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

Annotation Manual

Debopam Das and Izabela Czerniak
Åbo Akademi University
debopam.das@abo.fi, izabela.czerniak@abo.fi

Version 1.1 (July 2025)

1 Introduction

The ICLE-RC is a corpus of learner English texts annotated for relative clauses and related phenomena. Relative clauses (henceforth RCs) are a type of subordinate clauses that typically modify nouns or noun phrases, as in (1).

(1) [The man] who came to dinner turned out to be from my hometown.

RCs also modify adjectives, adverbs, PPs, VPs, and even entire clauses, as respectively illustrated below.

(2) Pat is [beautiful], which, however, many consider her not.

(3) He moved [abroad] where he found a good job.

(4) He found a body [under the bridge] where nothing grows.

(5) She told me to [design it myself], which I simply can't.

(6) [Alex bought a mansion], which made him bankrupt.

Phenomena related to RCs include 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 RCs.

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 over 900 RCs, with respect to a wide array of lexical, syntactic, semantic, and discourse features. These texts are also annotated for about 400 related phenomena, which we call *other constructions* (henceforth OCs).

This annotation manual is structured as follows: In Section 2, we outline the motivation for the development of the ICLE-RC. Section 3 describes data selection and the setup of the corpus. The annotation frameworks for RCs and OCs are detailed in Section 4 and Section 5, respectively. We include additional notes for annotation in Section 6. Finally, Section 7 provides an example of the ICLE-RC annotation.

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[1]. 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 RC-related phenomena 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. 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[2]. 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.

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[3]), which are further balanced for the gender of the writer[4], whenever possible. The distribution of the essays in the ICLE-RC is provided in Table 1.

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

[2]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.*

[3]The only exception was made for Urdu (see Table 1). Only four essays were available from Govt College for Women Jhang. In order to balance that out, a total of 12 essays were compiled from GC University Faisalabad.

[4]The classification follows from the ICLE.

L1	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 1: Distribution of the essays in the ICLE-RC

#	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 2: Primary categories of relative clause annotation

4 Annotation framework for RC

The relative clauses (RCs)[5] in the ICLE-RC are annotated for seven broad features (as listed in Table 2), representing a wide range of lexical, syntactic, semantic, and discourse characteristics of RCs. The complete taxonomy of the annotation features is provided in Table 3. These features and sub-features are defined and exemplified, as follows.

4.1 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[6].

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

(8) 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]

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

4.2 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.

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

[6] **Conventions for RC 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).

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 3: Taxonomy of features for RC annotation

4.2.1 Subject

The referent is the subject in the matrix clause. This includes three sub-features, which are defined and exemplified below.

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

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

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

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

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

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

4.2.2 Direct Object

The referent is the direct object in the matrix clause. This includes three sub-features, which are defined and exemplified below.

4.2.2.1 direct-object-head-n: The head noun of the direct object NP is the referent. (If there is any complement, modifier, or adjunct within that NP, the whole NP is considered as the referent.)

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

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

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

4.2.2.3 in-dir-obj-adjunct: An (NP which is part of an) adjunct within the direct object NP is the referent.

(15) 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]

4.2.3 Indirect Object

The referent is the indirect object in the matrix clause. This includes three sub-features, which are defined and exemplified below.

4.2.3.1 indir-obj-head-n: The head noun of the indirect object NP is the referent. (If there is any complement, modifier, or adjunct within that NP, the whole NP is considered as the referent.)

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

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

(17) 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]

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

(18) John sent his application to the professor of history with 100 publications, *some of* ***which*** *are quite remarkable.* [our example][7]

4.2.4 Predicative Complement

The referent is the predicate complement in the matrix clause. This includes four sub-features, which are explained below.

4.2.4.1 pred-comp-np: The referent is an NP functioning as the predicative complement in the matrix clause. This further includes three sub-features, which are defined and exemplified below.

`pred-comp-head-n:` The head noun of the predicative complement NP is the referent. (If there is any complement, modifier, or adjunct within that NP, the whole NP is considered as the referent.)

