ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (CS–431)
Chappelier, J.-C., Rajman, M. & Bosselut, A.

INTRODUCTION TO NATURAL LANGUAGE PROCESSING (CS–431)

Fall 2022 Examination

Thursday, January 26th, 2023.

NAME: Hanon Ymous

SCIPER: 000000

Seat: 0

Instructions:

You have three hours (9:15–12:15) for this exam, which consists of six independent questions with different weights. For each question, the points are indicated and the total number of points is 115 ($29 + 24 + 17 + 9 + 22 + 14$).

All documents are allowed. However, no electronic device (computer, cell phone, calculator, etc.) is authorized.

Write your answers directly on the exam sheets. If you run out of space, ask for additional official exam sheets. Do not use your own sheets; they will not be considered.
Notice that, if needed, you have an extra answering page at the end of this exam. Please do not use that page as draft paper.

QUESTION I : The Daily Gazette [29 pt]

You have been publishing a daily column for the Gazette over the last few years and have recently reached a milestone — your 1000th column! Realizing you'd like to go skiing more often, you decide it might be easier to automate your job by training a story generation system on the columns you've already written. Then, whenever your editor pitches you a title for a column topic, you'll just be able to give the title to your story generation system, produce the text body of the column, and publish it to the website!

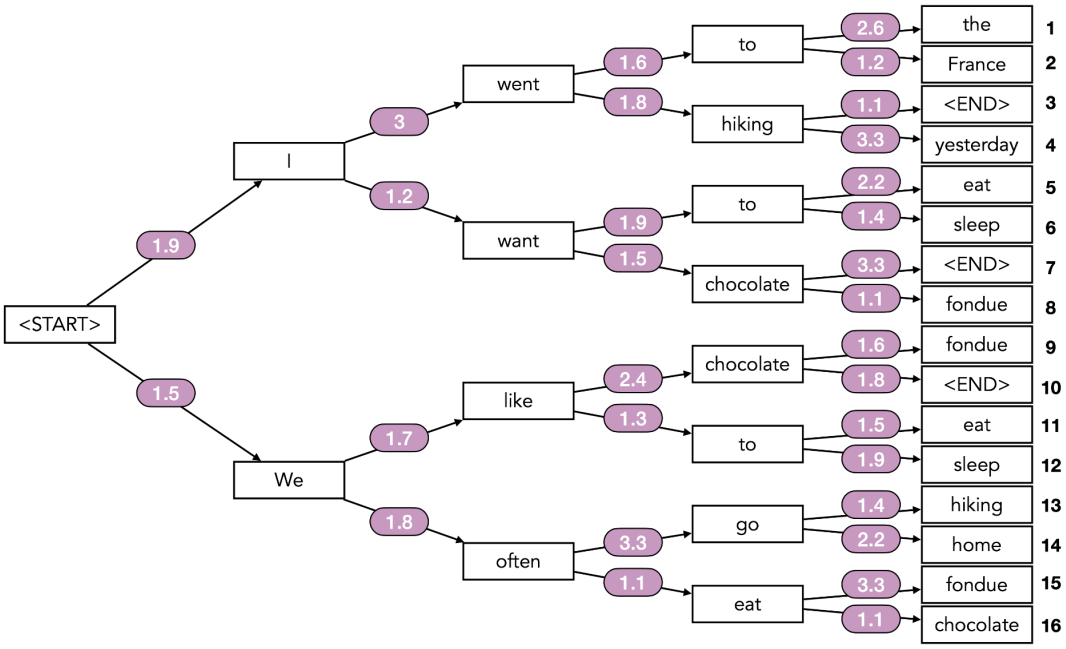
① **[2 pt]** You consider using either a transformer or a recurrent neural network (RNN) as the underlying model for your text generator. Assuming there are no practical issues with selecting either one (such as the amount of data available), which one would you choose for this task? Give **two** reasons why.

② **[1 pt]** Given that you have published 1000 columns at the Gazette, you have around 800 training examples that you can use for your system. Given the size of your dataset, do you think it would be helpful to pretrain your model on other text? Why or why not?

③ **[1 pt]** Would you use a causal language modeling or masked language modeling training objective to train your model? Why?

