# ICLE-RC: International Corpus of Learner English for Relative Clauses

**Debopam Das**
Åbo Akademi University
Tehtaankatu 2
20500 Turku, Finland
debopam.das@abo.fi

**Izabela Czerniak**
Åbo Akademi University
Tehtaankatu 2
20500 Turku, Finland
izabela.czerniak@abo.fi

**Peter Bourgonje**
University of Potsdam
Karl-Liebknecht-Str. 24-25
14476 Potsdam, Germany
bourgonje@uni-potsdam.de

## Abstract

We present the ICLE-RC, a corpus of learner English texts annotated for relative clauses and related phenomena. The corpus contains a collection of 144 academic essays from the International Corpus of Learner English (ICLE; Granger et al., 2020), representing six L1 backgrounds – Finnish, Italian, Polish, Swedish, Turkish, and Urdu. These texts are annotated for over 900 relative clauses, with respect to a wide array of lexical, syntactic, semantic, and discourse features. The corpus also provides annotation of over 400 related phenomena (it-clefts, pseudo-clefts, existential-relatives, etc.). Here, we describe the corpus annotation framework, report on the IAA study, discuss the prospects of (semi-)automating annotation, and present the first results from our corpus analysis. We envisage the ICLE-RC to be used as a valuable resource for research on relative clauses in SLA, language typology, World Englishes, and discourse analysis.

## 1 Introduction

Relative clauses (henceforth RCs) are a type of subordinate clauses that typically modify nouns or noun phrases, and sometimes also adjectives[1], adverbs[2], PPs[3], VPs[4], and even entire clauses[5]. RCs in English (and beyond) have extensively been studied for a wide range of themes, such as syntactic and typological variation (Comrie, 1998; Grosu, 2012), semantic features (Cornish, 2018), discourse functions (Brandt et al., 2009), diachronic development (Fajri and Okwar, 2020), FLA/SLA (Doughty, 1991), parsing (Goad et al., 2021), and processing (Reali and Christiansen, 2007), to name but a few of more recent work.

[1] *Pat is [beautiful], which, however, many consider her not.*
[2] *He moved [abroad] where he found a good job.*
[3] *He found a body [under the bridge] where nothing grows.*
[4] *She told me to [design it myself], which I simply can't.*
[5] *[Alex bought a mansion], which made him bankrupt.*

In this paper, we present the ICLE-RC, a new corpus of English RCs and related phenomena. The latter includes constructions such as it-clefts, pseudo-clefts, and existential-relatives that employ words like *that*, *which*, or *who*, which are otherwise known as relative markers, frequently used to introduce relative clauses. The ICLE-RC uses a subset of the International Corpus of Learner English (ICLE; Granger et al., 2020). The first version of the ICLE-RC contains 144 ICLE texts, covering six L1 backgrounds – Finnish, Italian, Polish, Swedish, Turkish, and Urdu – with 24 texts from each. These texts are annotated for 924 RCs, with respect to a wide array of lexical, syntactic, semantic, and discourse features. These texts are also annotated for 407 related phenomena, which we call *other constructions* (henceforth OCs).

The paper is structured as follows: Section 2 outlines the motivation behind the creation of the ICLE-RC. The composition of the corpus is described in Section 3. We describe the annotation framework for RCs and OCs in Section 4 and Section 5, respectively. Section 6 reports on an IAA study, and highlights challenges in our RC annotation. The prospects of (semi-)automating the RC annotation is discussed in Section 7. We present the first results from our corpus analysis in Section 8. Related work is briefly described in Section 9. Section 10 concludes the paper with an outlook on the future work and applications of the corpus.

## 2 Motivation

The development of the ICLE-RC stems from a number of reasons. First, the corpus would provide real language data to assess English learners' use of RCs against the standard rules of English grammars (e.g., the use of *which* for a human referent, or the use of a comma for integrated RCs). Second, the six L1 backgrounds covered in the ICLE-RC represent six different language families (Pereltsvaig,

2023) – Finnish: Uralic; Italian: Romance; Polish: Slavic; Swedish: Germanic; Turkish: Turkic; and Urdu: Indo-Aryan[6]. This would allow identifying typological patterns for certain RC features potentially resulting from cross-linguistic influence (e.g., the use of extraposed RCs). This would also offer significant implications for research in World Englishes, in comparison to native varieties of English (e.g., by comparing the ICLE-RC with comparable corpora such as ICNALE (Ishikawa, 2023) as well as those of native academic English such as LOCNESS (Granger, 1998)). Third, the corpus would help us explore English learners' use of other constructions as alternative strategies of information structuring, in addition to RCs. Finally, although corpus-based studies exist for English RCs, they have mostly used small-size data sets designed to tackle very specific RC-oriented issues (see Section 9). To our knowledge, there is no large-scale corpus of English RCs with a feature-rich annotation framework. The ICLE-RC is designed to accommodate a wide variety of English texts, and support the annotation of RCs therein with a comprehensive coverage of linguistic features pertaining to lexical, syntactic, semantic, and discourse domains.

## 3 Data selection and setup of the corpus

The ICLE-RC derives from the ICLE (Granger et al., 2020), which is a corpus of academic essays written by undergraduate students from a given set of topics. These students are intermediate or advanced learners of English, coming from different L1 backgrounds such as Chinese, Dutch, Finnish, French, German, Greek, Hungarian, Italian, Japanese, Polish, Russian, Spanish, Swedish, Turkish, and Urdu. The data collection for the ICLE was initiated in the late 1990s, and has since been coordinated by Sylviane Granger at the Centre for English Corpus Linguistics at the University of Louvain. The corpus has grown over the years as a result of close collaboration with a large number of partner universities around the world. The most recent version of the corpus (ICLEv3) includes over 5.5 million words covering 25 L1 backgrounds[7].

The ICLE-RC includes 144 ICLE essays (100K+ words), which are equally distributed into 24 essays from six L1 backgrounds, namely Finnish, Italian, Polish, Swedish, Turkish, and Urdu. These 24 essays for each language are compiled from three institutions (with 8 essays from each), which are further balanced for the gender of the writer[8], whenever possible. The detailed distribution of the essays in the ICLE-RC is provided in Table 9 in the Appendix.

## 4 Annotation framework for RC

The relative clauses (RCs)[9] in the ICLE-RC are annotated for a wide range of lexical, syntactic, semantic, and discourse features. These features are grouped into seven primary categories, as listed in Table 1. The complete taxonomy of the annotation features is provided in Table 10 in the Appendix.

