Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Fuzzy Boundaries: A Distributional Approach to POS Overlap Across Typologically Diverse Treebanks

Ioana-Madalina Silai
Modyco, Université Paris Nanterre
imsilai@parisnanterre.fr

ABSTRACT

Part-of-speech (POS) categories are often treated as discrete silos, yet linguistic reality suggests a syntactic continuum. Universal Dependencies (UD) standardisation often obscures functional overlaps across languages. We propose a distributional approach to quantify these "fuzzy" boundaries by constructing morphosyntactic lexeme profiles. Purity and entropy help identify boundary crossing "bridge words" and the features driving these shifts. A French-Chinese contrastive analysis reveals that distributional overlap stems from distinct contexts: verbal modifiers in French versus predicative flexibility of adjectives in Chinese. This work establishes a quantitative diagnostic for annotation bias, enhancing model interpretability within the UD framework.

RÉSUMÉ

Frontières floues : une approche distributionnelle du chevauchement des POS dans des treebanks typologiquement divers

Les catégories de parties du discours (POS) sont souvent traitées comme des silos isolés, malgré un continuum syntaxique réel. Si Universal Dependencies (UD) normalise ces étiquettes, cette standardisation masque les chevauchements fonctionnels des langues. Cet article propose une approche distributionnelle quantifiant ces frontières « floues » en construisant des profils de lexèmes basés sur leurs contextes morphosyntaxiques. Grace à des mesures de pureté et d'entropie, nous identifions les « mots-ponts » et les traits favorisant ces glissements catégoriels. L'étude contrastée des treebanks français et chinois montre que ce chevauchement provient de contextes distincts : modifieurs verbaux en français versus flexibilité prédicative des adjectifs en chinois. Ce travail fournit un diagnostic quantitatif des biais d'annotation, améliorant l'interprétabilité des modèles UD.

KEYWORDS: Categorical overlap, Universal Dependencies, Distributional typology, Functional equivalence.

MOTS-CLÉS : Chevauchement catégoriel, Universal Dependencies, Typologie distributionnelle, Équivalence fonctionnelle.

1 Introduction

Determining where one part-of-speech (POS) ends and another begins has long been a central challenge in linguistic theory (Aarts, 2006). While traditional grammars often treat categories as discrete silos, modern linguistics views them as points on a continuum. This concept, known as syntactic gradience, suggests that words can possess varying degrees of membership within a category (Aarts, 2007). This "fuzziness" is especially prevalent in adjectives (Croft, 2007), whose morphosyntactic encoding ranges from noun-like to stative verb-like behaviours (Bhat, 1994; Dixon

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 28–40.

& Aikhenvald, 2004).

Universal Dependencies (UD) standardises these categories to facilitate cross-linguistic NLP (de Marneffe *et al.*, 2021), yet this can impose Indo-European biases and obscure language-specific functional realities (Croft, 2016; Haspelmath, 2010).

We address this tension by using the "noise" within UD treebanks to map language typology. Rather than accepting POS tags as ground truth, we treat them as anchor points to explore the distributional space around them. We propose a methodology to quantify "categorical drift" in French and compare it to the structurally distinct environment of Chinese, focusing specifically on adjectives. Our contributions are twofold: (1) quantitative metrics for purity and entropy, to identify "bridge words" that traverse POS boundaries, and (2) a tool for both linguists and NLP researchers to diagnose annotation bias, highlighting where the UD schema struggles to capture the nuances of non-Western languages. By viewing categorical overlap not as annotation error, but as a reflection of linguistic reality, this work provides a data-driven bridge between formal typology and computational linguistics.

2 Related Work

The computational counterpart to the linguistic gradience tradition begins with Manning (2011), who argues that probabilistic representations are better suited than discrete categories to phenomena like "transitive adjectives" (e.g., worth) whose mixed properties make categorical classification "artificial [...], whatever its convenience for current part-of-speech tagging technology" (p. 186). The Penn Treebank itself permitted disjunctive POS tags such as `{VERB, ADJ}` for indeterminate cases, though this option was used in only 0.01% of tokens (Marcus *et al.*, 1993; Manning, 2011). Empirical evidence for gradient POS membership has since accumulated from two directions: inter-annotator disagreement, where Plank *et al.* (2014) show that POS disagreement patterns are systematic across domains and languages, pointing to genuine ambiguity rather than annotator error; and distributional analysis, where Kutuzov *et al.* (2016) train classifiers on word2vec embeddings against UD POS labels and find that prediction errors cluster around well-known fuzzy classes. We differ from the latter in using interpretable UD morphosyntactic features rather than opaque embeddings, identifying bridge words via purity and entropy, and contrasting typologically distinct languages.

Our framing aligns most closely with the recent gradient typology work (Levshina, 2022, 2023; Levshina *et al.*, 2023), which argues, with extensive support from UD treebanks and using entropy as a primary measure, that typological categories are best modelled as continuous rather than discrete variables. Where Levshina and colleagues apply this view to word order, we extend it to lexical category. A close approach to ours belongs to Yih & Dai (2023), who examine UPOS x DEPREL cross-tabulations across 20 PUD treebanks to identify "transcategorization" mismatches that parallel our bridge words. However, their analysis is qualitative and combination-typed, whereas ours is continuous and lexical-unit-typed.