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

`in-pred-comp-np-comp:` An NP which is part of a complement within the predicative complement NP is the referent.

(20) It is the story of a seventeen-year-old boy ***who*** *has been rejected by his parents ...* [Italian; ITTO-1012]

`in-pred-comp-np-adjunct:` (An NP which is part of) an adjunct within the predicative complement NP is the referent.

(21) Professor Burns is a scholar of history with many publications, *many of* ***which*** *are published by Blackwell* [our example]

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

4.2.4.2 pred-comp-adjp The referent is an AdjP functioning as the predicative complement in the matrix clause. This further includes three sub-features, which are defined and exemplified below.

`pred-comp-head-adj:` The head adjective of the predicative complement AdjP is the referent.

(22) Pat is beautiful, ***which****, however, many consider her not.* [our example, repeated from (2)]

`in-pred-comp-adj-comp:` An NP which is part of a complement within the predicative complement AdjP is the referent.

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

`in-pred-comp-adj-adjunct:` (An NP which is part of) an adjunct within the predicative complement AdjP is the referent.

(24) John is fond of the new teacher with cool eyeglasses, ***which*** *are now sold on Amazon.* [our example]

4.2.4.3 pred-comp-pp: This applies to a PP that functions as the predicative complement in the matrix clause. This includes two sub-features, which are defined and exemplified below.

`pred-comp-head-n:` The head preposition of the PP is the referent.

(25) The spare chairs are downstairs, ***where*** *the children fear to go.* [our example][8]

`in-pred-comp-pp-comp:` An NP which is part of a complement within the predicative complement PP is the referent.

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

4.2.5 Adjunct

The referent is an adjunct phrase in the matrix clause. This includes two sub-types.

`adjunct:` The whole adjunct phrase serves as the referent.

(27) 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 is the referent.

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

4.2.6 Clause

The whole matrix clause is the referent of the RM.

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

[8] Words such as *downstairs*, *abroad*, and *outside* occurring after a *be* verb are considered prepositions rather than adverbs. For more information, see Huddleston et al. (2021, p.182-186).

4.3 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`.

4.3.1 Subject

The relativised item functions as the subject in the RC.

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

4.3.2 Direct Object

The relativised item functions as the direct object of the main verb in the RC.

(31) The biggest crime ***that*** *a society could commit to itself* would be to forgive its criminals. [Polish; POPZ-1004]

4.3.3 Indirect Object

The relativised item functions as the indirect object of the main verb in the RC.

(32) The internal-combustion engine is **a very useful invention** ∅ *we owe a lot to* ... [Italian; ITBO-2003]

4.3.4 Predicative Complement

This feature has two sub-types.

4.3.4.1 pred-comp-full: The relativised item represents the whole predicative complement in the RC.

(33) ...they should be able to co-exist in a highly civilized society, ***which*** *we like to think our own is.* [Swedish; SWUG-2007]

4.3.4.2 in-pred-comp: The relativised item represents a part of the predicative complement in the RC.

(34) There is not a single sound reason to worry that the advancing industrialization wreaks havoc on **the formidable might** ∅ *the human mind is naturally endowed with.* [Polish; POPZ-1001]

4.3.5 Genitive Subject Determiner

The relativised item (*whose*) is the genitive determiner in the subject NP of the RC.

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

4.3.6 Predicate

The VP-complement of the *do*-support verb is relativised.

(36) They advised me to call the police, ***which*** *I did [do] immediately.* [our example]

4.4 Complement of Auxiliary Verb

The VP complement of the main auxiliary verb is relativised.

(37) He told me to design it myself, ***which*** *I simply can't.* [our example]

4.4.1 Head of a *to*-infinitival VP

The VP complement of a TP (headed by inf-*to*) is relativsed.

(38) He has asked me to go with him, ***which*** *I'd much like to.* [our example]

4.4.2 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*.

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

4.5 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*).

(40) The emphasis should be put on integration, since all cultures must be considered equal, and they should be able to co-exist in a highly civilized society, ***which*** *[we like to think] our own is.* [Swedish; SWUG-2007][9]

4.6 Extraposition

Extraposition occurs when an RM does not immediately follow its referent. Instead, there are some intervening elements between the RM and its referent.