**RELATIVE MARKER (RM):** RMs are words that introduce an RC. RMs include the subordinator *that* and relative pronouns such as *which*, *who*, or *whose*. In the ICLE-RC, the RM feature includes three sub-features: `that`, `wh-word`, and `zero` (i.e., the absence of an overt RM for bare-relatives). These categories are exemplified below[10].

(1) Our duty should be to select programmes and to see only things ***that*** *open our mind*. [Italian; ITRS-1002]

(2) Those, ***who*** *cannot afford advertising campaigns led on a large scale*, have no chances of achieving success in any kind of business. [Polish; POLU-1006]

(3) **The status** ø *English has acquired today* is so dominant that it seems unlikely that the situation could ever change. [Finnish; FIJO-1003]

**REFERENT FUNCTION:** This feature identifies the grammatical function of the referent of the RM in the matrix clause. It includes seven categories: `subject`, `direct object`, `indirect object`, `predicative complement`, `adjunct`, and `clause`. Each category (except `clause`) further includes sub-categories; for example, `direct object`,

[6]The selection yields four Indo-European and two non-Indo-European languages.

[7]For specimen essays, check out the ICLE500 dataset.

[8]The classification follows from the ICLE.

[9]We only annotate full RCs, and exclude reduced RCs on grounds of parsing and processing difficulties (Acuña Fariña, 2000; McKoon and Ratcliff, 2003).

[10]**Conventions for examples:** The RC is in italics; the RM is in bold; the referent is underlined. In case of RM-zero, there is no overt RM, and the referent is marked in bold instead. The text inside the square brackets lists the L1 background and the file number of the source text. **Note:** Some examples contain grammatical/spelling errors (as written by L2 students).

| # | feature | examples (of sub-features) | feature type |
|---|---|---|---|
| 1 | relative marker (RM) | *that*, *which*, *who*, zero | lexical/syntactic |
| 2 | grammatical function of referent | subject, object, predicative complement | syntactic |
| 3 | grammatical function of RM | subject, object, adjunct | |
| 4 | embedding of RC | embedded, non-embedded | |
| 5 | extraposition of RC | extraposed, non-extraposed | |
| 6 | type of referent | human, abstract entity | semantic/discourse |
| 7 | restrictiveness | integrated, supplementary | syntactic/discourse |

Table 1: Primary categories of relative clause annotation

which refers to the direct object in the matrix clause, has three subtypes:

**`direct-object-head-n:`** The head noun of the direct object NP is the referent, as in (4). (If there is any complement and/or adjunct within that NP, the whole NP is considered as the referent.)

(4) ... they watch programms [sic] of cartoons ***which*** *are mostly in Hindi ...* [Urdu; PALW-1014]

**`in-dir-obj-comp:`** An NP which is part of a complement within the direct object NP is the referent, as in (5).

(5) The main objection is the fact that it creates the demand for things ***that*** *people do not need.* [Polish; POLU-1006]

**`in-dir-obj-adjunct:`** An (NP which is part of an) adjunct within the direct object NP is the referent, as in (6).

(6) According to that great king ... people ... should be punished by imposing on them the penalty equal in quality to **the criminal offences** ø *those people were charged with.* [Polish; POSI-1001]

**MARKER FUNCTION:** This feature identifies the grammatical function of the relativised item (represented by the RM) in the RC. It comprises nine categories, largely adapted from Huddleston and Pullum (2002): `subject`, `direct object`, `indirect object`, `predicative complement`, `genitive subject determiner`, `predicate complement of auxiliary verb`, `head of a to-infinitival VP`, and `adjunct`. For illustration, we here define and exemplify only three of those categories (for information about all categories and sub-categories, see Table 10 in the Appendix).

**`subject:`** The relativised item functions as the subject in the RC, as in (7).

(7) These teachers ***who*** *want to prevent cheating* were once students. [Turkish; TRCU-1004]

**`genitive subject determiner`**: The relativised item (*whose*) is the genitive determiner in the subject NP of the RC, as in (8).

(8) ... his proposal is not only urgent but necessary as well for a democracy ***whose*** *purpose consists of controlling any political power.* [Italian, ITRS-1004]

**`adjunct:`** The relativised item functions as an adjunct or part of an adjunct in the RC. For adjuncts, the RC is usually introduced by *which*, *when*, or *where* (as in (9)).

(9) ... the newspapers have talked about child-porno and the right to have in one's possession videos or photos ***where*** *children are being exploited.* [Finnish; FIJY-1006]

**EMBEDDING:** This feature concerns whether the RC (and also its host clause) is embedded within a more superordinate matrix clause. The embedding clause is usually an attributive clause (e.g., *he said*) or a similar clause with a cognitive verb (e.g., *I think*), as in (10)[11]. Embedding rarely occurs in the ICLE-RC.

(10) The emphasis should be put on integration, since all cultures must be considered equal, and they should be able to co-exist in

[11] The embedder clause is marked by square brackets.

a highly civilized society, ***which*** *[we like to think] our own is*. [Swedish; SWUG-2007]

**EXTRAPOSITION:** Extraposition occurs when an RM does not immediately follow its referent. Instead, there are some intervening elements between the RM and its referent, as in (11). Unlike German which frequently allows extraposition of RCs (Gamon et al., 2002), the use of such constructions is found to be marginal in English (Levy et al., 2012), and also in the ICLE-RC.

(11) The once mighty state-churches have mostly diminished into mere baptizing-, wedding-, and funeral-organizers, ***whose*** *congregations rarely even believe in God*. [Finnish; FIHE-1015]

**REFERENT TYPE:** This represents a semantic/discourse category. The referent can be an entity, an abstract entity, or a proposition (a full clause). Furthermore, an entity can either be human or non-human. Examples of human, non-human, and abstract entity are given in (2), (9), and (10), respectively. (12) illustrates the proposition category.

(12) ... the product not advertised does not exist for customers, ***which*** *means it brings no profits*. [Polish; POLU-1006]

**RESTRICTIVENESS:** This feature identifies whether an RC is integrated or supplementary[12]. An integrated RC is an integral part of the referent NP that contains it. A supplementary RC, by contrast, is characterised by a weaker link to its referent or surrounding structures. In writing, the difference is often marked by putting a comma before the supplementary RCs. (13) and (14) exemplify integrated and supplementary RCs, respectively.