3 Data

We use Universal Dependencies (UD) treebanks (de Marneffe *et al.*, 2021), which provide standardised morphosyntactic annotations to facilitate comparisons between 180 typologically diverse languages (Zeman *et al.*, 2025). However, this consistency may impose an Indo-European structural bias on other languages, which we address by not simply comparing the behaviour of lexemes annotated as adjectives. Instead, we analyse the syntactic contexts of all lexemes to identify functional overlaps hidden behind universal labels. To test our distributional metrics across disparate grammatical architectures, we selected the largest available treebanks for two typologically distinct languages:

French (UD_French-GSD), an inflectional, fusional system containing approximately 400k tokens across 16k sentences (Guillaume *et al.*, 2019), and Chinese (UD_Chinese-GSD) an isolating, analytic system, consisting of 123k tokens and 4.9k sentences (Shen *et al.*, 2025).

4 Methodology

If categorical boundaries are gradient rather than discrete, then this gradience must be observable in the distributional behaviour of words. Our methodology operationalises this intuition. Rather than treating POS tags as ground truth, we use them as reference points and measure how strongly individual lexical items gravitate towards or away from their annotated category in syntactic space. This is achieved in four steps: (1) representing each word as a distributional syntactic profile, which is used to create a feature vector characterising all its occurrences in the treebank; (2) identifying its closest neighbours using cosine similarity; (3) measuring the homogeneity of its neighbourhood using purity and entropy; (4) aggregating these statistics across all words in the treebank, and visualising the stability and overlap of categories.

4.1 Lexical Units and Syntactic Profiles

To capture distributional behaviour while avoiding homographic ambiguity, we define a *lexical unit* as the pair `(<lemma>, <pos>)`. By distinguishing between units such as *(close, ADJ)* and *(close, VERB)* for example, we ensure that our analysis captures distinct grammatical functions rather than conflating homographs.

We characterise each lexical unit by its contextual environment within the dependency tree. Following Herrera *et al.* (2024), we define a search space for any given node that includes its parent, its children, and its immediate linear neighbours. Crucially, we exclude the morphosyntactic features of the node itself and the dependency relation linking it to its parent. This prevents circular results where units cluster together purely due to UD annotation conventions (e.g., the `amod` relation for adjectives or the `Tense` feature for verbs). By removing the node itself from the search space, we base its syntactic profile purely on contextual distribution.

To minimise noise, we exclude all lexical units with fewer than 10 occurrences, and to ensure comparability between languages, for each of the nodes in the search space, we restrict features to the universal attributes defined by UD[1].

The resulting feature matrix represents each lexical unit as a vector of percentages. For instance, if the unit *(manger, VERB)* occurs 100 times with a noun parent in 7 instances and a singular child in 30, the features `parent:upos=NOUN` and `child:Number=Sing` are 0.07 and 0.3 respectively. This normalisation ensures that the profile reflects the shape of the word's distribution rather than its raw frequency, allowing us to compare high-frequency and low-frequency words on equal footing.

4.2 Quantifying Categorical Proximity

To identify overlaps, we examine the neighbours of each lexical unit in the feature space using cosine similarity, which captures functional orientation regardless of frequency. For a lexical unit L, we identify its k nearest neighbours $L_1, L_2, ..., L_k$ subject to two constraints: a similarity threshold of above 0.7 and a neighbourhood cap of $k = 20$.

[1]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.

This ensures we only consider strong distributional alignments while preventing high-frequency categories such as nouns from disproportionately occupying a unit's vicinity. The cut-off of 0.7 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.

Once the neighbourhood of a lexical unit L has been identified, we determine its distribution across categories. Let $UPOS$ denote the set of universal POS categories in the treebank, and let $pos(L)$ be the annotated category of L. Each neighbour L_i contributes to the neighbourhood proportionally to its similarity with L. We assign a weight w_i such that $w_i = similarity(L, L_i)$, ensuring that highly similar neighbours influence the distribution more strongly than marginal ones.

For any given category $C \in UPOS$ we define the neighbourhood proportion for L as:

$$p(L, C) = \frac{\sum_{i=1}^{k} w_i \cdot \mathbb{I}(pos(L_i)=C)}{\sum_{i=1}^{k} w_i}$$

where $\mathbb{I}$ is 1 if the neighbour belongs to category C, and 0 otherwise. Intuitively, $p(L, C)$ represents how much of L's neighbourhood is occupied by category C.

However, raw neighbourhood proportions are influenced by global category size. Larger categories naturally appear more often amongst nearest neighbours, so we normalise these probabilities by the overall frequency of each category. Let $freq(C)$ denote the relative frequency of category C amongst all lexical units. We define:

$$p_{norm}(L, C) = \frac{p(L,C)/freq(C)}{\sum_{C' \in \text{UPOS}} (p(L,C')/freq(C'))}$$

All subsequent measures are computed using this normalised proportion, $p_{norm}(C)$.

