ORMC Advanced 2: Linguistics II

Sunny Liang

Here are six problems from International Linguistics Olympiad (IOL) and North American Computational Linguistics Open Competition (NACLO), arranged in approximate order of difficulty. Feel free to do the problems in any order as they interest you. Some of these problems are very hard, don't be afraid to collaborate with your peers and ask for help from instructors!

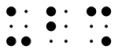
1 Tenji Karaoke (NACLO 2009, Patrick Littel)

Braille is a tactile writing system, based on a series of raised dots, that is widely used by the blind. It was invented in 1821 by Louis Braille to write French, but has since been adapted to many other languages. English, which uses the Roman alphabet just as French does, required very little adaptation, but languages that do not use the Roman alpha- bet, such as Japanese, Korean, or Chinese, are often organized in a very different manner!

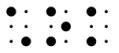
karaoke 

Above is a Japanese word written in the tenji ("dot characters") writing system. The large dots represent the raised bumps; the tiny dots represent empty positions.

The following tenji words represent *atari*, *haiku*, *katana*, *kimono*, *koi*, and *sake*. Which is which? You don't need to know either Japanese or Braille to figure it out; you'll find that the system is highly logical.

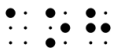
a. _____ 

b. _____ 

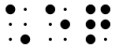
c. _____ 

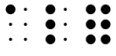
d. _____ 

e. _____ 

f. _____ 

What are the following words?

g. _____ 

h. _____ 

Write the following words in Tenji:

- *miso*
- *samurai*

2 What's the time in Tallinn? (NACLO 2014, Babette Newsome)

Tallinn is the capital of Estonia, where about 1 million people speak Estonian, a non-Indo-European language closely related to Finnish. Here are six times and how an Estonian might refer to them:

(a) Kell on üks.

(b) Kell on kaks.

(c) Veerand kaks.

(d) Pool neli.

(e) Kolmveerand üksteist.

(f) Viis minutit üks läbi.

Here are some numbers in Estonian:

6 kuss	7 seitse	8 kaheksa	10 kümme

Translate the following times to or from Estonian:

1. Kakskümmend viis minutit üheksa läbi.
2. Veerand neli.
3. Pool kolm.
4. Kolmveerand kaksteist.
5. Kolmkümmend viis minutit kuus läbi.
6. 8:45
7. 4:15
8. 11:30
9. 11:05
10. 12:30

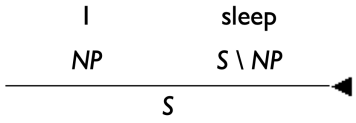
3 CCG (NACLO 2014, J. Kummerfeld, A. Blackwell, P. Littell)

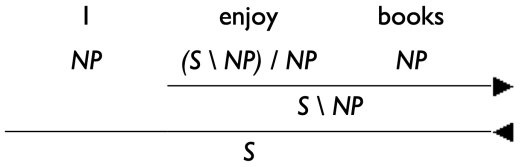
One way for computers to understand language is by forming a structure that represents the relationships between words using a technique called Combinatory Categorical Grammar (CCG). Computer scientists and linguists can use CCG to parse sentences (that is, try to figure out their structure) and then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

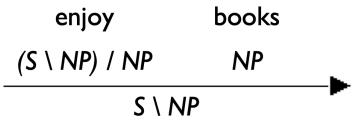
I	NP
books	NP
sleep	$S \setminus NP$
enjoy	$(S \setminus NP) \,/\, NP$

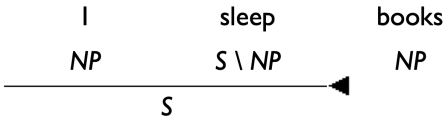
These categories are then combined in systematic ways. We will not explain how, but we will give you two successful parses...

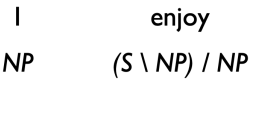


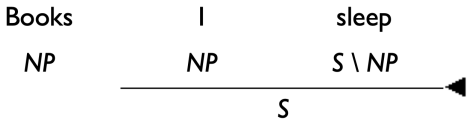


..and four unsuccessful parses...









If a parse is successful, the sentence is declared "grammatical"; if not, the sentence is declared "ungrammatical Using the above examples as evidence, figure out how CCG parses sentences.