(13) The people ***who*** *happened to fall victim to this shameful disease* were persecuted. [Polish; POLU-1007]

(14) ... I haven't mentioned about inequality in the social life, ***which*** *is the extension of inequality in the family life*. [Turkish; TRCU-1003]

[12]The integrated-supplementary division of RCs corresponds to the distinction between restrictive and non-restrictive RCs (hence the feature name is 'restrictiveness'). For the differences between these two dichotomies, see Huddleston and Pullum (2002).

**ADDITIONAL META-FEATURES:** The essays are also marked for three additional features: `native language` (L1 background), `institution` (the source institution and also the country), and `gender` (of the writer; male or female). An example of the ICLE-RC annotation is provided in Table 11 in the Appendix.

## 5 Annotation framework for OC

In addition to RCs (and their linguistic features), the texts in the ICLE-RC are also annotated for a wide range of OCs (other constructions). OCs either resemble RCs (particularly because of the use of words such as *that* and *which*) but are not RCs proper, or they are a special type of RCs. OCs comprise six types, as defined and exemplified below.

**IT-CLEFT:** In a cleft construction, a single clause is split up into two clauses, each containing its own verb. An it-cleft construction begins with a dummy *it*, which is typically followed by a copula and an NP. The information in the *it*-clause is emphasised for the listener (foregrounded information). The clause that follows the *it*-clause is introduced by *that* (sometimes also *which* or *who*), and it contains information that is already understood (backgrounded information).

(15) It is the threat of a punishment that prevents us from committing felonies and offences. [Finnish; FIJO-1022]

**PSEUDO-CLEFT:** Pseudo-cleft constructions, like *it*-clefts, also configure themselves in terms of backgrounded and foregrounded information. Pseudo-clefts are typically introduced by *what*.

(16) What we learn in our schools today are not words of wisdom. [Swedish; SWUL-1003]

**RELATIVE-THERE:** This feature refers to existential clauses (introduced by the dummy pronoun *there*) that are followed by an RC.

(17) There are many reasons which leads to the failure of a marriage. [Urdu; PAGJ-1010]

**FUSED RELATIVE:** Fused relatives are a special type of RC in which the referent and the relativised element are fused together instead of being expressed separately as in regular RCs. Fused

relatives are introduced by a wide range of RMs (otherwise used in regular RCs), such as *who(ever)*, *what(ever)*, *which(ever)*, or *where(ever)*.

(18) A student should think and try to draw conclusions on whichever lesson he is taking. [Turkish; TRME-3001]

**SO:** This feature identifies [*so* + ADJ + (*that*)] constructions, which usually present a reason-claim relation.

(19) Nowadays we are so used to television that we find difficult to think that it did not exist before... [Italian; ITRS-1001]

**SUCH:** This feature, like the previous SO feature, identifies [*such* + ADJ + (*that/which*)] constructions, which usually present a reason-claim relation.

(20) ... it can make people dependent on it to such an extent that they finally neglect their health, family and other vital things. [Polish; POSI-1002]

## 6 Reliability of annotation

The ICLE-RC is aimed to offer gold-standard data, and is entirely created from human annotation. The possibility of pre-annotating the source texts using heuristics based on (dependency or constituency) parsing output from parsers was excluded due to their limited success on learner English data[13]. The ICLE essays typically contain grammatical errors, missing words, truncated or incomplete sentences, and non-standard usages, and our preliminary experiments based on SpaCy dependency parses were not sufficiently satisfatory.

The RCs and OCs in the ICLE-RC were annotated by two annotators (two of the authors), who have many years of experience with various kinds of linguistic annotation. On average, the annotators took between 30 minutes and one hour to annotate a single essay (including revisions). The annotators used the UAM CorpusTool (version 2.8.16) (O'Donnell, 2008) to perform the annotation. A screenshot of an RC-annotation in UAM CorpusTool is provided in Figure 1 in the Appendix.

[13]For an overview of applying (UD) parsers to learner data, see Hashemi and Hwa (2016) and Huang et al. (2018).

In order to test the reliability of the corpus, we conducted an IAA study. The annotators independently annotated all 24 texts for the Polish part of the corpus. Given our multi-layered, feature-rich annotation scheme (Table 10), we calculated agreement only for the seven broad RC features: RM, REFERENT FUNCTION, MARKER FUNCTION, EMBEDDING, EXTRAPOSITION, REFERENT TYPE, and RESTRICTIVENESS.

It was found that the two annotators individually identified 163 RCs and 157 RCs, respectively, while both identified 151 common RCs[14]. According to Cohen's kappa (Landis and Koch, 1977), agreement was almost perfect for REFERENT FUNCTION and MARKER FUNCTION (0.86, 0.80), substantial for RM and REFERENT TYPE (0.77, 0.73), and moderate for RESTRICTIVENESS (0.58), as shown in Table 2. For the remaining two features, EMBEDDING and EXTRAPOSITION, prevalence prevented the calculation of meaningful $\kappa$-values. The agreement score was 89.35% for both features.

| feature | type | $\kappa$-value |
|---|---|---|
| RM | lexical/syntactic | 0.77 |
| referent function | syntactic | 0.86 |
| marker function | | 0.80 |
| referent type | semantic/discourse | 0.73 |
| restrictiveness | syntactic/discourse | 0.58 |

Table 2: Inter-annotator agreement for five features

Importantly, the variation in agreement can be interpreted as indicative of the relative complexity of the annotation task for a target feature type. First, syntactic features (e.g., REFERENT FUNCTION, MARKER FUNCTION), in comparison to other feature types, are relatively more objective in nature. Hence, their identification is quite straightforward, which caused a very high degree of agreement. Second, the identification of RM (a lexical/syntactic feature) is quite uncomplicated when it is explicitly marked by *that* or a *wh*-word, but not necessarily the same when there is no overt RM (for bare-relatives). In our IAA study, the annotators also agreed overwhelmingly more on the presence of an RM than on their absences, which resulted in a higher degree of substantial agreement. Third, the identification of REFERENT TYPE operates on a semantic/discourse level, which brings subjectivity into analysis. This is evidenced by a lower degree

[14]The task of identifying RCs can sometimes pose considerable challenges due to the absence of an overt RM for bare-relatives, or the similarity between RCs and OCs.

of substantial agreement between the annotators. For instance, (21) presents such a case in which the referent 'a merciful God' was annotated as `entity` by the first annotator, but as `abstract-entity` by the second annotator.

