982), Functional Unification Grammar (FUG) (Kay 1984), Generalized Phrase Structure Grammar (GPSG) (Gazdar et al. 1985), Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag 1994), and Combinatory Categorial Grammar (CCG) (Steedman 2000). All these frameworks shared a central assumption with theoretical linguistics: that syntactic structure underlies semantic interpretation.

Over time, the scope of collaboration expanded beyond syntax. Advances in computational power, the growing availability of large-scale datasets, and the increasing interdisciplinarity of language research opened up new opportunities for joint work. Prominent areas include discourse processing (Grosz and Sidner 1986; Walker, Joshi, and Prince 1998) and the development of linguistic resources (Marcus, Santorini, and Marcinkiewicz 1993; Fellbaum 1998; Prasad et al. 2008). Within the broader landscape of the language sciences, significant intersections also emerged with lexicography (Boguraev and Briscoe 1989; Wilks, Slator, and Guthrie 1996), psycholinguistics (Crocker 1996; Dijkstra and de Smedt 1996), and corpus linguistics (McEnery and Wilson 2001; Sampson 2001). Across all these domains, computational methods enriched linguistic inquiry, while linguistic theory continued to inform the design of computational models and resources.

Over time, however, this initially close relationship between computational linguistics and linguistics gradually weakened. Several factors contributed to this divergence. On the computational side, there was a growing predominance of methodologies oriented toward NLP, a shift reinforced by the rise of empirical, data-driven approaches. These developments marginalized linguistically informed models in favor of statistical and machine-learning techniques, and the resulting emphasis on application-driven research was reflected in the adoption of terms such as *Natural Language Engineering* to describe the field. Meanwhile, in mainstream linguistic theory, corpus-based evidence continued to play a relatively minor role, limiting opportunities for integration with computational approaches. The consequence was a widening gap between research in computational linguistics, on the one hand, and linguistics, on the other. Such a divide is memorably, if provocatively, captured in Frederick Jelinek's well-known quip: "Every time I fire a linguist, the system performance improves".

Over the past fifteen years, however, this trend has reversed. The historical synergy between computational linguistics and linguistics has experienced a marked revival, driven by multiple converging factors. On the CL side, the growing maturity of language technologies and the increasing availability of large-scale digital resources - particularly textual corpora - have been key enablers. In linguistics, new scientific paradigms have emerged, exemplified by the notion of "quantitative turn" proposed in Kortmann (2021), which emphasizes the central role of empirical, quantitative, and statistical methods in linguistic research. As a result, corpus linguistics, probabilistic approaches, and usage-based models have gained prominence, becoming an integral part of contemporary linguistic inquiry.

Empirical evidence for this shift comes from Kortmann's analysis of 380 research articles published in the journal *English Language and Linguistics* between 1997 and 2019, which docuemnts a sharp decline in purely qualitative studies alongside a steady increase in the use of statistical methods, a trend also observed in other major linguistics journals. The combination of extensive linguistic resources with advances in computational research - particularly in machine learning, NLP and statistical data analysis - has further reinforced this methodological reorientation, fostering a "computational turn" in linguistics. This turn reflects a growing commitment to the exploitation and development of computational methods and resources to address core research questions, thereby reestablishing fertile ground for collaboration between the two disciplines.

Numerous editorial initiatives and publications attest to this renewed synergy. Since 2008, the journal *Linguistic Issues in Language Technology* (LiLT)[2] has focused on the mutual enrichment between linguistic insights and language technology. Special issues have explicitly addressed the interaction between linguistic theory and computational methods, such as *Implementation of Linguistic Analyses Against Data* (2010) and *Interaction of Linguistics and Computational Linguistics* (2011), alongside volumes devoted to more specific themes, including theoretical vs. computational morphology (2014) and the relationship between annotated corpora and linguistic research (2012, 2019). Other significant contributions come from scholars such as Emily M. Bender, whose works on linguistic fundamentals for NLP have become widely cited references, see e.g. Bender (2013), Bender and Lascarides (2019).

Collaborative spaces have also been fostered by workshops and conferences explicitly devoted to bridging these fields. Notable examples include the *Statistical Parsing of Morphologically Rich Languages* workshop series (since 2010)[3], which provides a forum for discussing the challenges in parsing languages where substantial syntactic information is encoded at the word level, and the more recent SIGTYP workshop series (since 2019)[4], a dedicated venue for typology-related research and its integration into multilingual NLP.

This synergy has been further extended through Digital Humanities initiatives and workshops, including:

- the workshop series *Computational Linguistics for Cultural Heritage, Social Sciences, Humanities, and Literature, SIGHUM* (since 2007)[5];
- the workshop series *Natural Language Processing for Digital Humanities, NLP4DH* (since 2021)[6];
- the workshop series *Language Technologies for Historical and Ancient Languages, LT4HALA* (held on 2020, 2022, and 2024)[7], which has hosted evaluation campaigns such as *EvaLatin*, for the evaluation of NLP tools for Latin, and *EvaHan* for Ancient Chinese Information Processing.

2 Published by the University of Colorado Boulder and licensed under CC BY 4.0, ISSN 1945-3604, `https://journals.colorado.edu/index.php/lilt/issue/archive`.
3 `https://aclanthology.org/venues/spmrl/`
4 `https://aclanthology.org/venues/sigtyp/`
5 `https://sighum.wordpress.com/`
6 `https://www.nlp4dh.com/`
7 `https://circse.github.io/LT4HALA/`

In addition to these interdisciplinary events, special issues of CL journals, such as *Language Resources and Evaluation, LRE* (Hinrichs et al. 2019), and the *Italian Journal of Computational Linguistics, IJCoL* (Nerbonne and Tonelli 2016) have addressed the intersection of language technologies and Digital Humanities. Monographs such as Piotrowski (2012) and Jenset and McGillivray (2017) have further contributed to this debate by exploring the use of NLP tools for historical and classical texts.

2.2 Computational Linguistics and Linguistics Today

We have seen that, in recent years, CL and linguistics have become increasingly intertwined. Linguistic theory provides essential guidance for the development, evaluation and interpretation of NLP systems, offering insights from syntax, semantics, and pragmatics. Conversely, advances in computational methods, from traditional statistical feature-based models to deep learning and Large Language Models (LLMs), equip linguists with powerful tools for large-scale corpus analyses, automatic linguistic annotation, and empirical testing of theoretical hypotheses. This bidirectional interplay, on the one hand, enhances the capabilities of computational models and, on the other hand, opens new avenues for theoretical and experimental research in linguistics.

A key distinction lies in the nature and scope of traditional machine learning models versus LLMs. Traditional supervised models rely on structured, labeled datasets with explicitly defined features, enabling precise control over input, interpretability, and reproducibility. These characteristics make them particularly suited for hypothesis-driven, fine-grained linguistic analyses. LLMs, by contrast, typically use an end-to-end paradigm, directly mapping linguistic input to output without intermediate linguistic modules or explicit linguistic (e.g. syntactic) annotation: they are trained on vast, unstructured text corpora using self-supervised objectives, learning to predict the next token or sequence in context, with significant implications for their relationship with theoretical linguistics.

