A Distributional Diagnostic of Categorical Stability in a Romanian Treebank

Ioana-Madalina Silai, Sylvain Kahane
Modyco, Université Paris Nanterre, CNRS
{imsilai, skahane}@parisnanterre.fr

Relevant UniDive working groups: WG 1, WG 2

1 Introduction

The standardisation of part-of-speech (POS) tags within the Universal Dependencies (UD) framework is essential for cross-linguistic NLP (de Marneffe et al., 2021). However, treating these categories as discrete silos often masks the functional overlaps and "fuzziness" inherent in linguistic reality. (Croft, 2016; Haspelmath, 2010). While traditional grammars and universal schemas aim for clear-cut labels, modern linguistics suggests a syntactic continuum where certain words possess varying degrees of membership across categories (Aarts, 2007).

As an outcome of the UniDive activities regarding annotation divergences, we propose a distributional approach to identify and analyse fuzzy boundaries between categories.

The strength of our methodology lies in its category-agnostic nature: rather than pre-defining which POS tags are problematic, we use distributional metrics to reveal categorical overlap. By constructing morphosyntactic lexeme profiles based on their contextual environments, we calculate the *purity* (how well a lexeme aligns with its assigned tag) and *entropy* (the diversity of its distributional neighbourhood). This provides a quantitative diagnostic tool that can be applied to lexemes in any language or category to detect annotation bias or inherent linguistic ambiguity.

For this study, we apply this diagnostic to the Romanian treebank UD_Romanian-RRT (Barbu Mititelu and Irimia, 2016). Romanian serves as a critical case study for fuzzy boundaries due to the complex interplay between its particles (PART), auxiliaries (AUX), and subordinating conjunctions (SCONJ). In Romanian, elements like the subjunctive marker "să" or the infinitive marker "a" sit at the crossroads of these categories, often leading to annotation inconsistencies or "categorical drift".

By focusing on these specific categories, we demonstrate how our method captures the functional profile of a language and highlights where universal labels may obscure language-specific syntactic realities, providing a data-driven bridge between formal typology and computational linguistics.

2 Methodology

We operationalise categorical drift by treating POS tags annotated in treebanks as reference points rather than ground truth. Our methodology follows four stages:

1. **Distributional Syntactic Profiles:** To avoid heterosemy, we represent words as *lexical units*, defined as a lemma and its POS tag. We represent each lexical unit as a high-dimensional vector based on the universal features[1] of its parent, children, and linear neighbours. We calculate the frequency of these features across all the occurrences of the lexical unit in the treebank. To avoid circularity, the features of the unit itself are excluded, and to minimise noise, we exclude lexical units with fewer than 10 occurrences.

2. **Weighted Neighbourhood Mapping:** In this vector space, we identify up to 20 neighbours with a cosine similarity higher than 0.7. This cut-off was selected as a conservative threshold for localised similarity, while the $k = 20$ cap is ideal in high-dimensional spaces (Radovanović et al., 2010). Exploratory analysis with similar thresholds yielded qualitatively stable results, confirming the robustness of these parameters. Each neighbour contributes to the lexical unit's neighbourhood profile proportionally to its similarity score. This ensures that highly similar distributional neighbours influence the results more strongly than marginal ones.

[1]Universal features defined by UD: Aspect, Animacy, Case, Clusivity, Definite, Deixis, DeixisRef, Evident, Negation, Number, Gender, Degree, ExtPos, Foreign, Mood, NounClass, NumType, Person, Polarity, Polite, Poss, PronType, Reflex, Tense, VerbForm, Voice.

3. **Frequency Normalisation and Aggregation:** To account for global frequency bias (e.g. the prevalence of nouns), raw neighbourhood proportions are normalised by the relative frequency of each category. Finally we aggregate these lexical-level statistics into category stability scores. These are visualised as a heatmap, where each row i sums to 1.0. Each cell (i, j) represents the similarity-weighted and frequency-normalised probability that members of category i are distributionally attracted to category j. Consequently, diagonal values ($i = j$) quantify internal category stability, while off-diagonal values ($i \neq j$) measure the distributional drift towards rival classes.

