Big Homework 04

Human Language for Artificial Intelligence
11-324, 624, 724

Due 11:59pm Sunday, Nov 24th

1 Introduction

The data in this problem was provided by TA Ben Lafond. Your job is to write an NLTK grammar that parses the grammatical sentences and fails to parse the ungrammatical sentences. See hints below with the data. **What to submit:**

- Your NLTK grammar.
- A writeup of your solutions to the two NACLO problems at the end of this pdf.

2 Intransitive Sentences

Considerations:

- In this language, a sentence can consist of only a verb phrase. The subject is not obligatory. Since the parser will only declare success when it finds an S, you should write a rule like this: S $\rightarrow$ VP.
- In the dataset below, morphemes are segmented with hyphens. However, the words in your lexical productions should be input without hyphens.
- You should include all morphological features on each word except for person. Since all data is in the third person, you may disregard this as a feature.
  - **Example:** The lexical production for "boy" can be IV $\rightarrow$ mal′čik [gend=masc, num=sg, case=nom].
- A reminder about case marking:
  - **Nominative (NOM)** is a case marker that says "I am the subject."
  - **Accusative (ACC)** is a case marker that says "I am the object."

The adventurous student may wish to create a LEXC file to produce features and then format them automatically into lexical productions for NLTK. However, this is not required.

1. Xod-it.
   walk-3sg.pres
   'He/she walks.'

2. Xod-jat.
   walk-3pl.pres
   'They walk.'

3. Devušk-a xod-it.
   girl-fem.nom.sg walk-3sg.pres
   'The girl walks.'

4. Mal′čik-∅ xod-it.
   boy-masc.nom.sg walk-3sg.pres
   'The boy walks.'

5. Devušk-i xod-jat.
   girl-fem.nom.pl walk-3pl.pres
   'The girls walk.'

6. Mal′čik-i xod-jat.
   boy-masc.nom.pl walk-3pl.pres
   'The boys walk.'

7. *Devušk-i xod-it.
   girl-fem.nom.pl walk-3sg.pres
   'The girl walks.'

8. *Devušk-u xod-it.
   girl-fem.acc.sg walk-3sg.pres
   'The girl walks.'

9. *Devušk-a xod-jat.
   girl-fem.nom.sg walk-3pl.pres
   'The girl walks.'

10. *Devušk-a xod-it malʹčik-a.
girl-fem.nom.sg walk-3sg.pres boy-masc.anim.acc.sg
'The girl walks the boy.'

3 Transitive Sentences

Considerations:

- The masculine accusative suffix and the feminine nominative suffix are identical in form. That is, *-a* means [case=nom] with a feminine noun and it means [case=acc] for a masculine noun.
- Similarly, the accusative ending for the masculine inanimate noun dom is identical to the nominative ending for the animate masculine noun *malʹčik*.
- The masculine animate accusative is the same as the masculine animate *cika*, one accusative genitive. There should be two lexical productions for *malʹ* ˇand one genitive.
- Please ignore sentences 11 and 12 unless you have elected to use LEXC. Those sentences just show that you can't put the animate accusative ending on an inanimate noun, and you can't put the inanimate accusative ending on an animate noun.
- A reminder about case marking:
  - **Genitive (GEN)** is a case marker that says "I am a possessor". For example, *boy* is genitive in sentence 14 because the boy possesses the house.
  - **Dative (DAT)** is a case that is most commonly used for recipients like *to the girl* in *I gave a gift to the girl.*
- For sentences containing possessive noun phrases like *the boy's house*, you should use a recursive rule like: NP →N NP[case=gen].
- Sentences 16-21 show a verb whose subject must be in the dative case and whose object must be in the nominative case. The verb agrees with the object. This is not so different from Hindi, where the subject was in the ergative case and the verb agreed with the object.
- Sentences 21-25 feature ditransitive verbs, which is the name given to verbs that take a direct object (also called patient) **and** an indirect object (or recipient).
  - The subjects of these verbs must be in nominative case.
  - The patients must be in accusative case.
  - The recipients must be in dative case.

