INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Language, Typology and Universals

Prof Dipti Misra Sharma

Academic Year 2020 - 2021

Relative Clause Construction

Author
Tanishq Goel
Chayan Kochar

IIIT HYDERABAD
LTU

Abstract

The aim of this project is to do a comparative study between three different types of languages for a specific type of construction. This type of comparative research helps us to capture different features across the languages of the world. The three languages taken into consideration are English, Hindi, and Telugu. English being one of the most general languages all over the world provides us with rich insights into the language. The rich amount of training dataset for English is another plus factor. Hindi belongs to the class of Indo-Aryan Language and Telugu belongs to the typological class of Dravidian Language. This paper comments on the typological classification based on research of Relative Clause Construction. We have tried to come up with implicational rules regarding the same. We have analyzed multiple patterns occurring in the different types of languages and tried to formulate out some interesting observations.

Note: Both the author are from the Northern side of India and hence are not well versed with the Dravadian Languages. Research data and manual annotation is done with help of batch mates and friends.

Chapter 1

Relative Clause

1.1 Definition

A relative clause is a grammatical structure that is “embedded” somewhere inside a sentence. The relative clause cannot stand on its own. Instead, it is contained by another sentence constituent, usually a noun phrase. Like all clauses, a relative clause must have at least a subject and a verb and may have an object and other grammatical phrases as well.

In the following sentence the underlined relative clause has a subject (which), a verb (has), and an object (a solid state image sensor).

A camera **which has a solid-state image sensor** is a digital camera.

The relative clause is contained inside the main clause portion, **A camera... is a digital camera.** Specifically, it is contained inside the noun phrase **a camera** to form a larger noun phrase, **a camera which has a solid-state image sensor.**

1.2 Relativization

The relative clause and its head noun form the relative construction. Languages use different strategies to encode the relative construction; we will refer to these as relativizing strategies.

There are different perspectives from which relativizing strategies can be studied. Thus, from the point of view of the linear order of the head noun and the relative clause, we can distinguish prenominal, postnominal.

Reason for taking postnominal and prenominal into account:
Consider the intuition / universal/ tendency reagrding the ease of relativization based on hierarchy of :

”subject > direct object > indirect object > possessor”

On the one hand - one could simply say that the universal is a tendency, rather than an absolute : the number of exceptions is small relative to the over-all sample, moreover the fact that most of the exceptions belong to a single genetic and areal grouping serves only to accentuate their exceptional nature. The alternative would be to try and reformulate the universal, effectively weakening it, so that the counterexamples are no longer counterexamples; this is the strategy adopted by Keenan and Comrie in the work cited. They argue that, if one distinguishes different strategies of forming relative clauses, in particular if one distinguishes between prenominal and postnominal.

For the purposes of the present study, we classify our sample languages according to the mechanisms by which the language in question expresses the syntactic-semantic role of the head noun in the relative clause, whereby we consider only formally expressed morphosyntactic means. In (1), for instance, the head noun serves as subject of the relative clause, and this is marked in English by use of the nominative relative pronoun who.

Languages may employ different morphosyntactic means, that is, different relativizing strategies, for different syntactic-semantic roles of the head noun. In the English sentence in (1), the head noun has the subject role, and it is relativized by means of a relative pronoun. If the same head noun, the girl, has the role of the object, one of the ways in which it may be relativized is by not using any morphosyntactic element at all, i.e. by means of a “gap”. Noun reduction and pronoun retention are the rest two strategies.

Chapter 2

Execution

2.1 Dataset

For English and Hindi, we gathered our dataset from `https://www.cfilt.iitb.ac.in/~parallelcorp/iitb_en_hi_parallel/`
From the above link we procured the en-hi parallel dev and test datasets. We then did some preprocessing to shortlist the sentences as discussed later and then final ran our code on those sentences.

For Telugu, we gathered dataset from `https://github.com/joshua-decoder/indian-parallel-cor`
`blob/master/te-en/` We felt the data was not upto the mark, hence we instead converted our english sentences into telugu and worked on them.

