Syntax of referents of relative markers: Evidence from a corpus of learner English

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

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

Abstract

We investigate the referents of relative markers of English relative clauses, focusing on their syntactic role in the matrix clauses. The referents, unlike relative markers and related features, have compratively remained understudied. We examine the syntactic environments of the referents as part of a larger project, which develops the ICLE-RC, a corpus of learner English texts annotated for relative clauses and related phenomena (it-/pseudo-clefts, existential-relatives, etc.). The corpus derives from the International Corpus of Learner English (ICLE; Granger et al., 2020), and contains 144 academic essays, representing six L1 backgrounds – Finnish, Italian, Polish, Swedish, Turkish, and Urdu. We annotate those texts for over 900 relative clauses (and over 400 related phenomena), with respect to a wide array of lexical, syntactic, semantic, and discourse features. Results from our analysis show that the relativisation of referents varies according to their syntactic functions. The referents are also observed to interact with other RC-features, yielding systematic variations across different L1 backgrounds, (some of) which can potentially be attributed to the typological properties of the associated L1.

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 constitute a rich body of research, addressing themes such as syntactic and typological variation (Comrie, 1998; Grosu, 2012), semantic features (Cornish, 2018), discourse functions (Brandt et al., 2009), FLA/SLA (Diessel and Tomasello, 2005; Doughty, 1991), parsing (Goad et al., 2021), processing (Reali and Christiansen, 2007), historical usage (Suárez-Gómez, 2006), diachronic development (Leech et al., 2009; Fajri and Okwar, 2020), corpus-based analysis (Biber et al., 1999; Weichmann, 2015), and World Englishes (Suárez-Gómez, 2015). Despite the depth and breadth of previous research, the scope of these studies have largely remained confined to the analysis of RCs alone and associated features found therein.

We strive to extend the scope of RC analysis, by examining the larger syntactic environment in which RCs occur. In particular, we investigate the referents of relative markers of English RCs, focusing on their syntactic role in their respective matrix clauses. We examine RC-referents as part of a larger project, the ICLE-RC, which builds a corpus of learner English annotated for RCs and related phenomena (it-/pseudo-clefts, existential-relatives, etc.). The corpus builds on a subset of the International Corpus of Learner English (ICLE; Granger et al., 2020), and contains 144 academic essays, representing six L1 backgrounds – Finnish, Italian, Polish, Swedish, Turkish, and Urdu. In this paper, we present our multi-layered, feature-rich annotation framework for RC(-referent) analysis, and report on our corpus analysis of RC-referents and their interaction with other RC-features.

This paper is structured as follows: We outline the previous work on RC-referents in Section 2. Section 3 introduces our large-scale corpus project, and describes the annotation schemes. We present the general results and those for referent functions in Section 4 and Section 5, respectively. Section 6 discusses the findings, and Section 7 concludes the paper, outlining some future research directions.

2 Previous work

One of the most influential work on RCs (and RC-referents) is offered by Keenan and Comrie's

[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.*

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 97–110

(1977) NP accessibility hierarchy (NPAH):

(1) subject > direct object > indirect object > oblique > genitive > comparative

NPAH stipulates that languages that relativise on one position on the hierarchy will also relativise on the positions above it. According to this scale, the subjects of the matrix clause are most prone to be relativised, followed by the direct objects, which is then to be followed by the indirect objects, and so on. The validity of NPAH is supported by numerous studies on RCs across languages, rendering it one of the few putative typological universals[6].

Besides NPAH, there exist some studies that considered RC-referents an important RC-feature. For example, Fox and Thompson (1990) investigated the syntactic and discourse properties of the head NPs in the matrix clause and their interaction with RCs in conversations. They observed that the structuring of RCs is crucially shaped by the formulation of the referents according to many interactive and cognitive factors of the communicative situations.