(41) 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]

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

4.7 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, abstract entity, and proposition are listed below respectively.

(42) 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]

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

(44) The emphasis should be put on integration, since all cultures must be considered equal, and they should be able to co-exist in a highly civilized society, ***which*** *we like to think our own is.* [Swedish; SWUG-2007]

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

4.8 Restrictiveness

This feature identifies whether an RC is integrated or supplementary[10]. 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. The examples of integrated and supplementary RCs are listed below respectively.

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

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

4.9 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).

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 four main types, as defined and exemplified below.

[10]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	sub-feature	example
RM	that	...it is the actions of the people living now ***that*** *will decide the future for the earth.* [Swedish; SWUG-2006]
	wh-word	...it is our spouses ***who*** *know us best.* [Polish; POPZ-1001]
	zero	It is not only **nature** ø environmental groups *defend...* [Swedish; SWUL-1006]
grammatical category of foregrounded element	NP	It is the lust of money ***that*** *is creating a great destruction in our society.* [Urdu; PALW-1005]
	PP	It may be partly for this reason ***that*** *it has gained comparatively little international importance.* [Finnish; FIJO-1002]
	AdjP	It wasn't **green** ø *I told you to paint it.* [Huddleston et al. (2021, p.1419)]
	AdvP	...it is not often ***that*** *people encounter environmental destruction themselves.* [Swedish; SWUG-2006]
	clause	It is only after stars fall ***that*** *new nations are born.* [Finnish; FIHE-1011]
marker function	subject	...it is the teacher ***who*** *speaks and who is the most active in a class.* [Turkish; TRKE-2014]
	direct object	...it is something ***that*** *you look forward to, dream about, and plan for.* [Swedish; SWUV-3005]
	indirect object	It is Alex ***whom*** *I sent this letter.* [our example]
	gen subj det	It is Pat ***whose*** *cat has gone missing.* [our example]
	adjunct	It is for this reason ***that*** *BICE...has arranged a Conference...* [Italian; ITTO-1004]
ref type	entity human	It is the house wife ***who*** *has to maintain all the household affairs...* [Urdu; PAGF-1017]
	entity non-human	It is a smart phone ***which*** *I bought last month.* [our example]
	abstract entity	It is the threat of a punishment ***that*** *prevents us from committing felonies and offences.* [Finnish; FIJO-1022]
	proposition	It is only after stars fall ***that*** *new nations are born.* [Finnish; FIHE-1011]

Table 4: Annotation scheme for it-clefts

5.1 *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).

(48) It is the threat of a punishment ***that*** *prevents us from committing felonies and offences.* [Finnish; FIJO-1022][11]

The annotation scheme for *it*-clefts is provided in Table 4.

5.2 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*.

[11] **Conventions for it-cleft examples:** The RC is in italics; the RM is in bold; the foregrounded element is underlined. In case of RM-zero, there is no overt RM, and the foregrounded element is marked in bold instead. The text inside the square brackets lists the L1 background and the file number of the source text.

feature	sub-feature	example
grammatical category of foregrounded element	NP	*What we learn in our schools today* are not words of wisdom. [Swedish; SWUL-1003]
	AdjP	*What her father is*, if you want my view, is arrogant, dogmatic, and pig-headed. [Huddleston et al. (2021, p.1422)]
	clause	*What all idealistic points of view have in common* is that it is not feasible to put them into practice. [Polish; POPZ-1004]

Table 5: Annotation scheme for pseudo-clefts

(49) *What we learn in our schools today* are not words of wisdom. [Swedish; SWUL-1003][12]

The annotation scheme for pseudo-clefts is provided in Table 5.

5.3 Relative-*there*

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

(50) There are <u>many reasons</u> ***which*** *leads [sic] to the failure of a marriage.* [Urdu; PAGJ-1010][13]

The annotation scheme for relative-*there* instances is provided in Table 6.

5.4 Fused Relative

Fused relatives are a special type of RCs 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)*.