Due to their large-scale training and deep architectures, LLMs are emerging as valuable instruments for linguistic research. They have been shown to perform well on behavioral tasks such as grammaticality judgments (Hu et al. 2024) or long-distance agreement prediction (Linzen, Dupoux, and Goldberg 2016; Gulordava et al. 2018; Nastase et al. 2024). They can be probed to test theoretical hypotheses (Tenney, Das, and Pavlick 2019; Hewitt and Manning 2019), metalinguistic abilities (Beguš, Dabkowski, and Rhodes 2025), or the learnability of typologically plausible vs implausible languages (Xu et al. To appear). LLMs can also be employed for linguistic annotation (Tan et al. 2024). Given their human-like communication abilities, LLMs also raise new methodological and theoretical questions about what it means for a machine to "know" language (Bender and Koller 2020; Chomsky, Roberts, and Watumull 2023; Futrell and Mahowald 2025).

However, when it comes to fine-grained theoretical analysis, that is, metalinguistic reasoning which lies at the core of linguistic theory, LLMs still display limited robustness and internal consistency, despite recent evidence of improvement (Beguš, Dabkowski, and Rhodes 2025). Their responses often fluctuate depending on prompt formulation or contextual framing, revealing a lack of stable underlying representations of linguistic principles. In contrast, supervised machine learning methods remain more reliable for linguistically grounded research, as they enable explicit control over input features, training data, and evaluation metrics. Such approaches are therefore better suited to hypothesis-driven investigations that require transparent modeling choices and reproducible empirical validation.

The two approaches can advance both computational and theoretical linguistics in complementary ways. The differences have important methodological and theoretical implications. While supervised models remain the preferred choice for tasks requiring precision, interpretability, and controlled hypothesis testing, LLMs provide opportunities for exploratory analyses, large-scale simulations, and probing complex linguistic patterns that were previously inaccessible. By carefully selecting the computational framework in accordance with the specific research questions being investigated, linguists can take advantage of the complementary strengths of the two approaches.

3. Language Typology between Linguistics and NLP

The case study presented in this paper focuses on word order, a foundational topic in typology that has been extensively studied since the field's inception and continues to attract sustained scholarly interest. In the following, we provide the theoretical background from the dual perspective of linguistics and NLP.

Cross-linguistic investigations of word order have traditionally relied on categorical typological data, emphasizing correlations between different ordering patterns; see Greenberg (1963), Vennemann (1974), Lehmann (1978), Dryer (1992) to name just a few pioneer studies on the topic. Initiated by Greenberg, this approach led to the formulation of implicational universals, i.e., statistically dominant tendencies observed across languages. Typically, such studies adopt a *type-based* approach, treating each language as a discrete data point and assigning a single dominant word order type: e.g., Subject-Verb-Object or SVO, SOV, VSO, ecc. This methodology underpins influential contemporary typological resources, such as *The World Atlas of Language Structures* or WALS[8] (Dryer and Haspelmath 2013) and *The Syntactic Structures of the World's Languages* or SSWL[9] (Koopman 2012-), which rely heavily on categorical classifications to enable large-scale cross-linguistic comparisons.

However, type-based approaches fail to adequately represent languages exhibiting variable word order, which is ubiquitous in language. As Levshina (2019) observes, treating a language as having a single dominant order tends to underestimate intra-linguistic variation, since many languages employ multiple word orders depending on discourse context, information structure, register, or different performance pressures. Intra-linguistic word order variation thus becomes typologically relevant. To address these limitations, recent linguistic research increasingly adopts a *token-based* approach, in which generalizations and classifications are derived from the distribution of tokens in actual language use. From this perspective, Haspelmath (2019) emphasizes that "what we compare across languages is not the grammars (which are incommensurable) but the languages at the level at which we encounter them, namely in the way speakers use them".

By incorporating the observed frequencies of words and constructions, token-based typology allows researchers to define classification criteria that are difficult or impossible to identify using traditional categorical methods. It also enables the use of aggregate variables, such as entropy, complexity, and average dependency length.

Unlike type-based approaches, token-based methods yield continuous variables, allowing scholars to quantify intra-linguistic variation and avoid biases toward canonical patterns. Levshina (2019, 2022) advocates for corpus-based approaches, showing

8 https://wals.info/
9 SSWL is now integrated into the TerraLing database, https://www.terraling.com/groups/7.

that intra-linguistic variation is typologically significant and emerges from competing influences in language processing, including probabilistic constraints and performance pressures. More recently, Levshina et al. (2023) describe this as a gradient approach, in which word order patterns are modeled as continuous rather than discrete variables.

The shift from type-based to token-based typology has been facilitated by the growing availability of multilingual corpora, particularly those annotated according to the Universal Dependencies (UD)[10] initiative (Nivre et al. 2016; de Marneffe et al. 2021). UD, developed since 2014, ensures consistency in annotation between languages, enabling meaningful cross-linguistic comparisons. Recent corpus-based typological studies on word order using UD include, among others, Futrell, Mahowald, and Gibson (2015), Nivre (2016), Croft et al. (2017), Gerdes, Kahane, and Chen (2019, 2021), Guzmán Naranjo and Becker (2018).

In recent years, linguistic typology has emerged as a growing area of interest within NLP, as evidenced by the increasing number of publications on the topic and the establishment of a dedicated Special Interest Group (SIGTYP) within the international Association for Computational Linguistics[11]. Large-scale typological resources are expected to offer valuable guidance for multilingual NLP, particularly for low-resource languages. For a comprehensive overview focusing on the empirical usefulness of typological information for NLP, see Ponti et al. (2019).

A considerable amount of work in this area focuses on the prediction of typological features, typically carried out within the framework of typological databases such as WALS. This task has been argued to be beneficial for several reasons, ranging from addressing data sparsity and incompleteness characterizing these databases - especially for under-documented languages - to improving the performance of multilingual NLP models, for example by guiding cross-lingual transfer (Lent et al. 2024).

However, an open question remains as to whether the prediction of typological features truly meets the needs of both NLP researchers and typologists. As noted by Bjerva (2024), the answer appears to be partly negative for both communities. On the NLP side, only limited benefits have been observed across standard tasks and languages with annotated features (Naseem, Barzilay, and Globerson 2012; Täckström, McDonald, and Nivre 2013; de Lhoneux et al. 2018). Moreover, previous work has shown that typological information can often be learned implicitly as a by-product of model training (Bjerva and Augenstein 2021). On the other hand, from the linguistic perspective the automatic inference of typological features for undocumented languages in WALS provides expected results, due to its relying on well-known correlations, either between features or between typologically similar languages. A recent survey conducted among experts in linguistic typology confirmed that this approach is not aligned with the current needs of the field, which – as seen above - increasingly favors a gradient, token-based typology.

To improve alignment between NLP and linguistic typology, Bjerva (2024) proposes different directions for future work, ranging from making predictions explainable to grounding them in real rather than structured categorical data. As he argues, "with improved alignment, a deeper and more comprehensive understanding of both the structure and function of language can be achieved, fostering novel scientific insights". This ongoing debate reflects a long-standing issue in NLP - often described as a pendu-

10 https://universaldependencies.org/
11 https://sigtyp.github.io/

lum swinging between linguistic and engineering perspectives (Church and Liberman 2021).

4. Word Order Variation in VERB-SUBJECT Constructions. A Case Study

The case study illustrated in this section is aimed at showing how the paradigm shift outlined above advances typological insight beyond the constraints of categorical type-based models, with particular attention to the opportunities offered by large annotated corpora such as those provided by UD as well as NLP-based methods. In particular, it exemplifies how a token-based approach makes it possible to capture continuous gradience and variability both within and across languages. To illustrate the potential of token-based analyses in linguistics, I concentrate on the relative ordering of Verb and Subject (i.e. Subject-Verb or SV versus Verb-Subject or VS), a key parameter of typological classification, among many others.