1. Devušk-a vid-it mal′čik-a.
   girl-fem.nom.sg see-3sg.pres boy-masc.anim.acc.sg
   ‘The girl sees the boy.’

2. Mal′čik-∅ vid-it devušk-u.
   boy-masc.nom.sg see-3sg.pres girl-fem.acc.sg
   ‘The boy sees the girl.’

3. Devušk-a vid-it dom-∅.
   girl-fem.nom.sg see-3sg.pres house-masc.inan.acc.sg
   ‘The girl sees the house.’

4. Mal′čik-∅vid-it dom-∅.
   boy-masc.nom.sg see-3sg.pres house-masc.inan.acc.sg
   ‘The boy sees the house.’

5. Vid-it mal′čik-a.
   see-3sg.pres boy-masc.anim.acc.sg
   ‘He/she sees the boy.’

6. Vid-it devušk-u.
   see-3sg.pres girl-fem.acc.sg
   ‘He/she sees the girl.’

7. Vid-it dom-∅. see-3sg.pres house-masc.inan.acc.sg
   ‘He/she sees the house.’

8. Vid-jat mal′čik-a.
   see-3pl.pres boy-masc.anim.acc.sg
   ‘They see the boy.’

9. Vid-jat devušk-u.
   see-3pl.pres girl-fem.acc.sg
   ‘They see the girl.’

10. Vid-jat dom-∅.
    see-3pl.pres house-masc.inan.acc.sg
    ‘They see the house.’

11. *Devušk-a vid-it mal′čik-∅.
girl-fem.nom.sg see-3sg.pres boy-masc.inan.acc.sg
‘The girl sees the boy.’

12. *Devušk-a vid-it dom-a.
girl-fem.nom.sg see-3sg.pres house-masc.anim.acc.sg
‘The girl sees the house.’

13. *Devušk-u vid-it mal′čik-a.
girl-fem.acc.sg see-3sg.pres boy-masc.anim.acc.sg
‘The girl sees the boy.’

14. Devušk-a vid-it dom-∅ mal′čik-a.
girl-fem.nom.sg see-3sg.pres house-masc.inan.acc.sg boy-masc.gen.sg
‘The girl sees the boy’s house.’

15. Mal′čik-∅ vid-it dom-∅ devušk-i.
boy-masc.nom.sg see-3sg.pres house-masc.inan.acc.sg girl-fem.gen.sg
‘The boy sees the girl’s house.’

16. Devušk-e nrav-it-sja mal′čik-∅.
girl-fem.dat.sg like-3sg.pres-refl boy-masc.nom.sg
‘The girl likes the boy.’

17. Mal′čik-u nrav-it-sja devušk-a.
boy-masc.dat.sg like-3sg.pres-refl girl-fem.nom.sg
‘The boy likes the girl.’

18. Devušk-am nrav-it-sja mal′čik-∅.
girl-fem.dat.pl like-3sg.pres-refl boy-masc.nom.sg
‘The girls like the boy.’

19. Mal′čik-am nrav-it-sja devušk-a.
boy-masc.dat.pl like-3sg.pres-refl girl-fem.nom.sg
‘The boys like the girl.’

20. Devušk-e nrav-jat-sja mal′čik-i.
girl-fem.dat.sg like-3pl.pres-refl boy-masc.nom.pl‘
The girl likes the boys.’

21. Mal′čik-u nrav-jat-sja devušk-i.
boy-masc.dat.sg like-3sg.pres-refl girl-fem.nom.pl‘
The boy likes the girls.’

22. Devušk-a dar-it podarok-∅ mal′čik-u.
girl-fem.nom.sg give-3sg.pres gift-masc.inan.acc.sg boy-masc.dat.sg
‘The girl gives a present to the boy.’

23. Mal′čik-∅ dar-it podarok-∅ devušk-e.
boy-masc.nom.sg give-3sg.pres gift-masc.inan.acc.sg girl-fem.dat.sg
‘The boy gives a present to the girl.’

24. Babušk-i devušk-i dar-jat podarok-∅ mal′čik-u.
grandmother-masc.nom.pl girl-fem.gen.sg give-3pl.pres
gift-masc.inan.acc.sg boy-masc.dat.sg
‘The girl’s grandmothers give a present to the boy.’

25. Dar-jat podarok-∅ babušk-am devušk-i.
give-3pl.pres gift-masc.inan.acc.sg grandmother-masc.dat.pl girl-fem.gen.sg‘
They give a present to the girl’s grandmothers.’

4 Past Tense

Considerations:
In the past tense, the verb agrees in gender with the subject (the gender of the verb matches the gender of the subject.)

1. Devušk-a vid-ela mal′čik-a.
girl-fem.nom.sg see-pst.fem boy-masc.anim.acc.sg
‘The girl saw the boy.’

2. *Devušk-a vid-el mal′čik-a.
girl-fem.nom.sg see-pst.masc boy-masc.anim.acc.sg
‘The girl saw the boy.’

3. *Devušk-a vid-eli mal′čik-a.
girl-fem.nom.sg see-pst.pl boy-masc.anim.acc.sg
‘The girl saw the boy.’