④ **[1 pt]** You initialize your model with a vocabulary V with $|V|$ tokens. Given a vector of scores $S = [s_1, \ldots, s_i, \ldots, s_{|V|}]$ output by your model for each token in your vocabulary, write out the softmax function to convert score s_1 to a probability mass $P(s_1)$:

⑤ **[7 pt]** Now that you've trained your model on your dataset, you can produce text from it. You pre-compute the step-by-step probability distributions over all tokens for four steps. Below, we show the top-2 highest probability tokens in these distributions at each step (along with their **negative log probability**):



For each of the following sub-questions a, b and c, provide your answer in a form similar to:
<START> I went hiking yesterday

a) [1 pt] What sequence would be produced using argmax decoding?

b) [4 pt] What sequence would be produced using beam search with a beam size of 2? **Justify** your answer by *annotating* the above graph with choices and calculations.

c) [2 pt] What is the optimal sequence?

continues on back ☞

⑥ **[4 pt]** You are given a probability distribution $P(y_t|y_0, \ldots, y_{t-1})$ over 100 possible next tokens to generate by your model. The distribution has the following characteristics:

- 20% of the probability mass is on the most probable token;
- 10% of the probability mass is on each of the next 4 most probable tokens;
- 1% of the probability mass is on each of the next 20 most probable tokens;
- the remaining mass is uniformly distributed across the remaining 75 tokens.

a) [2 pt] In top-k sampling, if $k = 15$, how much probability mass will be included in the set of tokens you sample from?
Fully justify your answer.

b) [2 pt] In top-p sampling, if $p = 0.75$, how many tokens will be included in the set of tokens you sample from?
Fully justify your answer.

⑦ **[4 pt]** The outputs of which decoding algorithm would be changed by increasing the temperature hyperparameter of the softmax calculation? Assume the seed of your random number generator remains the same. Select all that apply and **justify** your answer for each:

[ ] Argmax decoding:

[ ] Beam search decoding:

[ ] Top-k sampling:

[ ] Top-p sampling:

⑧ **[4 pt]** To evaluate your system, you decide to hold out some of the columns you have previously written and use them as an evaluation set. After generating new columns using the same titles as these held-out columns, you decide to evaluate their quality.

a) [1 pt] What would be an advantage of using a content overlap metric?

b) [1 pt] What would be a disadvantage of using a content overlap metric?

c) [1 pt] What would be an advantage of using a model-based metric?

d) [1 pt] What would be a disadvantage of using a model-based metric?

continues on back ☞

⑨ **[3 pt]** Your column generation system has become quite successful and you've managed to automate most of your job simply by typing your editor's title pitches into your model to produce your column every day. Two years later, during the COVID–25 pandemic, your editor proposes to use your system to generate an information sheet about the pandemic for anyone looking for information about symptoms, treatments, testing sites, medical professionals, etc. Given the similarity to a previous pandemic many years before, COVID–19, you train your model on all news articles published about COVID–19 between the years of 2019–2022. Then, you generate the information page from your trained model.

Give an example of a potential harm that your model could produce from the perspective of:

- leaking private information;
- disinformation;
- human interaction harms.

⑩ **[2 pt]** In your writings on the topic of medicine, you typically quoted two professionals, Dr. John Smith (around 60% of the time) and Dr. Virginia Jones (around 40% of the time). Would you expect the proportion of quotes by Dr. John Smith to be less than, greater than, or around 60% in the generated stories your model produces? Why?

QUESTION II : Pulsed lasers [24 pt]

Consider the following sentence:

High-energy pulsed laser beams are used in soft-tissue surgery.

① **[1 pt]** Using a tokenizer that splits on whitespaces and punctuation (including hyphens (-)), what is the token sequence?

② **[1 pt]** Using a 1-gram language model and the same tokenizer, what is the probability of the above sentence? Provide your answer as a formula, but clearly explaining each variable.

③ **[2 pt]** Same question but using a 2-gram language model:

④ **[4 pt]** Tokenization is now enhanced with Named Entity Recognition (NER) specialized on technical and medical terms.

a) **[1 pt]** How is your answer to question ② modified? **Fully justify** your answer.

b) **[3 pt]** What would be the advantage of doing so? What would be the major drawback? Justify your answers.

continues on back ☞

Consider now the following toy learning corpus[1] of 59 tokens[2], out of a possible vocabulary of $N = 100$ different tokens:

> Pulsed operation of lasers refers to any laser not classified as continuous wave, so that the optical power appears in pulses of some duration at some repetition rate. This encompasses a wide range of technologies addressing a number of different motivations. Some lasers are pulsed simply because they cannot be run in continuous wave mode.