The analysis focuses on four Indo-European languages representing three different genera according to the WALS classification: Bulgarian (Slavic, BUL), English (Germanic, ENG), Italian and Spanish (Romance, ITA and SPA). All four languages are classified as Subject–Verb–Object (SVO) languages (Dryer 2013b). At first glance, such a selection may appear limited from a traditional type-based typological perspective, given the close genealogical relationship among the languages and their shared basic word order. However, as discussed in Section 3, contemporary approaches to linguistic typology increasingly emphasize the analysis of token-based distributions of linguistic structures rather than the mere identification of categorical types - an orientation captured by the notion of *Distributional Typology* (Bickel 2015). Within this framework, focusing on closely related languages is not a limitation but a deliberate methodological choice: precisely because of their linguistic proximity, systematic differences are more difficult to detect and therefore particularly informative. The present case study thus offers a valuable testing ground for assessing the explanatory power of a distributional approach to typological variation.

4.1 From Type-based to Token-based analysis

Let us introduce the case study by looking at how typological databases report information about the relative ordering of Verb and Subject, in other words, how type-based typology encodes this specific feature.

Table 1
The VERB-SUBJECT construction in typological databases

Typological DB	*Order*	*BUL*	*ENG*	*ESP*	*ITA*
WALS	Subject-Verb	NDO	+	NDO	NDO
	Verb-Subject		-		
SSWL	Subject-Verb	+	+	+	+
	Verb-Subject	+	-	+	+
URIEL (AVG)	Subject-Verb	1.00	1.00	1.00	1.00
	Verb-Subject	0.33	0.00	0.67	0.67

Table 1 reports - for the four languages - the different orderings of lexical (i.e. non pronominal) subjects with respect to the verb as resulting from the WALS and SSWL databases. It can be noted that the two provide a slightly different picture for the four languages, following from the fact that whereas WALS records the "dominant" order, SSWL testifies to "productive" word order patterns. According to WALS, no dominant VERB-SUBJECT order (NDO in the table) exists for Bulgarian, Italian and Spanish (Dryer 2013a), whereas both SV and VS are considered productive word orders for all of them in SSWL. The table also reports the situation resulting from the URIEL knowledge base (Littell et al. 2017), which combines information from various sources, including WALS and SSWL, togheter with other typological, phylogenetic, and geographical databases. URIEL adopts different aggregation methods to report feature information. The score reported in the table was calculated using the average aggregation method, where each value corresponds to the mean across all available sources. Values other than 0 or 1 indicate disagreement among sources. Notably, in URIEL, BUL is characterized differently from ESP and ITA, in contrast to WALS and SSWL. Beyond factual inaccuracies, such inconsistencies - well documented in the literature - arise from the varying criteria used to define feature values (see above) as well as from missing values. As a result, the reliability of URIEL is limited, at least from the perspective of linguistics.

The partly contradictory picture just reported can be refined and unified if we take into account actual frequencies of occurrence extracted from corpora, namely, if we adopt a token-based perspective. Consider below the picture emerging from the Bulgarian, English, Italian, and Spanish "gold" treebanks. Specifically, we considered the following UD treebanks[12]: UD_Bulgarian-BTB (156.149 tokens and 11.138 sentences) (Simov et al. 2004); English Web Treebank (254.830 tokens and 16.622 sentences) (Silveira et al. 2014); Spanish UD treebank (547.680 tokens and 17.680 sentences) (Alonso and Zeman 2016); Italian Stanford Dependency Treebank (278.429 tokens and 14.167 sentences) (Bosco, Montemagni, and Simi 2013).

Table 2
Percentage distribution of SV and VS orders involving all versus lexical subjects

	All Subjects			Lexical Subjects		
Language	*SV*	*VS*	*H*	*SV*	*VS*	*H*
BUL	77.09	22.91	0.54	72.28	27.72	0.59
ENG	95.49	4.51	0.18	89.08	10.92	0.34
ESP	83.03	16.97	0.46	78.23	21.77	0.52
ITA	75.75	24.25	0.55	70.69	29.31	0.60

Table 2 reports, for the four treebanks, the percentage distribution of different orders involving all and lexical (i.e. non-pronominal) Subjects. Note that, in BUL, ESP and ITA lexical subjects correspond respectively to 78.76%, 75.03% and 77.34% of the cases (due to the shared pro-drop feature), whereas in ENG they cover only 37.84% of the explicitly expressed subjects. The four languages turned out to prefer the Subject-Verb order, but with significant differences: namely, the Bulgarian, Italian, and Spanish treebanks are characterized by a much higher percentage of left-headed subjects than English. For all languages, relatively higher ordering flexibility is reported for lexical subjects. Note that in both the WALS and SSWL databases the values reported for this feature refer

12 The UD treebanks employed in this case study correspond to version 2.3, released in November 2018.

to lexical subjects: this makes the discrepancy between the type-based and token-based pictures more evident.

Token-based typology can also use aggregate variables derived from the distributions of usage tokens, such as entropy, complexity, average dependency length, etc. Word order freedom has already been discussed from a corpus-based perspective by Liu (2010), Futrell, Mahowald, and Gibson (2015), Guzmán Naranjo and Becker (2018), Buljan (2023). They used several aggregate measures based on conditional entropy of whether a head is to the right or left of a dependent. In their study, each language was represented by only one aggregate score of each type. In contrast to this "black box" approach, we follow Levshina (2019), who computes Shannon's entropy (Shannon 1948) for syntactic dependencies individually. Table 2, under the H columns, reports the entropy values observed for the VERB-SUBJECT construction. English shows lower values with respect to the other languages, but still high values if compared with percentage distributions: with entropy, a small amount of variation is sufficient to obtain relatively high values.

Both word order measures reported above are purely descriptive and data-driven, i.e. based on attested word order patterns. Their most important advantage is that they are objectively defined, unlike such notions as word order "freedom" or "flexibility", which usually reflect intuitions that may be of substantial interest but which need to be weighted to be used for investigating word order variation intra- and cross-linguistically based on actual usage (e.g. as testified by treebanks).

4.2 NLP-based Typological Analysis

We have seen that linguistic resources, in particular gold treebanks annotated with a shared scheme, make it possible to reconstruct and quantify word order patterns, defined in terms of both linear order and degree of flexibility. The open issue that needs to be investigated concerns the factors driving the preference for one or the other order, both intra- and cross-linguistically. NLP-based methods and techniques can help identify the factors underlying the variation reported above and define their interaction. Their potential contribution is exemplified here through the findings of the case study presented by Alzetta et al. (2018, 2019), aimed at exploring how structural and contextual factors shape the typological behavior both within and across languages by using an NLP-based methodology.

4.2.1 Data and Method

The case study relies on two main components: linguistically annotated multilingual corpora, and a modeling algorithm used to qualitatively and quantitatively assess variation in the same construction within and across languages.

