

Subject: NLP
Experiment No:5

Aim: Perform Morphological Analysis on a word.

Theory:

Morphemes are considered as smallest meaningful units of language. These morphemes can either be a root word(play) or affix(-ed). Combination of these morphemes is called morphological process. So, word "played" is made out of 2 morphemes "play" and "-ed". Thus finding all parts of a word(morphemes) and thus describing properties of a word is called "Morphological Analysis". For example, "played" has information verb "play" and "past tense", so given word is past tense form of verb "play".

A morphological analyzer is a program that analyzes the morphology of an input word. It uses rules to identify the root and grammatical features of given words. It splits a given word into its root, lexical category, gender, number, person, case, case marker or tense aspect modality(TAM), suffix and other required grammatical features as given below.

1. root : Root of the word (e.g. ladZake word has root ladZakA)
2. cat : Category of the word (e.g. Noun=n, Pronoun=pn, Adjective=adj, verb=v, adverb=adv, post-position=psp and avvya=avy)
3. gen : Gender of the word (e.g. Singular=sg, Plural=pl, dual, and any)
4. num : Number of the word (e.g. Singular=sg, Plural=pl, dual, and any)
5. per : Person of the word (e.g. 1st Person=1, 2nd Person=2, 3rd Person=3, and any)
6. case : Case of the word (e.g. direct=d, oblique=o and any)
7. tam : Case marker for noun or Tense Aspect Mood(TAM) for verb of the word
8. suff : Suffix of the word

Analysis of a word :

बच्चों (bachchoM) = बच्चा(bachchaa)(root) + ओं(oM)(suffix)
(ओं=3 plural oblique)

A *linguistic paradigm* is the complete set of variants of a given lexeme. These variants can be classified according to shared inflectional categories (eg: number, case etc) and arranged into tables.

Paradigm for बच्चा

case/ num	singular	plural
direct	बच्चा(bachcha a)	बच्चे(bachche)
oblique	बच्चे(bachche)	बच्चों (bachchoM)

Algorithm to get बच्चों(bachchoM) from बच्चा(bachchaa)

1. Take Root बच्च(bachch)**आ(aa)**
2. Delete **आ(aa)**
3. output बच्च(bachch)
4. Add **ओं(oM)** to output
5. Return बच्चों (bachchoM)

Therefore आ is deleted and ओं is added to get बच्चों

Add-Delete table for बच्चा

Dele te	Add	Numb er	Ca se	Variants
आ(a a)	आ(aa)	sing	dr	बच्चा(bachchaa)
आ(a a)	ए(e)	plu	dr	बच्चे(bachche)
आ(a a)	ए(e)	sing	ob	बच्चे(bachche)
आ(a a)	ओं(o M)	plu	ob	बच्चों(bachcho M)

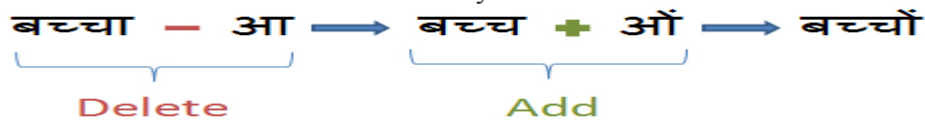
Paradigm Class

Words in the same paradigm class behave similarly, for Example लड़क is in the same paradigm class as बच्च, so लड़का would behave similarly as बच्चा as they share the same paradigm class.

बच्चा	-औं
लड़का	-औं
play	-ed
want	-ed

Conclusion:

Words can be analysed morphologically if we know all variants of a given root word. We can use an 'Add-Delete' table for this analysis.



Code:

```
import codecs
import re

#read the input file
filepath = 'hindi.txt'
f = codecs.open(filepath, encoding = 'utf-8')
text = f.read()

sentences=text.split(u"|") #since hindi sentences end with '|'
words_list = list()
for sentence in sentences:
    words = sentence.split(' ') #words are seperated by a space in hindi
    words_list += words

suffixes = {
    1:
    [u"ाएगी",u"ाएगा",u"ाओगी",u"ाओगे",u"ेंगी",u"ेंगे",u"ेंगे",u"ूंगी",u"ूंगा",u"ाती",u"नाओं",u"नाएं",u"ताओं",u"ताएं",u"ियाँ",u"ियों",u"ियां"],
    2: [u"ो",u"े",u"ू",u"ु",u"ीय",u"ि",u"ा"],
    3:
    [u"कर",u"ाओ",u"िए",u"ाई",u"ाए",u"ने",u"नी",u"ना",u"ते",u"ीं",u"ती",u"ता",u"ाँ",u"ां",u"ों",u"ें"],
    4:
    [u"ाकर",u"ाइए",u"ाई",u"ाया",u"ेगी",u"ेगा",u"ोगी",u"ोगे",u"ाने",u"ाना",u"ाते",u"ाती",u"ाता",u"ती",u"ाओं",u"ाएं",u"ुओं",u"ुएं",u"ुओं"],
    5: [u"ाएंगी",u"ाएंगे",u"ाऊंगी",u"ाऊंगा",u"ाइयाँ",u"ाइयों",u"ाइयां"],
}
```

```
stems=list()
for word in words_list:
    for L in range(1,5):
        if len(word) >= L + 1:
```

```

        for suffix in suffixes[L]:
            if word.endswith(suffix):
                word=word[:-L] #stripping the suffix from the word
                try:
                    if word[-1] == u"ि":
                        word = word[:-1] + u"ी"
                except:
                    print(word)

    if word:
        stems.append(word)

filename='stems_generated.txt'
f=codecs.open(filename,"w",encoding='utf-8') #open in write mode
for stem in stems:
    f.write(str(stem))
    f.write(u"\u0020")
f.close()