4.3 Purity and Entropy

To identify the lexical units responsible for the overlap between two categories we use two complementary metrics: purity and entropy.

Purity measures the degree to which a lexical unit remains aligned with its annotated category. Formally, we define it as the normalised probability mass assigned to the category in which the unit is annotated:

$$Pur(L) = p_{norm}(L, pos(L))$$

A purity score of 1 means every close neighbour shares the same category, while a score of 0.5 means only half of the neighbours do. A score of 0 means the unit is completely surrounded by elements of a different category than its own.

Purity alone does not indicate whether a lexical unit's drift away from its annotated category is directed towards a single rival category or dispersed across several categories. To capture this distinction, we compute the entropy of the neighbourhood distribution:

$$H(L) = -\sum_{C \in \text{UPOS}} p_{norm}(L, C) \log(p_{norm}(L, C))$$

The interpretation is that entropy is 0 if every neighbour belongs to the same category, while a high entropy indicates that neighbours are from many different categories.

The interaction between the two metrics is particularly informative. A lexical unit with low purity and low entropy, is a "bridge word" moving away from its annotated category towards one specific rival category. By contrast, a lexical unit with low purity and high entropy is not being pulled towards a single category, but is instead surrounded by a diverse set of categories.

4.4 Global Category Overlap

The measures above describe individual lexical units. To understand category-level behaviour, we aggregate these local statistics across the entire treebank. We can extend $p_{norm}(L, C)$ for the whole category of L. For a given category C_0, we calculate $p_{norm}(C_0, C)$ as the average of $p_{norm}(L, C)$ for all $L \in C_0$. We interpret this as the probability that given a lexical unit in category C_0, its closest neighbours belong to category C. We visualise this using two complementary heatmaps, which together provide a global map of categorical fuzziness within the treebank:

- The Stability Matrix: The diagonal values $p_{norm}(C, C)$ measure the internal cohesion of each category. High diagonal values indicate that lexical units annotated as C tend to cluster together in the feature space.
- The Intruder Matrix: To examine the functional drift more closely, we remove the diagonal and renormalise the remaining values to sum to 1. This produces a matrix that amplifies off-diagonal signals. This visualisation reveals which rival category exerts the strongest pull when units deviate from their annotated class.

4.5 Bridge Words and Categorical Pull

To extract the words driving a high overlap between two categories, we identify the units with low purity and low entropy, designated as bridge words. These are lexical items that the annotation schema places in one category, but the distributional data pulls toward another.

A clear example is the participle in French: several verbs such as *isoler* (to isolate) or *extraire* (to extract) are almost always in the participle form in the French treebank, which makes their behaviour identical to that of prototypical adjectives. Our method detects this shift by comparing the vector of the bridge word (in this example, a verb) to the centroid (the median vector) of the rival category (in this example, adjectives).

By identifying the features where the value for the bridge word and that of the rival centroid are closest, we can pinpoint the exact morphosyntactic features responsible for the overlap. For example, if both the verb and the prototypical adjective have a similar value for the feature `parent:upos=NOUN`, as their parent is a noun in a similar proportion of their occurrences, we have empirical evidence of the behaviour that drives the overlap.

4.6 Null Baseline

To distinguish genuine distributional structure from normalisation artefacts, we compare each observed cell of $p_{norm}(C_0, C)$ to its expected value under a null in which neighbours are drawn independently at random, weighted by global category frequency. Under this null, two effects cancel: a random neighbour is more likely to fall in a large category, but each neighbour's contribution is then divided by that same frequency. The size advantage is exactly offset, so the expected value of every cell is $1/N$ for N categories. For our two treebanks ($N = 13$) this baseline is approximately 0.08; all reported diagonal values and the principal off-diagonal entries in Section 5 fall well above it. The observed structure therefore reflects genuine distributional signal rather than a side effect of the normalisation pipeline.

5 Results

While this approach is applicable to any category of any language with a treebank, we are focusing here on French, and more particularly on the case of adjectives. A comparison with Mandarin Chinese is proposed.

5.1 French Case Study

Figures 1 and 2 present the Stability Matrix and the Intruder Matrix respectively. Together they provide a snapshot of the categorical stability and fuzziness within the UD_French-GSD treebank.

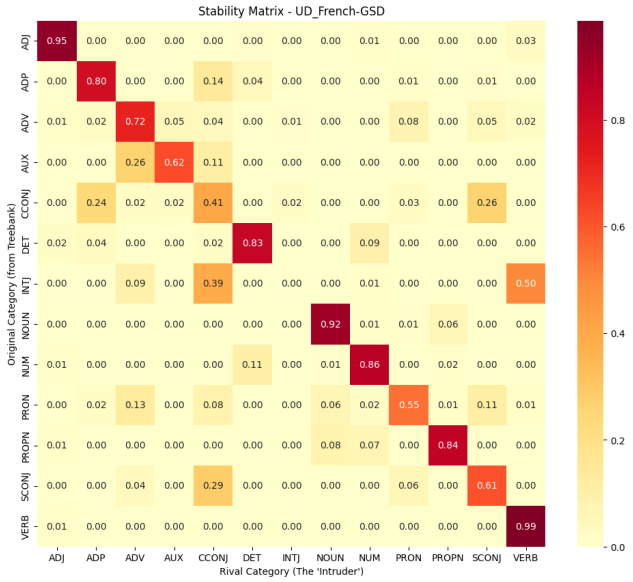

Figure 1: Stability matrix for French, showing category internal cohesion on the diagonal

