===
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 17:10 October 2017
UGC Approved List of Journals Serial Number 49042
==

Additive Constructions in Kannada Language Using Ancorra Annotation Scheme

RoopasriNagathi R., Research Scholar, HCU, Hyderabad

===

Abstract

The present study investigates the annotation of Kannada additive constructions based on Ancorra annotation scheme/ guidelines. Ancorra annotation scheme is an annotation scheme based on Paninian Grammar Formalism, which is developed to annotate Indian languages. It makes easy to annotate Indian languages and said to be an extensive application on both free word-order languages and positional languages. It can be placed in Dependency approach, one of the syntactic approaches, which currently became a mainstream research area in both linguistics and computer science disciplines.

Additive constructions are coordinating constructions that use ‘and’ in English sentences. In Kannada, other than conjunct words such as *mattu* and *hāgu* ‘and’, other similar constructions use comma and alternatively the inclusive clitic ‘*ū’* etc. to indicate additive constructions. The main aim is to identify various issues while annotating the additive constructions in the Kannada language. The present study of Kannada additive constructions is based on a corpus of nearly hundred annotated sentences. The purpose of the article is an attempt to exhibit the annotation of the additive conjuncts and follow the labels that have given in Ancorra guidelines. The article presents the application that extends the Paninian Grammar Formalism to Kannada.

Keywords: Annotation, Additive Conjunct, Kannada, Ancorra Guidelines

Introduction

The present article is an effort to develop a Treebank for the Kannada language. A Treebank is a collection of syntactically annotated sentences in which the annotation has been manually checked. These syntactically annotated sentences are described and often come in the form of tree structures. The annotated linguistic resource is essential to develop the Natural Language Process (NLP). It (linguistics resources) can be used in both basic as well as advanced applications, such as Machine Translation, Parser, etc.

There are various Treebanks are available in preparing such annotated linguistic resources, most notably the Penn Treebank (PTB) for English, Prague Dependency Treebank (PDT) for Czech, Alpino for Dutch, and English, TUT for Italian, TIGER for German, Latin Dependency Treebank (LDT) for Latin, Hyderabad Dependency Treebank etc.

Lack of such linguistics resources has been a major drawback for developing natural language processing tools. These resource-rich Treebank has been developed by using phrase structure approach or dependency approach.

According to literature, the use of Phrase Structure (PS) is not suited for free-word-order languages rather than, the dependency approach which would be better suited (A. Bharathi and R. Sangal, 1993; A. Bharathi., V. Chaitnya and R. Sangal, 1995,). The effort described in this paper follows the Paninian grammatical framework (PGF) which has been successfully used for Hindi, Telugu, Bangla, Malayalam, Urdu, etc., languages. This Framework also worked well for positional language like English (A. Bharathi, V. Chaitnya and R. Sangal 1997; A. Vaidya, S. Husain, P. Mannem, and D.M. Sharma. 2009; Chaudhry, H., & Sharma, D. M. 2011)

Since Kannada lacks such resource (based on available literature) to develop a good natural language processing tool for Kannada, one has to develop linguistics resources. Hence, it is an initial effort to come up with linguistics resources. The Kannada language falls into Free-Word order (FWO) constructions, the annotation is based on dependency annotation. Here, this paper introduces the Ancorra Guidelines/ Hyderabad Dependency Treebank for data annotation that is developed based on PGF. The annotation can be done at various levels of linguistics. Since the coordinate structure comes under syntactic level, the article focuses on the syntactic level. An annotation is a process of description or explanation or comment applied to the raw data.

Kannada is one of the major Dravidian languages with relatively free word-order, spoken in South India. It has its own script. It is also morphologically very rich. A significant amount of discussion on Kannada is available but the discussion about coordination structure is limited in both linguistically as well as computational linguistically. So it is an attempt to bring out the structure of coordination especially additive constructions in Kannada and the annotation of them based on PGM.

Additive Conjunct in Kannada

There are only a small number of coordinating conjunctions in Kannada. Coordination refers to the combination of the similar syntactic unit into some larger group of the same category that attached through coordination conjunctions like and, or, but. Additive construction is one of such coordinating sentences in which meaning "and" is expressed (S. N. Sridhar 1990). One of the most frequently expressed coordination in Kannada (S. N. Sridhar 1990) is additive conjunct where addition is described as one of the most basic forms of cohesion (Halliday M. A. K. and Hasan R. 1976).