```

hindi - Notepad

File

Edit

View

भारतीय लड़किया लड़कियां अफगानिस्तान के यंग क्रिकेटर बहीर शाह ने इसी सीजन में अपना फर्स्ट क्लास डेब्यू किया और पहले ही मैच में उन्होंने डबल सेन्चुरी (256* रन) लगाकर कमाल कर दिया। बता दें कि अपनी पहली धमाकेदार पारी के साथ शाह डेब्यू मैच में सबसे अधिक रन बनाने वाले दूसरे बल्लेबाज बन गए हैं। फर्स्ट क्लास डेब्यू में सबसे अधिक रन बनाने का रिकॉर्ड अभी तक मुंबई से खेलने वाले अमोल मजूमदार (260) के नाम दर्ज है। बहीर ने अपने फर्स्ट क्लास करियर में सिर्फ चार मैच खेले हैं। जिनमें वो 831 रन बना चुके हैं, जो कि सबसे ज्यादा रन बनाने का नया वर्ल्ड रिकॉर्ड है। इस अफगान क्रिकेटर ने अपने पहले मैच में 256*, दूसरे में 34 और 11, तीसरे में 111 और 116 और चौथे में नॉटआउट 303 रनों की इनिंग खेलकर ये रिकॉर्ड अपने नाम किया। बहीर से पहले ये रिकॉर्ड ऑस्ट्रेलिया के बिल पोसफोर्ड के नाम था। जिन्होंने 94 साल पहले 1923 में अपने पहले चार फर्स्ट क्लास मैचों में 741 रन बनाए थे। अपने फर्स्ट क्लास करियर के चौथे मैच में बहीर ने नाबाद 303 रन की धमाकेदार पारी खेली। इसके साथ ही वो सबसे कम उम्र में फर्स्ट क्लास ट्रिपल सेन्चुरी लगाने वाले दुनिया के दूसरे बल्लेबाज भी बन गए हैं। बहीर ने 18 साल 251 दिन की उम्र में ये कमाल किया। सबसे कम उम्र में तिहरा शतक लगाने का रिकॉर्ड पाकिस्तान के जावेद मियांदाद के नाम है, जिन्होंने 17 साल 310 दिन की उम्र में तिहरा शतक ठोका था।

