

Cross-Lingual Transfer Efficiency in Low-Resource Morphosyntactic Parsing: A Contrastive Annotation-Theoretic Framework for Dependency Treebank Calibration

Pat Evans

Professor

Utrecht University

Heidelberglaan 8, 3584 CS Utrecht, Netherlands

Jesse Smith

Associate Professor

Seoul National University

1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

Jamie Smith

PhD

McGill University

845 Sherbrooke Street West, Montreal, Quebec H3A 0G4, Canada

Abstract. The calibration of dependency treebanks for low-resource languages remains a critical bottleneck in cross-lingual transfer learning pipelines, particularly when morphosyntactic divergence between source and target languages is pronounced. This study proposes a contrastive annotation-theoretic framework that systematically reconciles Universal Dependencies (UD) schema inconsistencies across typologically distant language pairs. Drawing on a multi-corpus evaluation spanning six agglutinative and fusional languages, we operationalize a weighted inter-annotator agreement metric sensitive to head-attachment ambiguity and functional category misalignment. Results demonstrate a statistically significant improvement in labeled attachment score (LAS) of 6.3–9.1 percentage points over baseline transfer models. The framework further offers reproducible guidelines for schema harmonization, establishing a scalable methodological benchmark applicable to under-documented language documentation initiatives and second language acquisition corpus research.

Keywords: cross-lingual transfer learning, dependency treebank calibration, Universal Dependencies schema, morphosyntactic parsing, low-resource languages, contrastive annotation theory, labeled attachment score, agglutinative language typology, inter-annotator agreement metrics

Introduction

The accelerating proliferation of large-scale multilingual natural language processing (NLP) architectures has cast renewed theoretical scrutiny upon a foundational, yet persistently underexplored, problem in applied corpus linguistics: the structural incompatibility of syntactic annotation schemas across typologically heterogeneous

languages. As of 2024–2026, over 7,500 languages remain severely under-resourced with respect to annotated syntactic corpora, constraining the downstream applicability of transformer-based parsing models that rely upon transfer learning from high-resource pivot languages such as English, Mandarin Chinese, or Modern Standard Arabic. Within this landscape, the Universal Dependencies (UD) initiative, while representing a landmark achievement in cross-lingual annotation standardization, has been increasingly critiqued for its latent anglocentric bias in the resolution of head-directionality conflicts, null-subject constructions, and polysynthetic morphological boundaries — deficiencies that systematically inflate cross-lingual parsing error rates and obscure the empirical validity of comparative syntactic studies. The problem is not merely computational but fundamentally epistemological: when annotation guidelines instantiated for high-resource, analytic languages are transplanted onto agglutinative or polysynthetic morphosyntactic systems without rigorous theoretical reconciliation, the resulting treebanks encode annotation artifacts rather than authentic grammatical structures. These artifacts propagate through training pipelines, producing parsing models whose apparent accuracy masks systematic misrepresentation of morphological argument structure, cliticization phenomena, and non-projective long-distance dependencies. For applied linguists engaged in second language acquisition (SLA) corpus research, learner language analysis, or cross-linguistic influence modeling, such methodological contamination directly undermines the interpretive integrity of empirical findings. Prior approaches to this challenge have addressed schema divergence through heuristic post-hoc conversion rules or language-specific treebank extensions; however, no unified, theoretically grounded framework has been proposed that simultaneously accounts for inter-annotator reliability calibration, typological distance weighting, and functional category misalignment correction within a single reproducible pipeline. The present study addresses this gap by introducing a Contrastive Annotation-Theoretic (CAT) framework designed to optimize treebank harmonization protocols across source–target language pairs exhibiting maximal morphosyntactic divergence. The article is structured as follows. Section 2 reviews the theoretical foundations of dependency grammar formalisms and existing UD schema critique literature. Section 3 details the multi-corpus data selection methodology and the operationalization of the weighted inter-annotator agreement coefficient. Section 4 presents the CAT framework architecture and its algorithmic components. Section 5 reports empirical evaluation results across six target languages, with comparative LAS benchmarking against established transfer baselines. Section 6 discusses implications for low-resource language documentation, SLA corpus design, and future directions in annotation-theoretic research.

This is a preliminary version. To read the full version of the article, please purchase a subscription.

References

1. Гурко, О. В. (2023). Іntenція персуазивності в мовленні В. Зеленського. Англістика та американістика: зб. наук. пр./Дніпр. нац. ун-т ім. Олеся Гончара. Вип, 20, 13-17.
2. Gojayeveva, L. MODERN METHODS OF TEACHING LITERATURE.