⑤ **[3 pt]** Using a 2-gram language model, what are the values of the parameters corresponding to "*continuous wave*" and to "*pulsed laser*"...

a) [1 pt] ...using Maximum-Likelihood estimates?

b) [2 pt] ...using estimation smoothed by a Dirichlet prior with parameters all equal to 0.01?

Justify your answers.

[1] [excerpt from `https://en.wikipedia.org/wiki/Pulsed_laser`]
[2] The same tokenizer was used, but *without* any NER.

Consider now the following shorter phrase:

laser used for surgery process

and an order-1 HMM for PoS tagging with the following parameters (not exhaustive, but no missing information to solve the question):

- `N`: 0.1, `V`: 0.15, `Adj`: 0.2, `Prep`: 0.05, ...
- *laser*: `Adj`: $4 \cdot 10^{-4}$, `N`: $5 \cdot 10^{-4}$
  used: `Adj`: $8 \cdot 10^{-4}$, `V`: $6 \cdot 10^{-4}$
  for: `Prep`: $9.5 \cdot 10^{-4}$
  surgery: `N`: $7.3 \cdot 10^{-4}$
  process: `N`: $7 \cdot 10^{-4}$, `V`: $5 \cdot 10^{-4}$
	Adj	N	V	Prep
Adj	0.2	0.5	0.1	0.15
N	0.4	0.2	0.3	0.1
V	0.1	0.4	0.15	0.3
Prep	0.02	0.45	0.51	0.01

⑥ **[7 pt]** What is the most probable sequence of tags for the above sentence?

Fully justify your answer.

continues on back ☞

Finally, we'd like to do some sentence topic classification using a Naive-Bayes model.

Consider the following toy learning corpus, where each sentence has been assigned a topic, either "Medical" or "Computer":

- **Medical**: plastic surgery process initial consultation can be scheduled by sending an email to the administration.
- **Medical**: in the process, the laser beam comes into contact with soft tissues.
- **Medical**: laser eye surgery process reshapes parts of the cornea by removing tiny amount of tissues.
- **Computer**: the team behind the laser based quantum computer includes scientists from the US, Australia and Japan.
- **Computer**: the optical laser barcode scanner was plugged on the USB port.
- **Computer**: cdrom laser lens cleaning process starts with opening the tray.
- **Computer**: laser was a computer trademark

The parameters are learned using some appropriate additive smoothing with the same value for all parameters. In the above learning corpus, there are 42 token occurrences in "Medical" documents and 42 token occurrences in "Computer" documents (punctuation is ignored).

⑦ **[6 pt]** How would the following short sentence:

pulsed laser used for surgery process

be classified by this model?

Fully justify your answer and mathematically support your claim.

QUESTION III : Systems and queries [17 pt]

Consider the following evaluation of two IR systems made on a toy corpus of four queries, where ✔ denotes a retrieved relevant document and ✗ denotes a retrieved non-relevant document (the two systems may not retrieve the same document at each rank):

query q_1:

	rank							
	1	2	3	4	5	6	7	8
system 1	✔	✔	✔	✗	✗	✗	✔	✔
system 2	✗	✔	✗	✔	✔	✔	✔	✔

query q_2:

	rank							
	1	2	3	4	5	6	7	8
system 1	✔	✔	✔	✗	✗	✔	✔	✔
system 2	✗	✔	✔	✔	✔	✔	✔	✔

query q_3:

	rank							
	1	2	3	4	5	6	7	8
system 1	✔	✗	✔	✗	✔	✗	✔	✗
system 2	✔	✔	✗	✗	✔	✔	✗	✗

query q_4:

	rank							
	1	2	3	4	5	6	7	8
system 1	✗	✔	✔	✔	✗	✗	✔	✔
system 2	✔	✗	✔	✗	✔	✔	✗	✗

In the above results, we assume that, for each query, at least one of the two systems retrieved **all** the relevant documents; and that the missing relevant documents are never retrieved.

① **[1 pt]** For each of the four queries, what is the total number of relevant documents?