For each language, a large Reference Corpus of automatically dependency parsed sentences was processed using LISCA (Dell'Orletta, Venturi, and Montemagni 2013a), an algorithm designed to build a probabilistic model (LISCA-LM), which - using a metaphor by Jakobson (1973) - can be seen as encoding the DNA of the language being analyzed. Each Reference Corpus consists of a collection of texts from the news and Wikipedia domains of around 40 million tokens, constituting a set of examples large enough to reflect the actual distribution of phenomena in the specific language. These corpora were morpho-syntactically annotated and dependency parsed using the

UDPipe annotation pipeline (Straka, Hajič, and Straková 2016; Straka and Straková 2017).[13]

To construct the computational model, LISCA collects statistics about a wide set of linguistically motivated features, both local and global, extracted from the automatically dependency-parsed reference corpus. Local features include, for instance, dependency length, the associative strength between the grammatical categories involved, the dependency type, and the relative linear order of head and dependent. Global features capture the position of each relation within the overall sentence structure, such as the dependent's distance from the root or from leaf nodes in the dependency tree, as well as the number of "sibling" and "child" nodes occurring to its left or right in the sentence.

The LISCA-LM is then used to assign a score to each dependency relation (DR) instance in a Target Corpus, represented here by the UD treebanks for the selected languages. In particular, every dependency relation instance is assigned a context-sensitive, frequency-based score. Following Tusa et al. (2016), the LISCA score - when computed against gold annotations - is interpreted as reflecting the "markedness" degree of a specific dependency instance. Haspelmath (2006) discusses 12 different senses of the widely debated "markedness" notion: among them, we refer here to "Markedness as Abnormality", and in particular to Markedness as *rarity in texts* and Markedness as *restricted distribution*. The LISCA score reflects both these micro-facets of the notion: on the one hand, the frequency of occurrence in actual language use, and on the other hand, the occurrence limited to specific contexts. High LISCA scores are assigned to relations that occur in "unmarked" (i.e. frequent, typical, likely) contexts according to the Reference corpus, whereas low scores mark more unusual or contextually restricted constructions.

The final LISCA output is a ranked list of the Target Corpus relations ordered by decreasing prototypicality. This ranking can be used to infer quantitative evidence about the gradual transition from unmarked (or prototypical) to marked (or less canonical) DR instances. For further details about the adopted methodology and techniques the interested reader is referred to Alzetta et al. (2019). The goal here is to show how NLP-enabled results, such as those returned by LISCA, can contribute to typological linguistic studies.

4.2.2 Verb-Subject Dependencies in the LISCA ranking

Alzetta et al. (2019) move beyond frequency counts of gold treebanks and integrate into the analysis the local and global features associated with individual relation instances. This multi-factorial approach allows them to detect subtle differences in subject positioning both within and across languages. The comparison of the ranked lists obtained for the different languages can shed light on typological similarities and differences. To carry out this comparative analysis, each ranked list of dependency relations in the Target Corpus has been divided into 10 groups of equal size, henceforth "bins": the first bins contain relations presenting a high LISCA score and, conversely, the last bins contain relations characterized by low LISCA scores.

In what follows, we examine the distribution of Verb-Subject dependency relations in the LISCA ranking of each Target treebank. Specifically, we concentrate on

13 To have an idea of the quality of the automatic annotation carried out with UDPipe trained on the UD 2.3 version of the used treebanks, the interested reader is referred to `https://github.com/jwijffels/udpipe.models.ud.2.3/blob/master/inst/udpipe-ud-2.3-181115/README#L317-L319`.

dependencies involving nominal subjects, labeled as `nsubj` in the Universal Dependencies (UD) relation inventory[14].

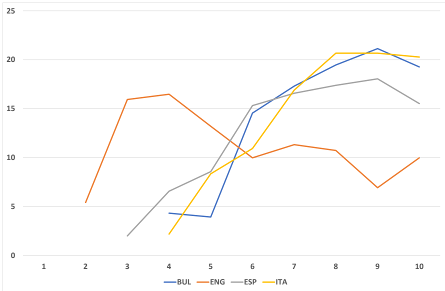

Figure 1
Distribution of `nsubj` relations across the LISCA bins

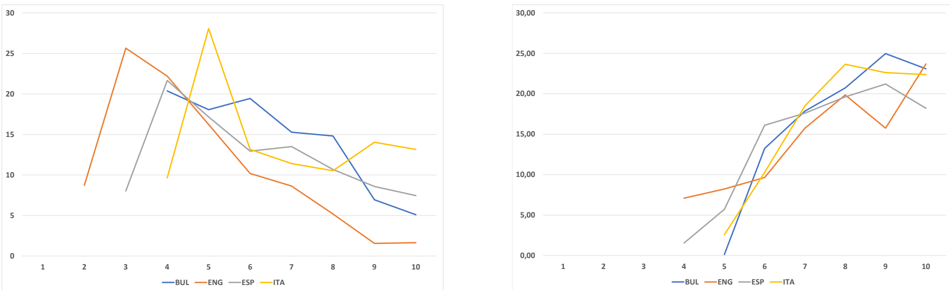

Figure 2
Distribution of pronominal (left) vs lexical (right) subjects across the LISCA bins

Figure 1 displays, for each language, the distribution of all subject types (i.e. pronominal, proper nouns, and common nouns) across the LISCA bins. It can be noted that Bulgarian, Italian, and Spanish share a similar distribution: `nsubj` instances first appear between the 3rd (Spanish) and 4th (Bulgarian and Italian) bins, with most of them concentrated in the second half of the bins. For English, `nsubj` relations appear earlier, already in the 2nd bin, and are characterized by a descending trend in the second half of the distribution.

Yet, if we compare the distribution of pronominal vs lexical subjects, the four languages show a similar distribution, as illustrated in Figure 2 where the left graph reports the distribution of pronominal subjects, and the right one that of lexical subjects. Pronominal subjects show a similar distribution in all languages: they appear earlier (between the 2nd and the 4th bin) with respect to lexical ones, mainly concentrate in the first half of the bins and show a descending distribution toward the tail of the ranked list. Despite this, they exhibit quite different frequencies, as already pointed

14 `https://universaldependencies.org/u/dep/nsubj.html`

out in Section 4.1: as pro-drop languages, Bulgarian, Italian, and Spanish permit the omission of subjects, whereas English obligatorily requires an explicit subject. Lexical subjects, on the other hand, appear between the 4th (English and Spanish) and the 5th (Bulgarian and Italian) bins, and are all characterized by an ascending trend across the bins, with only minor cross-linguistic variation.

Beyond language-specific differences, the POS-based distribution of subjects is similar across the languages. Interestingly, the fact that pronominal subjects turned out to be more prototypical than lexical subjects (including both proper and common nouns) aligns with the "referential hierarchy" first proposed by Silverstein (1976), providing a framework for understanding cross-linguistic patterns. In this hierarchy, prototypical subjects are represented by pronouns (which denote a referent that is already present in the discourse) and are easier to process than those low in the hierarchy; they are followed by proper nouns and then by common nouns. This is a topic worth further investigating, which was mentioned here just to show that the NLP-enabled LISCA ranking of subjects can provide useful evidence for a cross-linguistic study of the referential hierarchy.

Consider now the distribution of pre- vs post-verbal lexical subjects across the bins for the four languages, shown in Figure 3. For ENG, post-verbal subjects mainly occur in the tail of the ranking, for BUL, ITA and SPA almost all bins combine both of them. In English, post-verbal subjects appear in the 7th bin and increase progressively, going from 2.70% of subjects in the 7th bin to 16.22% in the last bin. In the other languages, post-verbal subjects appear already in the first bins[15] and do not show an increasing distribution as in the ENG case. In each bin, their percentage ranges from 23.24% to 33.96% for BUL, from 13.11% to 25.78% for ESP, and from 17.50% to 38.89% for ITA.