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

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.

4 Transcendental Algebra (IOL 2003, Ksenia Guiliarova)

In 1916 the Russian scholar Jacob Linzbach invented a universal writing system intended to transcend local language and culture, hence 'transcendental.' It borrows conventions from mathematics to convey analogous ideas, hence 'algebra.'

1.	$(\frac{\dot{\Lambda}\dot{\Delta}\dot{\imath}\dot{\Delta}}{\dot{\Delta}\dot{\imath}\dot{\Delta}} + \frac{\dot{\imath}\dot{\Delta}}{\dot{\Delta}})^{\leqslant}$	The father and the brother are talking.
2.	$\dot{\pi}(> \dot{\mathrm{I}})^{[\vdash - \mathrm{t}}$	The giants are working without haste.
3.	$(\frac{\dot{\imath}\dot{\Delta}(-\dot{\Lambda}\dot{\Delta})}{(-\dot{\Lambda}\dot{\Delta})})^{\not{}} = \boxtimes$	The orphans are writing a letter.
4.	$(-\dot{\pi}\dot{\mathrm{I}}_1)^{\not{}} - \mathrm{t} = \dot{\mathrm{I}}_2$	It wasn't us who wrote about you (sg.).
5.	$\boxtimes^{\sqrt{\not{}}} - \mathrm{t} = -\dot{\Delta}_3$	It was not by her that the letter was written.
6.	$(\frac{\dot{\Lambda}\dot{\Delta}\dot{\imath}\dot{\Delta}}{\dot{\Delta}\dot{\imath}\dot{\Delta}})^{-\smile} = [\vdash$	The father doesn't like the work.
7.	$((> \dot{\mathrm{I}}) - \smile)^{\frown} - \mathrm{t} = \frac{\dot{\Lambda}\dot{\Delta}\dot{\imath}\dot{\Delta}}{\dot{\imath}\dot{\Delta}}$	The wicked giant ate the parents.
8.	$\dot{\Delta}_3^{-\mathrm{t}}$	She is not in a hurry.

Translate the following four sentences to English. Can you also come up with a sentence of your own and write it in Transcendental Algebra?

$\dot{\mathrm{I}}_3^{\smile - \sqrt{\smile}}$

$(\frac{\dot{\Lambda}\dot{\Delta}\dot{\imath}\dot{\Delta}}{\dot{\Lambda}\dot{\Delta}\dot{\imath}} - \leqslant)^{\not{}} + \mathrm{t} = \frac{\dot{\Lambda}\dot{\Delta}\dot{\imath}\dot{\Delta}}{\dot{\Delta}\dot{\imath}\dot{\Delta}} + \frac{\dot{\Lambda}\dot{\Delta}\dot{\imath}\dot{\Delta}}{\dot{\Lambda}\dot{\imath}\dot{\Delta}}$

$\dot{\Delta}_2^{[\vdash + \mathrm{t} - \leqslant} - \mathrm{t}$

$\boxtimes^{\sqrt{\frown}} - \mathrm{t} = \frac{\dot{\imath}\dot{\Delta}}{\dot{\imath}} - \frown$

5 Counting in the Roon (NACLO 2023, Riley Kong)

Roon is an Austronesian language with ∼1000 speakers in Indonesia. Roon's terms for numbers have changed over the years. This problem investigates numbers in Roon at three points in time: the years 1855, 1955, and 2012. Some number terms have remained unchanged in this time:

Number	**1855**	**1955**	**2012**
2	nuru	nuru	nuru

Some have changed once:

10	onemerim	safur	safur

However the majority have changed twice:

7	onemenuru	rimenuru	fik
32	arzus safur nuru	aresoyosier safur nuru	ares kior beberin nuru

Below are some more numbers or expressions in Roon, representing the same value in different years. + and × represent addition and multiplication respectively. Fill in the missing cells. Your answers may not contain mathematical operations.

Note: Some numbers have been slightly simplified. ŋ is pronounced like the *ng* in *sing*.

Hint 1: "Onem/wonem" is 6. "Yoser/yosier" is 1.

Hint 2: Strangely, for the year 1855, the words for the numbers 7 through 10 have 1 subtracted from their apparent value, hence, "onemenuru" is 7, even though onem (6) + nuru (2) = 8. This is not otherwise the case.