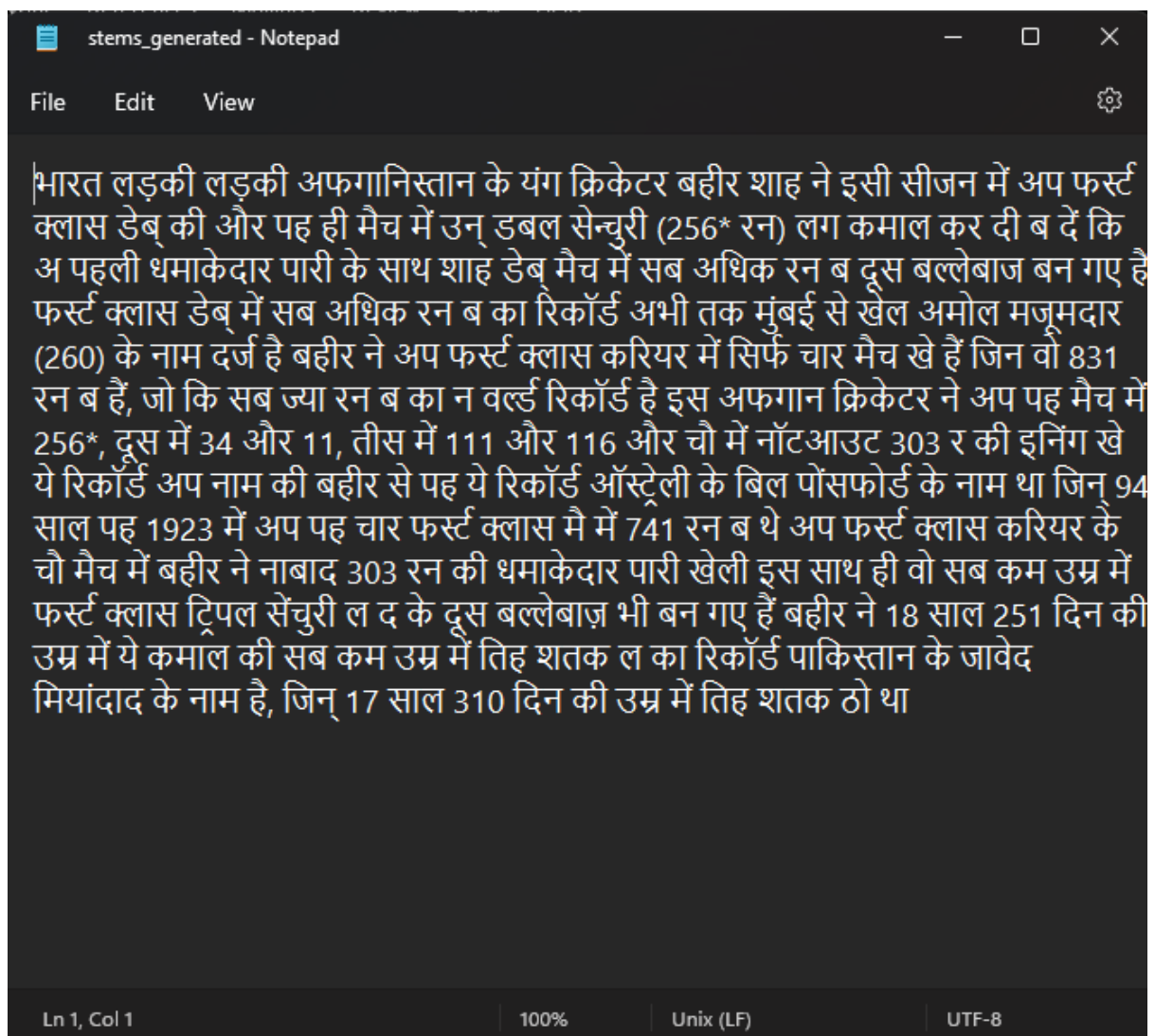
Ln 1, Col 1

100%

Unix (LF)

UTF-8

Output:



```
stems_generated - Notepad
File Edit View
|भारत लड़की लड़की अफगानिस्तान के यंग क्रिकेटर बहीर शाह ने इसी सीजन में अप फर्स्ट
क्लास डेब की और पह ही मैच में उन् डबल सेन्चुरी (256* रन) लग कमाल कर दी ब दें कि
अ पहली धमाकेदार पारी के साथ शाह डेब मैच में सब अधिक रन ब दूस बल्लेबाज बन गए है
फर्स्ट क्लास डेब में सब अधिक रन ब का रिकॉर्ड अभी तक मुंबई से खेल अमोल मजूमदार
(260) के नाम दर्ज है बहीर ने अप फर्स्ट क्लास करियर में सिर्फ चार मैच खे है जिन वो 831
रन ब है, जो कि सब ज्या रन ब का न वर्ल्ड रिकॉर्ड है इस अफगान क्रिकेटर ने अप पह मैच में
256*, दूस में 34 और 11, तीस में 111 और 116 और चौ में नॉटआउट 303 र की इनिंग खे
ये रिकॉर्ड अप नाम की बहीर से पह ये रिकॉर्ड ऑस्ट्रेली के बिल पोंसफोर्ड के नाम था जिन् 94
साल पह 1923 में अप पह चार फर्स्ट क्लास मै में 741 रन ब थे अप फर्स्ट क्लास करियर के
चौ मैच में बहीर ने नाबाद 303 रन की धमाकेदार पारी खेली इस साथ ही वो सब कम उम्र में
फर्स्ट क्लास ट्रिपल सेचुरी ल द के दूस बल्लेबाज भी बन गए है बहीर ने 18 साल 251 दिन की
उम्र में ये कमाल की सब कम उम्र में तिह शतक ल का रिकॉर्ड पाकिस्तान के जावेद
मियांदाद के नाम है, जिन् 17 साल 310 दिन की उम्र में तिह शतक ठो था
Ln 1, Col 1 | 100% | Unix (LF) | UTF-8
```

Conclusion: Hence, we have successfully performed morphology on a hindi text.