In English, the VS order turned out to be highly marked, i.e. quite rare and typically restricted to parenthetical or journalistic inversion contexts (e.g., *"...." said Trump*), where pragmatic emphasis licenses the departure from canonical order. In contrast, Bulgarian, Italian and Spanish exhibit both SV and VS orders across the full LISCA spectrum. This pattern suggests a moderate and gradual markedness for the VS constructions which are permissible in a wider range of syntactic environments, including unaccusative verbs, passive or impersonal constructions, topicalized or heavy constituents that delay the subject. Although in all languages left-headed subjects are marked, they differ in the degree of markedness. This can be seen as useful evidence towards a gradual rather than boolean markedness notion, which can be usefully exploited for a typological description of languages.

4.2.3 Behind VERB-SUBJECT Order Patterns

In the previous section, the intra-linguistic and cross-linguistic variability of `nsubj` dependencies has been illustrated with a specific view to how it relates to different degrees of prototypicality, or - from a complementary perspective - of markedness. Traditional word order typology has generally concentrated on those dependencies that exhibit relatively low intra-linguistic variability but high cross-linguistic variation: the VERB-SUBJECT construction is one of those, as it is often taken as a diagnostic pattern for identifying canonical word order types.

15 Note that the 5th bin for both BUL and ITA contains a negligible number of subjects, all occurring in pre-verbal position: they cover 0.12% and 2.57% of all cases, against the ENG pre-verbal subject only bins which cover 24.96% of the cases.

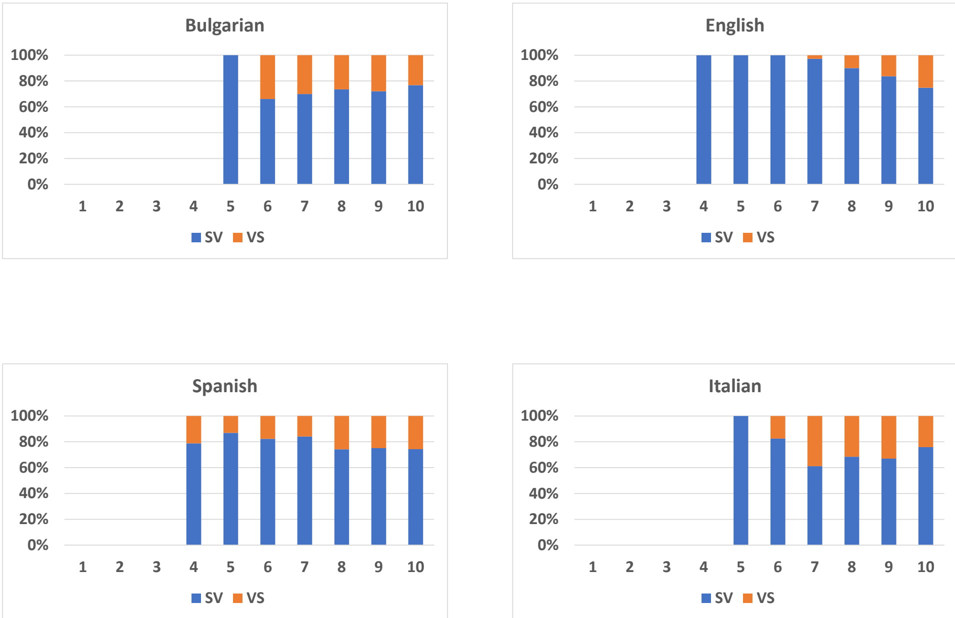

Figure 3
Distribution of pre- vs post-verbal lexical subjects across the bins for the four languages

By focusing on the dominant orders only, the risk is overlooking other ordering phenomena that, although less salient in terms of cross-linguistic differentiation, nonetheless provide crucial insights into language universals. A well-documented example is the preference for pronominal subjects to precede the verb in most languages, which can be explained in terms of the interaction between discourse organization and cognitive processing demands (Levshina 2019). This observation connects with the broader framework of processing typology, which posits that recurrent structural preferences across languages are shaped, at least in part, by general cognitive pressures (Diessel 2017). In this perspective, word order patterns are not fixed categorical templates but fluid, probabilistic outcomes of competing forces that emerge from the interplay between memory limitations, communicative needs, and processing efficiency.

By adopting NLP-enabled methods, it becomes possible to test these hypotheses on a large scale and to provide empirical evidence for the parallels between processing preferences and typological distributions. In what follows, this connection is explored more closely by focusing specifically on the relationship between the length[16] of `nsubj` relations and the dependency direction, to investigate whether and how processing constraints may condition cross-linguistic variation. Table 3 reports, for the Target treebanks, the average length of lexical subjects, as well as of pre- and post-verbal ones.

16 Dependency length is the linear distance between two syntactically related words, i.e. a head and its dependent. The length, defined as the number of words intervening between the two, is typically seen as reflecting sentence complexity and cognitive load (Hawkins 2014; Futrell, Levy, and Gibson 2020).

Table 3
Lexical subject length

	BUL	ENG	ESP	ITA
Avg `nsubj` length	3.43	3.44	4.62	4.68
Pre-verbal	4.03	3.51	5.24	5.52
Post-verbal	1.98	2.84	2.38	2.65

Besides global cross-linguistic differences in average `nsubj` length, a closer look at the data reveals interesting asymmetries between pre- and post-verbal subjects. In Bulgarian, Spanish, and Italian, the average dependency length of post-verbal subjects is almost half with respect to pre-verbal ones, whereas English exhibits a less pronounced difference, which appears to be consistent with its more rigid syntactic configuration and the limited occurrence of subject–verb inversion. In other words, for BUL, ITA and SPA, significantly longer dependencies are associated with the dominant SV order, while this tendency is less prominent in English. This observed pattern suggests that languages with a higher degree of word order flexibility tend to minimize dependency length when a marked order is chosen. Specifically, the use of the marked VS order is accompanied by shorter dependencies, which reduce processing costs in line with Dependency Length Minimization (DLM) principles (Hawkins 2014).

These findings not only highlight dependency length as a significant structural factor, but also strengthen the connection between typological distributions and processing constraints. In line with the processing typology perspective, they suggest that the availability and frequency of particular word order options are conditioned by the cognitive pressures associated with memory and comprehension, and that cross-linguistic variation in VERB-SUBJECT ordering can be interpreted as a probabilistic adaptation to these pressures.

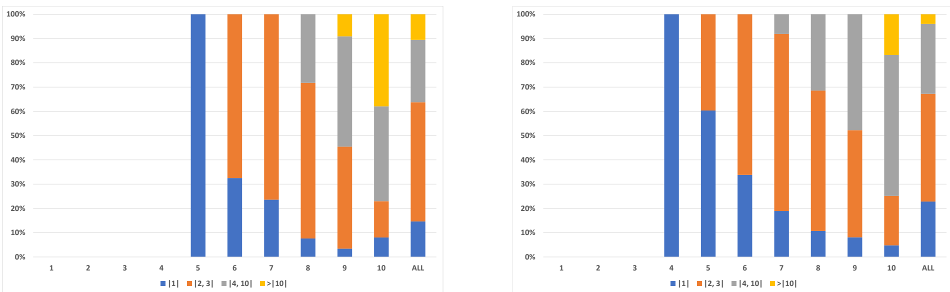

Figure 4
`nsubj` length across the LISCA bins: Italian (left), English (right)

Consider the distribution of lexical subjects by dependency length across the LISCA bins, as reported in Figure 4. Here, we focus on Italian and English only. In Italian, `nsubj` relations with a length greater than ten tokens appear in the ninth and tenth bins, covering 9% and 38% of instances in those bins, respectively. By contrast, in English such long subject relations are confined to the final bin, where they account for 17% of the cases. Regarding short subjects (length = 1), the two languages display a comparable distribution only in the earliest relevant bin (the fifth for Italian and the fourth for English). In English, the proportion of short subject dependencies decreases gradually and consistently from the fourth to the final bin, while in Italian the decline is sharper, with short subjects falling below 10% of the cases from the eighth bin onwards. Overall, the main cross-linguistic differences concern the distribution of the shortest (length = 1) and longest (length $>$ 10) dependencies: the former account for 15% of the cases in Italian and 23% in English, while the latter cover 11% and 4%, respectively.