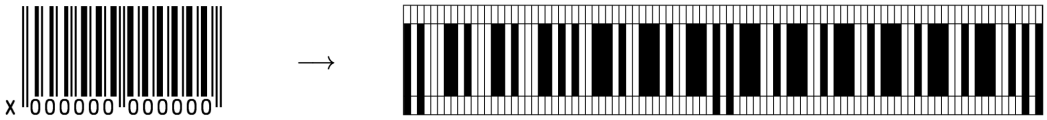
Hint 3: Roon is base-20 in 1855 and 1955, but base-10 in 2012.

Hint 4: Keep in mind how languages often mix addition and multiplication. For example, the French say 92 by saying "quatre-vingt-douze", literally, "four-twenty-twelve".

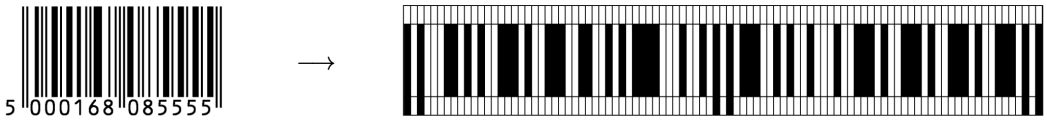
Number	**1855**	**1955**	**2012**
	nuru ŋokor	rimeyosier	yosier + rim
	onem × fak	iŋokor × rimiŋokor	ares nuru beberin fiak
	safur onem + onemefak	aresoyosier rim	(sio × nuru) + fik
	arzus di nuru yoser + safur lim	safur nuru + aresonuru fak	ares rim beberin wonem
	fak	fak	fiak
3			kior
8			war
			safur fik
21			
79			

6 EAN-13 (IOL 2011, Hugh Dobbs)

The barcode language EAN-13 (or GTIN-13) is used in almost every country in the world, yet nobody speaks it. It has 10 main "dialects", but this problem is not concerned with dialect 0, which is effectively the same as the older language UPC(A).



This is not a barcode: it belongs to a possible dialect of EAN-13 which is not in use. (On the right the machine-readable part of the code has been enlarged and transferred onto a grid for ease of observation.)



This is a barcode: it belongs to dialect 5. This barcode is from a packet of biscuits from the UK, and the number starts with the country code or system number for the UK, which is 50. Usually the first part of the code (5-000168) identifies the producer and the next part (08555) is chosen by the producer and identifies the product. The last digit is always a checksum.

Here are some more system numbers:

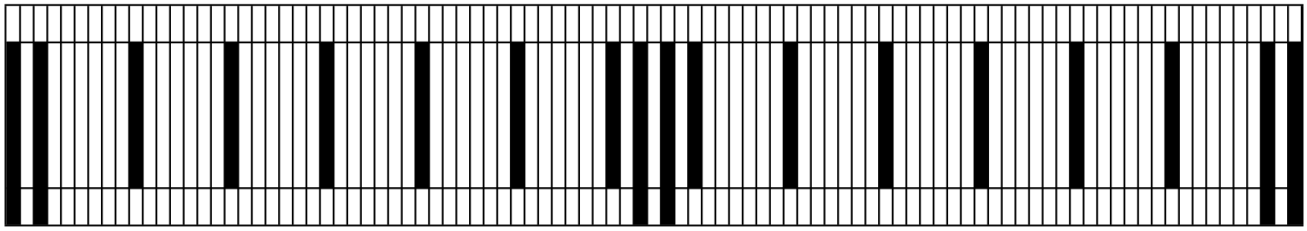
20-29	in-store functions	539	Ireland	84	Spain
30-37	France	64	Finland	978	ISBN (books)
40-44	Germany	73	Sweden	??	Norway

Hint 1: The first digit determines the dialect.
Hint 2: Each digit has three different encodings, but which one you use depends on the position and the dialect.
Hint 3: A scanner needs to be able to tell if a barcode is upside-down.