Additive conjunct is the most basic forms because it is acquired early by children than other conjunct types. In Kannada, additive conjunct can form through conjunct words or suffixes. Those are lexical item *mattu* and *hāgu* ‘and’, ‘inclusive clitic *–ū*’, and as *dhīrgās* ‘lengthened vowels’, and also putting the comma or without a comma after each coordinate element in a (written) sentence. Few model sentences have mentioned below:

(1) rādha   kālēji   -ge,   rāNi   skūl   -ge   hōdaru.

Radha   college   -dat   Rani   school   -dat   go-past-3p,pl,

‘Radha went to the college and Rani went to the school.’

(2) rādha   kālēji   -ge   mattu   rāNi   skūl   -ge   hōdaru.

Radha   college   - dat   and   Rani   school   -dat   go-past-3p,pl,

‘Radha went to the college and Rani went to the School.’

(3) rādha   kālēji   -g -ū   mattu   rāNi   skūl   -g-ū   hōdaru.

Radha   college   - dat-inc   and   Rani   school   -dat-inc   go-past-3p,pl,

‘Radha went to the college and Rani went to the School.’

The sentences mentioned above giving same meaning with different coordinate expressions. The sentence (1) is one of the additive constructions that expressed by simply juxtaposition (S. N. Sridhar 1990). The sentences express the coordination through with or without a comma (note: wherever the comma occurs, it can be optional. It is just a writing convention). Similarly the sentence (2) conveys the coordination through the word *mattu* ‘and’, where as the sentence (3) expresses the same through inclusive clitic *–ū*. The information of conjunction is explicit in sentences (2) and (3).

An additive conjunction can also form by using the word *hāgu*. More or less it has the meaning of ‘and'. It can be used as synonymy of *mattu*. The word *hāgu* can be used in a similar fashion that has mentioned above to conjoin the elements in the sentence (2). Hence, the word *mattu* can be replaced by the word *hāgu.* The same has mentioned below in (4) and (5). Only the destination differs between (4) and (5).The sentence (4) states that there are two different destinations: college and school which have occurred with two different proper nouns or agent that they prefer to go has conjoined with conjunct word, where as in sentence (5), school only the destination for two different agents that conjoined with conjunct word.

(4) rādha kālēji -ge hāgu rāNi skūl -ge hōdaru.

Radha college - dat and Rani school -dat go-past-3p,pl,

‘Radha went to the college and Rani went to the School.’

(5) rādha hāgu rāNi skūl -ge hōdaru.

Radha and Rani school -dat go-past-3p,pl,

‘Radha and Rani went to the School.’

Another set of coordination construction in Kannada is expressed through *dhīrgās* which can include conjunct two elements in a sentence. *dhīrgās* in sentences occur by lengthening the vowel at the end of the lexical items. In such constructions, the information of conjunction is implicit.

(6) rāju siLLe hākuttā negiyuttā manege banda.

Raju whistle put-n.past-ptpl jump- n.past-ptpl house-dat come-past-3p,sg,m

‘Raju came to home by whistling and dancing.’

The kind of construction that has mentioned in (6) also comes under additive constructions. In such constructions, the *dhīrgās* indicates the multiple actions of the single agent at the given time.

In sentence (6), dance, as well as whistling has done at the same time by a single agent, Raju. In such construction, the *karta* will be obligatorily deleted in every conjunct except in the first conjunct (i.e., subject phrase) and the verbs in the sentence will turn in to participle (non-past) except final verb. Hence, the verbs in such constructions are in non finite that indicated by *dhīrgās*.

The sentences that mentioned above in (1), (2), (3), and (4), have only one action *hōdaru* "go" and two coordinate elements *rādha kālēgige* and *rāNi skūlge*. What happens when more than two elements are coordinated? Let's look at few more examples:

(7) rādha kālēji -ge, rāNi skūl -ge, raju āpīsi -ge hōdaru.

Radha college -dat, Rani school -dat, Raju office -dat go-past-3p,pl,

'Radha went to the college; Rani went to the School and Raju went to the Office.'

(8) rādha kālēji -g -ū rāNi skūl -g -ū raju āpīsi -g -ū

Radha college -dat -inc Rani school -dat -inc Raju office -dat -inc

hōdaru.

go-past-3p,pl,

'Radha went to the college; Rani went to the School, and Raju went to the Office.'

(9) rādha kālēji -g -ū rāNi skūl -g -ū mattu raju āpīsi

Radha college -dat -inc Rani school -dat -inc and Raju office

-g -ū hōdaru.