Tagliamonte et al. (2005) examined relative markers in vernacular varieties of British English, and observed the prevalence of *that* and zero marker, instead of *wh*-forms. The authors identified the type of the referent (e.g., definite or indefinite NPs) as one of the determining factors behind the marker preference. More particularly, indefinite referents (along with sentence structure) entailed the use of the zero-variant in RCs.

Hinrichs et al. (2015) investigated the changing trends in the use of restrictive relativisers, examining the shift from *which* to *that* in written standard English. The authors conducted a multivariate analysis on a large collection of RCs (16K+) from the Brown corpus[7], and used a number of independent variables which included, among other features, a set of referent-features, such as the POS, number, length, and definiteness of the referent. The study concluded that the shift (*which* → *that*) took place largely under the influence of American English and was regulated by various prescriptivism-related factors.

[6]Nevertheless, counter-evidence to NPAH (e.g., the uniformity of the subject-object asymmetry) has been provided by some later studies. For an overview, see Kidd (2011).

[7]https://varieng.helsinki.fi/CoRD/corpora/BROWN/

3 The ICLE-RC project

We have developed the ICLE-RC to investigate RCs and related phenomena (it-/pseudo-clefts, existential-relatives, etc.) in learner English. The corpus builds on a subset of the International Corpus of Learner English (ICLE; Granger et al., 2020). The ICLE is a corpus of academic essays written by undergraduate students from a set list of topics[8]. These students are intermediate or advanced learners of English, coming from different L1 backgrounds. The first version of the ICLE-RC contains 144 ICLE texts (100K+ words), covering six L1 backgrounds – Finnish, Italian, Polish, Swedish, Turkish, and Urdu – with 24 texts from each[9]. 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)[10].

The ICLE-RC is designed to serve a number of purposes. First, the corpus provides 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 ICLE-RC covers six L1 backgrounds representing six different language families (Pereltsvaig, 2023) – Finnish: Uralic; Italian: Romance; Polish: Slavic; Swedish: Germanic; Turkish: Turkic; and Urdu: Indo-Aryan[11]. This would allow identifying typological patterns for certain RC features as well as highlighting those which potentially result 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

[8]Some of the ICLE essay topics are as follows: (1) *The prison system is outdated.*, (2) *No civilised society should punish its criminals: it should rehabilitate them.*, (3) *Feminists have done more harm to the cause of women than good.* For specimen essays, check out the ICLE500 dataset.

[9]The detailed distribution of the essays in the ICLE-RC is provided in Table 11 in the Appendix.

[10]OCs either resemble RCs (particularly because of the use of words such as *that*, *which*, or *who*) but are not RCs proper, or they are a special type of RCs. OCs comprise four major types, as follows:

it-cleft: *It was only last year that he got his tenure.*
pseudo-cleft: *What I need is a long cool drink.*
relative-there: *There was one man that kept interrupting.*
fused-relatives: *The dog ate what I had left on my plate.*

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

as those of native academic English such as LOCNESS (Granger, 1998)). Finally, the corpus would help us explore English learners' use of OCs as alternative strategies of information structuring, in addition to RCs.

3.1 Main annotation framework

In the ICLE-RC, we have annotated the RCs[12] for a wide range of lexical, syntactic, semantic, and discourse features, as listed in Table 1. The complete taxonomy of the annotation features is provided in Table 12 in the Appendix.

Here, we first outline the main annotation features, except the grammatical functions of the referent (REFERENT FUNCTION), which is described in greater detail in the next sub-section[13].

RELATIVE MARKER (RM): RMs include the subordinator *that* and *wh*-words (e.g., *which*, *who*, *whose*) that introduce an RC. An additional feature *zero* is recognised to mark the absence of an overt RM for bare-relatives. These categories are exemplified below[14].

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

(3) 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]

(4) **The status** ø *English has acquired today* is so dominant that it seems unlikely that ... [Finnish; FIJO-1003]

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 types (for more information about all categories and sub-categories, see Table 12).