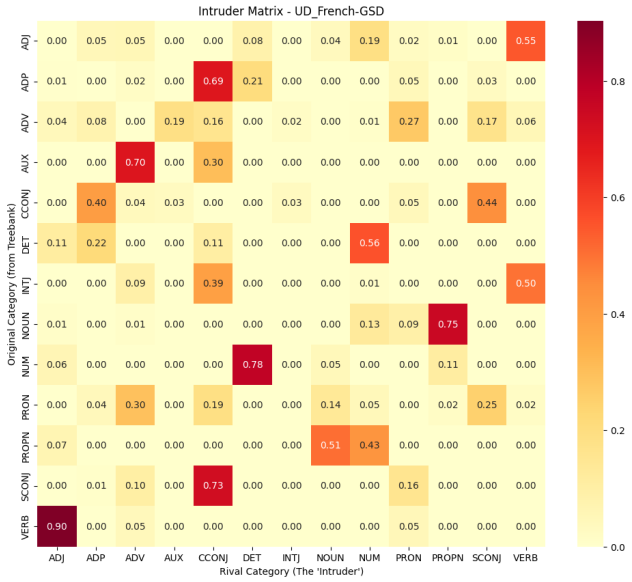

Figure 2: Intruder matrix for French, showing the functional drift more closely

5.1.1 Global Stability

The Stability Matrix displays on the diagonal the probability of a lexeme having same-category neighbours, reflecting internal cohesion. Core French categories, nouns, verbs and adjectives, are highly stable (>90%) occupying well-defined, discrete clusters within the syntactic feature space. Conversely, functional categories exhibit significant drift. Interjections are very heterogenous with a 0% stability and 50% of their neighbours annotated as verbs, revealing their role as stand-alone utterances. Coordinative conjunctions are similarly unstable, splitting neighbours between subordinative conjunctions (26%) and prepositions (24%) to form a broad “relator” cluster.

Pronoun fragmentation, drifting towards adverbs (13%), conjunctions (11% and 8%), and nouns (6%), highlights the diverse functional roles pronouns play, ranging from clitic-like behaviour near verbs to acting as full nominal substitutes. For instance, *rien* shares syntactic slots with negative adverbs (*pas, jamais*), while quantifiers like *beaucoup* shift between adverbial and nominal domains. To capture these shifts, we use the External POS feature as the POS label, allowing us to distinguish between *beaucoup* as an adverb and its pronominal use in partitive constructions (*beaucoup de...*), isolating the functional shift from the lexical root.

5.1.2 The Intruder Matrix

The Intruder Matrix (Figure 2) serves as a diagnostic tool to zoom in on these categorical drifts by isolating only the neighbours that belong to a different category than the annotated one. By re-normalising the off-diagonal signal, we can interpret these "intruders" as a measure of contextual interchangeability. The behaviour of French numerals (NUM) provides an example of this interpretive logic.

When a numeral is located in a neighbourhood dominated by non-numeric categories, it drifts toward determiners in 78% of cases. This high value reflects a near-total distributional overlap in quantification slots (*deux enfants* vs. *les enfants*). Smaller drifts toward proper nouns (11%) and adjectives (6%) further highlight how numerals occupy prepositional complements (*en 1952* vs *en France*) or pre-nominal modifier slots (*les deux enfants* vs *les beaux enfants*). By re-normalising these "intruders", the matrix demonstrates that while POS tags provide discrete labels, the distributional reality is one of significant, measurable overlap.

5.1.3 French Adjectives

While the methodology can diagnose overlaps across the entire treebank, we now narrow our focus to the adjective category, due to their pull by nominal and verbal domains.

While French adjectives are highly stable (95%), their 5% drift is directed mainly towards verbs, which account for 55% of non-adjectival neighbours. Conversely, when verbs exhibit fuzziness, the intruder is an adjective in 90% of instances. This reciprocal pull visible in Figure 2 reveals a functional symmetry between the two categories.

	ADJ → VERB (55%)	**ADJ → NUM (19%)**
Top Features	1. parent:position=before 2. next:upos=ADP 3. child:rel=cop	1. parent:upos=NOUN 2. parent:position=after 3. next:upos=NOUN
Top Bridge Words	*issu* 'issued/from'(0.13 | 0.39) *originaire* 'originating' (0.32 | 1.02) *censé* 'supposed' (0.26 | 0.57) *responsable* 'responsible' (0.32 | 0.63) *présent* 'present' (0.32 | 0.63)	*nombreux* 'numerous' (0.21 | 0.52) *divers* 'different' (0.34 | 0.64) *multiple* 'multiple' (0.39 | 0.67) *occidental* 'occidental' (0.55 | 0.69) *autre* 'other' (0.56 | 0.69)

