

AnnCorra : An Introduction

OUTLINE

1. BACKGROUND
2. GUIDELINES
 - 2.1. AN ILLUSTRATION
 - 2.1.1. WHAT TO MARK
 - 2.1.2. HOW TO MARK
3. DEFAULT CONVENTIONS
4. GRAMMATICAL MODEL FOLLOWED
5. TAGSETS
 - 5.1. TAGSET-1
 - 5.2. TAGSET-2
6. DOs AND DON'Ts
 - 6.1. DOs #CORRECTIONS ??
 - 6.2. DON'Ts
7. SAMPLE INPUT
8. SAMPLE ENTRY

1.BACKGROUND

AnnCorra is a project that was decided to be taken up for developing Lexical Resources for Indian Languages(LERIL), at the "Workshop on Lexical Resources for Natural Language Processing", 5 - 8 Jan 2001, held at IIIT Hyderabad.

The name AnnCorra, shortened for "Annotated Corpora", is for an electronic lexical resource of annotated corpora. The purpose behind this effort is to fill the lacuna in such resources for Indian languages. It will be an important resource for the development of Indian language parsers, machine learning of grammars, lakshancharts (discrimination nets for sense disambiguation) and a host of other tools.

This is a project of LERIL (Lexical Resources for Indian Languages), an open-source, collaborative initiative of several groups

(and individuals) to create shareable resources for Indian languages. Another project, TransLexGram, is already underway

The AnnCorra effort is being started based on the electronic corpora available freely for various Indian languages. One such resource is the English-Hindi Electronic Dictionary developed through a voluntary collaborative effort Co-ordinated by Language Technologies Research Centre, Indian Institute of Information Technology, Hyderabad. Another resource is an electronic corpus of Hindi developed by Ministry of Information Technology, Government of India.

Like TransLexGram the present task is also a collaborative effort among individuals and institutions. The resources so developed will be available as a "free" resource under GPL (General Public License).

The effort is being coordinated by a steering committee coordinated by the natural language technology team at NCST, Mumbai. If you wish to join the effort, send an email to <leril@ncst.ernet.in>.

2. GUIDELINES

The effort requires you to do the following

1. Analyse the sentences
2. Mark the tags expressing the analysis.(tagset is provided)
3. If a sentence is generally ambiguous, but has a single meaning in a given context, then only that meaning should be marked.

The task can be better understood by looking at some examples.

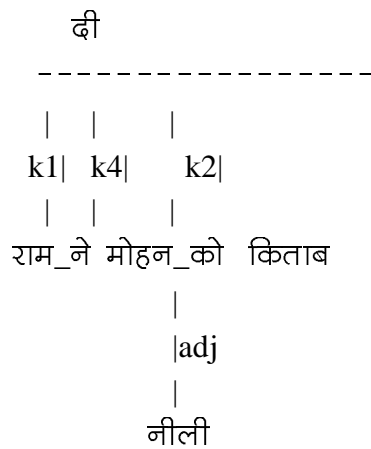
2.1. AN ILLUSTRATION

Here is a sentence from Hindi

0:: राम ने मोहन को नीली किताब दी

Tree-1 is a representation of verb, argument relationship within

the various constituents of the sentences -



Tree-1

2.1.1. WHAT TO MARK

Since the objective of tagging is to explicitly mark the relationships between various components of a sentence, therefore, verbs and their arguments have to be marked. If needed, some relationships between nouns and other grammatical categories such as adjectives are also to be marked.

2.1.2. HOW TO MARK

The information represented in the tree above can also be represented in a linear fashion. The tags showing the branch can be marked after the constituent they refer to.

- First, the elements forming a certain relationship should be bracketted within square brackets ([]),

राम_ने मोहन_को [नीली किताब] दी

- Next mark the appropriate tag markers

राम_ने/ मोहन_को/ [नीली किताब]/ दी::

NOTE - Symbol '/' denotes an ARC tag and '::' denotes a NODE tag (explained in greater details under TAGSETS)

- Then type in the required tagname

राम_ने/k१ मोहन_को/k4 [नीली किताब]/k२ दी::v

Following tags (most of which are based on Paninian grammatical model) have been used above. (A more comprehensive list of tags is given under TAGSETS):

k१ : कर्ता

k२ : कर्म

k४ : सम्प्रदान

V : क्रिया

The idea here is to mark only the specific grammatical information.

Certain DEFAULT CONVENTIONS are left unmarked. For example the adjective 'नीली' of 'किताब' has been left unmarked in the above example.

Following DEFAULT CONVENTIONS will save unnecessary typing efforts.