`subject`: The relativised item functions as the subject in the RC, as in (5).

(5) 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 (6).

(6) ... 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, as in (7).

(7) ... 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]

REFERENT TYPE: 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 (3), (7), and (6), respectively. (8) illustrates the proposition category.

(8) ... 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[15]. An integrated RC is an integral part of the referent NP that contains it, as in (9). A supplementary RC, by contrast, is characterised by a weaker link to its referent or surrounding structures, as in (10).

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

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

[13] We exclude from the description two main RC-features, EMBEDDING and EXTRAPOSITION, as they are not central to the RC-referent analysis (and also because of the space constraint). For the same reason, we also do not include the annotation framework for OCs. For the detailed annotation guidelines, visit the project website.

[14] **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).

[15] 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).

#	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 RC annotation

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

3.2 The referent function sub-scheme

The REFERENT FUNCTION feature identifies the grammatical function of the referent of the RM in the matrix clause. It includes seven broad categories and fifteen specific sub-categories, as shown in Table 2[16]. These sub-categories are described below.

category	sub-category
subject	subj-head-n
	in-subj-comp
	in-sub-adjunct
direct object	dir-obj-head-n
	in-dir-obj-comp
	in-dir-obj-adjunct
indirect object	indir-obj-head-n
	in-indir-obj-comp
	in-indir-obj-adjunct
predicative complement	pred-comp-np
	pred-comp-adj
	pred-comp-pp
adjunct	adjunct
	in-adjunct
clause	clause

Table 2: REFERENT FUNCTION sub-scheme

subj-head-n: The head noun of the subject NP of the matrix clause is the referent. (If there is any complement and/or adjunct within that NP, the whole NP is considered as the referent.)

(11) **The third type of advertisement** ø *I do not like* is concerned to the tobacco business. [Italian; ITBO-1001]

in-subj-comp: An NP which is part of a complement within the subject NP is the referent.

(12) A secret to a slim figure, ***which*** *is a dream of many*, surely does not lie in fast food. [Polish; POLU-1008]

in-subj-adjunct: (An NP which is part of) an adjunct within the subject NP is the referent.

(13) All the informations are [sic], even the minor ones ***that*** *are seen unimportant*, are the chains of each other. [Italian; TRME-3006]

dir-obj-head-n: The head noun of the direct object NP in the matrix clause is the referent.

(14) We must look into ourselves and forget all the boring scientific theories ***which*** *have taken hold of our sense of reality* ... [Swedish; SWUL-1005]

in-dir-obj-comp: An NP which is part of a complement in the direct object NP is the referent.

(15) 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 in the direct object NP is the referent.

(16) 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]

indir-obj-head-n: The head noun of the indirect object NP in the matrix clause is the referent.

[16]Each feature under predicative complement is divided into further sub-types. For the complete annotation scheme, see Table 12 in the Appendix.

(17) If only done properly, mining and timbering... bring lots of revenue to **the state** ø *they live in*. [Swedish; SWUL-1006]

`in-indir-obj-comp`: An NP which is part of a complement within the indirect object NP is the referent.

(18) Thomas Sternes Eliot published 'The Waste Land' in 1922 and owes its final shape to the collaboration of Ezra Pound ***who*** *actually corrected it* ... [Italian; ITRS-1030]

`in-indir-obj-adjunct`: (An NP which is part of) an adjunct within the indirect object NP is the referent.

(19) John sent his letter to the professor of history with 100 publications, *some of* ***which*** *are quite remarkable*. [our example][17]

`pred-comp-np`: The referent is (part of) an NP that serves as the predicative complement in the matrix clause.

(20) Unfortunately, life is not a situation comedy ***where*** *every problem is happily solved*. [Italian; ITTO-1002]

`pred-comp-adjp`: The referent is (part of) an AdjP that serves as the predicative complement in the matrix clause.