② **[1 pt]** What is the *recall* of system 1 for query q_1?

Provide your answer in the form of a fraction and **justify** your answer.

③ **[2 pt]** What is P@6 for each of the two systems for query q_1?

Provide your answers in the form of fractions and **justify** your answers.

④ **[4 pt]** What is the R-Precision for each of the **two** systems?

Provide your answers as simple arithmetic expressions involving only a few fractions and **justify** your answers.

continues on back ☞

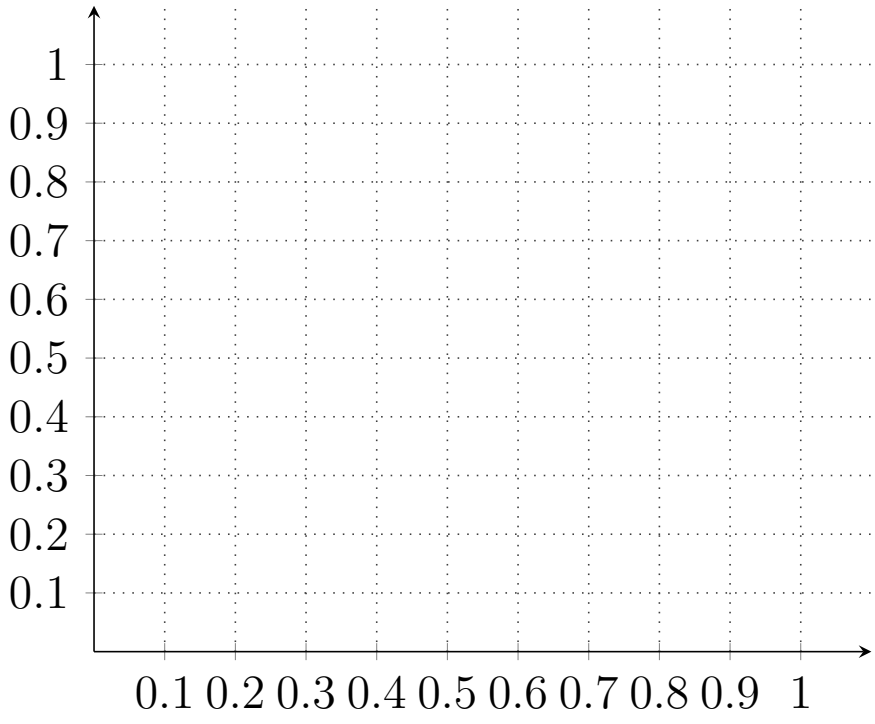
⑤ **[5 pt]** What is the MAP for system 1?

Provide your answers in the form: $\frac{1}{n}(A_1 + A_2 + ...)$ by providing the value of n and expressing each of the A_i as a simple arithmetic expression involving only a few fractions.

Then **justify** your answers.

⑥ **[4 pt]** Draw the P-R curve for system **2** and query q_1 only (using P@k).

Justify your answer and **explain** what each of the two axis represents.



A few numerical approximations:

$\frac{1}{6} \simeq 0.15$	$\frac{5}{6} \simeq 0.83$
$\frac{1}{7} \simeq 0.14$	$\frac{3}{7} \simeq 0.42$
$\frac{5}{7} \simeq 0.71$	$\frac{6}{7} \simeq 0.85$
$\frac{1}{8} \simeq 0.12$	$\frac{3}{8} \simeq 0.37$
$\frac{5}{8} \simeq 0.62$	$\frac{7}{8} \simeq 0.87$

QUESTION IV : About planes and balloons [9 pt]

You have been provided with the following definitions for the possible meanings of the words "balloon" and "plane":

```
balloon:
- meaning 1:
  balloon --(hyponym)--> inflatable

- meaning 2:
  balloon --(hyponym)--> transport
```

```
plane:
- meaning 1:
  plane --(hyponym)--> transport
  plane --(holonym)--> wing

- meaning 2:
  plane --(hyponym)--> surface
```

① **[2 pt]** How would you express the provided four meanings in plain English?

② **[1 pt]** What type of approach has been used to produce this type of semantic representations? What principle does it rely on?