3. DEFAULT CONVENTIONS

1) Within paranthesis, right most element is the Head.

Example - in the constituent [नीली किताब] noun 'किताब' would be the Head.

2) In case the noun is followed by a postposition (vibhakti etc.) it should be included in the parantheses and the unit noun_vibhakti remains the head.

Example - [नीली किताब_में],

The the noun 'kitAba' is followed by a postposition 'meM', the head, in this case, is 'kitAba_meM' and not 'meM' alone.

3) If the number of elements within parantheses is more than one, then by default all of them are to be taken as modifiers of the head.

Example - [मेरी नीली किताब], both 'merI' and 'nIII' are modifying the Head 'kitAba'.

4) In case the number of elements within paranthesis is more than two(Head plus two) and one or more of them do not modify the head then it should be marked.

Example - [हल्की नीली किताब],

Here, 'halkI' can qualify both 'nIII' and 'kitAba'. In case it is modifying 'kitAba', say, in terms of light weight, then it should be left unmarked. But if it modifies 'nIII', in terms of light shade, then it SHOULD be marked. Mark this by adding 'ᳵ' on the right of 'halkI' [हल्कीऐ नीली किताब].

Symbol 'ᳵ' indicates that the element immediately on its right is modified.

5) Karakas attach to the nearest verb on the right (inflected-kriya or kridanta).

[राम_ने/k१ खाना/k२ खाकर::Kr पानी/k२ पिया::v]<s>

There are two 'k2s', in the above example and two verbs (Vkr and V). By default, therefore, first '[khAnA]/k2' will attach itself to

[khAkar]::vkr' (the nearest verb) and [pAnI]/k2' to [piyA]::v'.

- 6) Karta karaka (k1) has a special default rule. If there is only one k1' and more than one verbs, then the default is that k1' should attach to the main verb. Example -
[राम_ने/k१ खाना/k२ खाकर::Kr पानी/k२ पिया::v]<s>

Though, semantically the element [rAma ne]/k1' is the agent for both 'khAkar' and 'piyA' but sense agreement and its vibhakti are controlled by the second verb therefore, it will be considered as attaching itself to the main verb 'piyA'.
Example - [राम खाना खाकर पानी पीता है]<s>
[सीता खाना खाकर पानी पीती है]<s>

- 7) kridanta attaches to immediately succeeding noun or verb (depending on the type of kridanta)

For example in Hindi 'kar-kridanta' attaches itself to another verb, eg. [राम_ने/k१ खाना/k२ खाकर::Kr:i पानी/k२ पिया::v:i]<s>.

'i' in the above example indicates that 'KAkar' attaches itself to the other 'i' element, 'V:i'.

But since it is a default it need not be marked. So the entry would be, [राम_ने/k१ खाना/k२ खाकर::vkr पानी/k२ पिया::v]<s>

However, participle form 'A_huA', in Hindi, can modify both, a noun and a verb. For example take the Hindi sentence -

[मैंने/k१ दौड़ते_हुए::Kr घोड़े_को/k२ देखा::v]<s>

This is an ambiguous sentence having two senses

- a) [मैंने [दौड़ते हुए]:i घोड़े:i को देखा]<s> ; as in,
[मैंने दौड़ता_हुआ घोड़ा देखा]<s>
- b) [मैंने [दौड़ते हुए]:i घोड़े को देखा:i]<s> ; as in,
[मैंने दौड़ते_हुए घोड़ा देखा]<s>

As earlier, symbol ':i' in the above sentences (a and b) indicates the element to which the participle form 'दौड़ते_हुए' attaches itself.

In a) the meaning is 'I saw a horse while the horse was running'
and in b) the sense is 'While I was running I saw the horse'.

Therefore, by default, a) will not be marked but b) WILL BE MARKED.

Please NOTE that in such sentences (which are ambiguous in isolation), the user should judge the correct meaning in the given context and mark appropriately.

4. GRAMMATICAL MODEL FOLLOWED

Paninian grammatical model has been chosen here for sentence analysis, hence for the tagnames as well. Preference for this model is based on the following factors -

- 1) Being based on analysis of an Indian language it can deal better with the type of constructions Indian languages have. Therefore is more appropriate for Indian language analysis.
- 2) The model not only offers a mechanism for SYNTACTIC analysis but also incorporates the SEMANTIC information (nowadays called dependency analysis). Thus making the relationships more transparent.

5. TAGSETS

The tagsets used here have been divided into two categories -

- 1) TAGSET-1 - Tags which express relationships are marked by a preceding '/'.
For example kArakas are grammatical relationships, thus they are marked '/k1', '/k2', '/k3' etc.