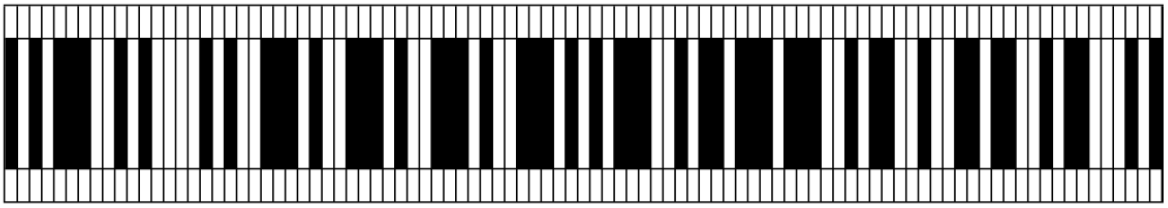
(a) Barcodes A–I, are provided at the end of this handout. Match the following to their corresponding barcode and fill in the question marks:

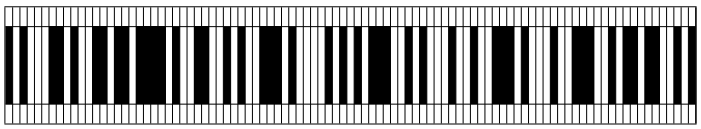
1. toilet paper (Spain) is barcode E;
2. smoked salmon (Ireland), product code = 02661, checksum = ?;
3. *The Lost Symbol* (ISBN book);
4. pork steak (packed in the store), cost = 4 euros and 16 cents;
5. mop head (from where?), full code = 4-023103-075702;
6. cholesterol-lowering spread (Finland);
7. sirloin steak (packed in the store), cost = ?;
8. Korsordboken (puzzle magazine, Sweden), full code = ?;
9. Mots Codés (puzzle magazine, France).

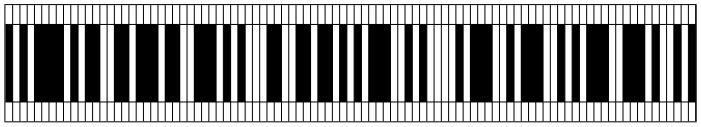
(b) Draw the (imaginary) barcode 1-453927-348790 in the grid below. Some of it has been filled in to help you.

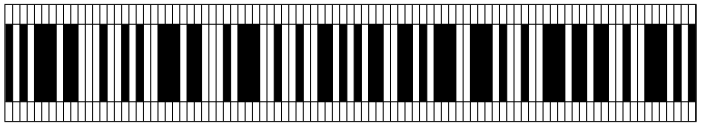


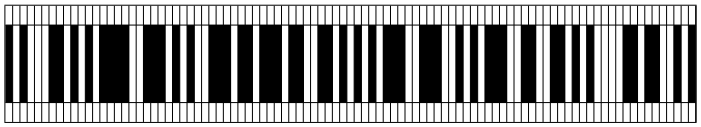
(c) The barcode below is from Dagbladet, a newspaper from Norway. Write out the full code. What is the system number or country code for Norway?

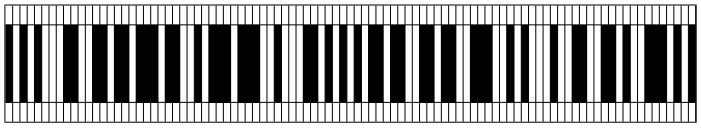


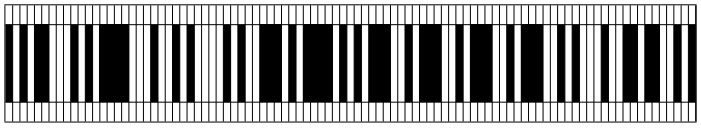
A. 

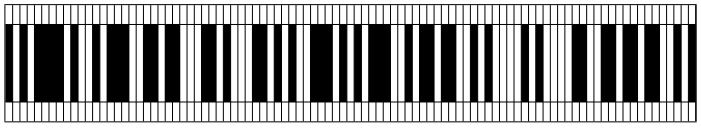
B. 

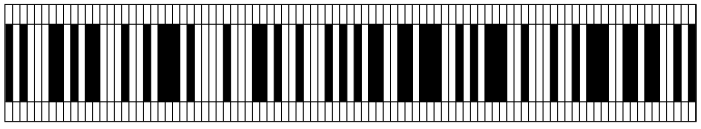
C. 

D. 

E. 

F. 

G. 

H. 

I.