4. Devušk-a vid-ela dom-∅.
girl-fem.nom.sg see-pst.fem house-masc.inan.acc.sg

‘The girl saw the house.’

5. Vid-ela mal′čik-a.
   see-pst.fem boy-masc.anim.acc.sg
   ‘She saw the boy.’

6. Vid-ela dom-∅.
   see-pst.fem house-masc.inan.acc.sg
   ‘She saw the house.’

7. Mal′čik-∅ vid-el devušk-u.
   boy-masc.nom.sg see-pst.masc girl-fem.acc.sg
   ‘The boy saw the girl.’

8. Mal′čik-∅ vid-el dom-∅.
   boy-masc.nom.sg see-pst.masc house-masc.inan.acc.sg
   ‘The boy saw the house.’

9. Vid-el devušk-u.
   see-pst.masc girl-fem.acc.sg
   ‘He saw the girl.’

10. Vid-el dom-∅.
    see-pst.masc house-masc.inan.acc.sg
    ‘He saw the house.’

11. Devušk-i vid-eli mal′čik-a.
    girl-fem.nom.pl see-pst.pl boy-masc.anim.acc.sg
    ‘The girls saw the boy.’

12. Mal′čik-i vid-eli devušk-u.
    boy-masc.nom.pl see-pst.masc.pl girl-fem.acc.sg
    ‘The boys saw the girl.’

13. Vid-eli mal′čik-a.
    see-pst.pl boy-masc.anim.acc.sg
    ‘They saw the boy.’

14. Vid-eli devušk-u.
    see-pst.pl girl-fem.acc.sg
    'They saw the girl.'

15. Vid-eli dom-∅.
    see-pst.pl house-masc.inan.acc.sg
    'They saw the house.'

5 Negatives

Considerations:

- Existential sentences (*There is/are* sentences) can start with a prepositional phrase PP consisting of a preposition, P, and a noun phrase, NP. Note the locative (location) case on the NP.
- *Jest′* is an existential verb and *njet* is a negative existential verb. The noun phrase after *jest′* is nominative, and the noun phrase after *njet* is genitive. (It would be like saying "There isn't of butter" in English, which is not a normal thing to say, or "Il n'y a pas de beurre" in French', which is a normal thing to say.)
- Starting with sentence 9, the sentences are not existential. With a negative verb, the object can be accusative or genitive, but with different meanings. We have not demonstrated how to make a meaning representation in NLTK, so just parse both sentences (accusative object and genitive object).

1. V dom-e jest′masl-o.
   in house-masc.loc.sg exist butter-neut.nom.sg
   'There is butter in the house.'

2. *V dom-e jest′masl-a.
   in house-masc.loc.sg exist butter-neut.gen.sg
   'There is butter in the house.'

3. V dom-e njet masl-a.

   in house-masc.loc.sg exist.neg butter-neut.gen.sg
   'There is no butter in the house.'

4. *V dom-e njet masl-o.
   in house-masc.loc.sg exist.neg butter-neut.nom.sg

‘There is no butter in the house.’

5. V dom-e jest′blinčik-i.
   in house-masc.loc.sg exist pancake-masc.nom.pl
   ‘There are pancakes in the house.’

6. *V dom-e jest′blinčik-ov.
   in house-masc.loc.sg exist pancake-masc.gen.pl
   ‘There are pancakes in the house.’

7. V dom-e njet blinčik-ov.
   in house-masc.loc.sg exist.neg pancake-masc.gen.pl
   ‘There are no pancakes in the house.’

8. *V dom-e njet blinčik-i.
   in house-masc.loc.sg exist.neg pancake-masc.nom.pl
   ‘There are no pancakes in the house.’

9. Devušk-a ne vid-it mal′čik-a.
   girl-fem.nom.sg neg see-3sg.pres boy-masc.anim.acc.sg
   ‘The girl does not see the boy.’

10. Devušk-a ne vid-it mal′čik-a.
    girl-fem.nom.sg neg see-3sg.pres boy-masc.gen.sg
    ‘The girl does not see a boy.’

11. Devušk-a ne vid-it mal′čik-ov.
    girl-fem.nom.sg neg see-3sg.pres boy-masc.gen.pl
    ‘The girl does not see (the) boys.’

12. Devušk-a ne vid-it dom-.
    girl-fem.nom.sg neg see-3sg.pres house-masc.inan.acc.sg
    ‘The girl does not see the house.’

13. Devušk-a ne vid-it dom-a.
    girl-fem.nom.sg neg see-3sg.pres house-masc.gen.sg
    ‘The girl does not see a house.’

