M.A. Project Report

ANAPHORA RESOLUTION IN BHOJPURI USING DMDX

A project submitted in the partial fulfillment of the requirements for the degree of Masters of Arts in Computational Linguistics

Submitted by:

Nishi Singh

Project Supervisor:

Dr. Atreyee Sharma

DEPARTMENT OF COMPUTATIONAL LINGUISTICS
SCHOOL OF LANGUAGE SCIENCES
THE ENGLISH AND FOREIGN LANGUAGES UNIVERSITY,
HYDERABAD
June 2021

Certificate

I hereby declare that this project entitled **Anaphora Resolution in Bhojpuri using DMDX** is a bona fide work done by **Nishi Singh** during the period of her study under my supervision and it has not previously formed the basis for award of any degree, diplomas or other similar title.

Project Supervisor:

Dr. AtreyeeSharma

Hyderabad School of LanguagesSciences

21.06.2021 The English and Foreign LanguagesUniversity

DECLARATION

I, Nishi Singh (Reg No: H00MACL201900025), do hereby declare that this project entitled **Anaphora Resolution in Bhojpuri using DMDX** is a bona fide work done by me during the period of my study under the supervision of **Dr. Atreyee Sharma** and it has not previously formed the basis for award of any degree, diplomas or other similar title.

Nishi Singh

Hyderabad H00MACL201900025

21.06.2021 MA ComputationalLinguistics

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. I would like to express my sincere thanks to all of them.

I am highly grateful to my guide Dr. Atreyee Sharma for her guidance and constant supervision as well as providing necessary information regarding the project and also her support in completing the project.

I would like to express my sincere gratitude to my family for their kind cooperation and encouragement which helped me in the completion of this project.

My thanks and appreciations also go to my friends in developing the project and to the people who have willingly helped me out with their abilities.

CONTENTS

Abstract

The main objective of the paper is to try and resolve ambiguity for anaphoric ambiguous sentences in Bhojpuri by its native speakers through the measurement of reaction times and the possible interpretations of the ambiguous sentences. We enable this through the use of a DMDX experiment. The test sentences are ambiguous and have been designed to have equal interpretations.

In the DMDX experiment, the test subjects are presented with a mix of 21 unambiguous and ambiguous sentences and we compare the reaction times for both of them. We are testing to see if the reaction times for ambiguous sentences are higher than that of their unambiguous counterparts when given the possible interpretations.

PART I: Introduction

Bhojpuri

Bhojpuri is an Indo-European language and belongs to the Eastern Indo-Aryan group group of the Indo-Aryan languages. The Magahi and Maithili languages of Eastern Indo-Aryan group are closest living relatives of Bhojpuri Bengali and Assamese are also closely related. Bhojpuri along with Magahi and Maithili, are grouped together as the Bihari languages. Together with the other branches of Eastern Indo-Aryan, the Bihari languages are considered to be direct descendants of the Magadhi Prakrit. Bhojpuri is spoken by 37.8 million people in India, primarily in the western part of the state of Bihar and eastern part of the state of Uttar Pradesh and some adjoining areas of Madhya Pradesh. In Nepal, Bhojpuri is spoken by 1.7 million as a first language plus by another 74,000 as a second-language. Bhojpuri is spoken by 336,000 people in Mauritius. The language is a minority language in Fiji, Guyana, South Africa, Suriname and Trinidad and Tobago.There are four major dialects of Bhojpuri which are mutually intelligible. Although the full range of variation is not firmly established, the differences among the dialects appear to be primarily lexical and phonological.

1. Northern Standard considered to be the most prestigious dialect of the language prevalent in the western Tirhut division in Bihar, and Gorakhpur division and Basti division in Uttar Pradesh. It is also spoken in Nepal.