(51) A student should think and try to draw conclusions on ***<u>whichever</u>*** *lesson he is taking.* [Turkish; TRME-3001][14]

The annotation scheme for fused relatives is provided in Table 7.

6 Additional notes on annotation

In addition to the main guidelines (as described above), a few extra rules are followed for RC and OC annotation. They are listed below:

[12] **Conventions for pseudo-cleft examples:** The backgrounded element is in italics. The text inside the square brackets lists the L1 background and the file number of the source text.

[13] **Conventions for relative-there 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.

[14] **Conventions for fused relative examples:** The RC is in italics; the RM is in bold and underlined. The text inside the square brackets lists the L1 background and the file number of the source text.

feature	sub-feature	example
RM	that	Is there a solution for a way of punishment ***that*** *would serve all purposes and be equal to all human beings*? [Finnish; FIJO-1015]
	wh-word	There are many reasons *which leads [sic] to the failure of a marriage.* [Urdu; PAGJ-1010]
	zero	There are **plenty of treatments** ∅ *we still do not have, for example, for HIV and cancer.* [Polish; POLU-1009]
grammatical category of referent	head-n	...there are advertisements ***which*** *have positive effects.* [Polish; POLU-1006]
		In universities, there are many social activities for students ***that*** *they can enjoy in their free time.* [Turkish; TRMW-3006] (Note: If there is any complement and/or adjunct within that NP, the whole NP is considered as the referent.)
	comp-n	There are medicines for many illnesses ***which*** *used to be fatal...* [Polish; POSI-1004]
	adjunct-n	Down the street, there is a house with a dense garden, ***which*** *grows beautiful flowers throughout the year.* [our example]
marker function	subject	There are programmes ***that*** *should be forbidden because you could lose your intelligence watching them.* [Italian; ITRS-1002]
	direct object	...there are **things** ∅ *I do not like about magazines...* [Italian; ITBO-1001]
	indirect object	There are many people in the town *who John owes money to.* [our example]
	gen subj det	There are also many people ***whose*** *way of life is bound up with the tropical rain forests.* [Swedish; SWUV-3002]
	adjunct	...there are special places in Poznan ***where*** *waste products are purchased.* [Polish; POPZ-1005]
restrictiveness	integrated	...now there are not many students ***that*** *see the university as the home of teaching science and education.* [Turkish; TRME-3005]
	supplementary	...in each culture there are some criteria of beauty *according to* ***which*** *a man must be tall and muscular, while a woman must be a fragile blonde.* [Polish; POLU-1006]
ref type	entity human	There are also many people ***whose*** *way of life is bound up with the tropical rain forests.* [Swedish; SWUV-3002]
	entity non-human	...there are special places in Poznan ***where*** *waste products are purchased.* [Polish; POPZ-1005]
	abstract entity	There is one period of our history ***that*** *wasn't revealed to us until now.* [Finnish; FIJY-1006]

Table 6: Annotation scheme for relative-*there*

feature	sub-feature	example
RM	wh-word	A person should be able to distinguish ***what*** *is true and what is fiction...* [Italian; ITRS-1002]
referent function	subject	*...**whoever** will think about the twentieth century* will consider the television one of its mains features and elements. [Italian; ITRS-1002]
	direct object	Life would have to slip back to ***where*** *it came from – the ocean.* [SWUL-1008]
	indirect object	...the media attracts our attention to ***what*** *is happening ground us.* [Turkish; TRKE-2001]
	pred comp	Today the signs of Zodiac are ***what*** *an ordinary person attributes astrology with.* [Polish; POLU-1003]
	adjunct	This will only happen ***when*** *the students will be well aware of the practical implementation of their knowledge besides the theoretical significance.* [Urdu; PAGF-1012]
marker function	subject	...the chief...has the right to punish ***whoever*** *breaks a rule obeyed by the rest of his subjects.* [Finnish; FIJO-1022]
	direct object	We are not allowed to do ***what*** *we want.* [Finnish; FIJY-1006]
	indirect object	Today the signs of Zodiac are ***what*** *an ordinary person attributes astrology with.* [Polish; POLU-1003]
	pred comp	Without the development of science, technology and industrialization this world would be radically different from ***what*** *it is now.* [Finnish; FIJO-1029]
	adjunct	Most of the people commit suicide ***when*** *they are hopeless and desperate.* [Turkish; TRCU-1006]
ref type	entity human	...the chief...has the right to punish ***whoever*** *breaks a rule obeyed by the rest of his subjects.* [Finnish; FIJO-1022]
	entity non-human	Media shouldn't interfere...***what*** *they wear...* [Turkish; TRKE-2012]
	abstract entity	...any nutritionist will tell you that developing healthy eating habits ***when*** *you are young* is the best investment you can make for the future. [Polish; POLU-1008]