4. **Diagnostic Interpretation (Bridge Words):** To identify the specific lexemes driving categorical overlap, we calculate each lexical unit's purity (Pur) and entropy (H). For a given lexical unit w with an annotated category c_w, let $p_{norm}(w, c)$ represent the similarity-weighted, frequency-normalised probability of category c appearing in its neighbourhood. We formally compute purity as the probability mass assigned to its baseline annotation: $Pur(w) = p_{norm}(w, c_w)$. Purity ranges from 1 (perfect alignment) to 0 (complete drift). Because purity cannot distinguish whether a unit's drift is directed or scattered, we measure the categorical diversity of the neighbourhood using Shannon entropy over the set of all UPOS categories:

$$H(w) = -\sum_{c \in \text{UPOS}} p_{norm}(w, c) \log_2 p_{norm}(w, c) \quad (1)$$

Lexemes exhibiting both low purity and low entropy signify a concentrated, directed drift away from their annotated class towards a single rival category. We explicitly define these items as "bridge words, a concept that computationally operationalises long-standing structural frameworks of *syntactic gradience* (Aarts, 2007) and *category squishes* (Ross, 1972). Finally, we isolate the specific morphosyntactic features causing the overlap by comparing the feature vector of each bridge word to the centroid (median vector) of its target rival category.

This gives us a categorial profile for a treebank, which we can then use to identify particularly unstable categories, and see to what extent the UD annotation schema mirrors the internal structure of the language.

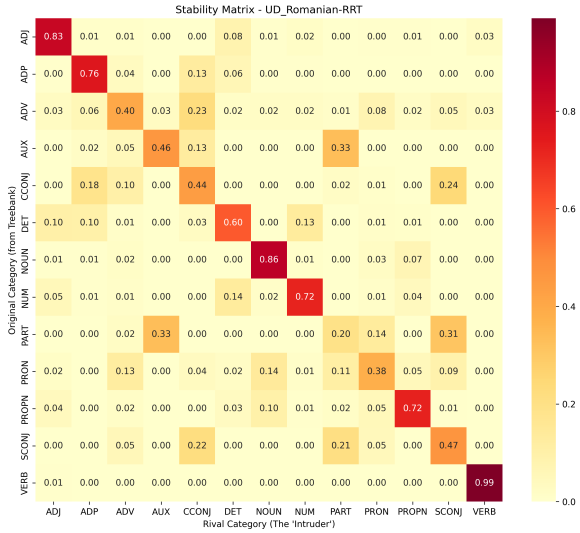

Figure 1: Stability Matrix

3 Results

The application of our methodology to the UD_Romanian-RRT treebank reveals a clear divide between stable content classes and functional categories characterised by high degrees of gradience. The Stability Matrix in Figure 1 shows that while categories such as nouns and verbs exhibit high internal cohesion (88% and 86% respectively), the diagonal values for functional categories indicate significant instability, most notably for particles (20%), pronouns (38%) or adverbs (40%).

The stability matrix highlights a significant three-way overlap between particles, auxiliaries and subordinating conjunctions. The overlap between particles and auxiliaries is notably symmetrical as the matrix indicates a 33% pull from particles towards auxiliaries, and vice-versa. This symmetry suggests a high degree of distributional syncretism, where the UD schema imposes distinct categorical labels on lexemes that inhabit nearly identical syntactic slots.

