UKLO Round 2 2012

1. A fox among the h (15 marks)

Spencer is working on an online biology reference, and he is currently working on the information retrieval program, so that people can type in things like "What do whales eat?" or "How much does a bee weigh?" and get the relevant answers.

Part of this task involves a process called stemming – taking text and figuring out what the "stem" of each word is. (The stem is the form of the word without any prefixes or suffixes, so dance is the stem of dancing, happy is the stem of unhappiness, etc.) The program needs the stem so that it can determine that a request about "walruses" needs data from the article WALRUS, and that one about "fungi" needs data from the article FUNGUS).

So Spencer writes a series of rules for determining the singular form of plural nouns. He writes a rule "Remove final S" to handle cats → cat, a rule "Replace final I with US" to handle fungi → fungus, a rule "Remove final E" to handle vertebrae → vertebra, and so on. He ends up with the following rules:

Remove final S

Replace final ICE with OUSE

Replace IES with Y

Remove final E

Remove final EN

Replace final A with UM

Replace final I with US

When he applies his program to a series of real words, however, it doesn't always work. Here are some outputs of his program:

INPUT	INTENDED OUTPUT	ACTUAL OUTPUT
cats	cat	cat
dogs	dog	dog
walruses	walrus	walrus
foxes	fox	fox
oxen	ox	ox
bacteria	bacterium	bacterium
fungi	fungus	fungus
horses	horse	hors
chimpanzees	chimpanzee	chimpanze
algae	alga	algum
guppies	guppy	guppi
hens	hen	h
mice	mouse	mous

Q1.1. What output would Spencer’s program produce for the following words?

a. bees
b. kiwis
c. flies
d. fleas
e. geese

Q1.2. What went wrong with the program?

Q1.3. What can you determine about the order in which Spencer’s program applied the rules?

Q1.4. Could putting the rules in a different order cause the program to work? If so, what is the order?

Q1 A fox among the h.

1.1.	a.	b.	
c.	d.	e.	
1.2.			
1.3.			
1.4.			

Q1 A fox among the h.

17

<table>
<tr><td>1.1.</td><td>a. BE</td><td>b. KIWUS</td><td rowspan="2">5</td></tr>
<tr><td>c. FLI</td><td>d. FLEUM</td><td>e. GEES</td></tr>
<tr><td colspan="3">1.2. Spencer's program doesn't stop applying rules when one of them succeeds – it keeps looking for applicable rules and then applies them to the output of the previous rules. This gives the right output for "walruses" and "foxes" – it removes the "s", then continues on and removes the "e" – but goes very wrong with "horses", "hens", etc.</td><td>2</td></tr>
<tr><td colspan="3">1.3. "Remove S" must come before "Remove E"; otherwise, we would get WALRUSE, FOXE, MOU, etc. instead.
"Remove S" must come before "Remove EN"; if it came after, we would get HEN instead.
"Remove S" must come before "Replace IES with Y"; if it came after, we would get GUPPY instead.
"Remove S" must come before "Replace I with US"; if it came after, we would get FUNGU instead.
"Replace I with US" must come before "Remove E"; if it came after, we would get GUPPUS instead.
"Remove E" must come before "Replace A with UM"; if it came after, we would get ALGA instead.
"Remove E" must come before "Replace ICE with OUSE"; if it came after, we would get MIC instead.

Rules do not apply twice – that is, Spencer’s program probably applies each rule in the list to any relevant words exactly once and only goes through the list once. Otherwise, we would get things like CHIMPANZ or WALRU, etc.</td><td>5</td></tr>
<tr><td colspan="3">1.4. There is no one order of rules that will make Spencer's program work, for several reasons:

1. Applying "Remove S" before "Remove E" is necessary to get "walrus" and "fox" correct, but it's exactly this interaction that produces "hors" and "chimpanze.”

2. No order will correctly produce "mouse.” Consider the two rules (A) "Remove E" and (B) "Replace ICE with OUSE". If A comes before B, we get "mic"; if B comes before A, we get "mous" (or even "mou").</td><td>5</td></tr>
</table>