14. Devušk-a ne vid-it dom-ov.
girl-fem.nom.sg neg see-3sg.pres house-masc.gen.pl
'The girl does not see (the) houses.'

15. *Devušk-a vid-it dom-a.
girl-fem.nom.sg see-3sg.pres house-masc.gen.sg
'The girl sees a house.'

16. Mal′čik-∅ ne vid-it devušk-u.
boy-masc.nom.sg neg see-3sg.pres.sg girl-fem.acc.sg
'The boy does not see the girl.'

17. Mal′čik-∅ ne vid-it devušk-i.
boy-masc.nom.sg neg see-3sg.pres girl-fem.gen.sg
'The boy does not see a girl.'

18. *Mal′čik-∅ vid-it devušk-i.
boy-masc.nom.sg see-3sg.pres girl-fem.gen.sg
'The boy sees a girl.'

YOUR NAME: REGISTRATION #

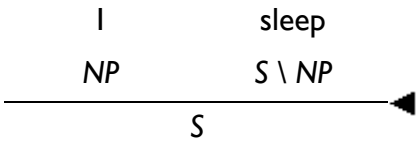
(O) CCG (1/2) [5 points]

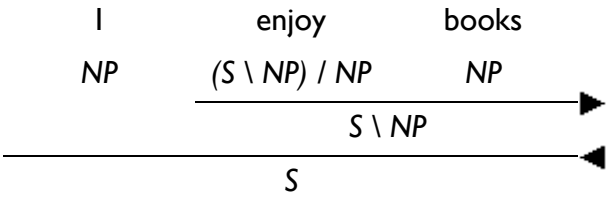
One way for computers to understand language is by forming a structure that represents the relationships between words using a technique called Combinatorial Categorial Grammar (CCG). Computer scientists and linguists can use CCG to parse sentences (that is, try to figure out their structure) and then extract meaning from the structure.

As the name suggests, Combinatorial Categorial Grammar parses sentences by combining categories. Each word in a sentence is assigned a particular category; note that / and \ are two different symbols:

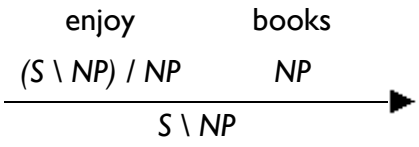
I	NP
books	NP
sleep	S \ NP
enjoy	(S \ NP) / NP

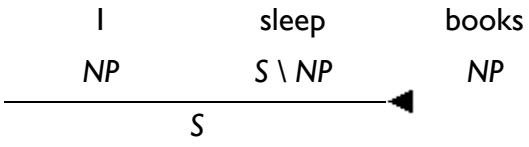
These categories are then combined in systematic ways. We will not explain how, but we will give you two successful parses…

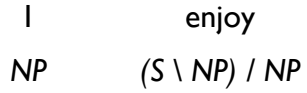


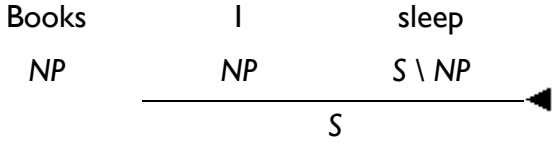


...and four unsuccessful parses…

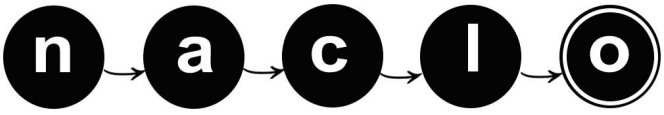








If a parse is successful, the sentence is declared "grammatical"; if not, the sentence is declared "ungrammatical".



YOUR NAME: REGISTRATION #

(O) CCG (2/2)

O1. Using the above examples as evidence, figure out how CCG parses sentences, and describe it briefly here:

We start with S.

Category/ies must cancel so that they equal the category directly below it.

A \ B means: there must be category B to the left, to form the category A.
A / B is similar but to the right.

Example: S \ NP left-cancels NP to form S. (S \ NP) / N right-cancels N to form S \ NP.

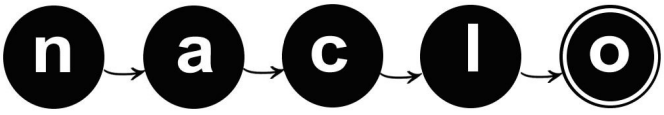
O2. In the sentence "I enjoy long books", list all of the categories that, if assigned to "long", make the sentence have a successful parse.

