Report

Language Typology and Universals

Aravapalli Akhilesh 2019114016
Suyash Vardhan Mathur 2019114006

Construction

The construction to be analyzed in the project is **Subject.** It is difficult to define the notion of subject cross-linguistically. It is generally considered as a syntactic role that follows certain properties and has a close relation to the semantic elements of agent and topic of a sentence. While it is difficult to say whether a noun IS or ISN'T a subject across languages, we can define a degree of a noun for being a subject.

Subject can occur in many different ways:
English:

- **It** is raining.(expletive)
- **John** killed Jack(Noun)
- **Big John** has bullied Jack. (Noun phrase)
- **To read** is easier than to write(to-infinitive)
- **His constant hammering** was annoying.(Gerund)
- **That he had travelled the world** was unknown to everyone.(That-clause)
- **He** went to the bar.(Pronoun)
- **Whatever he did** was always suspicious.(Free Relative clause)
- **"I love you"** is not heard these days.(Direct quotation)

Telugu:

- వాన పడుతుంది. (Noun)
- రాము సురేష్ ని కొట్టాడు. (Noun)
- చదవడం రాయడం కన్న సులువు. (Gerund)
- ఆయన కొట్టు కి వెళ్ళాడు. (Pronoun)
- "నమస్తే" ఈ రోజుల్లో చాలా తక్కువగా వినపడుతుంది. (Direct Quotation)
- అతని వంట చాలా బాగుంటుంది. (Gerund)
- అతను ఏమి చేసిన అనుమానంగా ఉంది. (Free Relative Clause)

Hindi:

- जॉन ने जैक को मारा। (Noun)
- बड़े जॉन ने जैक को मारा। (Noun phrase)
- पढ़ना लिखने से बेहतर है। (Gerund)

- उसकी बड़ी बड़ी बातें लुभावनी थी। (Gerund)
- वह बार में गया। (Pronoun)
- जो भी वो कर रहा था अजीब था। (Free Relative Clause)
- “मैं प्यार करता हूँ” कम सुनाई देता है। (Direct Quotation)

Languages used

English, Hindi & Telugu

Corpora

For Hindi and English:

- Hindi English Parallel corpus iitb
- https://github.com/joshua-decoder/indian-parallel-corpora

For telugu

- https://github.com/joshua-decoder/indian-parallel-corpora

Method

We used the following methods to find the observations and the conclusions:

- Use automated tools(Stanza) to mark SUBJECT, VERB and OBJECT in the sentences.(The script can be found with the name ‘<language_name>_annotater.py’)
- Manually search for all ways how a subject can occur in a sentence in all the languages.
- Code to annotate the SUBJECT, VERB and OBJECT for various features(Gender/Number/Person/Case/Tense/Aspect/Modality).
- Check the different annotations to obtain statistics about agreements.
- Make rules about occurrence of Subject based on different patterns and agreement.
- Discover implicational rules for conditions on how a subject can occur.

Type of Processing Done/Proposed Tools

(Pre)processing of text

- Using Nltk to remove special characters and characters from other languages.
- Tokenization
- Parsing

1. Stanza

Annotation done

- We used tools(Stanza) to annotate the subject, verb and object occurring in the sentence.
- We used scripts to observe patterns in the sentences from all languages.
- Cross check some parsing outputs.
- Manually annotated sentences in Telugu.