(21) We treat it like a valuable gift from a merciful God ***who*** *enabled us to use our skills and abilities* ... [Polish; POSI-1002]

Finally, RESTRICTIVENESS presents an interesting case. RESTRICTIVENESS distinguishes integrated and supplementary RCs, and is determined based on syntactic cues; e.g., use of a comma for supplementary RCs, or the non-use of *that* for supplementary RCs (according to standard English grammars). RESTRICTIVENESS is also conveyed through discourse meaning, i.e., whether the RC presents an integral part of the meaning of the matrix clause, or as a separate, additional unit of information. In the ICLE(-RC), which is a corpus of L2 English student essays, the students did not seem to have strictly adhered to the standard grammatical rules for marking integrated and supplementary RCs. (22) presents such a case (an RC with *who)*, where the annotators disagreed on identifying the RESTRICTIVENESS value.

(22) ... we can point out to the case of Oscar Wilde ***who*** *was tried for being a homosexual* ... [Polish; POLU-1007]

In those circumstances, the ICLE-RC annotators had to rely only on the available discourse meaning, which invited a greater amount of subjectivity in the interpretation. The challenge of determining restrictiveness has also been addressed in the RC literature (Bache and Jakobsen, 1980; Hundt et al., 2012). Ambiguities of this kind probably caused only a moderate degree of agreement between the annotators.

## 7 (Semi-)automating annotation

In order to assess the feasibility of automating our annotation procedure, we implemented a classifier based on `distilroberta-base` (Sanh et al., 2019). We annotated markers as spans in plain text, but for classification purposes, we tokenised[15] the entire corpus and mapped the span annotations onto words, resulting in IO (inside-outside) tags. We first trained a binary classifier, predicting whether or not a word is (part of) an RM. We use the first 76 files of the corpus as training data, and the remaining 20 files as test data. This results in 52,034 words in the training split and 11,663 words in the test split, of which only 144[16] are annotated as (being part of) an RM. We are thus dealing with a heavily unbalanced data set and therefore focus on the macro-averaged scores. The results for this binary classification set-up are included in Table 3.

| | p | r | f1 | support |
|---|---|---|---|---|
| none | 0.99 | 1.00 | 1.00 | 11,519 |
| relcl | 0.83 | 0.36 | 0.50 | 144 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.91 | 0.68 | 0.75 | 11,663 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11,663 |

Table 3: Binary classification results.

The same classification set-up is used to train and predict the values on the second level of the taxonomy in Table 10. We already face a severe class imbalance in the binary case (114 words labeled as (part of a) relative clause vs. 11,519 unlabeled words) and this only increases in multi-class classification set-ups where labels are further split up into different classes. This is reflected by the macro-averaged f1-scores: 0.46, 0.17, 0.38, 0.50, 0.50, 0,49, and 0.59 for RM, REFERENT FUNCTION, MARKER FUNCTION, EMBEDDING, EXTRAPOSITION, REFERENT TYPE, and RESTRICTIVENESS, respectively. The classification reports are included in Tables 12 to 18 in the Appendix.

Based on these results, we conclude that automatically suggesting RM spans with a binary classifier, which has a comparatively high precision, would be a feasible way to semi-automate the annotation procedure. In order to automatically provide candidate labels for the more fine-grained task of feature assignment, we consider the performance too low, and perhaps more training examples can further improve performance. Alternatively, using an LLM for this task might be a feasible strategy. Generative foundation models are not necessarily designed for text span annotation tasks, but recent studies have shown promising results (Kasner et al., 2025) and we consider this an important piece of future work.

[15] Using spaCy's `en_core_web_sm` model.

[16] The test split contains 119 RMs, resulting in on average 1.2 words per marker for the test split.

## 8 First results

The essays from different L1 backgrounds in the ICLE-RC vary with respect to the number of words and sentences, as shown in Table 4. For example, on average the students with Finnish L1 produced the lengthiest essays (867.04 words per essay) while the students with Swedish L1 produced the shortest essays (664.29 words per essay)[17], although both groups produced sentences of almost equal length (about 22 words per sentence).

| language | # avg words | # avg sentences | # avg words per sentence |
|---|---|---|---|
| Finnish | 867.04 | 39.38 | 22.02 |
| Italian | 718.33 | 27.21 | 26.40 |
| Polish | 705.92 | 33.17 | 21.28 |
| Swedish | 664.29 | 29.34 | 22.61 |
| Turkish | 786.75 | 39.25 | 20.04 |
| Urdu | 711.29 | 43.29 | 16.43 |
| AVG | 742.27 | 35.27 | 21.46 |

Table 4: General statistics for essays in the corpus

Table 5 shows the distribution of RCs for different L1 backgrounds, their rate and percentage of occurrence with respect to sentences. RCs are found to be a high-frequency feature for Italian: RCs occur in every 3.23 sentences, or 30.93% of the sentences contain an RC. By contrast, RCs occur least frequently for Urdu (only in every 11.81 sentences or in 8.47% of all sentences).

| language | # RCs | # sentences | rate | % |
|---|---|---|---|---|
| Finnish | 187 | 945 | 5.05 | 19.79 |
| Italian | 202 | 653 | 3.23 | 30.93 |
| Polish | 163 | 796 | 4.88 | 20.48 |
| Swedish | 147 | 705 | 4.80 | 20.85 |
| Turkish | 137 | 942 | 6.88 | 14.54 |
| Urdu | 88 | 1039 | 11.81 | 8.47 |
| TOTAL | 924 | 5080 | 5.50 | 18.19 |

Table 5: Distribution of RCs

Similarly, Table 6 shows the distribution of OCs for different L1 backgrounds, their rate and percentage of occurrence with respect to sentences. OCs are found to be used most frequently by the Polish and Finnish students, and least frequently by the Urdu students.