NP / NP

((S \ NP) / NP) \ ((S \ NP) / NP)

O3. Not every grammatical sentence of English will be declared "grammatical" by the process above. Using only the words "I", "books", "sleep", and "enjoy", form a grammatically correct English sentence that will fail to parse given the categories above. You don't have to use all four of the words.

I enjoy sleep



YOUR NAME: REGISTRATION #

(P) Combining Categories in Tok Pisin (1/2) [15 points]

This problem is a follow-up to problem O and has to be solved after that problem. Tok Pisin (also referred to as New Guinea Pidgin or Melanesian Pidgin) is a creole language spoken in the northern mainland of Papua New Guinea and surrounding islands. It is an official language and the mostly widely used language in the country, spoken by over 5 million people.

Many Tok Pisin words come originally from English – its name comes from "talk" and "pidgin"[1] -- but Tok Pisin isn't just English. It has a distinct grammar and uses these words in different (but systematic!) ways.

P1. Below are sentences in Tok Pisin with a scrambled list of English translations. Match each sentence to its English equivalent.

1.	Brata bilong em i stap rit.	E
2.	Ol i stap dringim wara.	H
3.	Ol i ken ritim buk bilong mi.	C
4.	Em i ritim buk pinis.	A
5.	Em i laik rit.	G
6.	Susa bilong em i ken rait.	D
7.	Susa bilong mi i boilim wara.	B
8.	Wara i boil pinis.	F

A.	He has read the book.
B.	My sister boils the water.
C.	They can read my book.
D.	His sister can write.
E.	His brother is reading.
F.	The water has boiled.
G.	He wants to read.
H.	They are drinking water.

P2. Translate the following Tok Pisin sentence into English:

Brata bilong mi i stap ritim buk bilong susa bilong mi.

My brother is reading my sister's book.

P3. Translate the following English sentence into Tok Pisin:

Their sister wants to write a book.

Susa bilong ol i laik raitim buk.

[1] A pidgin language is a communicative system developed by two or more groups of people who do not share a common language. Tok Pisin started out as a pidgin but has since developed into a creole, a complex language in its own right.

YOUR NAME: REGISTRATION #

(P) Combining Categories in Tok Pisin (2/2)

P4. Describing these words in terms of their CCG categories (introduced in Problem O) highlights that these aren't English words combined according to English rules, but are Tok Pisin words combined according to Tok Pisin rules.

Match each Tok Pisin word to its CCG category. Some categories will be used more than once. The symbol S_b is short for 'Bare Clause'.

1.	bilong	B
2.	brata	A
3.	boil	D
4.	boilim	E
5.	buk	A
6.	dringim	E
7.	em	A
8.	i	C
9.	ken	G
10.	laik	G

11.	mi	A
12.	ol	A
13.	pinis	F
14.	stap	G
15.	raitim	E
16.	rit	D
17.	ritim	E
18.	susa	A
19.	wara	A

A.	NP
B.	(NP \ NP) / NP
C.	(S \ NP) / (S_b \ NP)
D.	(S_b \ NP)
E.	(S_b \ NP) / NP
F.	(S_b \ NP) \ (S_b \ NP)
G.	(S_b \ NP) / (S_b \ NP)

P5. Explain your answer.

- There is only one way to form S, which is S \ NP (from C)
- A: brata, buk, susa, wara, mi, ol, em can be subjects (and they are to the left)
  so we assume it is NP, which combines with S \ NP to make S.
- B: bilong needs an owner to the right, and forms a new noun phrase (NP \ NP) / NP fits,
  because in buk bilong mi, bilong mi cancels to get (NP \ NP), which cancels with buk to get a NP
- C: analyzing "Em i laik rit", em is NP, so "i laik rit" must be S \ NP. Parsing it as (i, (laik, rit)), (laik,rit) must be (Sb \ NP). Thus assigning C for i gets us S \ NP which is what we needed
- D: "laik rit" must combine to be Sb \ NP. It seems that simple present verbs like rit always appear after "i" or "laik/ken/stap", so we assign D as they combine with "i" and "laik/ken/stap" to form S\NP and Sb \ NP respetively-
- E: "ritim, boilim, raitim" seem to take an object (transitive verb) and succeeds "i", so it must consume NP, and return Sb \ NP. E is (Sb \ NP) / NP.
- F: "pinis" happens after a verb. F works because "i boil pinis" - boil pinis must be (Sb \ Np). Boil is (Sb \ Np) so pinis must consume (Sb \ Np) and return (Sb \ Np). F works.
- G: "laik, ken, stap" happen before verbs. G works - same reasoning as F except G must consume (Sb \ Np) to the right.