Note: For Telugu, as described later, we had to work manually, hence we could observe quite lesser sentences.

2.2 Text-Preprocessing

From the given dataset, we shortlisted those sentences which contained relative clauses. We did using the dependency tags by stanza and checking if the tag 'acl:relcl' occurs.
This generated duplicate sentences, if a sentence had more than one instance of relative clause. Hence we again worked on that and finally produced the final shortlisted data containing.

2.3 Implementation

To Annotate:

```
def get_ann_tags(doc, n):
    """Get POS-tagged tokens in the format of a list of
    (token, POStag) pairs for all sentences in doc.
    Returns upos (Universal part-of-speech) tag only, not
    xpos (treebank-specific part of speech)"""

    def getann(i):
        tokens = []
        for token in doc.sentences[i].words:
            tokens.append((token.id, token.text, token.lemma, token.upos, token.xp
        return tokens

    return [getann(i) for i in range(n)]
```

Eg: Sentence 8: Former UP Minister, Rajaram Pandey, who was known for his controversial speeches, passed away late Thursday evening following a heart attack.

```
(1, ’Former’, ’former’, ’ADJ’, ’JJ’, ’Degree=Pos’, 3, ’amod’)
(2, ’UP’, ’UP’, ’PROPN’, ’NNP’, ’Number=Sing’, 3, ’compound’)
(3, ’Minister’, ’Minister’, ’PROPN’, ’NNP’, ’Number=Sing’, 16, ’nsubj’)
(4, ’,’, ’,’, ’PUNCT’, ’,’, None, 3, ’punct’)
(5, ’Rajaram’, ’Rajaram’, ’PROPN’, ’NNP’, ’Number=Sing’, 3, ’appos’)
(6, ’Pandey’, ’Pandey’, ’PROPN’, ’NNP’, ’Number=Sing’, 5, ’flat’)
(7, ’,’, ’,’, ’PUNCT’, ’,’, None, 3, ’punct’)
(8, ’who’, ’who’, ’PRON’, ’WP’, ’PronType=Rel’, 10, ’nsubj:pass’)
(9, ’was’, ’be’, ’AUX’, ’VBD’, ’Mood=Ind—Number=Sing—Person=3—Tense=Past—VerbForm=F
10, ’aux:pass’)
(10, ’known’, ’know’, ’VERB’, ’VBN’, ’Tense=Past—VerbForm=Part’, 3, ’acl:relcl’)
(11, ’for’, ’for’, ’ADP’, ’IN’, None, 14, ’case’)
(12, ’his’, ’he’, ’PRON’, ’PRP$’, ’Gender=Masc—Number=Sing—Person=3—Poss=Yes—PronTyp
14, ’nmod:poss’)
(13, ’controversial’, ’controversial’, ’ADJ’, ’JJ’, ’Degree=Pos’, 14, ’amod’)
(14, ’speeches’, ’speech’, ’NOUN’, ’NNS’, ’Number=Plur’, 10, ’obl’)
(15, ’,’, ’,’, ’PUNCT’, ’,’, None, 3, ’punct’)
(16, ’passed’, ’pass’, ’VERB’, ’VBD’, ’Mood=Ind—Tense=Past—VerbForm=Fin’,
0, ’root’)
(17, ’away’, ’away’, ’ADP’, ’RP’, None, 16, ’compound:prt’)
(18, ’late’, ’late’, ’ADJ’, ’JJ’, ’Degree=Pos’, 20, ’amod’)
(19, ’Thursday’, ’Thursday’, ’PROPN’, ’NNP’, ’Number=Sing’, 20, ’compound’)
(20, ’evening’, ’evening’, ’NOUN’, ’NN’, ’Number=Sing’, 16, ’obl:tmod’)
(21, ’following’, ’follow’, ’VERB’, ’VBG’, ’VerbForm=Ger’, 24, ’case’)
(22, ’a’, ’a’, ’DET’, ’DT’, ’Definite=Ind—PronType=Art’, 24, ’det’)
(23, ’heart’, ’heart’, ’NOUN’, ’NN’, ’Number=Sing’, 24, ’compound’)
(24, ’attack’, ’attack’, ’NOUN’, ’NN’, ’Number=Sing’, 16, ’obl’)
(25, ’.’, ’.’, ’PUNCT’, ’.’, None, 16, ’punct’)
```