(21) The world is full of ambitious and resolute persons ***who*** *are at the some time reliable and sensitive*. [Polish; POLU-1003]

`pred-comp-pp`: The referent is (part of) an PP that serves as the predicative complement in the matrix clause.

(22) It is like a chain process *in* ***which*** *better cures are required* ... [Polish; POSI-1004]

`adjunct`: The referent is an adjunct phrase in the matrix clause.

(23) Nobody is happy in a dictatorship ***where*** *violence and hypocrisy reigns* [sic]. [Swedish; SWUV-3003]

`in-adjunct`: An NP that is part of an adjunct in the matrix clause is the referent.

[17]No token for this category was found in our corpus.

(24) In a family, ***which*** *is made up by four people*, there are at least two cars. [Italian; ITBO-2001]

`clause`: The whole matrix clause is the referent.

(25) In some countries homosexual marriages have been recently legalised, ***which*** *of course gave rise to many protests*. [Polish; POLU-1007]

An example of the ICLE-RC annotation is provided in Table 13 in the Appendix.

4 General results

The purpose of developing the ICLE-RC is to offer gold-standard data, and hence, the corpus is entirely created from human annotation. The RCs and OCs in the ICLE-RC were annotated by two annotators (the authors), who have many years of experience with various kinds of linguistic annotation. The annotation was performed using the UAM CorpusTool (version 2.8.16) (O'Donnell, 2008), and is saved in a stand-off XML format.[18]

The reliability of the annotation was tested through an IAA study. The two annotators independently annotated all 24 texts for the Polish part of the corpus. Given our multi-layered, feature-rich annotation scheme (Table 12), we calculated agreement only for the seven broad RC features: RM, REFERENT FUNCTION, MARKER FUNCTION, EMBEDDING, EXTRAPOSITION, REFERENT TYPE, and RESTRICTIVENESS. 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)[19]. For the remaining two features, EMBEDDING and EXTRAPOSITION, prevalence prevented the calculation of meaningful κ-values. The agreement score was 89.35% for both features.[20]

The essays from different L1 backgrounds in the ICLE-RC vary with respect to the number of

[18]For more information about the prospects of pre-annotating the ICLE-RC for syntactic (dependency) parses and the feasibility of (semi-)automating the RC annotation, see Das et al. (to appear).

[19]Previous research (Bache and Jakobsen, 1980; Hundt et al., 2012) also addressed the challenge of determining restrictiveness.

[20]For a detailed discussion about the reliability of the ICLE-RC annotation, see Das et al. (to appear).

words and sentences, as shown in Table 3. 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)[21], although both groups produced sentences of almost equal length (about 22 words per sentence).

L1	# 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 3: General statistics for essays in the corpus

Table 4 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).

L1	# 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 4: Distribution of RCs

5 Results for referent functions

We begin with presenting the distribution of referent functions in the corpus, as shown in Table 5[22]. Overall, direct objects in the matrix clauses are found to be relativised most frequently in the RCs (32.25%, in the rightmost column), followed by adjuncts, predicative complements, and subjects. By contrast, the least frequently relativised items are (matrix) clauses and indirect objects.

The pattern, however, does not apply strictly on individual L1 backgrounds. For example, for Polish the pattern is less strongly pronounced (with the scores for the categories being close to each other), or for Swedish, adjuncts (instead of direct objects) in the matrix clauses are relativised most often, or for Urdu, subjects and predicative complements score higher than adjuncts.

Next, we examine the co-occurrence of referent functions and other RC features. First, Table 6 presents the distribution (in percentages) of the RM types for referent functions[23]. Overall, across all L1s *wh*-words (e.g., *which*, *who*, *whose*) constitute the most common RM type in the RCs, regardless of the referent functions. The students with L1 Urdu are found to use *wh*-words almost exclusively for RMs. This partially holds for Italian (only with the `subj` feature) and Polish (only with the `pred-comp` feature). Turkish almost never uses `zero` (bare relatives).