An important theme of investigation in our work is whether/how different RC features (and subfeatures) vary across languages. For the purpose of illustration, we only provide the distribution of two features: RM and RESTRICTIVENESS. First,

[17]The official ICLE instructions stipulate ca. 600 words.

| language | # OCs | # sentences | rate | % |
|---|---|---|---|---|
| Finnish | 100 | 945 | 9.45 | 10.58 |
| Italian | 58 | 653 | 11.29 | 8.88 |
| Polish | 86 | 796 | 9.26 | 10.80 |
| Swedish | 56 | 705 | 12.58 | 7.94 |
| Turkish | 76 | 942 | 12.39 | 8.07 |
| Urdu | 31 | 1039 | 33.52 | 2.98 |
| TOTAL | 407 | 5080 | 12.48 | 8.01 |

Table 6: Distribution of OCs

Table 7 presents the distribution of RMs[18]. The Urdu students are found to structure RCs almost exclusively with an overt RM (*that* or a *wh*-word). By contrast, the occurrence of bare-relatives (with a `zero` marker) is found to be a highly frequent feature exploited by both the Finnish and Swedish students (about 20% of all RCs). Furthermore, the distribution of the overt RMs vary across these languages. For example, the subordinator *that* is used more frequently for Finnish, Swedish, and Turkish. By contrast, Italian, Polish, and Urdu show a more frequent use of a *wh*-word. Furthermore, the distribution of the *wh*-words shows a consistent pattern across these languages, with *which* being the most frequent *wh*-word, followed by *who* and then *where* (albeit with a larger margin). The remaining *wh*-words (*when*, *whose*, or *whom*) occur rarely in the corpus.

Next, the distribution of RCs for RESTRICTIVENESS (in Table 8) also shows variation across languages and RMs. For example, the frequency of supplementary RCs is found to be high for Italian and Polish (ca. 40%), intermediate for Finnish and Urdu (ca. 28-32%), and low for Swedish and Turkish (ca. 23%). One consistent pattern to emerge from the data, however, is that supplementary RCs are introduced by *that* by the students from all L1 backgrounds (albeit in small numbers). Such usage, strictly speaking, is not sanctioned by the (prescriptive) grammars. This might result from the insufficient learning outcomes of the L2 learners of English rather than an exposure to L1 varieties of English (both standard and non-standard), in which the co-occurrence of supplementary RCs and *that* is observed, albeit rarely (for an overview, see Hillberg, 2012).

It might be the case that (some of) these observed variations originate from the ways RCs are structured in the corresponding L1s. This can be validated by thoroughly examining the RC-related

[18]The occurrence of 5 or fewer number of tokens for a category was excluded from the table.

| RM-type | RM | Finnish | Italian | Polish | Swedish | Turkish | Urdu | Total/Avg |
|---|---|---|---|---|---|---|---|---|
| that | *that* | 52 (27.81%) | 38 (18.81%) | 19 (11.66%) | 46 (31.29%) | 43 (31.39%) | 14 (15.91%) | 212 (22.94%) |
| wh-word | *which* | 49 (26.20%) | 65 (32.18%) | 70 (42.94%) | 35 (23.81%) | 43 (31.39%) | 38 (43.18%) | 301 (32.58%) |
| | *who* | 32 (17.11%) | 49 (24.26%) | 40 (24.54%) | 24 (16.33%) | 30 (21.90%) | 23 (26.14%) | 198 (21.43%) |
| | *where* | 12 (6.42%) | 13 (6.44%) | - | 8 (5.44%) | 6 (4.38%) | 7 (7.95%) | 49 (5.30%) |
| | *when* | - | - | - | - | - | - | 13 (1.41%) |
| | *whose* | - | - | - | - | - | - | 9 (0.97%) |
| | *why* | - | - | - | - | - | - | 8 (0.87%) |
| | *whom* | - | - | - | - | - | - | - |
| | *what* | - | - | - | - | - | - | - |
| | *how* | - | - | - | - | - | - | - |
| zero | zero | 37 (19.79%) | 28 (13.86%) | 21 (12.88%) | 29 (19.73%) | 9 (6.57%) | - | 128 (13.85%) |
| TOTAL | | 187 | 202 | 163 | 147 | 137 | 88 | 924 |

Table 7: Distribution of RMs

| restrictiveness | RM | Finnish | Italian | Polish | Swedish | Turkish | Urdu | Total/Avg |
|---|---|---|---|---|---|---|---|---|
| integrated | *that* | 41 (21.93%) | 25 (12.38%) | 16 (9.82%) | 41 (27.89%) | 38 (27.74%) | 9 (10.23%) | 170 (18.40%) |
| | *wh*-word | 56 (29.95%) | 67 (33.17%) | 60 (36.81%) | 44 (29.93%) | 59 (43.07%) | 46 (52.27%) | 332 (35.93%) |
| | zero | 37 (19.79%) | 28 (13.86%) | 21 (12.88%) | 28 (19.05%) | 8 (5.84%) | 4 (4.55%) | 126 (13.61%) |
| supplementary | *that* | 11 (5.89%) | 13 (6.44%) | 3 (1.84%) | 5 (3.40%) | 5 (3.65%) | 5 (5.68%) | 42 (4.55%) |
| | *wh*-word | 42 (22.46%) | 69 (34.16%) | 63 (38.65%) | 29 (19.73%) | 27 (19.71%) | 24 (27.27%) | 254 (27.49%) |
| TOTAL | | 187 | 202 | 163 | 147 | 137 | 88 | 924 |

Table 8: Distribution of RCs for RESTRICTIVENESS

grammar of each L1, and comparing these results against those grammars to see whether any cross-linguistic factors influence the patterning of the RC features. We leave this task for the next stage in our work.

## 9 Related work

Although there are no large-scale corpora exclusively annotated for RCs, there exists a rich body of corpus-based studies on RCs in English. Weichmann (2015) provides a detailed, usage-based analysis of RCs (in 500 texts, with 80,000 parse trees) in the British component of the International Corpus of English (ICE)[19]. Biber et al.'s (1999) corpus-based account of English grammar, among many other grammatical phenomena, describes the use and distribution of RCs in a variety of registers. More commonly, specific aspects of RCs have been subject to corpus-based scrutiny, such as modified entity (Fox and Thompson, 1990), type of modification (Tse and Hyland, 2010), relativisers and their functions (Keenan and Comrie, 1977), referents of RCs (Kjellmer, 2008), (non-)humanness (Fox and Thompson, 1990), restrictiveness (Cornish, 2018), and bare-relatives (Lehmann, 2002). A significant line of research involves the analysis of RCs in historical corpora (Nevalainen and Raumolin-Brunberg, 2002; Johansson, 2006; Suárez-Gómez, 2006; Allen, 2022) and diachronic changes in the use of RCs (Leech et al., 2009; Xu and Xiao, 2015; Fajri and Okwar, 2020). Yet another important theme in RC research concerns the usage and variation of RCs in regional varieties of L1 English (Lehmann, 2002; Tagliamonte et al., 2005; Szmrecsanyi, 2013) as well as in World Englishes (Suárez-Gómez, 2015a,b). Finally, corpus-based research also explored phenomena related to RCs (OCs),

[19] https://www.ice-corpora.uzh.ch/en.html

such as pseudo-cleft (Breivik, 1999) and relative-*there* (Maschler et al., 2023).