2) TAGSET-2 - Tags expressing type of node are marked by a preceding ':'

Verbs etc. are nodes, so they will be marked ':v'

NOTE : a) Items marked '***' in the lists below are OPEN FOR DISCUSSION.

b) More tags can be added as and when the need comes.

5.1. TAGSET-1 (Expressing relationship labels) Marked '/'

s : Sentence

Example - [राम ने खीर खायी]<s>

k१: कर्ता

Example - [राम_ने/k१ खीर खायी]<s>

k२: कर्म

Example - [राम_ने खीर/k२ खायी]<s>

k३: करण

Example - [राम_ने चम्मच_से/k३ खीर खायी]<s>

k४: सम्प्रदान

Example - [राम_ने मोहन_को/k४ खीर दी]<s>

k५: अपादान

Example - [राम_ने क१/रोरी_से/k५ चम्मच_से खीर खायी]<s>

h: हेतु

Example - [मोहन [व्यवसायिक लक्ष्य_से]/h काम करता है]<s>

t: तादर्थ्य

Example - [मोहन पढ़ने_के_लिये/t स्कूल जाता है]<s>

k७.१: कालाधिकरण

Examples - [कल/k७.१ पानी बरसा]<s>

[[उस जमाने_में]/k७.१ मँहगाई कम थी]<s>

[बचपन_में/k७.१ वह बहुत शैतान था]<s>

[पहले/k७.१ राम आया]<s>

k७.२: देशाधिकरण

Examples - [मेज़_पर/k७.२ किताब है]<s>

[हवा_में/k७.२ ३/४ डक है]<s>

k७.३: विषयाधिकरण या अन्य ??? ***

Examples - [बहुत से युवा [इस स्वतन्त्रता संग्राम_में]/

k७.३ हिस्सा लिया]<s>

[उन्होंने अपने शिष्य को आश्रम की सेवाओं से

[मुक्त करने_में]/k७.३ संकोच नहीं किया।.<s>

k1ud: कर्ता-उद्देश्य

Example - [धनिया/k1ud इतनी व्यवहारकुशल न थी]<s>

k1vid: कर्ता-विधेय

Example - [धनिया [इतनी व्यवहारकुशल]/k1vid न थी]<s>

k2ud: कर्म-उद्देश्य

Example - [राम मोहन_को/k2ud बुद्धिमान मानता है]<s>

k2vid: कर्म-विधेय

Example - [राम मोहन_को बुद्धिमान/k2vid मानता है]<s>

Vjt: ज्यों-त्यों/जब-तब समानकालिकत्व सम्बन्ध

Example - [ज्यों-ज्यों पुस्तक की कीमत बढ़_रही_है/Vjt:i

त्यों-त्यों पा३/४क की क्रय शक्ति घ1/2_रही_है ।]/Vjt:i]<s>

up: उपपद

Examples - [राम/k1:i मोहन_के_साथ/up:i गया]<s>

[राम ने किताब_के_साथ/up:i पेन/k2:i खरीदा]<s>

[पेड़_के_(c)पर/up पक्षी उड़ रहा है]<s>

[राम_के_प्रति/up मोहन को श्रद्धा है]<s>

sdr: सादृश्य

Examples - [पुत्र पिता जैसा<s>dr है]<s>

६: षष् ३/४ी

Examples - [सम्मान_का/६ भाव]

[पुस्तक_की/६ कीमत]

[पा३/४क_की/६ क्रय शक्ति]

k1udj: 'जो' वाक्य वाला उद्देश्य

Example - [[जिसने काम किया है] वह]/k1udj राम है]<s>

k1vidj: 'जो' वाक्य वाला विधेय

Example - [जिसने काम किया है] वह राम/k1vidj है]<s>

jovo : Relative clause modifiers

Example - [जिसने काम किया है]/jovo वह राम है]<s>

adj: विशेषण

Example - [किताब मैंने देखी नीली_सी/adj]<s>

Krv: क्रियाविशेषण

Examples - बराबर

तेजी_से

हल्के_से

?? : UNABLE TO DECIDE ***

Example - [फलतः/?? वह असफल हो गया]<s>

5.2. TAGSET-2 (for nodes) Marked ':'

v: क्रिया

Kr: कृदन्त

Examples - [मीरा के आते_ही::Kr मोहन चला गया]s

[खाना खाकर::Kr राम ने पानी पिया]s

vH: है

Example - राम अध्यापक है

nr: निर्धारण(superlatives में) ***

Example - [सबसे::nr [महत्वपूर्ण प्रश्न]]

vibh : विभक्त

Examples - [राम_से_ज्यादा/vibh मोहन बुद्धिमान है]s

Examples - [मोहन राम_से_कम/vibh बात करता है]s

qs: प्रश्नवाचक

Examples - [कौन/qs आया है?]s

[राम क्या/qs खा रहा है]s

inj : interjections

Examples - अरे!