	VERB → ADJ (90%)	**DET → ADJ (11%)**	**PROPN → ADJ (7%)**
Top Features	1. parent:position=before 2. parent:upos=NOUN 3. next:upos=ADP	1. parent:position=after 2. parent:upos=NOUN 3. prev:upos=ADP	1. parent:position=before 2. parent:upos=NOUN 3. child:rel_shallow=nmod
Top Bridge Words	*intituler* 'to title' (0.07 | 0.27) *réputer* 'to deem' (0.19 | 0.49) *isoler* 'to isolate' (0.24 | 0.55) *armer* 'to arm' (0.24 | 0.55) *extraire* 'to extract' (0.47 | 0.69)	*du* 'of the' (0.44 | 1.32) *aucun* 'none/any' (0.49 | 1.25) *un* 'a/an' (0.49 | 1.25) *tel* 'such' (0.44 | 1.03) *chaque* 'each' (0.49 | 1.03)	*B* (0.00 | 0.00) *Notre-Dame* (0.29 | 0.60)

Table 1: French adjective features and bridge words with purity/entropy scores $(Pur|H)$. Table headers represent the categorical overlaps and their scores, as identified in the Intruder Matrix.

The verb-adjective overlap emerges from two directions. From the adjectival side, bridge words like *issu* ($Pur = 0.13 \mid H = 0.39$) or *censé* ($Pur = 0.26 \mid H = 0.57$) exhibit low purity scores. A Pur value of 0.13 indicates that 87% of the word's distributional neighbours are annotated as a different category, in this case verbs. The low entropy suggests these adjectives are pulled consistently towards a single other category, here the verbal domain. Their high valency and predicative roles are captured by features like `next:upos=ADP` and `child:rel=cop`, where they function as clausal heads rather than simple modifiers.

From the verbal side, past participles of verbs like *intituler* ($Pur = 0.07 \mid H = 0.27$) or *réputer* ($Pur = 0.19 \mid H = 0.49$) lose eventive properties in passive or attributive contexts, and function as direct nominal modifiers (`parent:upos=NOUN`). This distributional evidence provides empirical weight to Tesnière (1959)'s theory of "transference", the process that transfers a verb into the adjectival domain, and to the more recent work of Abeillé & Godard (2021), who also treat participles as a distinct category with both verbal and adjectival properties (p. 126).

Fuzziness also occurs at the intersection of adjectives, numerals, and determiners. "Quantifying adjectives" (Abeillé & Godard, 2021) such as *nombreux* ($Pur = 0.21 \mid H = 0.52$) and *divers* ($Pur = 0.34 \mid H = 0.64$) spend much of their "distributional life" in the neighbourhood of numerals sharing the same pre-nominal environment (`next:upos=NOUN`). Conversely, items like *du* ($Pur = 0.44 \mid H = 1.32$) or *aucun* ($Pur = 0.49 \mid H = 1.25$) show higher entropy, indicating that the determiners are in a more diverse neighbourhood, with more than two categories overlapping.

Finally, our metrics surface a specific functional shift where proper nouns behave as adjectives in fixed environments. The bridge words *B* ($Pur = 0.00 \mid H = 0.00$) in *annexe B* and *Notre-Dame* ($Pur = 0.29 \mid H = 0.60$) in *l'église Notre-Dame* show the distribution of post-nominal adjectives. These Pur and H scores expose functional roles often masked by UD's focus on morphological form.

5.2 Contrastive Analysis: The Chinese Distributional Model

The UD_Chinese-GSD treebank reveals a categorical landscape that contrasts sharply with the French baseline. In this isolating language, the lack of morphological markers makes distributional profiles the primary signal for classification.

5.2.1 Global Stability

The Chinese Stability Matrix (Figure 3) identifies verbs as the most cohesive category (90%). This suggests that the verbal environment, defined by negation patterns and aspect marking, is remarkably distinct. In Chinese, verbal forms frequently perform functions that Indo-European languages assign to categories like prepositions or adjectives.

However, rather than the verb category being fragmented by its diverse functions, its dominance appears to destabilise peripheral functional categories. Unlike the relatively stable French functional classes, Chinese coordinating conjunctions show only 35% internal cohesion, drifting towards adverbs (13%), prepositions (12%), and auxiliaries (11%). This suggests that items traditionally glossed as "conjunctions" function as a broad class of discourse-level particles. Similarly, adverbs and determiners drift towards structural markers and numerals, forming a "functional particle" cluster where the modifier-marker boundary is significantly thinner than in French.

Crucially, while French adjectives are highly stable (95%), Chinese adjectives possess same-category neighbours in only 60% of cases. Their distributional proximity to verbs (7%) and determiners (10%) provides a quantitative reflection of the "stative verb" debate in Chinese linguistics (Chao, 1968).

In French, the adjective-verb overlap is an inflectional "transference" (e.g., participles). In Chinese, the overlap stems from predicative flexibility: adjectives can function as predicates without a copula, occupying intransitive verb slots. Our metrics capture this split identity, showing the Chinese adjective is distributionally divided between its role as a nominal modifier and its role as a verb-like predicate.