Table 7: Annotation scheme for fused relatives

6.1 Consideration of immediate syntactic context

The scope of RC annotation is confined only to the RC and its referent. The referent is strictly analysed in reference to the matrix clause that includes it and directly hosts the RC (and not any higher matrix clauses, even if available). Prototypically, the matrix clause is a full, finite, main clause. This is illustrated by (52), where the referent *the idea of a universe* is part of the main matrix clause (marked by square brackets) that hosts the RC.

(52) [Einstein has destroyed the idea of a universe] ***that*** *could be understood by using empirical fact as a measuring stick.* [Finnish; FIHE-1018]

In numerous cases, however, the matrix clause is a non-finite, subordinate clause. This is shown by (53), where the referent *one single nation* is part of the clause/VP headed by the verb *forming*, which is thus considered for annotation. The higher matrix clause "*When you really try to think of them...*" here falls outside the scope of annotation.

(53) When you really try to think of them [forming one single nation,] ***which*** *would eventually have even its own defence and foreign policy*, you can only come to one conclusion. [Finnish; FIHE-1005]

Still, on a few occasions the referent (+ the RC) functions as an apposition to a noun (or an NP) that precedes it. In such cases, the referent is not considered an integral part of the previous clause, but as a free-standing adjunct hosting the RC. This is illustrated by (54), where *a story* is the referent as well as the host of the following RC.

(54) The primitive people usually explained the phenomena they encountered as the handiwork of some supernatural being, or made up a story about the phenomenon's beginning, [a story] ***that*** *made sense to their idea of reality.* [Finnish; FIHE-1018]

6.2 Treatment of phrasal and prepositional verbs

Phrasal verbs (e.g., *carry out*, *find out*, *put up with*) and prepositional verbs (e.g., *look at*, *talk about*, *listen to*) are treated in the way postulated by Biber et al. (2021). These verbs are considered single semantic units, which can be replaced by a single verb: *carry out* → *perform/undertake*; *deal with* → *handle*; *look at* → *observe*; *find out* → *discover*; *talk about* → *discuss*.[15] Consequently, the annotation value of the relevant RC feature (`referent function` or `marker function`) is determined with reference to that replacement verb. For example, in (55) the phrasal verb *set up* can be replaced by its semantic equivalent *establish*, which is a transitive verb, and hence, the `referent function` is `direct object`.

(55) France has even set up [= establish] a law ***which*** *states that whenever a French word exists it must be used instead of an English word or expression.* [Finnish; FIJO-1003]

(56) I believe in [= value] justice and safety ***which*** *should be ensured by the law.* [Polish; POSI-1006]

(57) Travelling is another thing ***that*** *many people dream about* [= desire/wish]. [Swedish; SWUV-3005]

[15] Since the ICLE-RC comprises L2 data and represents different L1 backgrounds, the *meaning* of a word is taken to supersede its *form*, which might have different manifestations in different L1s.

(58) Very often it is understood as God's favour and protection which means it is **something holy** ø *we have to care about* [= mind] *and make use of* [= utilise], otherwise we will get punished. [Polish; POSI-1002]

6.3 Treatment of *have*

In the ICLE-RC, the verb *have*, when used as a lexical verb (and not as a non-modal auxiliary verb) is treated as a transitive verb. Consequently, the annotation value of the relevant RC feature (`referent function` or `marker function`) is determined to be `direct object`.

