Computational Linguistics-1

Summer 2021, IIIT Hyderabad

Assignment 1

Task A

Original Sentences

COVID-19 പ്രക്ഷേപണം തടയുന്നതിന് സംസ്ഥാന സർക്കാർ ജൂൺ 5 മുതൽ ജൂൺ 9 വരെ കർശനമായ ലോക്ക്ഡൗൺ ഏർപ്പെടുത്തും.
COVID-19 prakʂe:paɳam taṭayunnatin’ samstha:na sarkka:r ju:ɳ 5 mutal ju:ɳ 9 vare karʃanama:ya lo:kḍauɳ e:rpeṭuttum.

ഈ കാലയളവിൽ അവശ്യ സേവനങ്ങൾ മാത്രമേ ഭരണകൂടം അനുവദിക്കുകയുള്ളൂവെന്ന് മുഖ്യമന്ത്രി പിണറായി വിജയൻ വ്യാഴാഴ്ച ഉന്നതതല പാൻഡെമിക് അവലോകന യോഗത്തിൽ പറഞ്ഞു.
i: ka:laya|avil avaʃya se:vanaŋŋa| ma:trame: bharaɳaku:ṭam anuvadikkukayu||u:venn’ mukhyamantri piɳaRayi vijayan vya:zha:zhca unnatatala pa:nḍemik avalo:kana yo:gattil paRaŋŋu.

ജീവിതം, മൊബിലിറ്റി, റീട്ടെയിൽ, വാണിജ്യം എന്നിവയ്ക്കുള്ള ഇളവ് ജൂൺ 4 ന് വൈകുന്നേരം 7 മണിക്ക് അവസാനിക്കും.
ji:vitam, mobiliRRi, Ri:ṭṭeyil, vaɳijyam ennivaykku||a i|av’ ju:ɳ na:lin’ vaikunne:ram 7 maɳikk’ avasa:nikkum.

അവശ്യസാധനങ്ങൾ, മരുന്നുകൾ, നിർമ്മാണ സാമഗ്രികൾ, പാക്കേജിംഗ്, വ്യാവസായിക ഉൽപാദനത്തിനുള്ള അസംസ്കൃത വസ്തുക്കൾ എന്നിവ വിൽക്കുന്ന കടകൾക്ക് മാത്രമേ ബിസിനസിന് തുറക്കാൻ കഴിയൂ.
avaʃyasa:dhanaŋ:a|, marunnuka|, nirmma:ɳa sa:magrika|, pa:kke:jimg’, vya:vasa:yika ulpa:danattinu||a asamskrta vastukka| enniva vilkkunna kaṭaka|kk’ ma:trame: bisinasin’ turakka:n kazhiyu:.

സ്ക്രാപ്പ് കൈകാര്യം ചെയ്യുന്ന സ്റ്റോറുകൾക്ക് ജൂൺ 4 ന് വൈകുന്നേരം 7 മണി വരെ ബിസിനസ്സ് നടത്താം.
skra:pp’ kaika:ryam ceyyunna sRRo:Ruka|kk’ ju:ɳ na:lin’ vaikunne:ram 7 maɳi vare bisinass’ naṭatta:m.

സർക്കാറും അർദ്ധ സർക്കാറും, പൊതുമേഖയും സ്ഥാപനങ്ങൾ, കോർപ്പറേഷനുകൾ, വിവിധ കമ്മീഷനുകൾ എന്നിവ അടച്ചിരിക്കും.
sarkka:Rum, arddha sarkka:Rum potume:khalayum stha:panaŋŋa|, ko:rppaRe:ʂanuka|, vividha kammi:ʂanuka| enniva aṭaccirikkum.

ജൂൺ 10 മുതൽ 50% ൽ താഴെ തൊഴിലാളികളുമായി അവയ്ക്ക് തുറക്കാൻ കഴിയും.
ju:ɳ 10 mutal 50% il ta:zhe tozhila:|ika|uma:yi avaykk’ turakka:n kazhiyum.