Here we got annotations for English and Hindi which included the parameter of ’feats’. But in case of Telugu, which is a synthetic language, has lot of embedded meaning, or say, important things hidden in the same word in form of affixes - which is also the case for relative clauses. But stanza’s annotation was not able to show the ’feats’ of Telugu, hence we printed out what we could.

To Analyze

We analyzed the relative clause sentences on the basis of the occurrence of relative pronoun - whether it occurs or not, on the basis of prenominal and postnominal relative clauses. This was done by comparing the indexes of the referent and the relative pronoun(if any).

We had already stored the whole annotation in form of json file, so that we dont have to compute every time. using those json files, we analyzed about our data. So it was like the dictionary ’data’ had the annotations such that ”data[1] had another dictinary containing annotations of all words of sentence 1, and the sentence itself. These were denoted by the keys which were number for each word, and ”sentence” for viewing the sentence.

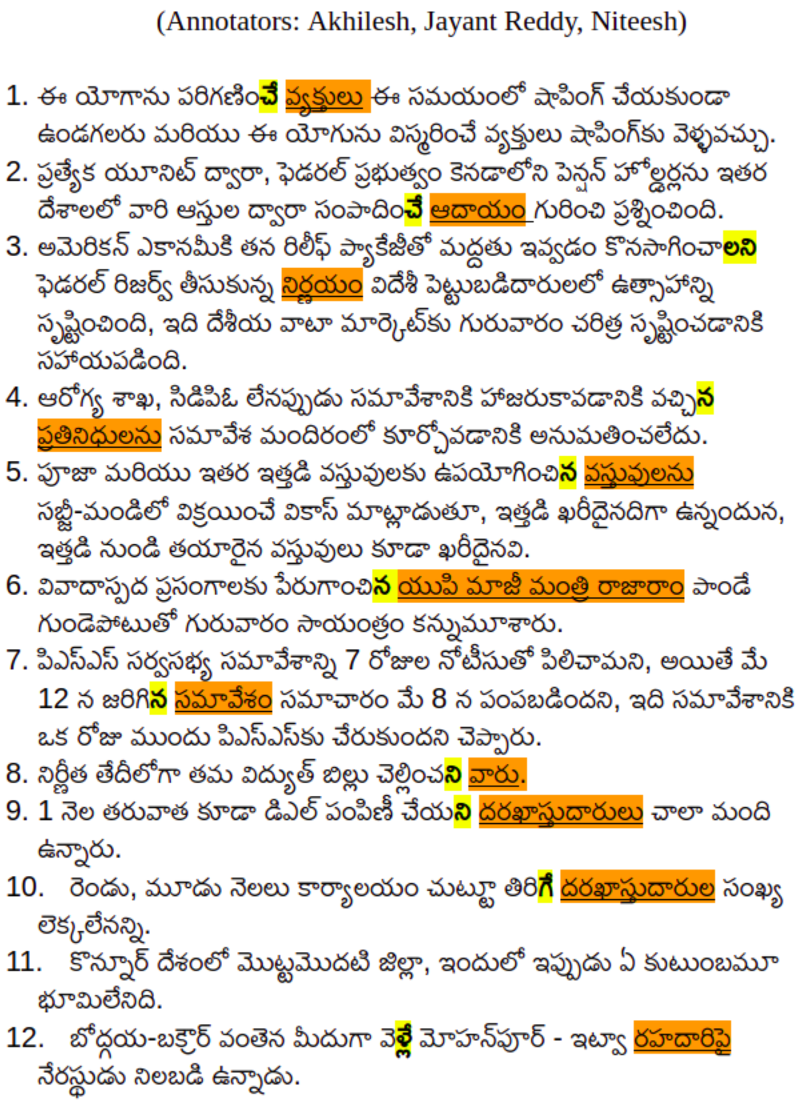
```
for j in data:
    for i in data[j]:
        if i == "sentence":
            continue
        if data[j][i][’deprel’] == ’acl:relcl’:
            pronoun_present = False
            headID = data[j][i][’head’]
            currentID = data[j][i][’id’]
            for k in data[j]:
                if k == ’sentence’:
                    continue
                if (data[j][k][’head’] == currentID):
                    if (data[j][k][’text’].lower() in relPronList):

                        if data[j][k][’id’] > headID:
                            nom_rel += 1
                            f1.write("Sentence "+str(j)+":")
                            f1.write(data[j][’sentence’])
                            f1.write("REL PRONOUN: \"" +
                                  data[j][k][’text’]+"\" with ID: "+str(data[j][k][’

                            f1.write(
                                "REFERNT: \"" + data[j][str(data[j][i][’head’])][’te

                        else:
                            rel_nom += 1
                            f2.write("Sentence "+str(j)+":")
                            f2.write(data[j][’sentence’])
                            f2.write("REL PRONOUN: \"" +
                                     data[j][k][’text’]+"\" with ID: "+str(data[j][k

                            f2.write(
                                "REFERNT: \"" + data[j][str(data[j][i][’head’])][’te
                pronoun_present = True
        if pronoun_present == False:
            f3.write("Sentence "+str(j)+":")
            f3.write(data[j]['sentence'])
            f3.write("NO PRONOUN\n")
            f3.write(
                "REFERNT: \"" + data[j][str(data[j][i]['head'])]['text'] + "\"
            red_rel += 1
```