③ **[2 pt]** Are the provided meaning representations formally correct? **Justify** your answer:

continues on back ☞

Assume that you are provided with a reference corpus consisting of a large amount of word definitions similar to the ones given above and that you are asked to design an evaluation metric aiming at assessing the quality of comparable definitions produced (for the same words) by various automated systems.

④ **[4 pt]** Do you think that an approach based on some kind of n-ary classification would be suitable? **Fully justify** you answer.

QUESTION V : Conjugation [22 pt]

The goal of this question is to illustrate how to use transducers to implement a simplified version of the conjugation of English verbs. We will restrict to the conjugated forms corresponding to the indicative mode and the present tense. Formally, this can be modeled as defining a transducer able to recognize associations such as:

`make+V+IndPres+1s`	make
`make+V+IndPres+2s`	make
`make+V+IndPres+3s`	makes
`make+V+IndPres+1p`	make
`make+V+IndPres+2p`	make
`make+V+IndPres+3p`	make

where "`V`" identifies the grammatical category "Verb", "`IndPres`" indicates that we are dealing with the Indicative mode and the Present tense, and "`1s`", "`2s`", "`3s`" (resp. "`1p`", "`2p`", "`3p`") refer to the first, second, and third person singular (resp. plural).

① **[1 pt]** In the above table, what do the strings in the first and the second column correspond to?

② **[1 pt]** Provide a detailed explanation of the interpretation of the strings present in the first column.

The idea is to build a transducer corresponding to the composition of three transducers:

- a transducer T_1 that defines the morphological paradigm, i.e. identifies the various cases to consider for conjugating a regular verb;
- a transducer T_2 that implements the identified cases in the form of transformation rules to be applied for the considered morphological paradigm;
- a transducer T_3 that handles all the exceptions to be implemented.

③ **[1 pt]** What is the number N of distinct cases to consider to conjugate *most* English verbs in present indicative? **Justify** your answer.

④ **[2 pt]** Describe in plain English the transformation rule associated with each of the N cases in the "present indicative" paradigm.

continues on back ☞

⑤ **[2 pt]** Is the provided number N valid for all English verbs (indicative mode, present tense)? If yes, explain why; if not, provide a simple counter-example.

⑥ **[2 pt]** Indicate in the table below the associations the transducer T_1 should recognize for the present indicative paradigm of the verb "to make" (each of the N cases mentioned in question ④ should be identified by a number between 1 and N; you can leave empty rows at the end of the table if necessary).

⑦ **[2 pt]** Provide a formal definition for transducer T_1:

⑧ **[3 pt]** Provide a formal definition of the transducer T_2 that implements the rule(s) identified in former question ④. **Fully justify** your answer.

⑨ **[1 pt]** What is the number M of (distinct) associations to be recognized by the transducer T_2 for any verb with a "present indicative" paradigm corresponding to N cases?

⑩ **[2 pt]** Indicate in the table below the associations the transducer T_2 should recognize to process the following three verbs: "to do", "to make" and "to try" (you can leave empty rows at the end of the table if necessary, but put the associations in *alphabetic order*).

⑪ **[1 pt]** Provide a formal definition for the transducer T_3 that should be used *if there would be no exceptions*:

⑫ **[1 pt]** Indicate in the table below the strings that should be associated by the transducer T_3 defined in previous question ⑪ to the outputs of T_2 indicated in the table of question ⑩, and, for each of them, indicate whether they are correct or not by writing "OK" or "notOK" in the second column (you can leave empty rows at the end of the table if necessary).

⑬ **[3 pt]** How should T_3 be modified to process correctly the cases identified as "notOK" in the table in question ⑫? Provide an updated formal definition for T_3.

QUESTION VI : Ambiguities **[14 pt]**

Consider the (toy) grammar G consisting of the following rules:

```
R1: S --> NP VP
R2: NP --> NN
R3: NP --> Det NN
R4: NN --> N
R5: NN --> NN NN
R6: NN --> NN PNP
R7: PNP --> Prep NP
R8: VP --> V
R9: VP --> Adv V
```

① **[2 pt]** Precisely define the type of grammar G is corresponding to (for that, consider at least the following aspects: dependency-based vs. constituency-based, position in the Chomsky hierarchy, and CNF); **justify** your answer for each of the aspects you will be mentioning.

② **[1 pt]** What type of rules does the provided grammar G consist of?
What type of rules should G be complemented with to be exploitable in practice?
What is the format of these missing rules?