## 10 Conclusions and outlook

The ICLE-RC is an extension of a subset of the ICLE, and it provides annotation for RCs and related phenomena, based on a comprehensive, multi-layered, feature-rich taxonomy. The first and present version of the ICLE-RC contains a collection of 924 RCs (and 407 OCs) from 144 academic essays, representing six L1 backgrounds and six corresponding language families. The annotations in stand-off XML format and the code for our classification experiments are available on GitHub[20]. The corpus is now in the post-production stage, and will soon be published as an open-access resource.

Our future work includes expanding the size and coverage of the corpus by adding more texts for the existing six languages as well as incorporating texts from other L1 backgrounds (from the ICLE), representing new (sub-)language families, such as Cantonese (Sino-Tibetan), Dutch (West Germanic), Greek (Hellenic), Japanese (Japonic), Farsi (Indo-Iranian), Russian (Slavic), and Tswana (Bantu). The extended corpus would enable us to employ statistical modeling on the data and draw reliable and comprehensive conclusions about the use of RCs by L2 English users.

We envisage that the ICLE-RC would be used as a valuable resource for research on RCs in various areas of linguistic analysis. In SLA and language typology, the corpus would help identifying varying patterns in the use of English RCs by L2 learners, and checking whether those patterns result from specific L1 backgrounds, or they, for example, conform to those stipulated by the NP accessibility hierarchy (Keenan and Comrie, 1977). The ICLE-RC can also be used to (re-)examine the properties of RCs in regional varieties of English, and validate or revise the resulting findings against the existing research in World Englishes. Furthermore, the corpus offers a rich repository of information-structuring devices (OCs, in addition to RCs), and this would aid research on discourse structure, supporting the analysis of fore-/back-grounding strategies, discourse referents, discourse segments, and discourse relations.

[20] https://anonymous.4open.science/r/law2025-relative-clause-classification-663F

## Limitations and ethical considerations

The annotators that contributed to the annotations were employed by their affiliated universities at the time of working on this project.

The classification experiments using `distilroberta-base` were done on a CPU/laptop with 32GB of RAM and in total amounted to approx. 10 hours of training and evaluating.


## Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback on the first submission version of the paper.


## References


J.C. Acuña Fariña. 2000. Reduced relatives and apposition. *Australian Journal of Linguistics*, 20(1):5–22.

C.L. Allen. 2022. Pronominally headed relative clauses in early english. *English Language and Linguistics*, 26(1):105–132.

C. Bache and L.K. Jakobsen. 1980. On the distinction between restrictive and non-restrictive relative clauses in modern english. *Lingua*, 52(3):243–267.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Limited.

S. Brandt, E. Kidd, E. Lieven, and M. Tomasello. 2009. The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics*, 20(3):539–570.

L.E. Breivik. 1999. On the pragmatic function of relative clauses and locative expressions in existential sentences in the LOB Corpus. In *Out of corpora: Studies in honour of Stig Johansson (Language and Computers: Studies in Practical Linguistics, 26)*, pages 121–135. Rodopi.

B. Comrie. 1998. Rethinking the typology of relative clauses. *Language Design*, 1:59–86.

F. Cornish. 2018. Revisiting the system of English relative clauses: structure, semantics, discourse functionality. *English Language and Linguistics*, 22:431–456.

C. Doughty. 1991. Second Language Instruction Does Make a Difference: Evidence from an Empirical Study of SL Relativization. *Studies in Second Language Acquisition*, 13(4):431–469.

M.S.A. Fajri and V. Okwar. 2020. Exploring a Diachronic Change in the Use of English Relative Clauses: A Corpus-Based Study and Its Implication for Pedagogy. *SAGE Open*, 10(4).

B.A. Fox and S.A. Thompson. 1990. A Discourse Explanation of the Grammar of Relative Clauses in English Conversation. *Language*, 66(2):297–316.

M. Gamon, E. Ringger, Z. Zhang, R. Moore, and S. Corston-Oliver. 2002. Extraposition: a case study in German sentence realization. In *Proceedings of COLING 2002*.

H. Goad, N.B. Guzzo, and L. White. 2021. Parsing Ambiguous Relative Clauses in L2 English: Learner Sensitivity to Prosodic Cues. *Studies in Second Language Acquisition*, 43(1):83–108.

S. Granger. 1998. The computer learner corpus: A versatile new source of data for SLA research. In S. Granger, editor, *Learner English on Computer*, pages 3–18. Addison Wesley Longman, London & New York.

S. Granger, M. Dupont, F. Meunier, H. Naets, and M. Paquot. 2020. The International Corpus of Learner English. Version 3.

A. Grosu. 2012. Towards a More Articulated Typology of Internally Headed Relative Constructions: The Semantics Connection. *Language and Linguistics Compass*, 6(7):447–476.

H.B. Hashemi and R. Hwa. 2016. An Evaluation of Parser Robustness for Ungrammatical Sentences. In *Proceedings of the 2016 EMNLP*.

S. Hillberg. 2012. Relativiser that in Scottish English news writing. In *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*.

Y. Huang, A. Murakami, T. Alexopoulou, and A. Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.

R. Huddleston and G.K. Pullum. 2002. *The Cambridge grammar of the English language*. CUP, Cambridge, UK.

M. Hundt, D. Denison, and G. Schneider. 2012. Relative complexity in scientific discourse. *English language and linguistics*, 16(2):209–240.

S. Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge.

C. Johansson. 2006. Relativizers in nineteenth-century English. In *Nineteenth-century English: Stability and change (Studies in English Language)*, page 136–182. Cambridge University Press.

Z. Kasner, V. Zouhar, P. Schmidtová, I. Kartáč, K. Onderková, O. Plátek, D. Gkatzia, S. Mahamood, O. Dušek, and S. Balloccu. 2025. Large language models as span annotators. *Preprint*, arXiv:2504.08697.