കേരളത്തിലേക്ക് പ്രവേശിക്കുന്ന ആളുകൾക്ക് COVID നെഗറ്റീവ് സർട്ടിഫിക്കറ്റ് ആവശ്യമാണ്.
ke:ra|attile:kk’ prave:ʃikkunna a:|uka|kk’ COVID negaRRi:v’ sarṭṭiphikkaRR’ a:vaʃyama:ɳ’.

മൂന്നാമത്തെ തരംഗത്തിന് സംസ്ഥാനം സ്വയം ശ്രമിച്ചു.
mu:nnamatte taramgattin’ samstha:nam svayam ʃramiccu.

ആദ്യ ഘട്ടമെന്ന നിലയിൽ, ഗസ്റ്റ് വർക്കർമാരെയും മറ്റ് ദുർബല വിഭാഗങ്ങളെയും ഭരണകൂടം പൂർണ്ണമായും കുത്തിവയ്ക്കും.
a:dya ghaṭṭamenna nilayil, gasRR' varkkarma:reyum maRR' durbala vibha:gaŋŋaḷeyum bharaṇaku:ṭam pu:rṇṇama:yum kuttivaykkum.

Tagged

COVID-19_NN1 പ്രക്ഷേപണം_NN1 തടയുന്നതിന്_VVGN-IF സംസ്ഥാന_JJ സർക്കാർ_NN1 ജൂൺ_NPM 5_MC മുതൽ_II ജൂൺ_NPM 9_MC വരെ_II കർശനമായ_JJ ലോക്ക്ഡൗൺ_NN1 ഏർപ്പെടുത്തും_VVF.
COVID-19_NN1 prakʂe:paṇam_NN1 taṭayunnatin'_VVGN-IF samstha:na_JJ sarkka:r_NN1 ju:ṇ_NPM 5_MC mutal_II ju:ṇ_NPM 9_MC vare_II karʃanama:ya_JJ lo:kḍauṇ_NN1 e:rpeṭuttum_VVF.

ഈ_DDP കാലയളവിൽ_NN1-II അവശ്യ_JJ സേവനങ്ങൾ_NN2 മാത്രമേ_RR ഭരണകൂടം_NN1 അനുവദിക്കുകയുള്ളൂ_VVF-O എന്ന്_CST മുഖ്യമന്ത്രി_NN1 പിണറായി_NP1 വിജയൻ_NP1 വ്യാഴാഴ്ച_NPD1 ഉന്നതതല_JJ പാൻഡെമിക്_NN1 അവലോകന_NN1 യോഗത്തിൽ_NN-II പറഞ്ഞു_VVD.
i:_DDP ka:laya|avil_NN1-II avaʃya_JJ se:vanaŋŋa|_NN2 ma:trame:_RR bharaṇaku:ṭam_NN1 anuvadikkukayu||u:_VVF-O enn'_CST mukhyamantri_NN1 piṇaRayi_NP1 vijayan_NP1 vya:zha:zhca_NPD1 unnatatala_JJ pa:nḍemik_NN1 avalo:kana_NN1 yo:gattil_NN-II paRaŋŋu_VVD.

ജീവിതം_NN1, മൊബിലിറ്റി_NN1, റീട്ടെയിൽ_NN1, വാണിജ്യം_NN1 എന്നിവയ്ക്കുള്ള_PH2NP-IF-EXJ ഇളവ്_NN1 ജൂൺ_NPM 4_MC ന്_II വൈകുന്നേരം_NPD1 7_MC മണിക്ക്_NN1-II അവസാനിക്കും_VVF.
ji:vitam_NN1, mobiliRRi_NN1, Ri:ṭṭeyil_NN1, vaṇijyam_NN1 ennivaykku||a_PH2NP-IF-EXJ i|av'_NN1 ju:ṇ_NPM na:lin'_MC-II vaikunne:ram_NPD1 7_MC maṇikk'_NN1-II avasa:nikkum_VVF.