At this stage, an open question is how VS ordering and dependency length interact and how they jointly influence the position of a given `nsubj` instance in the LISCA ranking; in other words, its prototypicality or - conversely - markedness. In particular, it is still unclear whether post-verbal subjects appearing in the final bins are always the longest ones. Figure 5, which shows the distribution of subjects by order and length, points to a more nuanced picture: short subjects can also be found in the final bins, while long subjects sometimes occur earlier in the ranking. This suggests that multiple factors contribute to determining the degree of markedness of `nsubj` relations. Moreover, the interaction between length and position is not uniform between the two languages. In English, left-headed subjects are almost entirely concentrated in the last three bins (with only 2.7% of cases in the preceding bin), and dependency length appears to be the dominant factor: in the final bin, 65.87% of cases are preverbal subjects with length $\geq$ 4. By contrast, in Italian, left-headed subjects already appear in the 6th bin and consistently account from the 17.50% to 38.89% of the cases. The highest percentage is observed in the 7th bin, while in the last bin they represent only 24.14% of `nsubj` instances. With respect to dependency length, in the 10th bin 77.10% of the cases involve subjects of length $\geq$ 4, 67.82% of which are preverbal.

Overall, both English and Italian show that dependency length and direction jointly affect the degree of markedness, though to different extents. In both languages, the internal composition of the last bin - assumed to represent the most marked `nsubj` instances - shows a similar predominance of long preverbal subjects. The key difference lies in the distribution of postverbal subjects: in Italian they occur throughout nearly all bins except the first, while in English they are concentrated at the end of the ranking.

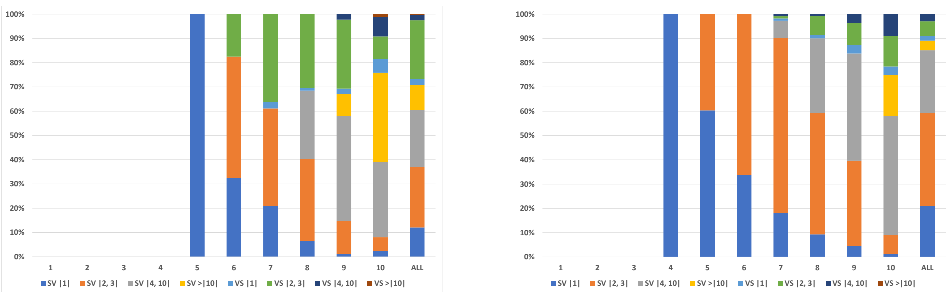

Figure 5
`nsubj` length and direction across the LISCA bins: Italian (left), English (right)

4.3 Discussion

This case study has shown how corpus-based typology provides important and useful evidence for the classification of languages and the identification of shared cross-linguistic trends. By revisiting the study by Alzetta et al. (2018, 2019), the added value of annotated corpora for typological investigations has been illustrated. Beyond the advantages of operating on treebanks through descriptive metrics such as frequency and entropy, the case study also demonstrates the advantages of NLP-enabled methods for uncovering finer-grained typological patterns.

In particular, the case study focused on word order variation in the VERB–SUBJECT construction in four different languages, all characterized by the SV order, which corresponds to the prototypical unmarked configuration. The main emerging cross-linguistic difference lies in the degree of word order flexibility: Romance languages (Italian and Spanish) and Bulgarian display a much higher proportion of post-verbal subjects compared to English. This general picture has been further refined by analyzing the LISCA ranking of relations for each Target treebank, which - since based on gold annotations - captures the relative prototypicality of the syntactic configuration represented by the specific dependency instance. This analysis highlighted both language-specific and shared tendencies with respect to:

- **Markedness** - In English, post-verbal subjects are strongly marked: as such, they are concentrated at the lower end of the LISCA ranking. In contrast, in Bulgarian, Italian, and Spanish, both pre- and post-verbal subjects occur across the whole ranking spectrum, albeit with different distributions;
- **Distributional constraints** - Post-verbal subjects occur in restricted contexts in all languages: in English, almost exclusively in parenthetical clauses; in Bulgarian, Italian and Spanish, in a wider range of constructions;
- **Grammatical category of the subject** - In all languages, pronominal subjects cluster in the top LISCA bins. In Bulgarian, Italian, and Spanish, this tendency holds irrespective of whether the pronoun precedes or follows the verb;
- **Dependency length** - Longer subject–verb dependencies consistently cluster at the lower end of the LISCA ranking across all languages, with English displaying overall shorter links than Bulgarian, Italian, or Spanish. In all languages, the dominant SV order tends to involve longer dependencies, whereas the marked VS order reduces dependency length. However, the extent of this reduction varies cross-linguistically: in English, the average dependency length of post-verbal subjects is only slightly shorter than that of pre-verbal ones, while in the other three languages the reduction is much more pronounced. Specifically, the ratio of post-verbal to pre-verbal subject length is about 0.5 in Bulgarian, Italian, and Spanish, compared to 0.8 in English.

This fine-grained analysis of word order variation, exemplified here through the VERB–SUBJECT construction in four different languages, has potentially far-reaching implications for linguistic typology. It lends support to a gradient, usage-based view of word order typology, in which markedness is contextually shaped rather than categorically fixed. At the same time, it underscores the need for multifactorial models that integrate frequency, structural constraints, and processing considerations.

The observed correlation between word order and dependency length can be interpreted in light of the proposal by Hahn, Jurafsky, and Futrell (2020), who suggest that the word order properties of languages result from a process of optimization for efficient communication. According to this view, linguistic systems balance two competing pressures: the need to reduce complexity and the need to minimize ambiguity, that is they must be "simple enough to allow the speaker to easily produce sentences, but complex enough to be unambiguous to the hearer".

From an NLP perspective, the findings are equally relevant. As noted by Ponti et al. (2019), the limitations of manually crafted typological databases can "be averted through the use of methods that allow typological information to emerge from the data in a bottom-up fashion, rather than being predetermined". The proposed methodology - once extended and tested on a broader set of languages and constructions to assess the generalizability of the results - could contribute to the development of more typologically informed models, particularly in multilingual settings where syntactic preferences affect parsing, generation, and alignment. In sum, the automatic induction of typological information and its integration into machine learning algorithms hold promise for addressing a major bottleneck in polyglot NLP.

Finally, the role of linguistic annotation deserves particular attention in light of the present case study, which relied on both automatically annotated and gold-standard corpora. The LISCA models draw on large, automatically dependency-parsed reference corpora, while the typological analysis was conducted on manually revised, or gold, target corpora. This combined use of automatic and manually verified annotation provides a valuable model for integrating breadth and precision in empirical research.