E. Keenan and B. Comrie. 1977. Noun Phrase Accessibility Hierarchy and Universal Grammar. *Linguistic Inquiry*, 8:63–99.

G. Kjellmer. 2008. "Troublesome Relatives": On *Whose Her* and Others. *English Studies*, 89(4):482–494.

R. Landis and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

G. Leech, M. Hundt, C. Mair, and N.I. Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge University Press.

H.M. Lehmann. 2002. Zero subject relative constructions in American and British English. In *New Frontiers of Corpus Research*, page 163–177. Rodopi.

R. Levy, E. Fedorenko, M. Breen, and E. Gibson. 2012. The processing of extraposed structures in English. *Cognition*, 122(1):12–36.

Y. Maschler, J. Lindström, and E. De Stefani. 2023. Pseudo-clefts: An interactional analysis across languages. *Lingua*, 291:103538.

G. McKoon and R. Ratcliff. 2003. Meaning Through Syntax: Language Comprehension and the Reduced Relative Clause Construction. *Psychological review*, 110(3):490–525.

T. Nevalainen and H. Raumolin-Brunberg. 2002. The rise of relative who in early Modern English. In *Relativisation on the North Sea Littoral*, page 109–121. Lincom Europa.

M. O'Donnell. 2008. The UAM CorpusTool: Software for corpus annotation and exploration. In Carmen M.. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*, pages 1433–1447. Almería, Universidad de Almería.

A. Pereltsvaig. 2023. *Languages of the World: An Introduction*, 4th edition. Cambridge University Press.

F. Reali and M.H. Christiansen. 2007. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 53:1–23.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

C. Suárez-Gómez. 2006. *Relativization in Early English (950–1050): The Position of Relative Clauses*. Peter Lang.

C. Suárez-Gómez. 2015a. Relative clauses in Asian Englishes. *Journal of English Linguistics*, 42(2):245–268.

C. Suárez-Gómez. 2015b. The places where English is spoken: adverbial relative clauses in World Englishes. *World Englishes*, 34(4):620–635.

B. Szmrecsanyi. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press.

S. Tagliamonte, J. Smith, and H. Lawrence. 2005. No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change*, 17:75–112.

P. Tse and K. Hyland. 2010. Claiming a territory: Relative clauses in journal descriptions. *Journal of Pragmatics*, 42(7):1880–1889.

D. Weichmann. 2015. *Understanding relative clauses: A usage-based view on the processing of complex constructions*. De Gruyter Mouton.

X. Xu and R.Z. Xiao. 2015. Recent changes in relative clauses in spoken British English. *English Studies*, 96(7):1–21.

# A Appendix

| language | institution | gender | # essays |
|---|---|---|---|
| Finnish (Uralic) | University of Helsinki | F | 4 |
| | | M | 4 |
| | University of Joensuu (now UEF) | F | 4 |
| | | M | 4 |
| | University of Jyväskylä | F | 4 |
| | | M | 4 |
| Italian (Romance) | University of Bergamo | F | 6 |
| | | M | 2 |
| | Sapienza University of Rome | F | 4 |
| | | M | 4 |
| | University of Turin | F | 4 |
| | | M | 4 |
| Polish (Slavic) | Maria Curie-Skłodowska University | F | 8 |
| | | M | 0 |
| | Adam Mickiewicz University | F | 4 |
| | | M | 4 |
| | University of Silesia in Katowice | F | 8 |
| | | M | 0 |
| Swedish (Germanic) | University of Gothenburg | F | 4 |
| | | M | 4 |
| | Lund University | F | 4 |
| | | M | 4 |
| | Växjö University | F | 6 |
| | | M | 2 |
| Turkish (Turkic) | Mersin University | F | 4 |
| | | M | 8 |
| | University of Mustafa Kemal | F | 2 |
| | | M | 2 |
| | University of Çukurova | F | 8 |
| | | M | 0 |
| Urdu (Indo-Aryan) | GC University Faisalabad | F | 4 |
| | | M | 8 |
| | Govt College for Women Jhang | F | 2 |
| | | M | 2 |
| | Lahore College for women university | F | 8 |
| | | M | 0 |
| TOTAL | | | 144 |