(59) The human creature has an abstract part ***which*** *complements the rational part* [Finnish; FIJO-1030]

(60) The characters of the programme may be **the only friends** ø *a lonely person has.* [Finnish; FIJY-1004]

For the non-modal auxiliary *have*, the annotation values are determined based on the main lexical verb which follows *have* in the construction.

(61) ...we have entered the era of 'designer babies', ***where*** *future parents 'order' a child with certain features.* [POLU-1001]

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

6.4 Dealing with language issues in L2 data

The ICLE(-RC) essays were written by L2 students of English, and they typically contain language errors and non-standard usages. The following strategies are followed to deal with those issues.

6.5 Missing words and extra words

In the event the absence of a word is clearly identified, the missing word is assumed in the construction for parsing and interpretation (but, no changes are made in the text itself).

(63) ...children are born well-structured to fit in **every situation** ø *they [+ find] around them* [Italian; ITRS-1004]

Grammatically, when there is an extra, redundant word, the construction is parsed and interpreted assuming the word is absent (but, no changes are made in the text itself).

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

6.6 Truncated or incomplete sentences

The ICLE(-RC) essays sometimes contain truncated or incomplete (ungrammatical) sentences, some of which are apparently used for stylistic purposes, such as (65) and (66), while the others seem to be clear construction errors, such as (67) and (68).

(65) How we should deal with immigrants is always a topic of discussion. Especially now ***when*** *the immigration is high.* [Swedish; SWUL-1007]

(66) Everyone strives to find himself a place in society, to get an education and make lots of money, for example as an engineer. An engineer ***who*** *spends his whole life trying to figure out new ways to produce new meaningless, mass-produced products that people don't need...* [Swedish; SWUL-1005]

(67) Could I live without television? Could the world live without television? **Two good questions** ø *I could hardly answer.* [Italian; ITRS-2001]

(68) ...the most important issue is terrorism. ***Which*** *has a negative impect [sic] on our economical, social and political system.* [Urdu; PAGF-1005]

Annotation of such instances is done on a case by case basis. For example, for (65), (66), and (67) the referent is not considered an integral part of the previous sentence, but as a free-standing adjunct hosting the RC. By contrast, the RC in (68) is considered to modify the word *terrorism* in the previous sentence, and hence, an integral part of it (albeit separate in terms of punctuation).

6.7 Minor grammatical errors, spelling mistakes, and typos

Minor grammatical errors (e.g., subject-verb agreement), spelling mistakes, and typos are disregarded (i.e., the correct forms are assumed for parsing and interpretation, but no changes are made in the text itself).

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

(70) At that time people distrust science but in general *what caused some changes* in [→ is] the tranquillity [→ tranquility] of their life. [Italian; ITB0-3002]

Sometimes, misspellings (or different writing/formatting conventions) result in different forms for an RM. Those forms are annotated as variant tokens of the same RM.

which → *Which*, *whch*, *Wwhich*, *whih*
whatever → *what ever*

7 Example of ICLE-RC annotation

An example of the RC annotation is provided in Table 8.

References

J.C. Acuña Fariña. Reduced relatives and apposition. *Australian Journal of Linguistics*, 20(1): 5–22, 2000. doi: M10.1080/07268600050003337.

D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. *Longman Grammar of Spoken and Written English.* John Benjamins, Amsterdam / Philadelphia, 2021.

S. Granger. 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, 1998.

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 8: Example of RC annotation

S. Granger, M. Dupont, F. Meunier, H. Naets, and M. Paquot. The International Corpus of Learner English. Version 3, 2020. URL `Mhttps://www.uclouvain.be/en/research-institutes/ilc/cecl/icle`.

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

R. Huddleston, G.K. Pullum, and B. Reynolds. *A Student's Introduction to English Grammar*. Cambridge University Press, 2 edition, 2021.

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

G. McKoon and R. Ratcliff. Meaning Through Syntax: Language Comprehension and the Reduced Relative Clause Construction. *Psychological review*, 110(3):490–525, 2003. doi: M10.1017/S027 2263120000285.

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