2. Western Standard prevalent in the areas of Varanasi, Azamgarh, Mirzapur, Sonbhadra, Sant Ravidas Nagar, and Bhadohi districts in Uttar Pradesh.
3. Southern Standard prevalent in the Shahabad district and the Saran region in Bihar, and the eastern Azamgarh and Varanasi regions in Uttar Pradesh.
4. Nagpuriais the southernmost popular dialect, found in the Chota Nagpur Plateau of Jharkhand, particularly parts of Palamau and Ranchi.

The grammar of Bhojpuri is similar to that of other Indo-Aryan languages. Grammatical relations are marked by inflectional suffixes. Bhojpuri morphology is fusional with a single ending representing several categories, which is typical of Indo-European languages. The basic vocabulary of Bhojpuri is Sanskrit in origin. It uses prefixes and suffixes to derive words from basic elements, as well as reduplication and compounding. Over the years Bhojpuri has borrowed words from Hindi, Bengali, and other neighboring Indo-Aryan languages, as well as from English.Bhojpuri is written in the Kaithi script which is widely used throughout North India, primarily in Uttar Pradesh and Bihar. Like other Indic scripts, Kaithi is a descendant of the Brahmi script. Kaithi script derives its name from the word Kayastha, one of the social groups of North India. The script can be traced back to the 16th century. It was widely used during the Mughal Empire. Today, the Kaithi script is being replaced by Devanagari.

DMDX

DMDX is a program created by Jonathan Foster designed to precisely time the presentation of text, audio, graphical and video material and to enable the measurement of reaction times to these displays with millisecond accuracy. It is compatible with x32 Windows architecture and is compatible with most versions of Windows.It is an extension of the DOS-only DMASTR program suite, designed originally by Rod Dickinson and Wayne Murray. DMDX provides greatly enhanced graphics over the previous DMASTR suite and it supported a wide variety of fonts, image and audio formats. The input to DMDX is a .rtf file that can be created in MS Word or Wordpad. This itemfile/script specifies what is to be displayed and how.

DMDX is used in psychological labs around the world to measure reaction times to audio-visual stimuli. It is considered to be very accurate and flexible meaning that the user can tailor the experiment down to the minute details. DMDX is also free and has a great community support despite only being available for windows systems. DMDX has a small learning curve associated the complex syntax required for designing the experiment but a lot of helpful tutorials are available online.

DMDX relies on RTF files in which parts of the stimuli, design, and procedure of an experiment are defined in a complicated syntax. Time DX is sets the software and hardware features for running DMDX, we can set the millisecond timer and, refresh rate and select the video and audio modes. It presents stimuli and records responses with ms-accuracy, based on a script that determines basic parameters of the experiment and the order and way in which items are displayed. Experiments can involve simple text displays or the script can call up additional media files with auditory or visual stimuli. DMDX allows for sound, video and picture files to be presented along with text at particular time windows. Video, sound and picture files are Stimulus files. All lines of text that represented to the participants, along with further details about their presentation and RT-measurements, are part of another file, namely the Item file.

ANAPHORA

Anaphora is the use of a pronoun or other linguistic unit to refer back to another word or phrase. A word that gets its meaning from a preceding word or phrase is called an anaphor. The preceding word or phrase is called the antecedent, referent, or head. The word anaphora was originated in the 16th century. It can be broken into two parts – ana (meaning: back) and phora (meaning carrying) which when combined means carrying back. It is used to avoid the repetition of a word or a phrase in a text document or in a spoken speech. An anaphora can be a pronoun, a verb, definite descriptions, a lexical modifier, a noun phrase, or a proper noun.

Types of Anaphora

1. Pronominal Anaphora: Pronominal anaphora are the most widely used anaphora. Here pronouns are used as anaphora to refer to the antecedent. Occurs at the level of Personal pronoun, processive pronoun, relative pronoun and reflexive pronoun.
   Example: [] John(1) went to school. He(2) forgot to take his books.
   In the sentence [], he(2) is anaphoric to John(1).

   Pronominal Anaphora can again be classified, such as:

- One Anaphora: In this type of anaphora, the pronoun ‘one’ is used to refer to the antecedent. Example: Sam bought a car. It was his childhood dream to buy one.