Second, Table 7 shows the co-occurrence of referent functions and marker functions. First of all, for all L1s the relativised items most often serve as the subject of the RCs, regardless of the referent functions. For specific L1s, some patterns are observed:

1. `subj` $\sim$ `subj`: When the referent is the subject in the matrix clause, the RM also tends to be the subject of the RC. This applies almost exclusively for Swedish, Turkish, and Urdu.
2. `dir-obj` $\sim$ `dir-obj`: When the referent is the direct object, the RM serves more often as a direct object (after the subject).
3. `pred-comp/adjunct` $\sim$ `dir-obj`: When the referent is a predicative complement or an adjunct, the RM is more often as an adjunct rather than a direct object (after the subject).

Third, we examine the co-occurrence of referent functions and referent types (e.g, human, abstract entity) in Table 8[24]. Overall, for all L1s the most common referent type is `abstract entity`, irrespective of the referent functions. However, the difference between the preference for `human` and `abstract entity` is less clear when the referent serves as the subject in the matrix clause. In fact, `human` outscores `abstract entity` in such a configuration for Polish, Turkish, and Urdu. By contrast, `non-human` (concrete) entities are rarely relativised in the RCs.

Finally, the co-occurrence of referent functions and restrictiveness in Table 9 shows that L2 English

[21]The official ICLE data collection instructions stipulate ca. 600 words per essay.

[22]The occurrence of fewer than 5 tokens for a category was excluded from all the tables.

[23]The `indir-obj` and `clause` features are excluded from the tables due to low frequency.

[24]The `proposition` feature was excluded from the table due to its low frequency.

type	Finnish	Italian	Polish	Swedish	Turkish	Urdu	avg
subj	32 (17.11%)	34 (16.83%)	34 (20.86%)	25 (17.01%)	22 (16.06%)	21 (23.86%)	168 (18.18%)
dir-obj	61 (32.62%)	69 (34.16%)	41 (25.15%)	49 (33.33%)	50 (36.50%)	28 (31.82%)	298 (32.25%)
indir-obj	-	-	-	-	-	-	14 (1.52%)
pred-comp	43 (22.99%)	38 (18.81%)	31 (19.02%)	15 (10.20%)	25 (18.25%)	18 (20.45%)	170 (18.40%)
adjunct	42 (22.46%)	57 (28.22%)	36 (22.09%)	51 (34.69%)	29 (21.17%)	13 (14.77%)	228 (24.68%)
clause	7 (3.74%)	-	17 (10.43%)	-	9 (6.57%)	7 (7.95%)	46 (4.98%)
TOTAL	187	202	163	147	137	88	924

Table 5: Distribution of referent functions

type	RM	Finnish	Italian	Polish	Swedish	Turkish	Urdu	avg
subj	*that*	18.75	-	17.65	24.00	27.27	-	18.45
	wh-word	56.25	82.35	73.53	52.00	63.54	80.95	68.45
	zero	25.00	-	-	24.00	-	-	13.10
dir-obj	*that*	36.07	24.64	17.07	34.69	36.00	-	28.19
	wh-word	49.18	56.52	68.29	44.90	58.00	82.14	57.38
	zero	14.75	18.84	14.63	20.41	-	-	14.43
pred-comp	*that*	30.23	18.42	-	60.00	28.00	-	25.29
	wh-word	41.86	71.05	77.42	33.33	60.00	77.78	60.59
	zero	27.91	-	-	-	-	-	14.12
adjunct	*that*	26.19	17.54	-	25.49	31.03	-	20.61
	wh-word	54.76	66.67	72.22	54.90	65.51	76.92	63.16
	zero	19.05	15.79	22.22	19.61	-	-	16.23