-dat -inc go-past-3p,pl,

'Radha went to the college; Rani went to the School, and Raju went to the Office.'

(10) rādha kālēji -g -ū rāNi skūl -g -ū raju āpīsi

Radha college -dat -inc Rani school -dat -inc Raju office

-g -ū hōdaru.

-dat -inc go-past-3p,pl,

‘Radha went to the college; Rani went to the School, and Raju went to the Office.'

(11) rādha kālēji -ge, rāNi skūl -ge, mattu raju āpīsi

Radha college -dat Rani school -dat and Raju office

-ge hōdaru.

-dat go-past-3p,pl,

‘Radha went to the college; Rani went to the School, and Raju went to the Office.'

Any number of clauses or phrases can be conjoined by using a comma and/or inclusive clitic *–ū* but the word *mattu* ‘and’ cannot occur after every element. The same is mentioned from the sentence (7) to (10). It can occur only at final element of coordination. The same reflected in the sentences (9) and (10), the word *mattu* ‘and’ has occurred only at the end. The one more point to notice in the (9) and (10) is that the word *mattu* ‘and’ can occur only at the final coordinate element with both a comma (10) and inclusive clitic (9) (S. N. Sridhar 1990).

Similarly, if the sentence has more sequential actions by a single agent (*karta*), then each coordinate element will be having either comma or inclusive clitic. But in such sentences, the *karta* will be obligatorily deleted in all the conjuncts except in first conjunct (i.e., subject phrase) and the verbs in the sentence will turn in to participle except final verb, which is similar to the constructions that has *dīrga (*sentence 6). See the examples below:

(12) rādha mane -ge hōg-i, baTTe badalāyisi, kelasa mugisi,

Radha house -dat go-pp clothes change-pp work complete-pp

malagidaLu.

sleep-past-3p,sg.f

‘Radha went home, changed her dress, finished her work and slept.’

(13)	rādha	mane	-ge	hōg-i,	baTTe- nū	badalāyisi,	kelasa- nū	mugisi,
	Radha	house	-dat	go-pp	clothes-inc	change-pp	work-inc	complete-pp

malagidaLu.

sleep-past-3p,sg.f

‘Radha went home, changed her dress, finished her work and slept.’

Since sentence (11) and (12) has multiple actions by a single agent, the word *mattu* ‘and’ has not occurred either in the last coordinate element or in each coordinate element. It is just opposite to sentences (9) and (10) which have karta for each action. And sentence (11) and (12) also differ with the sentence (6). Actions in the sentence (6) are simultaneous whereas in sentence (11) and (12), an actions took place one after the other. However, in both the cases, the agent got deleted except in first coordinate element.

Annotation

Nearly 100 sentences have been taken for annotation which consists of the coordinate element. The annotation is done using Sanchay tool (annotation interface), in Shakti Standard Format (SSF). The sentences have been run in the Sanchay Tool, where the sentence will split into tokens and then need to add information like Morphological Analyzer (MA), Parts Of Speech (POS) tag and chunk to the tokens. After that annotator can build a tree for the sentences. Dependency Relations is given to indicate the type of the relation that the tokens has. It is annotator’s job to annotate the dependency relations. Hence, by using Sanchay tool, one can have information from MA to dependency relation.

As mentioned, the annotation has done by following the Ancorra guidelines (A. Barathi, et al 2002; R. Begum, et al 2008) in which the Paninian Grammar has taken as a base. The grammar captures certain syntactic- semantic relations. In Paninian grammatical model, each word should belong to the list of either modifier or modified. The PGM has mentioned two kinds of dependency relations: karaka and non-karaka relations. Based on these relations, the labels have given in Ancorra guidelines. There are nearly 40 labels identified considering various sentence types. Along with considering the karaka and non-karaka relations, it is also considered some tags that don't have dependency relations. A few under specified tags of the type are vmod, nmod and ect.

Annotation of additive construction comes under non-karaka relations in Ancorra guidelines. It is also called as Hyderabad Dependency Treebank (HyDT). Based on PGM, HyDT is developed for Hindi. HyDT is developed as a part of LTRC project. It has provided a guidelines called Ancorra for annotation of Indian languages as well as other than Indian languages. The additive construction is tagged as ‘CCP’ in HyDT.