- Definite pronominal: This type of reference is definite since it refers to a single unique entity in the universe. Example: Jane loved the cake. It was decorated beautifully.
- Indefinite pronominal: In this type of reference, the pronoun refers to an entity or object which is not well-defined or well-specified. Example: The students decided to cancel the class even though many were against it.
- Adjectival pronominal: In this type of anaphora, there is reference to adjectival form of the entity which has occurred earlier. Example: Mary helped the blind man. She loves to help such people.
- Reflexive pronominal: In this type of anaphora, reflexives like ‘each-other’, ‘one-another’, ‘himself’ and others, are used to refer to the antecedent. Example: While the teacher was teaching, Ram and Shyam were talking to each other in the class.

2. Nominal Anaphora: Nominal anaphora occurs when a non-pronominal noun phrase is used as an anaphora to refer to the antecedent. Example: [] John bakes beautiful cakes. The town people love the baker.
3. Zero Anaphora: Zero anaphora, denoted by Ø, are ‘invisible anaphora’ which is often omitted in a sentence but nevertheless understood. It uses a gap in a phrase or a clause to refer back to its antecedent. Example: Ram went to the airport and Ø boarded the plane by evening.
4. Discontinuous sets (split anaphora): In this type of anaphora, the pronoun may refer to more than one antecedent. Example: Mary and Jane went for a walk. They went by the lake.
5. Cataphora: A cataphoric expression serves to point to entities which may succeed it. Example: If he wants to pass the exam, Ravi has to work hard.

ANAPHORA AMBIGUITY AND RESOLUTION

Anaphoric ambiguity occurs when the sentence offers two or more potential antecedents to which the anaphor can refer. Example:

‘The doctor talked to the patient’s brother who was tested positive for Covid’

Here, ‘who’ is the anaphor which can refer to either the ‘patient’ or the ‘patient’s brother’. Thus, it comes to one’s analysis of the sentence which might be done through some context, the center of attention or by other factors.

The centering theory is the most approachable solution for anaphoric ambiguity. At any given point of a discourse, the discourse participant’s attention is centered on a set of entities, a proper subset of all the entities being talked about in the discourse. For a given utterance, the discourse participant’s attention is centered on a singleton unit and the rest of the utterance makes a prediction about this entity. This is the notion of center of attention or the centering theory.

To disambiguate a sentence, the approach for anaphora resolution with this centering view is that the search for the referents of the anaphoric expression should be restricted to the set of centered entities. Here, the assumption being that in discourse, it is these entities that are most likely to talk about and refer to with the use of anaphoric expressions. Eye movements are of great help while trying to disambiguate an ambiguous sentence, especially in the absence of linguistic cues.

PART II: Experiment

Participants

A total of 15 test subjects were taken for the DMDX experiment. The ages ranged from 23-64 with a mean age of 38.3. All 15 participants were fluent native speakers of Bhojpuri and residents of Bihar. The experiment was conducted in an environment where no external disturbances or influences would come into play and the subject's decision is their own. The subjects are presented with an ambiguous sentence. In the following screens two equally accurate interpretations of the test sentence are displayed and the subject is asked to select the one they think is accurate. The setup consists of ten test sentences interspersed with eleven filler sentences to prevent biasing. These test sentences are independent of any context and thus the subjects choose the interpretation they prefer. These preferences are influenced by certain factors and through this experiment we lay the probabilistic factors that might influence the subjects' preferences.