Table 6: Co-occurrence of RMs and referent functions

type	m-function	Finnish	Italian	Polish	Swedish	Turkish	Urdu	avg
subj	subj	68.75	79.41	82.35	68.00	77.27	90.48	77.38
	dir-obj	18.75	-	14.71	-	-	-	12.50
	adjunct	-	14.71	-	-	-	-	8.93
dir-obj	subj	57.38	55.07	60.98	59.18	64.00	64.29	59.40
	dir-obj	26.23	28.99	17.07	26.53	22.00	17.86	24.16
	adjunct	14.75	15.94	19.51	12.24	10.00	17.86	14.77
pred-comp	subj	51.16	63.16	54.84	66.67	52.00	66.67	57.65
	dir-obj	30.23	13.16	19.35	-	16.00	-	18.82
	adjunct	18.60	18.42	25.81	-	32.00	33.33	22.35
adjunct	subj	57.14	59.65	50.00	52.94	65.52	76.92	57.89
	dir-obj	-	17.54	22.22	17.65	20.69	-	15.79
	adjunct	35.71	15.79	22.22	25.49	-	-	22.81

Table 7: Co-occurrence of marker functions and referent functions

users, regardless of their L1s, use integrated RCs more often than supplementary RCs. The pattern is more strongly pronounced for Finnish, Swedish, Turkish, and Urdu when the referent is the subject. This also holds true for Swedish, Turkish, and Urdu, when the referent is the predicative complement.

6 Discussion

In the ICLE-RC, the students (advanced L2 learners of English) were found to relativise all major constituents in the matrix clause in the RCs, but with varying degrees: direct objects > adjunct > predicative complement / subject > (matrix) clause > indirect object (in Table 5). This order is, however, not corroborated by NPAH (Keenan and Com-

type	ref-type	Finnish	Italian	Polish	Swedish	Turkish	Urdu	avg
subj	human	37.50	44.12	52.94	32.00	54.55	61.90	45.43
	non-human	-	-	-	-	-	-	2.98
	abstract	56.25	50.00	41.12	68.00	45.45	38.10	50.00
dir-obj	human	19.67	15.94	-	18.37	26.00	-	17.45
	non-human	-	14.49	24.39	18.37	-	-	10.74
	abstract	78.69	65.22	65.85	63.27	74.00	82.14	70.81
pred-comp	human	13.95	21.05	16.13	33.33	20.00	-	18.24
	non-human	-	13.16	-	-	-	-	7.65
	abstract	79.07	65.79	77.42	53.33	76.00	83.33	73.53
adjunct	human	14.29	28.07	33.33	15.69	20.69	38.46	23.25
	non-human	-	-	-	17.65	-	-	7.46
	abstract	80.95	66.67	61.11	66.67	72.41	53.85	68.42

Table 8: Co-occurrence of referent types and referent functions

type	restrictiveness	Finnish	Italian	Polish	Swedish	Turkish	Urdu	avg
subj	integrated	87.50	58.82	82.35	88.00	100.00	85.71	82.14
	supplementary	-	41.18	17.65	-	-	-	17.86
dir-obj	integrated	63.93	71.01	60.98	77.55	78.00	64.29	69.80
	supplementary	36.07	28.99	39.02	22.45	22.00	35.71	30.20
pred-comp	integrated	81.40	47.37	61.29	80.00	88.00	72.22	70.00
	supplementary	18.60	52.63	38.71	-	-	-	30.00
adjunct	integrated	73.81	57.89	61.11	74.51	68.97	69.23	67.11
	supplementary	26.19	42.11	38.89	25.49	31.03	-	32.89

Table 9: Co-occurrence of integrated/supplementary RCs and referent functions

rie, 1977). We find this quite intriguing, and call for a closer scrutiny of a larger variety of learner English. We also observed deviations from this pattern for Swedish (adjunct > direct object > subject > predicative complement) and Urdu (direct object > subject > predicative complement > adjunct). This raises an important question: Do these specific orders for constituents originate from the ways RCs are structured in those languages, or do they show influence of prior (institutionalised) learning? Unfortunately, as studies on referent functions are not abundant, we cannot directly compare our results to previous research, and we thus leave these inquiries for further investigation.