Chapter 3

Observations

3.1 Telugu

3.1.1 Manual Annotation



(Annotators: Akhilesh, Jayant Reddy, Niteesh)

1. ఈ యోగాను పరిగణిం**చే** వ్యక్తులు ఈ సమయంలో షాపింగ్ చేయకుండా ఉండగలరు మరియు ఈ యోగును విస్మరించే వ్యక్తులు షాపింగ్‌కు వెళ్ళవచ్చు.
2. ప్రత్యేక యూనిట్ ద్వారా, ఫెడరల్ ప్రభుత్వం కెనడాలోని పెన్షన్ హోల్డర్లను ఇతర దేశాలలో వారి ఆస్తుల ద్వారా సంపాదిం**చే** ఆదాయం గురించి ప్రశ్నించింది.
3. అమెరికన్ ఎకానమీకి తన రిలీఫ్ ప్యాకేజీతో మద్దతు ఇవ్వడం కొనసాగించా**లని** ఫెడరల్ రిజర్వ్ తీసుకున్న నిర్ణయం విదేశీ పెట్టుబడిదారులలో ఉత్సాహాన్ని సృష్టించింది, ఇది దేశీయ వాటా మార్కెట్‌కు గురువారం చరిత్ర సృష్టించడానికి సహాయపడింది.
4. ఆరోగ్య శాఖ, సిడిపిఓ లేనప్పుడు సమావేశానికి హాజరుకావడానికి వచ్చి**న** ప్రతినిధులను సమావేశ మందిరంలో కూర్చోవడానికి అనుమతించలేదు.
5. పూజా మరియు ఇతర ఇత్తడి వస్తువులకు ఉపయోగించి**న** వస్తువులను సబ్జీ-మండిలో విక్రయించే వికాస్ మాట్లాడుతూ, ఇత్తడి ఖరీదైనదిగా ఉన్నందున, ఇత్తడి నుండి తయారైన వస్తువులు కూడా ఖరీదైనవి.
6. వివాదాస్పద ప్రసంగాలకు పేరుగాంచి**న** యుపి మాజీ మంత్రి రాజారాం పాండే గుండెపోటుతో గురువారం సాయంత్రం కన్నుమూశారు.
7. పిఎస్ఎస్ సర్వసభ్య సమావేశాన్ని 7 రోజుల నోటీసుతో పిలిచామని, అయితే మే 12 న జరిగి**న** సమావేశం సమాచారం మే 8 న పంపబడిందని, ఇది సమావేశానికి ఒక రోజు ముందు పిఎస్ఎస్‌కు చేరుకుందని చెప్పారు.
8. నిర్ణీత తేదీలోగా తమ విద్యుత్ బిల్లు చెల్లించ**ని** వారు.
9. 1 నెల తరువాత కూడా డిఎల్ పంపిణీ చేయ**ని** దరఖాస్తుదారులు చాలా మంది ఉన్నారు.
10. రెండు, మూడు నెలలు కార్యాలయం చుట్టూ తిరి**గే** దరఖాస్తుదారుల సంఖ్య లెక్కలేనన్ని.
11. కొన్నూర్ దేశంలో మొట్టమొదటి జిల్లా, ఇందులో ఇప్పుడు ఏ కుటుంబమూ భూమిలేనిది.
12. బోధ్గయ-బక్రౌర్ వంతెన మీదుగా వె**ళ్లే** మోహన్‌పూర్ - ఇట్వా రహదారిపై నేరస్థుడు నిలబడి ఉన్నాడు.