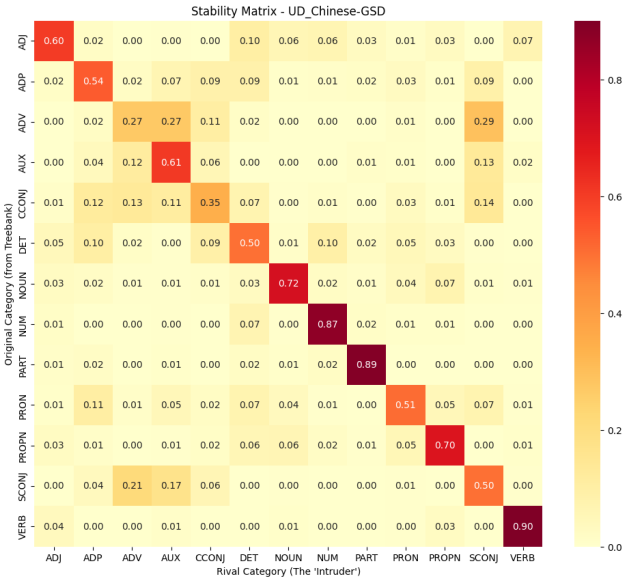

Figure 3: Stability matrix for Chinese, showing category internal cohesion on the diagonal

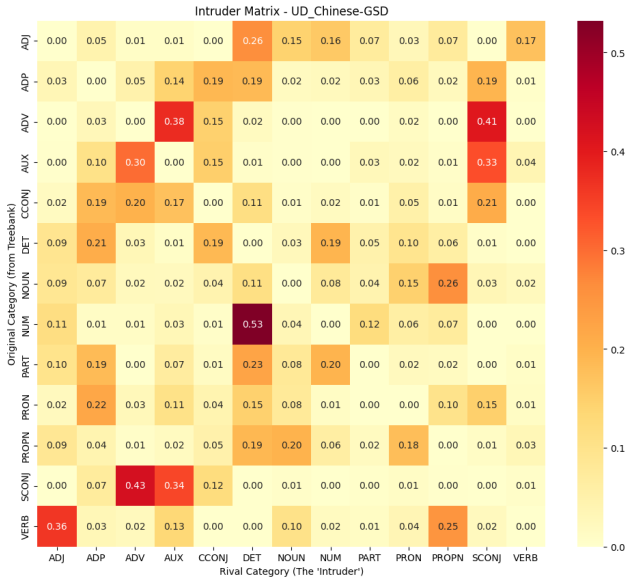

Figure 4: Intruder matrix for Chinese, showing the functional drift more closely

5.2.2 The Intruder Matrix

While the French Intruder Matrix revealed specific, relatively reciprocal bridges, the Chinese matrix (Figure 4) suggests a more diffused functional overlap. When Chinese verbs exhibit distributional drift, they gravitate towards adjectives (36%) and proper nouns (25%). While the adjectival overlap is the expected signature of stative predication, the drift toward proper nouns likely reflects nominalization in newspaper title contexts.

Unlike the French baseline, Chinese adjectives gravitate more strongly towards determiners (26%) and numerals (16%). This shift is uniquely driven by the Chinese classifier system: adjectives, numerals, and classifiers frequently compete for the same pre-nominal modifier slot. Although their internal syntax differs, their shared position relative to the noun head creates distributional similarity in a feature space sensitive to parent and neighbour relations.

5.2.3 The Chinese Adjective

Bridge words in Table 2 reveal a functional landscape fundamentally different from the French case. While French drift is often a focused shift between two categories (low entropy), Chinese adjectives exist in a state of distributional sprawl (high entropy), where multiple categories compete for the same syntactic slots.

The stative verb phenomenon is empirically confirmed by bridge words like 少 ('few') and 多樣 ('diverse'), which exhibit a purity of 0.00. This indicates that these annotated adjectives have zero neighbours tagged as adjectives, they are entirely surrounded by verbs. Unlike French participles, which use inflectional morphology to "bridge" the verb-adjective gap, these Chinese items rely on positional equivalence: they act as verbs because they occupy the verbal slot.

A major divergence from French is the high-entropy overlap between adjectives, determiners, and numerals. Quantifiers like 一些 ('some') and 少數 ('a few') show extremely low purity ($Pur \leq 0.07$) but high entropy ($H > 1.4$). This confirms they occupy a fuzzy space, shared by multiple cate-

	ADJ → DET (26%)	ADJ → VERB (17%)	ADJ → NUM (16%)	ADJ → NOUN (15%)
Top Features	1. parent:position=after 2. parent:upos=NOUN 3. prev:upos=VERB	1. parent:position=after 2. child:upos=NOUN 3. prev:upos=NOUN	1. parent:position=after 2. parent:upos=NOUN 3. prev:upos=NOUN	1. parent:upos=VERB 2. next:upos=PUNCT 3. parent:position=after
Top Bridge Words	一些 'some' (0.07 | 1.57) 少數 'a few' (0.06 | 1.42) 一般 'general' (0.18 | 1.63) 共同 'common' (0.16 | 1.47) 公開 'public' (0.05 | 1.21)	多樣 'diverse' (0.00 | 0.68) 小 'small' (0.21 | 0.52) 少 'few' (0.00 | 0.00) 豐富 'abundant' (0.36 | 0.65) 長 'long' (0.49 | 0.69)	超級 'super' (0.13 | 1.39) 私人 'private' (0.26 | 1.39) 高速 'high-speed' (0.24 | 1.34) 天然 'natural' (0.36 | 1.43) 高等 'advanced' (0.37 | 1.32)	不滿 'dissatisfied' (0.00 | 1.24) 般 'sort of' (0.00 | 0.65) 一樣 'the same' (0.00 | 0.00) 主 'main' (0.00 | 0.00) 好 'good' (0.55 | 1.15)