അവശ്യസാധനങ്ങൾ_NN2, മരുന്നുകൾ_NN2, നിർമ്മാണ_JJ സാമഗ്രികൾ_NN2, പാക്കേജിംഗ്_NN1, വ്യാവസായിക_JJ ഉൽപാദനത്തിനുള്ള_NN1-IF-EXJ അസംസ്കൃത_JJ വസ്തുക്കൾ_NN2 എന്നിവ_PH2NP വിൽക്കുന്ന_VVGJ കടകൾക്ക്_NN2-IF മാത്രമേ_RR ബിസിനസിന്_NN1-IF തുറക്കാൻ_VDI കഴിയൂ_VVF-O.
avaʃyasa:dhanaŋ:a|_NN2, marunnuka|_NN2, nirmma:ṇa_JJ sa:magrika|_NN2, pa:kke:jimg'_NN1, vya:vasa:yika_JJ ulpa:danattinu||a_NN1-IF-EXJ asamskrta_JJ vastukka|_NN2 enniva_PH2NP vilkkunna_VVGJ kaṭaka|kk'_NN2-IF ma:trame:_RR bisinasin'_NN1-IF turakka:n_VDI kazhiyu:_VVF-O.

സ്ക്രാപ്പ്_NN1 കൈകാര്യം_NN1 ചെയ്യുന്ന_VVGJ സ്റ്റോറുകൾക്ക്_NN2-IF ജൂൺ_NPM 4_MC ന്-IF വൈകുന്നേരം_NPD1 7_MC മണി_NPD1 വരെ_II ബിസിനസ്സ്_NN1 നടത്താം_VVP.
skra:pp'_NN1 kaika:ryam_NN1 ceyyunna_VVGJ sRRo:Ruka|kk'_NN2-IF ju:ṇ_NPM na:lin'_MC-IF vaikunne:ram_NPD1 7_MC maṇi_NPD1 vare_II bisinass'_NN1 naṭatta:m_VVP.

സർക്കാരും_NN1-CC അർദ്ധ_JJ സർക്കാരും_NN1-CC, പൊതുമേഖയും_NN1-CC സ്ഥാപനങ്ങൾ_NN2, കോർപ്പറേഷനുകൾ_NN2, വിവിധ_JJ കമ്മീഷനുകൾ_NN2 എന്നിവ_PH2NP അടച്ചിരിക്കും_VVF.
sarkka:Rum_NN1-CC, arddha_JJ sarkka:Rum_NN1-CC potume:khalayum_NN1-CC stha:panaŋŋa|_NN2, ko:rppaRe:ʂanuka|_NN2, vividha_JJ kammi:ʂanuka|_NN2 enniva_PH2NP aṭaccirikkum_VVF.

ജൂൺ_NPM 10_MC മുതൽ_II 50%_DB ൽ_II താഴെ_RL തൊഴിലാളികളുമായി_NN2-II അവയ്ക്ക്_PH2NR തുറക്കാൻ_VDI കഴിയും_VVF.
ju:ṇ_NPM 10_MC mutal_II 50%_DB il-II ta:zhe_RL tozhila:|ika|uma:yi_NN2-II avaykk'_PH2NR turakka:n_VDI kazhiyum_VVF.

കേരളത്തിലേക്ക്_NP1-II പ്രവേശിക്കുന്ന_VVGJ ആളുകൾക്ക്_NN2-IF COVID_NN1 നെഗറ്റീവ്_JJ സർട്ടിഫിക്കറ്റ്_NN1 ആവശ്യമാണ്_JJ-VBZ.
ke:ra|attile:kk'_NP1-II prave:ʃikkunna_VVGJ a:|uka|kk'_NN2-IF COVID_NN1 negaRRi:v'_JJ sarʈʈiphikkaRR'_NN1 a:vaʃyama:ɳ'_JJ-VBZ.

മൂന്നാമത്തെ_MD തരംഗത്തിന്_NN1-IF സംസ്ഥാനം_NN1 സ്വയം_PPX ശ്രമിച്ചു_VVD.
mu:nnamatte_MD taramgattin'_NN1-IF samstha:nam_NN1 svayam_PPX ʃramiccu_VVD.