Some other patterns, however, can potentially be explained with reference to the ways RCs function in different L1s. For example, the students with L1 Urdu overwhelmingly used an overt RM, particularly *wh*-words (in Table 6). A scrutiny of the Urdu grammar reveals that (finite) RCs in Urdu, like in many other Indo-Aryan languages, are introduced with a correlative construction: a demonstrative pronoun + a relative pronoun (Srivastav, 1991; Bhatt, 2003). This is illustrated by the *vo-jo* pair in (26), taken from Butt et al. (2007, p.113).

(26) ***vo*** *laṛki* ***[jo***
that.Dem. girl.F.Sg.Nom which.Rel.
k^haṛi *hɛ]* *lʌmbi*
stand-Perf.F.Sg. be.Pres.3.Sg tall.F.Sg.
hɛ
be.Pres.3.Sg.

'The girl who is standing is tall.'

This explicit (double-)marking of RCs might have a direct influence on the Urdu students for not preferring the use of bare-relatives in English.

The same reasoning apparently fails to apply to Turkish, however. Turkish does not employ an overt *wh*-element or complementiser to introduce RCs; rather, RCs are marked morphologically by certain particles (suffixes), as shown in (27), taken from Kornfilt (1997, p.29).[25]

[25] In fact, it has traditionally been argued that Turkish lack genuine RCs, and have only deverbal adjectives: *a running child* instead of *a child who is running* (Kornfilt, 2000, p.123).

(27) *[geçen* *yaz* *ada-da* *ben-i*
last summer island-Loc. I.Acc.
*gör-**en**]* *kişi-ler*
see.Part. person.Pl.

'The people who saw me on the island last summer'

Like Urdu, had we assumed an L1 effect of Turkish on the structuring of English RCs, we would have expected that the Turkish students would use mostly bare-relatives in English. Yet, we find counter-evidence in our data: The Turkish students (like Urdu students) have almost always used an overt RM for RCs in English. Slobin (1986) argues that Turkish RCs are not readily isolable since they are synthetic and even noncanonical to a Turkish clause. Furthermore, the processing of RCs in Turkish necessitates the use of more demanding strategies by children acquiring the language. By contrast, English RCs are analytic and canonical to an English clause. Based on this, we speculate that the Turkish students, when producing RCs in English, might have resorted to using the more distinguishable, canonical English RC structures involving the use of an overt RM. Alternatively, it might also be the case that since Turkish RCs are always marked, albeit by a particle, the Turkish students chose to always mark the English RCs by an overt RM rather than leave them unmarked (i.e., use bare-relatives). In any of these cases (and beyond), we believe that these conflicting results have important implications for research on the competing roles between L1 influence and the efficacy and success of L2 instructions, and require further exploration.

Next, the distribution of the marker functions (in Table 7) shows a clear ordering: subject > direct object / adjunct. This is validated by previous research (e.g., Tavakoli, 2013). Furthermore, the co-occurrence of referent functions and marker functions, however, shows some interesting patterns (see previous section). These patterns may well be determined based on the product of the relative complexity of each of the functions (Hundt et al., 2012), the distance between referent-heads and RM (Tagliamonte et al., 2005), or the level of RC-embedding (Karlsson, 2007). This, we feel, falls beyond the scope of the present study, and we intend to investigate it further in our future work.

For referent types, Fox and Thompson (1990) found that non-human subject heads in the matrix clause tend to co-occur with the objects in the RCs, and also that non-human object heads in the matrix clause do not tend to co-occur with the objects in the RC[26]. This is partially corroborated by our data, as we found evidence only for the second claim, but counter-evidence for the first one. The distribution of the relevant categories is provided in Table 10.