This tag ‘CCP’ is used for both coordinate and subordinate conjunctions. The conjunct is annotated as head in HyDT and takes the coordinating elements as its children. The usually the relationship between the coordinate elements and its children is named as ccof. The annotated tree is given below for the sentence (2) *rādha kālēgi-ge mattu rāNi skūl-ge hōdaru.*

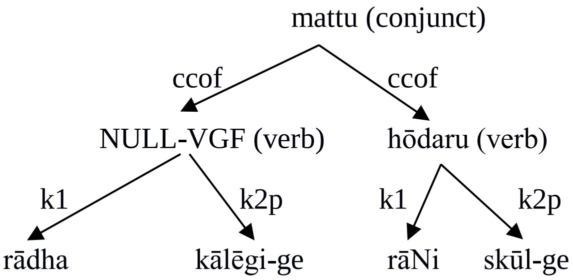

Fig-1

Here the word *mattu,* a simple coordinating conjunct gives explicit information of conjunct and tagged as CCP. The CCP is the head with a coordinate element as its children. In the fig-1, the relation between *mattu*, and the verb is tagged as ccof. It does not reflect the kāraka relation in HyDT. Hence, does not start by ‘k’ which indicates the kāraka relation. Similarly, the *mattu* can be replaced by the lexical word *hāgu* ‘and’, which can also expresses the coordination explicitly and will have similar tree.

There are sentences which mentioned in this article stands for implicit conjunct words i.e. the sentence does not take explicit additive conjunct *mattu* or *hāgu* but takes comma, inclusive clitic *–ū* and *dhīrgās*. which is mentioned above in sentence (1), (3), (4), (5), (6). In such cases, the conjunctive is annotated as NULL-CCP. The fig-2 gives the idea of implicit additive conjunct in a sentence. The NULL_CCP is a tag which is given by HyDT, should insert at chunk level in Sanchay tool[3]. The following tree represents same but with the different tag.

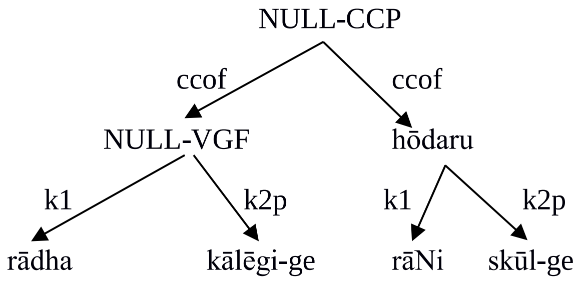

Fig-2

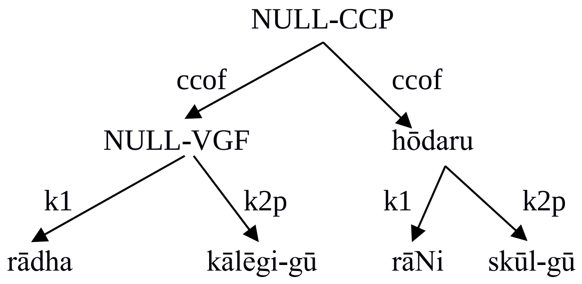

Fig-3

The NULL-CCP tag is used when the information of conjunction is implicit, such as mentioned in sentence (1) and (6). In the sentence (1) we have a comma and in (6) *dhīrgās* without having an overt conjunct word. The sentence (2) and (3) are an example for the explicit conjunct, is tagged as CCP at the tree level.

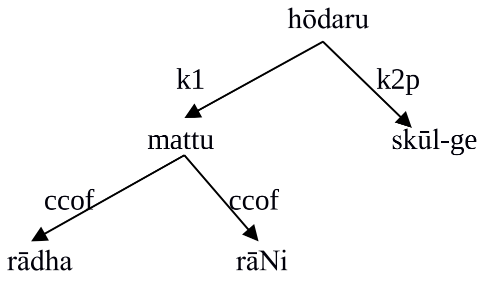

Fig-4

The tree in Fig-4 differs with previous trees by the position that placed in the tree. But the CCP has occurred under the verb nod and has k1 relation. Since verb is considered central to the sentence in PGM, it is the head and participants occur under it. Here the conjunct occurs between two agents. Since conjunct word corresponds to participants and attaches with the verb node, it gets actual karaka relation that supposed to get with a noun and the relation between conjunct and agent is said to be ccof. Similarly, if the conjunct is implicit in this manner, then we can use the NULL-CCP in the place of CCP. The tree would be like Fig-5.