	VERB → ADJ (36%)	NUM → ADJ (11%)	PART → ADJ (10%)	DET → ADJ (9%)	NOUN → ADJ (9%)
Top Features	1. parent:position=after 2. child:upos=SCONJ 3. parent:upos=NOUN	1. parent:position=after 2. parent:upos=NOUN 3. prev:upos=NOUN	1. parent:position=after 2. parent:upos=NOUN 3. parent:upos=PART	1. parent:position=after 2. parent:upos=NOUN 3. prev:upos=VERB	1. parent:position=after 2. next:upos=NOUN 3. prev:upos=NOUN
Top Bridge Words	非 'non-' (0.00 | 1.34) 領導 'lead' (0.12 | 1.11) 建設 'build' (0.08 | 0.28) 知 'know' (0.10 | 0.32) 現存 'existing' (0.06 | 0.24)	八 'eight' (0.84 | 0.60) *22* (0.88 | 0.46) *18* (0.90 | 0.40) 第三 'third' (0.93 | 0.25) *60* (0.93 | 0.25)	非 'non-' (0.17 | 1.35) 主 'main' (0.16 | 1.12)	所有 'all' (0.27 | 1.58) 任何 'any' (0.56 | 1.13) 各類 'each kind' (0.63 | 1.11)	大型 'large-scale' (0.04 | 1.68) 小型 'small-scale' (0.07 | 1.49) 高等 'high-level' (0.02 | 1.25) 黑人 'black person' (0.17 | 1.45) 飛行 'flight' (0.17 | 1.45)

Table 2: Chinese adjective features and bridge words with purity/entropy scores $(Pur|H)$

gories acting as pre-nominal modifiers (`parent:upos=NOUN`). For determiners, this quantitatively identifies the friction of enforcing the UD `DET` category on Chinese, as these items lack a distinct distributional profile.

While French boundaries are maintained by agreement patterns, the Chinese classifier system and lack of agreement cause these categories to converge. This is further evidenced by the adjective-numeral overlap (超級 'super', 高速 'high-speed'), where items appear in complex nominal chains, behaving less like qualitative descriptors and more like structural components of a noun phrase.

The noun-adjective drift surfaces a uniquely Chinese linguistic category: the distinguishing words (*qūbiécí*) (Zhu, 1982). Lexemes ending in 型 ('type/scale'), such as 大型 ('large-scale') and 小型 ('small-scale'), show near-zero purity and high entropy. Traditionally classified as non-predicate adjectives, our data shows their distribution is genuinely hybrid, spanning nominal and adjectival contexts.

Similarly, the verb-adjective overlap highlights modifiers like 非 ('non-'). While tagged as a verb or particle in UD, its Pur and H scores reveal that 非 distributionally mimics an adjective. This diagnostic confirms that UD's morphological tagging often conflicts with the actual functional role of "prefix-like" verbs in Chinese, a tension that is far less prevalent in the more morphologically transparent French treebank.

5.3 Implications for NLP and UD Annotation

The quantitative patterns observed here have implications for computational modelling. In languages with high category overlap, distributional separability in the UD feature space is reduced. We hypothesise that this structural proximity may constrain separability in strictly categorical POS tagging architectures, independent of model quality.

This does not indicate annotation error. Rather, it reflects the fact that cross-linguistically standardised POS distinctions may correspond to gradients in morphosyntactic behaviour in some languages. The UD framework provides consistent cross-linguistic categories; however, the distributional evidence suggests that these categories are not equally discrete across languages.

By quantifying overlap through entropy and purity measures, this study offers a method for identifying where categorical distinctions are robust and where they are diffuse. Such information may inform the development of tagging systems that incorporate probabilistic or gradient representations of category

membership, particularly in languages with high predicate flexibility or limited morphological marking of category.

6 Conclusion

This paper shifts the perspective on cross-linguistic NLP by treating UD "noise" not as an annotation deficit, but as a signal of linguistic gradience. By representing lexical units as distributional profiles and quantifying their categorical drift through purity and entropy metrics, we have demonstrated that part-of-speech categories are far from discrete silos.

While French demonstrates a “transference” model where morphological safeguards maintain categorical stability even during functional shifts, Chinese presents a model where categories like adjectives and stative verbs share a single distributional profile. These findings suggest that classifying languages based on functional overlap, rather than purely genetic or geographical criteria, can reveal deeper structural convergences. Furthermore, the automatic identification of “bridge words” provides a powerful diagnostic tool. By identifying bridge words, from the adjectival behaviour of French past participles to Chinese *qūbiécí*, our metrics quantify the tension between formal category and contextual function.