Test Sentences

1. Shivam aego laeki se millak okra sange hm school jaat rehni ha

2. Kehu uu abhinetri k beti k goli maar delak jekra k award milal rahe.
3. Hamaar cycle garage me rakhal baa je humra lage 10 saal se dher samay se baa
4. Hum abhi apun dost ke bhai ke dekhni ha je Switzerland me rahela
5. Doctor uu patient k bhai se bhet kaeli jekara k cancer rahe
6. Anushka uu kariya kapda wali laeki k sange aayil rhali je hamra se address puchath raheli
7. Photographer shubham k phone kaelak je ki khub khisyael rahe
8. Uu lalka suitcase m hamar kpda rahe jekra k hm fek deni
9. Ram aaur Shyam Sita aaur Gita k sange ghumat ghumat ek dusra se batiyaat rhe
10. Ham Ram k okra ghr me dekhni

Code

```
<azk><fd 100><t 90000><id “keyboard”><dbc 0><dwc
000255000><cr><rcot><nfb><d 0><n 70><vm desktop><fbp 0>

0 “Press spacebar to start”;

0 "Read each Bhojpuri sentence ";
```

0 “Two inferences based on the sentences will be displayed”;

0 "Choose the one that you think is correct”,

<ln 2> “by pressing **Left Shift** or **Right Shift** key”;

0 "Try to respond as quickly as possible”,

<ln 2>"but not so quickly that you make errors”;
0 "Press the SPACEBAR to begin the demo ";
0 “Press Right or left shift to display the choices once sentence has been displayed”;

+100*“Kaal aego aisen din rahe jon din sab kuch galat ho rahal rahe”;

+1*“ Kal jo bhi ho raha tha sab galat ho raha tha – Left Shift”,

<ln 2>“ Kal ka din galat tha – Right Shift”;

+101*“Uu laika bottle tod delak jekar baap aego doctor baare”;

+1*“ Uss ladke ne bottle tod di, uske pita ek doctor hai – Left Shift”,

<ln 2>” Uss ladke ke pita ne bottle tod di, uske pita ek doctor hai – Right Shift”;

+201*” Shivam aego laeki se millak okra sange hm school jaat rehni ha”;

+1*” Shivam ek ladki se mila, mai Shivam ke sath school jati thi – Left Shift”,

<ln 2>” Shivam jis ladki se mila mai uss ladki ke sath school jati thi – Right Shift”;

+102*“Laptop aego computer baa jekra leke ghumal jaa sakela”;

+1*” Laptop ek computer hai, laptop ko lekar ghuma jaa sakta hai – Left Shift”,
<ln 2>” Laptop ek computer hai, computer ko lekar ghuma jaa sakta hai – Right shift”;

+202*” Kehu uu abhinetri k beti k goli maar delak jekra k award milal rahe”;
+1*” Kisine uss abhinetri ki beti jisse award mila tha usse goli maar di – Left Shift”,
<ln 2>” Kisine uss abhinetri jisse award mila tha uski beti ko goli maar di – Right Shift”;

+103*“Hamaar cycle garage me rakhal baa je humra lage 10 saal se dher samay se baa”;
+1*” Meri cycle garage me rakhi hui hai, garage 10 saal se adhik samay se mere pas hai – Left Shift”,
<ln 2>” Meri cycle garage me rakhi hui hai, cycle 10 saal se adhik samay se mere pas hai – Right Shift”;

+203*”Hum abhi apun dost ke bhai ke dekhni ha je Switzerland me rahela”;
+1*” Maine abhi apne dost ke bhai ko dekha, mere dost ka bhai Switzerland me rehta hai – Left Shift”,
<ln 2>” Maine abhi apne dost ke bhai ko dekha, mera dost Switzerland me rehta hai – Right Shift”;

+104*“Uu chhot laiki jekar gudiya bhula gayil uu bahut dukhi biya”;

+1*” Woh chhoti ladki ki gudiya kho gayi thi, woh chhoti ladki bahot dukhi hai – Left shift”,
<ln 2> “Woh chhoti ladki ki gudiya kho gayi thi, uski gudiya bahot dukhi hai – Right Shift”;

+204*” Doctor uu patient k bhai se bhet kaeli jekara k cancer rahe”;
+1*” Doctor uss patient jisse cancer tha uske bhai se mili – Left Shift”,
<ln 2>” Doctor uss patient ke bhai jisse cancer tha se mili – Right Shift”;