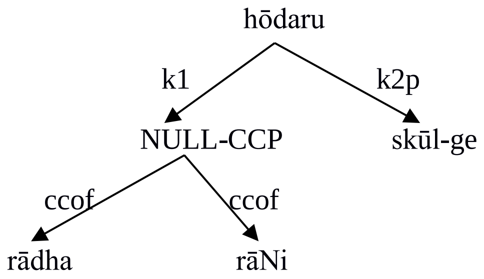

Fig-5

Conclusions and Future Work

The paper has introduced the annotation scheme and applied for coordination construction in the Kannada language. Coordination construction comes under non-karaka relations in PGM. In Kannada, coordinating constructions are marked by conjunct words or suffixes like lengthening, inclusive clitic, or comma. Hence, it is explicit in some cases and implicit in some other cases. For explicit, the HyDT/ Ancorra guidelines has given CCP tag and in implicit construction, one has to insert the node at chunk level and annotate as NULL_CCP, which is also given in HyDT. Kannada has used both tags for annotation.

Data can be further run in to malt parser or some other parser. The out come of the parser can be presented as a paper.

===

Note:

1. Panini is one of the great Indian Grammarians, worked on Sanskrit and used the concept of dependency 2500 years ago.

2. kāraka is the term, given by Panini for Sanskrit language. It refers to the relationship between a noun and a verb in a sentence.

3. Sanchay is an open source platform for working on languages, with components like a text editor with customizable support for languages and encodings, annotation, interfaces, etc.

4. The word Ancorra can also be used as AnnCorra in this paper.

===

References

[1]. A. Bharati, and R. Sangal. 1993. "Parsing Free Word Order Languages in the Paninian Framework." Proceedings of the 31st annual meeting of Association for Computational Linguistics Association for Computational Linguistics. June 1993, pp. 105-111.

[2]. A. Bharati, V. Chaitanya and R. Sangal. 1995. Natural Language Processing: A Paninian Perspective, New Delhi: Prentice-Hall of India.

[3]. A. Bharati, Medhavi Bhatia, V. Chaitanya, and R. Sangal. 1997. Paninian Grammar Framework Applied to English. *South Asian Language Review*

[4]. A. Bharati, R. Sangal, V. Chaitanya, A. Kulkarni, D. M.Sharma, and K. V. Ramakrishnamacharyulu. 2002. "AnnCorra: building tree-banks in Indian languages", Proceedings of the 3rd workshop on Asian language resources and international standardization-V.12 Association for Computational Linguistics, August 2002, pp. 1-8.

[5]. A. Bharati , A. Kulkarni. 2005. English from Hindi viewpoint: A Paaninian Perspective. Platinum Jubilee conference of Linguistic Society of India at HCU, Hyderabad.

[6]. A.Bharati, R.Sangal, and D.M. Sharma. 2006. Shakti- Analyser: SSF Representation. *IIT Hyderabad*.

[7]. A. Vaidya, S. Husain, P. Mannem, and D.M. Sharma. 2009. "A karaka-based dependency annotation scheme for English". CICLing LNCS, vol. 5449, A. Gelbukh, Eds. Springer, Heidelberg 2009, pp. 41–52.

[8]. Chaudhry, H., & Sharma, D. M. 2011. Annotation and issues in building an english dependency treebank. In *Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Chennai*.

[9]. Halliday M. A. K. and Hasan R. 1976. Cohesion in English. London: Longman.

[10]. Polguère, A. (Ed.). Dependency in Linguistic Description (Vol. 111). John Benjamins Publishing. 2009.

[11]. R. Begum, S.Husain, A. Dhwaj, D. M. Sharma, L. Bai, and R. Sangal. 2008. "Dependency Annotation Scheme for Indian Languages". Proceedings of IJCNLP-January 2008, pp. 721-726.

[12]. S. N. Sridhar, Kannada: A Descriptive Grammar. London: Routledge. 1990.

[13]. Vempaty, C., V. Naidu., S. Husain., R. Kiran., L. Bai., D. M. Sharma., and R. Sangal. 2010. "Issues in Analyzing Telugu Sentences towards Building a Telugu Treebank", International Conference on Intelligent Text Processing and Computational Linguistics, Springer Berlin Heidelberg. March 2010. vol 6008. pp. 50- 59.

14]. http://linguapress.com/grammar/conjunctions.htm accessed on 3rd April 2017.

===

RoopasriNagathi R.,
Research Scholar
Centre for Applied Linguistics and Translation Studies
School of Humanities
University of Hyderabad
Hyderabad – 500046
Telangana
India
roopasri.nagathi@gmail.com