Table 1 identifies the bridge words driving this instability. The most striking finding is the pair comprised of the auxiliary *vrea* (future marker) and the particle *a* (infinitive marker). Both units yield a Purity and Entropy score of 0.00, indicating that within the contextual features captured by our model (specifically the dependency on a following infinitive), these two units are distributionally in-

Table 1: Romanian particle features and bridge words with (Purity | Entropy) scores

	AUX → PART (33%)	SCONJ → PART (21%)
Top Features	1. parent:position=after 2. next:Tense=Pres 3. next:VerbForm=Inf	1. parent:position=after 2. parent:Mood=Sub 3. parent:Mood=Ind
Top Bridge Words	*vrea* 'will' (0.00 | 0.00)	*dacă* 'if' (0.38 | 1.43) *fără* 'without' (0.29 | 1.16) *fiindcă* 'because' (0.44 | 1.38) *ca* 'so that' (0.50 | 1.17)

	PART → AUX (33%)	PART → SCONJ (31%)
Top Features	1. parent:position=after 2. parent:VerbForm=Inf 3. prev:Case=Acc	1. parent:position=after 2. parent:VerbForm=Fin 3. parent:Tense=Pres
Top Bridge Words	*a* (0.00 | 0.00)	*nu* 'not' (0.27 | 1.16) *să* (0.32 | 1.19)

distinguishable. In other words, our results suggest that the distinction between PART and AUX for these lexemes is a matter of categorical convention rather than distinct syntactic behaviour.

Similarly, the overlap between SCONJ (21% pull towards PART) and PART (31% pull towards SCONJ) reflects the functional proximity of mood markers and subordinators. The subjunctive marker *să*, often classified as a conjunction in other Romance frameworks, is identified here as a bridge word ($Pur = 0.32$, $H = 1.19$). Its distributional similarity to subordinators like *ca* ('so that') or *fără* ('without') is driven by their shared requirement for subjunctive verbal parents. Conversely, the negative particle *nu* gravitates towards conjunctions like *fiindcă* ('because') as both are distributionally licensed by indicative parents.

4 Conclusion

These results demonstrate the efficacy of our methodology as a quantitative diagnostic for categorical overlap. By identifying bridge words through distributional profiles, we can empirically pinpoint where UD labelling conventions diverge from the syntactic realities of Romanian. While this analysis primarily addresses the functional triad of PART, AUX, and SCONJ, the significant instability observed in other categories provides a clear roadmap for future research into the fuzzy boundaries of categories in the Romanian treebank.

Furthermore, while the present study focuses on a single language to account for specific typological complexities, the category-agnostic nature of the approach allows for seamless application across the UD ecosystem. Comparing these distributional stress tests across diverse treebanks will directly contribute to the broader UniDive objective of quantifying and resolving cross-linguistic annotation divergences, ultimately enhancing the consistency and interpretability of universal schemas.

References

Bas Aarts. 2007. *Syntactic gradience: the nature of grammaticaI indeterminacy*. Oxford University Press, Oxford.

Verginica Barbu Mititelu and Elena Irimia. 2016. Linguistic data retrievable from a treebank. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016)*, pages 19–27, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Gheorghe Constantinescu-Dobridor. 1974. *Morfologia limbii române*. Editura Ştiinţifică, Bucureşti.

Cristina Corla Hanţ. 2015. The infinitive built prepositionally. *Journal of Humanistic and Social Studies*, 6(2):67–73. Published by Editura Universității Aurel Vlaicu.

William Croft. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2):377–393.

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

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Iorgu Iordan and Vladimir Robu. 1978. *Limba română contemporană*. Editura Didactică şi Pedagogică, Bucureşti.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar*, volume 1 of *Theoretical Prerequisites*. Stanford University Press, Stanford, CA.

Gabriela Pană Dindelegan, editor. 2010. *Gramatica de bază a limbii române*. Editura Univers Enciclopedic Gold, Bucureşti.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531.

Brett Reynolds. 2021. Quantifying the differences between lexical categories: The case of pronouns and determinatives in english. *Cadernos de Linguística*, 2(3):e399.

John Robert Ross. 1972. The category squish: Endstation Hauptwort. In *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*, pages 316–328, Chicago, IL. Chicago Linguistic Society.

John R. Taylor. 2003. *Linguistic Categorization*, 3rd edition. Oxford Textbooks in Linguistics. Oxford University Press, Oxford.