+105*”Vyshnavi aego bachcha ke nahwat rahli jon jhula jhulat rahe”;

+1 *“Vyshnavi ek bachche ko nahla rahi thi, bachcha jhula jhul rha tha – Left Shift”,
<ln 2>” Vyshnavi ek bachche ko nahla rahi thi, Vyshnavi jhula jhul rahi thi - Right Shift”;

+205*” Anushka uu kariya kapda wali laeki k sange aayil rhali je hamra se address puchath rahe”;
+1 *“ Anushka uss kaale kapde wali ladki ke sath aayi thi, Anushka mujhse address puch rahi thi – Left Shift”,
<ln 2>” Anushka uss kaale kapde wali ladki ke sath aayi thi, wo ladki mujhse address puch rahi thi – Right Shift”*;

+106*”Lane 26 me uu dukaan abhiyo baa jahan hum prachi se pahilka baar mille rehni”;
+1*” Lane 26 me woh chai ki dukan abhi bhi hai, chai ki dukan pe mai Prachi se mili thi – Left Shift”,
<ln 2>” Lane 26 me woh chai ki dukan abhi bhi hai, lane 26 me mai Prachi se mili thi – Right Shift”;

+206*” Photographer shubham k phone kaelak je ki khub khisyael rahe”;

+1*” Photographer ne Shubham ko phone kiya, photographer kafi gusse me tha – Left Shift”,

<ln 2>” Photographer ne Shubham ko phone kiya, Shubham kafi gusse me tha – Right Shift”;

+107*”Kaa tu humra ke uu slide bhej debu whatsapp pe jekra bare me sir batiyat rahle”;

+1*” Kya tum mujhe wo slide bhej sakti ho jiske bare me sir Whatsapp pe baat kar rahe the – Left Shift”,

<ln 2>” Wo slide jiske bare me sir baaat kar rahe the, kya tum mujhe wo Whatsapp pe bhej sakti ho – Right Shift”;

+207*” Uu lalka suitcase m hamar kpda rahe jekra k hm fek deni”;

+1*” Uss laal suitcase me mere kapde the, maine suitcase fek diya – Left Shift”,

<ln 2>” Uss laal suitcase me mere kapde the, maine kapde fek diye – Right Shift”;

+108*”Uu kutta paglaa gayil baa jekar maalik London me rahela”;

+1*“Woh kutte ka maalik London me rehta hai, maalik pagal ho gaya hai – Left Shift”,

<ln 2>” Woh kutte ka maalik London me rehta hai, kutta pagal ho gaya hai – Right Shift”;

+208*” Ka tu uu kitaab liyaa debu library me je ham padhat rahni ha”;

+1*” Kya tum wo kitaab library me laa sakti ho jo mai padh rahi thi – Left Shift”,

<ln 2>” Kya tum wo kitaab laa sakti ho jo mai library me padh rahi thi – Right Shift”;

+109*”EFL University jaha hamaar bahin pdheli uu Hyderabad me baate”;

+1*” EFL University Hyderabad me hai, meri behen Hyderabad me padhti hai – Left Shift ”,

<ln 2>” EFL University Hyderabad me hai, meri behen EFL University me padhti hai – Right Shift”;

+209*” Ram aaur Shyam Sita aaur Gita k sange ghumat ghumat ek dusra se batiyaat rhe”;

+1*” Ram or Shyam ek dusre se baat kar rhe the sita or gita k sath ghumte huye – Left Shift”,

<ln 2>” Ram or Shyam Sita or Gita se baat kr rhe the ghumte huye – Right Shift”;

+110*”Hum kamra ke aesan rang se rangal chahataani je aakarshak laage”;

+1*”Mai kamre ko aakarshak rang se rangana chahti hu – Left Shift”,

<ln 2>”Mai kamre ko aise rang se rangana chahti hu jisse kamra aakarshak ho – Right Shift”;

+210*” Ham Ram k okra ghr me dekhni”;