Properties used for Analysis

Using scripts we analysed the following:

- Subject-verb agreement(Annotate verb for Gender, number, person, TAM as well)
- Position of the S/O/V
- Morphological Case

Analysis

English

Gender:

All Agree: 0

Category - 1	Category - 2	Count
Subject	Object	1
Subject	Verb	0
Verb	Object	0

Case:

All agree: 0

Category - 1	**Category - 2**	**Count**
Subject	Object	0
Subject	Verb	0
Verb	Object	0

Number:

All agree: 20

Category - 1	**Category - 2**	**Count**
Subject	Object	98
Subject	Verb	134
Verb	Object	1

Person:

All agree: 0

Category - 1	**Category - 2**	**Count**
Subject	Object	1
Subject	Verb	5
Verb	Object	0

Hindi

Gender:

All Agree: 0

Category - 1	**Category - 2**	**Count**
Subject	Object	53
Subject	Verb	222
Verb	Object	35

Case:

All agree: 0

Category - 1	**Category - 2**	**Count**
Subject	Object	49
Subject	Verb	71
Verb	Object	0

Number:

All agree: 90

Category - 1	**Category - 2**	**Count**
Subject	Object	9
Subject	Verb	295
Verb	Object	15

Person:

All agree: 43

Category - 1	**Category - 2**	**Count**
Subject	Object	75
Subject	Verb	213
Verb	Object	2

Telugu(Out of 100 sentences)

Gender:

Category - 1	**Category - 2**	**Count**
Subject	Object	16
Subject	Verb	45
Verb	Object	14

Case:

Category - 1	**Category - 2**	**Count**
Subject	Object	11
Subject	Verb	-
Verb	Object	-

Number:

Category - 1	**Category - 2**	**Count**
Subject	Object	6
Subject	Verb	48
Verb	Object	7

Person:

Category - 1	**Category - 2**	**Count**
Subject	Object	15
Subject	Verb	100
Verb	Object	16

Subject - Verb - Object Occurrence:

	ENGLISH	HINDI	TELUGU
SOV	-	116	10
SVO	253	18	-
OSV	-	19	10
OVS	-	-	-
VOS	-	-	-
VSO	-	-	-

	English	Hindi	Telugu
SUBJECT OBJECT	253	134	10
OBJECT SUBJECT	-	19	10

Implication Rules:

English

- **Gender -** Verb in English isn't inflected with Gender Information, and so the agreement counts for Subject-Verb and Object-Verb are 0 in all the cases, and can't predict anything.
- **Case -** In English, Verb is not marked for case, despite subject and objects being marked for case. For this reason, there are 0 agreements in the table.
- **Number -** Subject-Verb agree in 134 cases, whereas Verb-Object agrees in 1 case, and all three agree in 20 cases. The all 3 agreeing case seems to be one where Verb-Subject are agreeing

and object is agreeing by simple coincidence. Thus, we can say that **the noun that agrees with Verb in terms of Number in English is the subject.**

- **Subject-object Ordering -** We can see that 253 cases follow SO order which can say that English is a SOV language and thus we can claim that the noun phrase occurring at the start is most likely to be the subject of the sentence.

Hindi

In Hindi, we can see that subject is agreeing with the verb in terms of

- **Gender** - Subject and Verb agree in absolute majority of the cases(**222** cases) as opposed to 53 cases of Subject-Object agreement and **35** cases of Verb-Object Agreement, which are in minority. Thus, we can say that **the noun that agrees with Verb in terms of Gender in Hindi is highly likely to be the subject.**
- **Case -** Subject-Verb agree in 71 cases, whereas Verb-Object agree in 0 cases. Subject-Object agree in terms of case in 49 cases, but this seems to be completely coincidental. Thus, we can say that **the noun that agrees with Verb in terms of Case in Hindi is the subject.**
- **Number -** Subject-Verb agree in 295 cases, which is much higher than Verb-Object(15). Thus, we can say **the noun that agrees with Verb in terms of Number in Hindi is the subject.**
- **Person -** Subject-Verb agree in Person in 213 cases, in comparison to which Verb-Object agreement is negligible(2). Thus, we can say that **the noun that agrees with Verb in terms of Person in Hindi is the subject.**
- **Subject-Object Ordering -** Hindi appears to be a Subject-Object ordered language, because it has 134 instances of SUBJECT OBJECT, but it also allows OBJECT SUBJECT ordering as well(19 instances). Thus, we can say that **the noun phrase that occurs first in the sentence is highly likely to be the subject.**
- **Subject-Object-Verb Ordering -** Hindi appears to be a SOV language, as it has 116 cases of SOV ordering. However, it also has 18 cases of SVO and 19 cases of OSV, which suggest that it

has SVO and OSV possible orders as well. This appears to be due to free-word ordering of Hindi. Thus, we can say that **the noun phrase that occurs before the verb is more likely to be the subject than the noun phrase that occurs after the verb in Hindi.**