Linguists have traditionally been cautious about relying on automatically annotated data, as linguistic inquiry has long depended on manually or semi-automatically annotated corpora. Such resources, however, are typically limited in size and inevitably capture only a restricted range of linguistic features. Yet, as observed by Montemagni (2013), results obtained from automatic annotation show a good degree of consistency with studies on diamesic and textual variation, see e.g. Voghera (2004, 2005), Cresti (2005); comparable evidence has also been reported for English by Dell'Orletta, Venturi, and Montemagni (2013b).

Currently, certain types of linguistic annotation - such as part-of-speech tagging and dependency parsing - can be performed automatically with high accuracy. Although a residual error rate persists, its extent depends on the linguistic level and the nature of the features under consideration. When critically assessed, automatically annotated data can thus provide reliable evidence for linguistic analysis. Manual annotation remains, however, essential for phenomena requiring interpretative judgment or involving subtle linguistic distinctions. The combined use of both annotation types, as illustrated in this study, allows researchers to balance the extensive coverage of large-scale data with the precision required for fine-grained linguistic analysis.

5. Conclusions

This paper has examined the evolving interplay between CL, NLP, and linguistics, showing how their interaction has gradually developed into a mutually enriching relationship. Yet, the potential of computational models for linguistic inquiry remains today far from fully realized.

The case study on linguistic typology presented here clearly illustrates this point. Despite the growing interest in typology within both the computational linguistics and linguistic research communities, their approaches often remain misaligned. In NLP, typological information is typically leveraged to enrich multilingual models, yet with limited success in capturing fine-grained linguistic variation. Linguistic typology, by contrast, aims to uncover the structural and functional principles underlying cross-linguistic diversity through token-based and gradient analyses. Bridging this methodological and conceptual divide promises substantial benefits for both fields.

More broadly, the gap between CL/NLP and linguistics persists, reflecting deeper asymmetries in epistemic priorities and methodological expectations. NLP research on language data is often conducted within computational frameworks that overlook the interpretative and theoretical dimensions central to linguistic inquiry. Conversely, linguists may underestimate the potential - or misunderstand the limitations - of computational approaches. Addressing this mutual misalignment requires sustained dialogue and genuinely collaborative practices, in which computational and linguistic expertise inform each other from the earliest stages of research design.

Unlocking the full potential of computational modeling for linguistic research demands equitable forms of collaboration between CL/NLP and linguistics. Ideally, such collaboration should advance both fields simultaneously. In some cases, however, innovation may occur primarily in one domain, with the other contributing as a methodological resource or as a provider of data and interpretive insight. Regardless of the direction of advancement, successful collaboration depends on a balanced distribution of roles: computational linguists should not be confined to the role of technical implementers, nor linguists to that of external interpreters of computational output. Rather, linguists should actively engage with NLP tools to formulate new research questions and perspectives, while CL/NLP research should design methods that are both technically robust and responsive to the needs of linguistic inquiry.

Key issues that emerged from the case study and that require continued attention in this dialogue concern (i) the choice of the computational approach, depending on the specific linguistic research question, and (ii) the type and degree of explicit data annotation, whenever required.

Selecting the most appropriate computational model for a specific research goal requires balancing multiple factors, including task complexity, resource availability, model interpretability, and desired performance. For highly specialized tasks with well-defined constraints, traditional NLP models may still offer the most suitable option, as their simpler architectures and reliance on explicit, interpretable features allow for greater experimental control and linguistic transparency, as demonstrated in the case study. In contrast, deep learning approaches based on word embeddings and LLMs have transformed computational linguistics, substantially improving performance across a wide range of tasks. However, they remain less reliable for metalinguistic analyses, which demand fine-grained, theory-driven reasoning. Nevertheless, ongoing developments suggest that future LLMs may begin to exhibit not only linguistic but also metalinguistic capabilities, potentially advancing linguistic theory in new and unexpected directions.

In addition to model selection, data-related factors present further challenges, especially regarding linguistic annotation. Linguists do not always trust the outcomes of automatic annotation processes, nor do they consistently recognize the benefits of adopting shared annotation standards. In the case study presented here, the LISCA computational model was trained on automatically annotated corpora and applied to gold-standard treebanks, both conforming to the Universal Dependencies framework, thus enabling systematic cross-lingual comparison. These differing perspectives often reflect broader methodological and epistemological divides between the linguistic and computational communities, underscoring the importance of sustained dialogue and collaboration in defining reliable, linguistically grounded annotation practices.

Addressing these challenges offers an opportunity to move toward a more unified framework for studying language, combining the empirical power of computational modeling with the theoretical depth of linguistic analysis. Future developments will likely focus on integrating explicit linguistic knowledge into large-scale models, improving interpretability, and enhancing their capacity for metalinguistic reasoning. At the same time, traditional models will continue to provide a valuable and reliable option for tasks requiring transparency, precise feature control, or resource-constrained scenarios. Together, these approaches can advance computational and theoretical linguistics in complementary ways, ensuring that models not only emulate linguistic performance but also contribute meaningfully to linguistic theory.

6. Acknowledgments

I gratefully acknowledge Chiara Alzetta, Felice Dell'Orletta and Giulia Venturi for their work on the case study, which constituted a key component of this research. Their careful data collection, analysis, and thoughtful discussions provided essential support and significantly enriched the study. I also thank the anonymous reviewers for their careful reading of the manuscript and for their insightful comments and suggestions.

References

Alonso, Héctor Martínez and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98.

Alzetta, Chiara, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2018. Universal Dependencies and quantitative typological trends. A case study on word order. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Alzetta, Chiara, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2019. Inferring quantitative typological trends from multilingual treebanks. A case study. *Lingue e Linguaggio*, XVIII(2):209–242.

Beguš, Gašper, Maksymilian Dabkowski, and Ryan Rhodes. 2025. Large linguistic models: Investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, 6(12):3453–3467.

Bender, Emily M. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In Timothy Baldwin and Valia Kordoni, editors, *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, March. Association for Computational Linguistics.

Bender, Emily M. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July. Association for Computational Linguistics.