Finally, the prevalence of integrated RCs in the ICLE-RC indicates that the RCs are used more often as an integral part of the referent NP rather than providing additional information or commentary about it. This implies that L2 English learners use RCs more as a syntactic device than a discourse one (i.e., RCs as discourse segments).

7 Conclusions and outlook

RC-referents in the ICLE-RC show variation for their syntactic functions across different L1 backgrounds. The variation seems even greater and multifarious when their co-occurrence with other RC-features is taken into account. In our future work, we would conduct a thorough examination of the RC-related grammar of each L1, and test our findings against them to see whether any cross-linguistic factors influence the patterning of referent functions in the English RCs.

The ICLE-RC is now in the post-production stage, and will soon be published as an open-access resource. Our future work would include expanding the size and coverage of the corpus by adding more texts for the existing six L1s as well as incorporating texts from other L1 backgrounds (from the ICLE), representing new (sub-)language families; e.g., Cantonese (Sino-Tibetan), Dutch (West Germanic), Greek (Hellenic), Japanese (Japonic), Farsi (Indo-Iranian), Russian (Slavic), Tswana (Bantu). This would facilitate large-scale studies on referent functions and many other RC-related phenomena.

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.

[26]This study is, however, not directly comparable to ours as it examined the use of RCs in (non-learner) conversations.

ref-type	r-function	m-function	#
non-human and abstract entity	subj	subj	53 (8.15%)
		dir-obj	20 (3.08%)
	dir-obj	subj	126 (19.39%)
		dir-obj	69 (10.62%)
TOTAL			650

Table 10: Co-occurrence of non-human, referent functions and marker functions

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.

R. Bhatt. 2003. Locality in correlatives. *Natural Language & Linguistic Theory*, 21:485–541.

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.

M. Butt, T.H. King, and S. Roth. 2007. Urdu Correlatives: Theoretical and Implementational Issues. In *Proceedings of LFG '07 Conference*, pages 107–127, Stanford, California. CSLI Publications.

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.

D. Das, I. Czerniak, and P. Bourgonje. to appear. ICLE-RC: International Corpus of Learner English for Relative Clauses. In *Proceedings of the 19th Linguistic Annotation Workshop*.

H. Diessel and M. Tomasello. 2005. A New Look at the Acquisition of Relative Clauses. *Natural Language & Linguistic Theory*, 81(4):882–906.

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.

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.

L. Hinrichs, B. Szmrecsanyi, and A. Bohmann. 2015. "WHICH"-HUNTING AND THE STANDARD ENGLISH RELATIVE CLAUSE. *Language*, 91(4):806–836.

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.

F. Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43(2):365–392.

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

E. Kidd. 2011. *The Acquisition of Relative Clauses: Processing, typology and function*. Benjamins, Amsterdam.

J. Kornfilt. 1997. On the Syntax and Morphology of Relative Clauses in Turkish. *Dilbilim Araştırmaları Dergisi*, 8:24–51.

J. Kornfilt. 2000. Some syntactic and morphological properties of relative clauses in Turkish. In *The Syntax of Relative Clauses*, pages 121–159. John Benjamins, Amsterdam/Philadelphia.

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.

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

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.

D.I. Slobin. 1986. The acquisition and use of Relative Clauses in Turkic and Indo-European Languages. In *Studies in Turkish Linguistics*, page 277–298. John Benjamins, Amsterdam.

V. Srivastav. 1991. The Syntax and Semantics of Correlatives. *Natural Language and Linguistic Theory*, 9(4):637–686.

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

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

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.

H. Tavakoli. 2013. *A dictionary of language acquisition: A comprehensive overview of key terms in first and second language acquisition*. Rahnama, Tehran.

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

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 11: Distribution of the essays in the ICLE-RC

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 12: Taxonomy of features for RC annotation

The sentence in which the RC features are to be annotated: 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 13: Example of RC annotation