Telugu

Information about the corpus and Telugu in general

- Verbs do not get inflected by case markings, instead noun and pronouns get inflected to represent case markings. We can see the stats that no verb agrees with subject/object regarding case in all sentences.
- Most of the cases, objects do not carry gender and is represented as neutral in the annotated dataset.
- In telugu, inanimate objects are grammatically considered as female. It can be observed when the OBJECT-VERB agrees with each other in gender even though the object is gender neutral
- In Telugu, when respect is shown to someone, verbs get inflected with plural numbers even though the recipient is singular. Most of the corpus is political and the names of political figures were represented with plural out of respect.
- Verbs tend to occur at the last even though there is a slight free word order possibility

Observations

- **Gender -** Subject-Verb agree in Gender in 45 cases, the places where Verb-Object agrees with on another on gender is where objects are considered as feminine gender and the verb is following the subject. Thus we can claim that noun phrase that agrees with verb in case of gender can be considered as subject.
- **Case -** Verb do not carry necessary information about case, only nouns,pronouns and noun phrases inflect for case marking. We can see that in the stats that no verb agrees with subject/object for case. Subject Object need not necessarily agree wrt case.

- **Number** - SV agree with number in 48 cases ,thus we can say that the subject agree with verbs in case of number. The cases where Subject Object agree with number is not necessary because they tend to carry different information about the sentence. Thus we can claim that noun phrase that agrees with verb in case of number can be considered as subject.
- **Person** - We can see that Subject and Verb agree 100 times wrt person. Thus, we can claim that the noun phrase which agrees with verb in terms of person can be considered as a subject

In telugu, the noun phrases that are occurring before the verb are most likely to become the subject of the sentence.

Typological systems

Subject - verb - object ordering is one of the linguistic typologies which uses the construction 'subject' while classifying languages. The categories that we found are:

- **Hindi:** Hindi appears to be a SOV language, as it has 116 cases of SOV ordering. However, it also has 18 cases of SVO and 19 cases of OSV, which suggest that it has SVO and OSV possible orders as well. This appears to be due to free-word ordering of Hindi.
- **Telugu:** As per data we chose to manually annotate, we cannot justify any ordering of Subject,Verb and Object. We can see that the number of SOV and OSV are same as per data chosen.
- **English:** English is a SOV language as per the given data, because it has 253 cases of SVO ordering, and none of any other ordering.

Subject-Object Ordering: Universal 1- "In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object."

- **Hindi:** Hindi appears to be a Subject-Object ordered language, because it has 134 instances of SUBJECT OBJECT, but it also allows OBJECT SUBJECT ordering as well(19 instances).

- **English:** English is a SUBJECT-OBJECT language, because it has 253 instances of SUBJECT-OBJECT, but 0 of the other order.
- **Telugu:** We cannot justify or condemn the universal mentioned above as we didnt annotate enough data.

The above analysis also follows the **Universal 32:** "Whenever the verb agrees with a nominal subject or nominal object in gender, it also agrees in number."

- In the case of Hindi, the Verb-Subject agree in gender, and the two agree in number as well.
- In the case of English, there is no agreement in terms of Gender, but only agreement in number.
- In the case of Telugu, the Subject-Verb agree in gender(48 times) and the two agree in number(45 times) as well.

Literature Reviewed

- Bernard Comrie chapter 5
- An Analysis of Subject Verb Agreement *Mampoi Irene Chele*