3.1.2 Analysis

So as mentioned earlier, 'stanza' did not include the morphology for telugu. Hence we could not do it computationally. Now, with the fact that both of us are non-Telugu speakers, having literally no prior knowledge about telugu, it was quite difficult for us to manage. We sought help from our friends who helped us in annotations and figuring out how the relative clause appears in telugu.

3.2 Hindi and English

3.2.1 Automated Annotation

Process of annotation was automated and implementation is explained above. Observation files, generated from the given data-set, can be found in the Observation directory on git. As both the authors are well versed with Hindi, we also discussed some of the possible implication rules which are discussed below. Manual analysis of English was also done up till some extent.

3.2.2 Analysis

Table 3.1: Some Observations

	Postnominal Instances	Prenominal Instances	Reduced Relative
English	381	1	128
Hindi	242	89	137

We observe the following things from this collected output:

1. We see that relative clauses in English always follow the pattern of referent -> relative. This also supports the fact that English is an SVO language with FIXED WORD ORDER.

2. The fact that Hindi has a good distribution among referent -> relative and relative -> referent proves that Hindi is a free word order language.

3. Both the languages having a good number of reduced relativization or gapping - which seems true. Eg:

4. This also reflects that no language falls to just one category when it comes to relativization. There is always an overlap. Here we observe that when it comes to relativization, both Hindi and English fall under the category of "relative pronouns" and "Gap relativization" as well.

   Additionally, we also observed in few of our sentences that the instances of prenominal, it included the instances where Hindi was showing characteristics of "Non Reduction relativization" - wherein the head noun is itself the part of relative clause.

5. Our findings are also in accordance with the **Greenberg Universal 24: "If the relative expression precedes the noun either as the only construction or as an alternate construction, either the language is postpositional, or the adjective precedes the noun or both."** So we see that both Hindi and Telugu outputs agrees to this fact.

3.3 Some more observations

3.3.1 What in the sentence or language structure implies the presence/pattern of your construction?

As we have discussed different languages use different elements to show presence of relative clause construction. For example, English is majorly a type of language which uses relative pronoun extensively for relativization. Relative Pronouns like who, which, whose etc. typically show the presence of relative clause construction in a sentence. Hindi however is a type of languge which utilizes multiple straegies to relativize a clause. Outputs also supplements our knowledge of Hindi which shows the occurence of non reduction type, relative pronoun, and gap type of relativization. So words like jo,vo etc can be used to identify the presence of relative clause construction. In telugu, suffixes like "Aina" shows the presence

3.3.2 What does the pattern of your chosen construction imply in that language?

3.3.3 What kind of typological classification does the pattern of your chosen construction imply?