+1*” Maine Ram ko Ram ke ghar me dekha– Left Shift”,

<ln 2>” Maine Ram ko kisi or ke ghar me dekha – Right Shift”;

0 “Press escape to exit.”;

PART III: RESULTS AND DISCUSSION

DMDX EXPERIMENT

The table below shows what percentage of the subjects selected an interpretation as correct one:

Test Sentence given in the code	Majority percentage	Chosen Majority
1	100%	Shivam jis ladki se mila mai uss ladki ke sath school jati thi.
2	73.3%	Kisine uss abhinetri jisse award mila tha uski beti ko goli maar di

3	73.3%	Meri cycle garage me rakhi hui hai, cycle 10 saal se adhik samay se mere pas hai
4	53.3%	Maine abhi apne dost ke bhai ko dekha, mera dost Switzerland me rehta hai
5	93.3%	Doctor uss patient jisse cancer tha uske bhai se mili
6	100%	Anushka uss kaale kapde wali ladki ke sath aayi thi, wo ladki mujhse address puch rahi thi
7	75%	Photographer ne Shubham ko phone kiya, photographer kafi gusse me tha
8	53.3%	Uss laal suitcase me mere kapde the, maine kapde fek diye
9	100%	Ram or Shyam Sita or Gita se baat kr rhe the ghumte huye
10	86.7%	Maine Ram ko kisi or ke ghar me dekha

The subjects were told to answer as fast as possible while being accurate. The results for the filler questions were as expected as there was no ambiguity present in them, and all the participants picked the right response. The reaction time results of the DMDX experiment for the test sentences are given below:

Test Sentences given in the code	Avg. RT taken by majority (in ms)	Avg. RT taken by the rest (in ms)
3	10209.35	0.0
5	9751.66	10599.33
6	9412.97	10284.01
7	11267.43	10103.83
9	8707.29	8943.87
11	11983.71	0.0
13	9552.79	10721.42
15	10649.74	12433.94
19	11075.26	0.0
21	8054.96	6980.73

Discussion

We have some interesting results on our hands. All the test sentences were made to be as ambiguous as possible. We find that two of the test sentences got a nearly

50% split which indicates that those sentences were very ambiguous. The other sentences were also designed to be completely ambiguous but there are several factors such as experience, socioeconomics, etc. that affect the participant's bias. However in this test, we are not concerned with those factors. One of the participants spoke the southern dialect of Bhojpuri. His responses differed slightly from the majority who speak the northern dialect.

The anaphoric ambiguity in Bhojpuri can be resolved by the centering theory. Majority of the participants considered the semantic analysis of the sentences over syntactic analysis and they centered one entity over the other as the antecedent. The test sentences and the fillers taken in this experiment are from everyday speech of the native speakers. While in a contextual discourse most people understand the meaning of the sentences but when presented isolated with 2 different interpretations, the participants made choices they weren't sure of. However from native fluency in the language, most participants had similar results.

Word order doesn't affect the meaning of the sentences neither the informational status. It is the fluency of the native speakers with which they identify the accurate meaning of the sentence. This accuracy, however, is debatable but is ultimately provided by the participants overall performance in the experiment through which we come to the general conclusion, open ended semantic inference which draws on world knowledge and inference procedures to identify the appropriate referent is too complex and extensive and is computationally unfeasible.

When we consider the reaction time taken by the majority and the rest, we find a vague pattern which shows that the subjects who chose the other interpretation from the majority, took longer. Also when we consider the results from the DMDX experiment and we find that the reaction times for the test ambiguous sentences are significantly higher than those of their unambiguous counterparts.

PART IV: CONCLUSION

When we look at the reaction time results of the test sentences from the DMDX experiment, we can see that they are much higher than the averages of the filler sentences present. When the subjects are parsing the sentences and their interpretations, they can observe that in most cases both the options are viable due the nature of phrasing in these test sentences, and they might be forced to rethink their initial assumptions subconsciously. Since all the participants were native speakers of Bihar, it is unlikely that other linguistic factors are coming into play.