Table 9: Distribution of the essays in the ICLE-RC

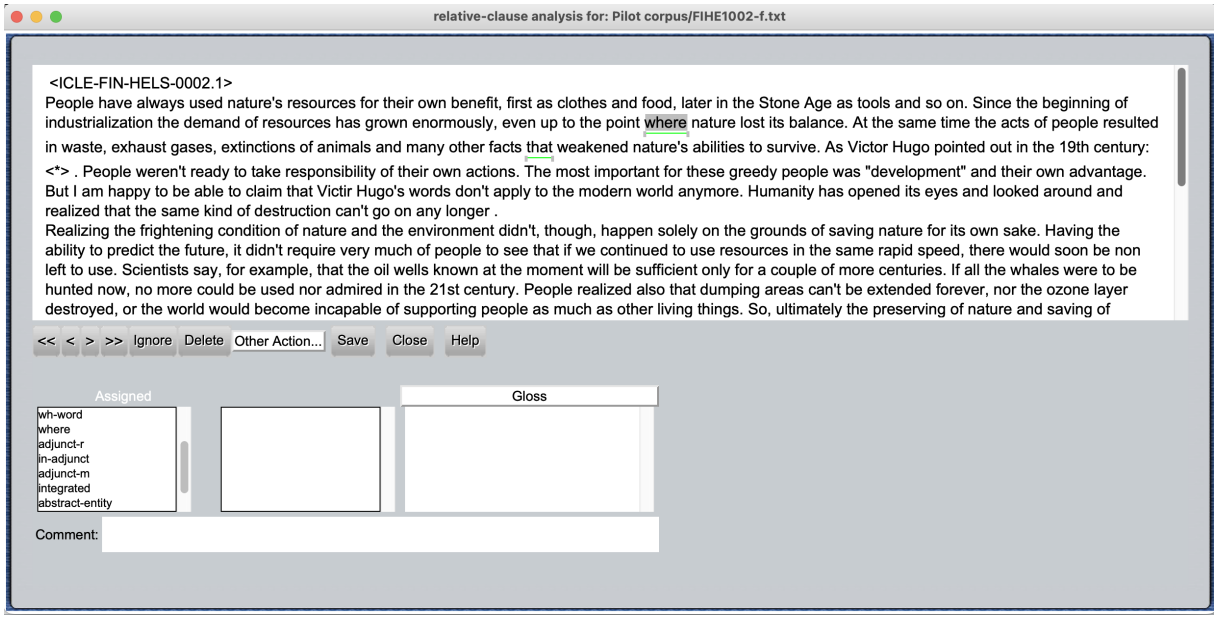


Figure 1: RC annotation in UAM CorpusTool

| RC annotation feature | | | |
|---|---|---|---|
| **level 1** | **level 2** | **level 3** | **level 4** |
| RM | that | | |
| | wh-word | *which*, *who*, *whose*, etc. | |
| | zero | | |
| referent function | subject | subj-head-n | |
| | | in-subj-comp | |
| | | in-subj-adjunct | |
| | direct obj | dir-obj-head-n | |
| | | in-dir-obj-comp | |
| | | in-dir-obj-adjunct | |
| | indirect obj | indir-obj-head-n | |
| | | in-indir-obj-comp | |
| | | in-indir-obj-adjunct | |
| | predicative complement | pred-comp-np | pred-comp-head-n |
| | | | in-pred-comp-np-comp |
| | | | in-pred-comp-np-adjunct |
| | | pred-comp-adjp | pred-comp-head-adj |
| | | | in-pred-comp-adjp-comp |
| | | | in-pred-comp-adjp-adjunct |
| | | pred-comp-pp | pred-comp-head-p |
| | | | in-pred-comp-pp-comp |
| | adjunct | adjunct | |
| | | in-adjunct | |
| | clause | | |
| marker function | subject | | |
| | direct obj | | |
| | Indirect obj | | |
| | predicative complement | pred-comp-full | |
| | | in-pred-comp | |
| | gen-subj-det | | |
| | predicate | | |
| | aux-comp | | |
| | head-to-inf-vp | | |
| | adjunct | | |
| embedding | yes | | |
| | no | | |
| extraposition | yes | | |
| | no | | |
| ref type | entity | human | |
| | | non-human | |
| | abstract | | |
| | proposition | | |
| restrictiveness | integrated | | |
| | supplementary | | |

Table 10: Taxonomy of features for RC annotation

| The sentence in which the RC features are to be annotated:<br>Unfortunately, life is not a situation comedy ***where*** *every problem is happily solved.* [Italian; ITTO-1002] | | |
|---|---|---|
| | | |
| meta-features | L1 | Italian |
| | institution | University of Turin |
| | gender | female |
| | | |
| RC features | RM | wh-word → *where* |
| | referent function | pred-comp → pred-comp-np → pred-comp-head-n |
| | marker function | adjunct |
| | embedding | no |
| | extraposition | no |
| | referent type | abstract entity |
| | restrictiveness | integrated |

Table 11: Example of RC annotation

| | p | r | f1 | support |
|---|---|---|---|---|
| none | 0.99 | 1.00 | 1.00 | 11,519 |
| that | 0.00 | 0.00 | 0.00 | 37 |
| wh-word | 0.83 | 0.90 | 0.86 | 58 |
| zero | 0.00 | 0.00 | 0.00 | 49 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.45 | 0.47 | 0.46 | 11,663 |
| weighted avg | 0.98 | 0.99 | 0.99 | 11,663 |

Table 12: Relative marker type classification results.

| | p | r | f1 | support |
|---|---|---|---|---|
| adjunct-r | 0.00 | 0.00 | 0.00 | 46 |
| clause-r | 0.00 | 0.00 | 0.00 | 2 |
| direct-obj-r | 0.23 | 0.16 | 0.19 | 51 |
| indirect-obj-r | 0.00 | 0.00 | 0.00 | 5 |
| none | 0.99 | 1.00 | 0.99 | 11,519 |
| pred-comp-r | 0.00 | 0.00 | 0.00 | 11 |
| subj-r | 0.00 | 0.00 | 0.00 | 29 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.17 | 0.17 | 0.17 | 11,663 |
| weighted avg | 0.98 | 0.99 | 0.99 | 11,663 |

Table 13: Referent function classification results.

| | p | r | f1 | support |
|---|---|---|---|---|
| adjunct-m | 0.50 | 0.18 | 0.27 | 22 |
| direct-obj-m | 0.00 | 0.00 | 0.00 | 47 |
| none | 0.99 | 1.00 | 0.99 | 11,519 |
| pred-comp-m | 0.00 | 0.00 | 0.00 | 3 |
| subject-m | 0.73 | 0.56 | 0.63 | 72 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.44 | 0.35 | 0.38 | 11,663 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11,663 |

Table 14: Marker function classification results.

| | p | r | f1 | support |
|---|---|---|---|---|
| embed-no | 0.81 | 0.38 | 0.52 | 137 |
| embed-yes | 0.00 | 0.00 | 0.00 | 7 |
| none | 0.99 | 1.00 | 0.99 | 11,519 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.60 | 0.46 | 0.50 | 11,663 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11,663 |

Table 15: Embedding classification results.

| | p | r | f1 | support |
|---|---|---|---|---|
| extrapose-no | 0.81 | 0.36 | 0.50 | 142 |
| extrapose-yes | 0.00 | 0.00 | 0.00 | 2 |
| none | 0.99 | 1.00 | 0.99 | 11,519 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.60 | 0.45 | 0.50 | 11,663 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11,663 |

Table 16: Extraposition classification results.

| | p | r | f1 | support |
|---|---|---|---|---|
| abstract-entity | 0.63 | 0.2 | 0.34 | 95 |
| entity | 0.82 | 0.49 | 0.61 | 47 |
| none | 0.99 | 1.00 | 0.99 | 11,519 |
| proposition | 0.00 | 0.00 | 0.00 | 2 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.61 | 0.43 | 0.49 | 11,663 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11,663 |

Table 17: Referent type classification results.

| | p | r | f1 | support |
|---|---|---|---|---|
| integrated | 0.62 | 0.18 | 0.28 | 118 |
| none | 0.99 | 1.00 | 1.00 | 11,519 |
| supplementary | 0.48 | 0.54 | 0.51 | 26 |
| accuracy | | | 0.99 | 11,663 |
| macro avg | 0.70 | 0.57 | 0.59 | 11,663 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11,663 |

Table 18: Restrictiveness classification results.