A limitation of the current approach is that it operates on morphosyntactic features alone, so bridge words that overlap purely on semantic grounds are not detected. Future work will involve enriching the representation of lexical units using embeddings that will add a semantic dimension. A second limitation is corpus size, which limits coverage of low-frequency lexical units. The approach would therefore be strengthened by enlarging the dataset with automatically parsed texts, as well as including a wider array of isolating and agglutinative languages. By bridging the gap between formal typology and computational methodology, we move closer to NLP practices that respect the inherent fluidity of grammatical categories, and ultimately of human language.

Acknowledgments

I would like to thank my supervisor, Sylvain Kahane, for his invaluable feedback and for the meticulous reading of multiple drafts of this paper.

References

AARTS B. (2006). Conceptions of categorization in the history of linguistics. *Language Sciences*, **28**(4), 361–385. DOI : 10.1016/j.langsci.2005.10.001.

AARTS B. (2007). *Syntactic gradience: the nature of grammaticaI indeterminacy*. Oxford: Oxford University Press.

ABEILLÉ A. & GODARD D., Éds. (2021). *La Grande Grammaire du français*. Arles: Actes Sud. In collaboration with Annie Delaveau and Antoine Gautier.

BHAT D. N. S. (1994). *The Adjectival Category*. Studies in Language Companion Series. Amsterdam, Netherlands: Benjamins (John) North America.

CHAO Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

CROFT W. (2016). Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, **20**(2), 377–393. DOI : 10.1515/lingty-2016-0012.

CROFT W. A. (2007). Beyond Aristotle and gradience: A reply to Aarts. *Studies in Language*, **31**(2), 409–430. DOI : https://doi.org/10.1075/sl.31.2.05cro.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. DOI : 10.1162/coli_a_00402.

DIXON R. M. W. & AIKHENVALD A. Y. (2004). *Adjective Classes: A Cross-Linguistic Typology*. Oxford University Press. DOI : 10.1093/oso/9780199270934.001.0001.

GUILLAUME B., DE MARNEFFE M.-C. & PERRIER G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]. *Traitement Automatique des Langues*, **60**(2), 71–95.

HASPELMATH M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, **86**(3), 663–687. DOI : 10.1353/lan.2010.0021.

HERRERA S., CORRO C. & KAHANE S. (2024). Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

KUTUZOV A., VELLDAL E. & ØVRELID L. (2016). Redefining part-of-speech classes with distributional semantic models. In S. RIEZLER & Y. GOLDBERG, Éds., *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, p. 115–125, Berlin, Germany: Association for Computational Linguistics. DOI : 10.18653/v1/K16-1012.

LEVSHINA N. (2022). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, **26**(1), 129–160. DOI : doi:10.1515/lingty-2020-0118.

LEVSHINA N. (2023). Word classes in corpus linguistics. In E. VAN LIER, Éd., *The Oxford Handbook of Word Classes*. Oxford University Press. DOI : 10.1093/oxfordhb/9780198852889.013.34.

LEVSHINA N., NAMBOODIRIPAD S., ALLASSONNIÈRE-TANG M., KRAMER M., TALAMO L., VERKERK A., WILMOTH S., RODRIGUEZ G. G., GUPTON T. M., KIDD E., LIU Z., NACCARATO C., NORDLINGER R., PANOVA A. & STOYNOVA N. (2023). Why we need a gradient approach to word order. *Linguistics*, **61**(4), 825–883. DOI : doi:10.1515/ling-2021-0098.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In A. F. GELBUKH, Éd., *Computational Linguistics and Intelligent Text Processing*, p. 171–189, Berlin, Heidelberg: Springer Berlin Heidelberg.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

PLANK B., HOVY D. & SØGAARD A. (2014). Linguistically debatable or just plain wrong? In K. TOUTANOVA & H. WU, Éds., *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 507–511, Baltimore, Maryland: Association for Computational Linguistics. DOI : 10.3115/v1/P14-2083.

RADOVANOVIĆ M., NANOPOULOS A. & IVANOVIĆ M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, **11**(86), 2487–2531.

SHEN M., MCDONALD R., ZEMAN D. & QI P. (2025). UD_Chinese-GSD. https://github.com/UniversalDependencies/UD_Chinese-GSD. Part of Universal Dependencies.

TESNIÈRE L. (1959). *Eléments de Syntaxe Structurale*. Paris: Klincksieck.

YIH T. & DAI Z. (2023). UPOS-DEPREL mismatches: Detecting annotation errors and improving UD guidelines based on linguistic knowledge. In C.-R. HUANG, Y. HARADA, J.-B. KIM, S. CHEN, Y.-Y. HSU, E. CHERSONI, P. A, W. H. ZENG, B. PENG, Y. LI & J. LI, Éds., *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, p. 46–58, Hong Kong, China: Association for Computational Linguistics.

ZEMAN D., NIVRE J. *et al.* (2025). Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

ZHU D. (1982). *Yǔfǎ jiǎngyì [Lectures on Grammar]*. Beijing: Commercial Press.