ആദ്യ_JJ ഘട്ടം_NN1 എന്ന_II നിലയിൽ_II, ഗസ്റ്റ്_NN1 വർക്കർമാരെയും_NN2-II-CC മറ്റ്_JJ ദുർബല_JJ വിഭാഗങ്ങളെയും_NN2-II-CC ഭരണകൂടം_NN1 പൂർണ്ണമായും_RR കുത്തിവയ്ക്കും_VVF.
a:dya_JJ ghaʈʈamenna_NN1 nilayil_II, gasRR'_NN1 varkkarma:reyum_NN2-II-CC maRR'_JJ durbala_JJ vibha:gaŋŋa|eyum_NN2-II-CC bharaɳaku:ʈam_NN1 pu:rɳɳama:yum_RR kuttivaykkum_VVF.

Extensions to CLAWS7 Tagset

The following are the extensions that were used in the above corpus, based on which a more complete extension is proposed in task B:

Verbs:
VVF: simple future/present habitual tense (suffix -um "-ും")
VVC: present continuous (suffix -unnu "-ുന്നു")
VVP: potential (suffix -a:m "-ാം")
VVGN: present gerund (noun)
VVGJ: present gerund (adjective)
EXJ: existential gerund (adjective)
-O: "only" (suffix -u: "-ൗ"; accompanies adverb ma:trame: "മാത്രമേ")

Conjunctions:
CST: quotative particle enn' "എന്ന്" (used as "that" in English)

Pronouns:
PH2NP: neuter 3rd person pronoun plural proximate
PH2NR: neuter 3rd person impersonal pronoun plural remote
PPX: reflexive pronoun (svayam "സ്വയം")

Note: Tags glossed as "prepositions" in CLAWS7 are treated as postpositional tags (for case markers) here.

Notes on Transliteration

Velar, dental and bilabial non-nasal stops are as in IAST and common transliteration.
The velar nasal is ŋ; the dental nasal is n; and the bilabial nasal is m.
The palatal consonants are c, ch, j, jh, ɲ.
The retroflex consonants are ʈ, ʈh, ɖ, ɖh, ɳ.
The semivowels common to Malayalam and Devanagari are transliterated the common way (y, r, l, v).
The sibilants are respectively ʂ (retroflex: Devanagari ष and Telugu ష), ʃ (palatal: Devanagari श and Telugu శ), s and h.
The retroflex lateral approximant is ɭ (Devanagari ळ and Telugu ళ) and the voiced retroflex approximant is zh.
The alveolar trill is R (as opposed to the alveolar tap which is r).

RR is a direct transliteration of the letter for R, doubled; it is pronounced as an alveolar voiceless stop (distinct from the retroflex and dental voiceless stops).

m occurring at the end of words or immediately before consonants represents the equivalent of Devanagari ं (अं) and Telugu ం (అం).

r occurring between two consonants represents the syllabic r (Devanagari ऋ and Telugu ఋ).

’ (apostrophe) at the end of words represents the close central unrounded vowel ɨ. In the script it is represented by either the equivalent of the halant or the symbol for ‘u’. It does not have an independent vowel sign (without a consonant).

Long vowels are marked by the ː symbol.

Task B

Question 1

No, the tags are not sufficient, since they are designed for English, an analytic language. For example, many tags which ought to be for verb forms in a synthetic language like Malayalam are given over to modals in English.

Question 2

The redundant groups of tags are:

- BCL: before-clause marker (e.g. “in order to”); Malayalam makes use of verb forms (usually non-finite) for these functions.
- CS: subordinating conjunction (e.g. if, because); see BCL.
- DD2: plural determiner (e.g. “these”); determiners have no distinction based on number.
  GE: Germanic genitive marker (’ or ’s): Malayalam has only one genitive marker, the case marker.
- IF: for (as preposition): The other uses of “for” have different translations. There is no confusion.
- IO: of (as preposition): see IF.
- IW: with, without (as preposition): see IF.
- JJR: general comparative adjective (e.g. “older”): The adjective is not modified; when “than” (equivalent) is present, it is understood.
- PNX1: reflexive indefinite pronoun (oneself): There is no equivalent of “one”.
- RA: adjective after nominal head (else, galore): There are no such adjectives.
- RPK: prep. adv., catenative (“about” in “be about to”): This verb form is not a catenative formation in Malayalam.
- RRR: comparative general adverb (e.g. “better”): see JJR.
- TO: infinitive marker (to): There is a specific verb form called the infinitive.
- VM: modal auxiliary (can, will, would): Malayalam uses inflections rather than modals.
- VMK: modal catenative (ought, used): see VM.
- VN: past participle (e.g. given): Malayalam has no equivalent to the past participle.

