

Cross-lingual expectation failure in definitional clauses: structural variation and prediction across five EU languages

Introduction. Definitional copular clauses (“*X*” means *NP*) function as expectation-anchoring discourse units: in legal and technical text, they establish a local discourse frame that resolves a term’s meaning for subsequent reference. A growing body of computational research shows that NLP systems build structural expectations about these constructions from training data, and that cross-lingual transfer of such expectations fails in predictable ways. We present corpus evidence from EU financial regulation — a uniquely controlled source where the same propositional content (financial term definitions) must be expressed across all 24 official EU languages — to show that structural expectations about definitional clauses differ substantially across languages, and to quantify the prediction cost of cross-lingual transfer.

We analyse ~120 definitional clauses per language across five typologically varied languages (English, German, Bulgarian, Finnish, Croatian) from three EU financial directives. The corpus controls for propositional content, allowing structural variation to be isolated from meaning.

Finding 1 — Structural expectations differ systematically across languages. English-trained systems expect the template [*quoted term*] + *means* + [*bare NP*], with the definiens as a nominative predicate DP. This expectation is violated in three of the five languages:

(1a) EN: ‘CCP’ means **a legal person** that interposes itself ... (nominative bare NP) (1b) HR: „strana” znači **pravnu osobu koja** ... (accusative DP — unexpected for copular parser) (1c) FI: tarkoitetaan ‘strana’+[ADESSIVE] **oikeushenkilöä**+[PARTITIVE] ... (impersonal passive, no subject)

Croatian (*znači*) assigns accusative case to the definiens DP — verified across all definitions in the corpus and confirmed as constructional, not lexical, by UD Croatian-SET treebank analysis (6,914 sentences; *znači* takes nominative 59% of the time in general language). German uses a single metalinguistic frame sentence (*bezeichnet der Ausdruck*) with no per-item verb, absent in nationally-drafted German law, confirming it is a EU-register convention rather than a general German structural expectation. Finnish uses an impersonal passive (*tarkoitetaan*) with no subject DP and partitive-marked definiens — a construction with no structural analogue in English.

Finding 2 — Expectation violations are quantifiable in NLP output. We validate the prediction-failure claim empirically using Stanza (UD Croatian-SET model) to parse 20 reconstructed Croatian definitional sentences. When the definiens is morphologically accusative, Stanza assigns obj (direct object) rather than nsubj (copular predicate): 11/20 (55%) definitions receive the transitive-VP parse. English-transferred definition-extraction systems that expect a nsubj/xcomp label miss 55% of Croatian definitions — an upper-bound recall of 45%. This is not a parser failure per se: Stanza correctly reads the accusative morphology; the failure is the mismatch between the structural expectation encoded in training data and the actual construction used.

Finding 3 — Within-language definiteness expectations track semantic type. Even within English, definiteness on the predicate DP functions as a structural expectation signal tied to the semantic type of the definition. Legal-person and instrument definitions consistently use an *indefinite* predicate DP (*means a legal person, means an agreement*), encoding kind-level / property predication. Process and unique-role definitions consistently use a *definite* DP (*means the process of, means the risk that*), encoding unique-referent predication. Chi-square tests confirm this split is non-random across all three languages with overt definiteness morphology (EN: $\chi^2=30.7$, $p<0.0001$; DE: $\chi^2=12.5$, $p=0.0004$; BG: $\chi^2=13.3$, $p=0.0003$). A model processing EU legal text can thus predict definiteness marking on the upcoming definiens DP from the semantic type of the term being defined — a within-clause expectation measurable before the DP is encountered.

Implications. The results show that definitional clauses, despite their uniform discourse function (term resolution), are realised through structurally diverse constructions across languages. This structural diversity generates systematic, quantifiable expectation mismatches when models trained on English process Croatian, Finnish, or German regulatory text. The findings speak directly to questions about how local syntactic expectations interact with global discourse function: the same discourse goal (definitional frame) can be realised without a copular verb (Finnish), with an accusative complement (Croatian), or with a metalinguistic subject (German), and each realisation predicts different downstream processing profiles. The parallel aligned corpus (~600 definitional clauses, publicly available via EUR-Lex) provides a benchmark for studying cross-lingual structural expectations in a tightly controlled definitional discourse genre.

References. Coppock & Beaver (2015) *Linguistics and Philosophy* 38. — Mikkelsen (2005) *Copular Clauses*, Benjamins. — Roy (2013) *Nonverbal Predication*, OUP. — Zovko Dinković (2010) *Suvremena lingvistika* 69.