बाप रे!

yo: योजक

Example - राम और/yo श्याम

yok२: वाक्यकर्म योजक ***

Example - [राम ने कहा कि/yok२ वह नहीं आ पाएगा]s

i : co-indexed

6. DOs AND DONTs

6.1. DOs #CORRECTIONS

a) In case auxiliary verbs, inflectional suffixes, vibhakti etc. are written leaving spaces in between (like in Hindi), fill the spaces by underscores. Example - जा रहा है should be conjoined by underscores, thus the final entry would be, जा_रहा_है. Similarly, in Hindi, the vibhakti should be conjoined by an underscore with the preceding noun, eg, 'राम ने' should be marked 'राम_ने'.

b) In case of Hindi, some people use the convention of attaching vibhakti to nouns. Example 'लड़केने'. For the sake of uniformity, please insert a '_' between the noun and its vibhakti. Therefore, 'लड़केने' should be changed to 'लड़के_ने' .

But please NOTE that vibhakti after a pronoun SHOULD NOT be changed.

For example - 'उसने' remains 'उसने'. DO NOT make it 'उस_ने'.

c) Correct errors relating to missing spaces between words.

For example - 'सहीसंख्या' should be corrected as 'सही संख्या'.

d) Emphatic markers such as 'ही', 'तो', 'भी' etc, in Hindi should be included within the parantheses of the preceding head and should be attached with an underscore.

For example -

[बड़े लड़कों ने ही किताब खरीदी] should be marked as

[[बड़े लड़कों_ने_ही]/k१ किताब/k२ खरीदी::v]<s>

e) In case you are not sure about the tag that a particular constituent should take mark it '??'

Example - फलतः/?? वह असफल हो गया

6.2. DONTs

incorrect except those mentioned in 6.1. Any corrections or

suggestions that you consider should be incorporated, write it in the COMMENT field provided after each sentence. (The only corrections being permitted in the sentence itself are listed in 6.1.).

7. SAMPLE INPUT

As extracted from the corpus as given to you.

SENT:: प्रकाशक व्यवसाय में सबसे महत्वपूर्ण प्रश्न पुस्तकों की बिक्री हैं।

COMMENT::

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

SENT:: प३/४न रुचि (रीडिंग हैबि १/२) का तो अभाव नहीं था किन्तु श्रेष् ३/४ पुस्तकों का प्रकाशन स्वल्प मात्रा में होता था।

COMMENT::

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

SENT:: मुद्रित पुस्तकोंकी सहीसंख्या की जानकारी तो स्थापित प्रकाशक भी नहीं देते।

COMMENT::

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

SENT:: फलतः ऐसे फसलीप्रकाशकों से लेखक अपनी रायल् १/२ से वंचित रह जाते हैं।

COMMENT::

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

8.SAMPLE ENTRY

(you will mark tags as shown here)

SENT:: [[प्रकाशक व्यवसाय_में]/k7.3 [सबसे::nr महत्वपूर्ण प्रश्न]/k1ud
[पुस्तकों_की/६ बिक्री]/k1vid हैं::vH |]<s>

COMMENT:: Verb 'हैं' is wrongly spelled. It should be 'है'.

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

SENT:: [[[प३/४न रुचि (रीडिंग- हैबि1/2)_का_तो] अभाव]/k1 [नहीं था]/VH
किन्तु/yo [[श्रेष्३/४ पुस्तकों_का]/६ प्रकाशन]/k1
[स्वल्प मात्रा_में]/k7.3 [होता था |]::v]<s>

COMMENT::

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

SENT:: [[[मुद्रित पुस्तकों_की]/6:i [सही संख्या की]/6:i जानकारी]/k२ तो
[स्थापित प्रकाशक_भी]/k1 [नहीं देते |]::v]<s>

COMMENT::

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

SENT:: ***फलत:/h [ऐसे फसली प्रकाशकों_से]/h लेखक/k1 [अपनी रायल्१/री_से]/k5
वंचित_रह_जाते_हैं |::v

COMMENT::

***OPEN FOR DISCUSSION

XXXXXXXXXXXXXXXXXXXXXXXXXXXX