Bender, Emily M. and Alex Lascarides. 2019. *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Bickel, Balthasar. 2015. Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine and Heiko Narrog, editors, *The Oxford handbook of linguistic analysis, 2nd edition*. Oxford University Press, pages 901–923.
Bjerva, Johannes. 2024. The role of typological feature prediction in NLP and linguistics. *Computational Linguistics*, 50(2):781–794, June.
Bjerva, Johannes and Isabelle Augenstein. 2021. Does typological blinding impede cross-lingual sharing? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online, April. Association for Computational Linguistics.
Boguraev, Bran and Ted Briscoe, editors. 1989. *Computational lexicography for Natural Language Processing*. Longman Publishing Group, USA.
Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August.
Buljan, Maja. 2023. What quantifying word order freedom can tell us about dependency corpora. In Owen Rambow and François Lareau, editors, *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, pages 54–67, Washington, D.C., USA, March. Association for Computational Linguistics.
Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The false promise of ChatGPT. *New York Times*.
Church, Kenneth W. and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4.
Cresti, Emanuela. 2005. La testualità parlata: alcuni dati dal corpus italiano di C-ORAL-ROM nella prospettiva del parlato romanzo. In Iørn Korzen, editor, *Atti del VIII Convegno internazionale SILFI*, pages 163–176, Frederiksberg-Copenhagen, Denmark, June 2004. Copenaghen Studies.
Crocker, Matthew W. 1996. *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Kluwer Academic Publishers, Dordrecht.
Croft, William. 2003. *Typology and Universals*. Cambridge University Press, Cambridge, second edition.
Croft, William, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In Markus Dickinson, Jan Hajic, Sandra Kübler, and Adam Przepiórkowski, editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), Bloomington, IN, USA, January 20-21, 2017*, volume 1779 of *CEUR Workshop Proceedings*, pages 63–75. CEUR-WS.org.
de Lhoneux, Miryam, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997, Brussels, Belgium, October-November. Association for Computational Linguistics.
de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June.
Dell'Orletta, Felice, Giulia Venturi, and Simonetta Montemagni. 2013a. Linguistically-driven selection of correct arcs for dependency parsing. *Computación y Sistemas*, 17(2):125–136.
Dell'Orletta, Felice, Giulia Venturi, and Simonetta Montemagni. 2013b. Unsupervised linguistically-driven reliable dependency parses detection and self-training for adaptation to the biomedical domain. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, John Pestian, and Jun'ichi Tsujii, editors, *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 45–53, Sofia, Bulgaria, August. Association for Computational Linguistics.
Diessel, Holger. 2017. Usage-based linguistics. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, New York.
Dijkstra, Ton and Koenraad de Smedt, editors. 1996. *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing*. Taylor & Francis, London, UK.
Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language*, 68:81–138.
Dryer, Matthew S. 2013a. Order of subject and verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.
Dryer, Matthew S. 2013b. Order of subject, object and verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Dryer, Matthew S. and Martin Haspelmath, editors. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
Fellbaum, Christiane, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.
Futrell, Richard, Roger Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.
Futrell, Richard and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, Advance online publication:1–98.
Futrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In Joakim Nivre and Eva Hajičová, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden, August. Uppsala University, Uppsala, Sweden.
Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA.
Gerdes, Kim, Sylvain Kahane, and Xinying Chen. 2019. Rediscovering Greenberg's word order universals in UD. In Alexandre Rademaker and Francis Tyers, editors, *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 124–131, Paris, France, August. Association for Computational Linguistics.
Gerdes, Kim, Sylvain Kahane, and Xinying Chen. 2021. Typometrics from implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, 6(1):17.
Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*. The MIT Press, Cambridge, Mass, chapter 5, pages 58–90.
Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, USA, June. Association for Computational Linguistics.
Guzmán Naranjo, Matías and Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104, Oslo, Norway, December. Linköping University Electronic Press.
Hahn, Michael, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.
Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of linguistics*, 42(1):25–70.
Haspelmath, Martin. 2019. How comparative concepts and descriptive linguistic categories are different. In Daniël Olmen, Tanja Mortelmans, and Frank Brisard, editors, *Aspects of Linguistic Variation*. De Gruyter Mouton, Berlin, Boston, pages 83–114.
Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford University Press.
Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4129–4138, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
Hinrichs, Erhard, Marie Hinrichs, Sandra Kübler, and Thorsten Trippel. 2019. Language technology for Digital Humanities. *Language Resources and Evaluation - Special Issue*, 53(4).
Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
Jakobson, Roman. 1973. *Essais de linguistique générale t. 2: rapports internes et externes du langage*. Les éditions de Minuit.
Jenset, Gard B. and Barbara McGillivray. 2017. *Quantitative historical linguistics: a corpus framework*. Oxford University Press, Oxford, UK.
Kaplan, Ronald and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical*

Relations. MIT Press, Cambridge, MA, pages 173–281.
Kay, Martin. 1984. Functional unification grammar: A formalism for machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 75–78, Stanford, California, USA, July. Association for Computational Linguistics.
Kay, Martin. 2005. ACL lifetime achievement award: A life of language. *Computational Linguistics*, 31(4):425–438.
Koopman, Hilda, editor. 2012-. *SSWL, The Syntactic and Semantic Structures of the World's Languages Database*.
Kortmann, Bernd. 2021. Reflecting on the quantitative turn in linguistics. *Linguistics*, 59(5):1207–1226.
Kučera, Henry. 1982. Markedness and frequency: A computational analysis. In Ján Horecký), editor, *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, pages 167–173, Prague, Czech Republic, July. North-Holland Publishing Company.
Lehmann, Winfred P. 1978. The great underlying ground-plans. In Winfred P. Lehmann, editor, *Syntactic typology: Studies in the phenomenology of language*. University of Texas Press, Austin, pages 58–90.
Lent, Heather, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.
Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.
Levshina, Natalia. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.
Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.
Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.
Littell, Patrick, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April. Association for Computational Linguistics.
Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.
Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press, Edinburgh, second edition.
Montemagni, Simonetta. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, Anno XLII(1):145–172.
Naseem, Tahira, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July. Association for Computational Linguistics.
Nastase, Vivi, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 631–643, Pisa, Italy, December. CEUR Workshop Proceedings.

Nerbonne, John and Sara Tonelli. 2016. Digital Humanities and Computational Linguistics. *Italian Journal of Computational Linguistics - Special Issue*, 2(2).
Nivre, Joakim. 2016. Universal Dependencies: A cross-linguistic perspective on grammar and lexicon. In Eva Hajičová and Igor Boguslavsky, editors, *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 38–40, Osaka, Japan, December. The COLING 2016 Organizing Committee.
Nivre, Joakim, 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, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).
Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, USA.
Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.
Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).
Sampson, Geoffrey. 2001. *Empirical Linguistics*. Continuum, London & New York.
Shannon, Claude Elwood. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.
Silveira, Natalia, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Robert M.W. Dixon, editor, *Grammatical categories in Australian languages*. Australian Institute of Aboriginal Studies, Canberra, pages 112–171.
Simov, Kiril, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2004. Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation. Special Issue*, 2:495–522.
Steedman, Mark. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA.
Straka, Milan, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).
Straka, Milan and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In Jan Hajič and Dan Zeman, editors, *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.
Täckström, Oscar, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA, November. Association for Computational Linguistics.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.

Tusa, Erica, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessitá. In Anna Corazza, Simonetta Montemagni, and Giovanni Semeraro, editors, *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 3–16, Napoli, Italy, December. Accademia University Press.

Vennemann, Theo. 1974. Theoretical word order studies: Results and problems. *Papiere zur Linguistik*, 7:5–25.

Voghera, Miriam. 2004. La distribuzione delle parti del discorso nel parlato e nello scritto. In Rika Van Deyck, Rosanna Sornicola, and Johannes Kabatèk, editors, *La variabilité en langue, I. Langue parlée et langue écrite dans le présent et dans le passé, II. Les quatre variations*. Gand, pages 261–284.

Voghera, Miriam. 2005. La misura delle categorie sintattiche. In Isabella Chiari and Tullio De Mauro, editors, *Parole e numeri. Analisi quantitative dei fatti di lingua*. Aracne, Roma, pages 125–138.

Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince, editors. 1998. *Centering Theory In Discourse*. Clarendon Press, Oxford.

Wilks, Yorick A., Brian M. Slator, and Louise Guthrie, editors. 1996. *Electric Words: Dictionaries, Computers, and Meanings*. The MIT Press, Cambridge, MA.

Xu, Tianyang, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. To appear. Can language models learn typologically implausible languages? *Transactions of the Association for Computational Linguistics*.