Question 3

The extra tags needed are:
VVF: simple future/present habitual tense (suffix -um “-ഉം”)

VVAP: present perfect (suffix -iʈʈuɳɖ’ “-ഇട്ടുണ്ട്”)
VVC: present continuous (suffix -unnu “-ഉന്നു” or -ukaya:ɳ “-ഉകയാണ്”)

VVP: potential (suffix -aːm “-ആം”)
VVI: conditional (suffix -aːl “-ആൽ” or -eŋkil “-എങ്കിൽ”)
VVO: interrogative (suffix -oː “-ഓ”)

VVCVI: primary causative infinitive “to cause to do” (variable suffix)
VVCV2I: secondary causative infinitive “to cause to cause to do” (variable suffix)
VVPVI: passive voice infinitive “to be done” (suffix -peṯuka “-പെടുക”)
VVPCVI: passive causative infinitive “to cause to be done” (suffix -peṯuttuka “-പെടുത്തുക”)
VVC1PVI: primary causative passive infinitive “to be caused to be done” (variable suffix)
VVC2PVI: secondary causative passive infinitive “to be caused to be done” (variable suffix)
[in this manner, each new combination of tense, aspect, mood and voice will need a new tag; this is not an exhaustive list]

VVGJ: present continuous gerund adjective (suffix -unna “-ഉന്ന”)
VVGN: present continuous gerund noun (suffix -unnat’ “-ഉന്നത്”)
[one gerund adjective and one gerund noun for any combination of tense, aspect, mood and voice]

[analogously, all verb tags for the existential uɳḍavuka “ഉണ്ടാവുക” and for the copula aːvuka “ആവുക”, which are irregular]

O: attached to verbs: “only”(-uː “-ഊ”)

PPIS2I: inclusive 1st person plural pronoun (nammaḷ “നമ്മൾ”)
PPIS2E: exclusive 1st person plural pronoun (ɲaŋŋaḷ “ഞങ്ങൾ”)

PPY1: 2nd personal casual singular pronoun (taːn “താൻ”)
PPY2: 2nd person informal singular pronoun (niː “നീ”)
PPY3: 2nd person informal plural/formal singular pronoun (niŋŋaḷ “നിങ്ങൾ”)
PPY4: 2nd person formal singular/plural pronoun (for royalty) (taːŋkaḷ “താങ്കൾ”)

PH2NP: neuter 3rd person pronoun plural proximate
PH2NR: neuter 3rd person pronoun plural remote
PHF: 3rd person formal pronoun (addeːham “അദ്ദേഹം”)

PPX: reflexive pronoun (svayam “സ്വയം”)

[for these and all existing pronominal tags, corresponding tags for all 7 cases; the accusative and genitive cases are covered in the existing tagset.]

Because there are so many TAM combinations and cases, the tagset might include an unwieldy number of tags for a synthetic language like Malayalam. The estimated number for a given verb is 144 according to the above categorisation; separate tags for the existential and the copula would more than triple the current size, which is 137.

It might therefore be more efficient to assign tags only to morphemes rather than entire words; thus case and TAM markers would be marked separately from their roots.
However, tokenisation would become significantly more complicated. Morpheme boundaries are fuzzy at best and non-existent at worst. There would have to be a tradeoff; possibly some of the categories could be tagged morpheme-wise and the remaining could have lexical tags (e.g. the tagset might include tags for all combinations of tense and aspect, while the morphemes for mood and voice are tagged separately).