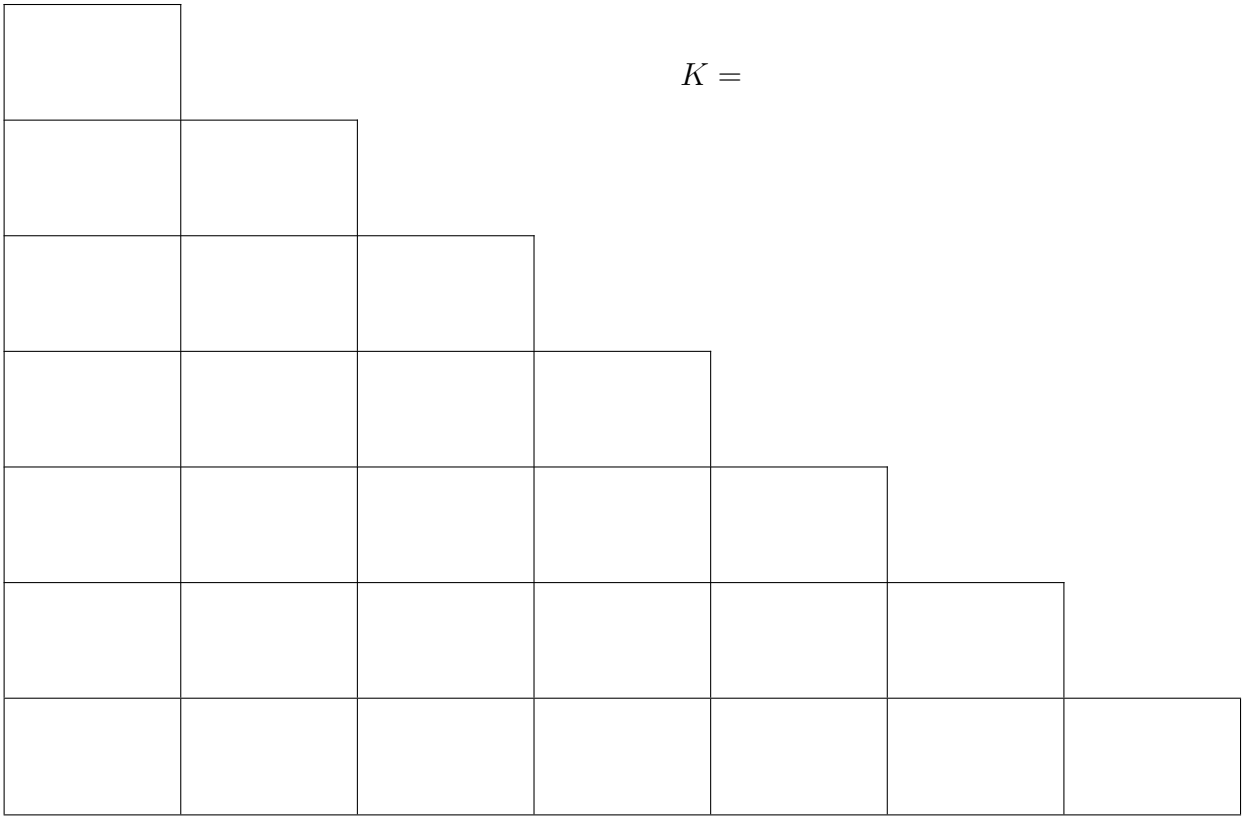
③ **[1 pt]** What is the number N of additional rules that should be added to G to make it applicable to any sequence of words from a set of 10 000 distinct words with an average syntactic ambiguity of 1.5? Justify your answer.

④ **[6 pt]** Indicate the number K of parse trees produced by G for the sequence:

the glass frames from shops often break

and describe the identified parse tree(s) in the form of a CYK chart where *only* the non-terminals and the links *required* for retrieving *full* parse tree(s) are represented:

$K =$



⑤ **[2 pt]** Indicate what type of constraints are (resp. are not) taken into account by the grammar G, and, for each constraint type mentioned, provide illustrative examples.

⑥ **[2 pt]** In how many rules should the 9 rules provided for G be expanded into to cope with simple number agreements? **Justify** your answer.

Extra space if needed to answer some question. Please do not use this page as draft paper.