Problem 1: Gelda's House of Gelbelgarg

A frequent problem in computational linguistics is that texts often contain words the computer simply doesn't have in its dictionary. Online slang evolves very fast, people use foreign words in English passages, people make typos and invent new abbreviations, etc. You could add new words to the dictionary as fast as you can find them and the next day the program could still be stumped by a new one!

But the program doesn't have to give up. Instead, it can try to work out as much as it can. Various clues can tell a program whether something is a noun or a verb, a person or an inanimate object – and you can even work out more!

Read the webpage where customers have rated their most recent experience at a restaurant called *Gelda's House of Gelbelgarg*. Even if you've never heard of any of these dishes, you can still figure out some things about them...

Question:

Based on the reviews below, attempt to categorize the following items as either:

   I: Individual, discrete food items

   L: Liquids, undifferentiated masses, or masses of uncountable small things

   C: Containers or measurements

You may not be able to categorize them with 100% certainty, but use the category that you think is most probable for each. Choose a *single* category for each word below. (Tick the appropriate cell.)

2 POINTS FOR EACH CORRECT ANSWER   TOTAL: _____ /8

	I	L	C
färsel-försel	✓		
gelbelgarg	✓		
gorse-weebel		✓	
rolse			✓
flebba			✓
göngerplose	✓		
meembel		✓	
sweet-bolger		✓	

Gelda's House of Gelbelgarg

✪✪✪ *based on 18 reviews*

1138 Euclid Ave.
Neighborhood: Lower Uptown
Category: Ethnic, Specialty
Price Range: $$
Hours: Mon-Fri. 10:00 a.m. - 9:00 p.m.
          Sat. 10:30 a.m. - 11:00 p.m.

---

mosfel2
Reviews: 2

Report this

A hidden gem in Lower Uptown! Get the färsel-försel with gorse-weebel and you'll have a happy stomach for a week. And top it off with a flebba of sweet-bolger while you're at it!

Food	✪✪✪✪
Service	✪✪✪
Atmosphere	✪✪✪✪
Value	✪✪

SanDeE*
Reviews: 2

Report this

The portions at this place are just too big! I'd rather have half the portions at a lower price – they just bring out too many göngerplose and too much meembel for me.

Food	✪✪✪
Service	✪✪
Atmosphere	✪✪✪✪
Value	✪✪

---

wndlHghs40
Reviews: 5

Report this

i took my nana here and she said it was just like she remembered from the old country. but the service was a bit lacking – nana ordered four gelbelgarg and the waitress only brought two!

Food	✪✪✪✪
Service	✪
Atmosphere	✪✪✪
Value	✪✪

---

xMandiee7x
Reviews: 4

Report this

I found the food confusing and disorienting. Where is this from? I randomly ordered the färsel-försel and had to send them back! Three words: weird, weird, and weird.

Food	✪
Service	✪✪✪
Atmosphere	✪✪✪
Value	✪

---

wrldTrvl1977
Reviews: 11

Report this

I went to Wolserl last year for a holiday, and this is the real thing. If you order the gelbelgarg, though, make sure you also get at least one rolse of sweet-bolger – it's how the locals like it!

Food	✪✪✪
Service	✪✪
Atmosphere	✪✪✪✪
Value	✪✪✪

money@home
Reviews: 103

User is on probation

the prices are steep, but i can afford them – i make up to $75/hr working at home! find out how i do it at http://bit.ly/grhCm

Food	✪✪✪
Service	✪✪✪
Atmosphere	✪✪✪
Value	✪✪✪

bu_zhidao
Reviews: 8

Report this

not a great date spot! i got a gelbelgarg and a rolse of meembel, but my date was so disoriented that she just ended up with some gorse-weebel. :/

Food	✪✪
Service	✪✪
Atmosphere	✪
Value	✪✪

wembley2000
Reviews: 2

Report this

The food was pretty good… But I would have liked more gorse-weebel and fewer göngerplose. You really feel like the chef is skimping on the good stuff..

Food	✪✪✪
Service	✪✪
Atmosphere	✪✪✪
Value	✪

Problem 2: SAY IT IN *ABMA*

> **Abma** is spoken by more than 8,000 people making it one of the largest indigenous languages of Vanuatu, a Pacific island nation that enjoys great linguistic diversity.

Carefully study these Abma sentences, then answer the questions which follow.

NOTE: There is no separate word for 'the' or 'he' in these Abma sentences.

ABMA	ENGLISH
Mwamni sileng.	He drinks water.
Nutsu mwatbo mwamni sileng.	The child keeps drinking water.
Nutsu mwegau.	The child grows.
Nutsu mwatbo mwegalgal.	The child keeps crawling.
Mworob mwabma.	He runs here.
Mwerava Mabontare mwisib.	He pulls Mabontare down.
Mabontare mwisib.	Mabontare goes down.
Mweselkani tela mwesak.	He carries the axe up.
Mwelebte sileng mwabma.	He brings water.
Mabontare mworob mwesak.	Mabontare runs up.
Sileng mworob.	The water runs.

Now here are some more words in Abma:

ABMA	ENGLISH
sesesrakan	teacher
mwegani	eat
bwet	taro
muhural	walk
butsu-kul	palm-tree

Question 1:

Based on your analysis of the Abma words and sentences given above, translate the following *seven* English sentences into Abma. Write in the space provided to the right of each English sentence.

2 POINTS FOR EACH CORRECT SENTENCE: 0.5 OFF FOR EACH ERROR (UP TO 2 FOR EACH SENTENCE)   TOTAL: _____ /14

	ENGLISH	ABMA
1.	The teacher carries the water down.	*Sesesrakan mweselkani sileng mwisib.*
2.	The child keeps eating.	*Nutsu mwatbo mwegani.*
3.	Mabontare eats taro.	*Mabontare mwegani bwet.*
4.	The child crawls here.	*Nutsu mwegalgal mwabma.*
5.	The teacher walks uphill.	*Sesesrakan muhural mwesak.*
6.	The palm-tree keeps growing downwards.	*Butsukul mwatbo mwegau mwisib.*
7.	He goes up.	*Mwesak.*

Question 2.

Now translate these *three* Abma sentences into English.

2 POINTS FOR EACH CORRECT SENTENCE: 0.5 OFF FOR EACH ERROR (UP TO 2 FOR EACH SENTENCE)   TOTAL: _____ /6
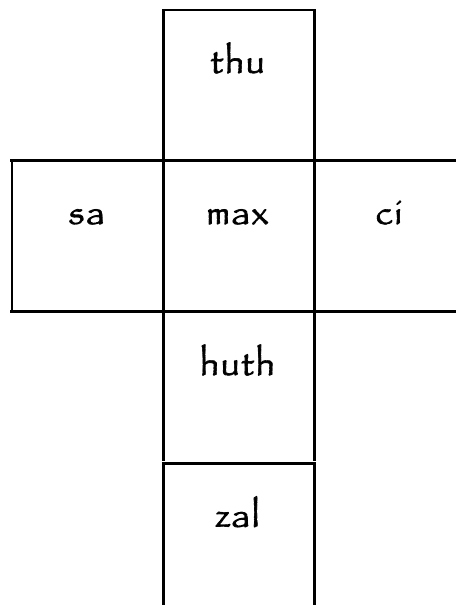
	ABMA	ENGLISH
1.	*Sesesrakan mweselkani bwet mwabma.*	**The teacher carries the taro here/in this direction.**
2.	*Sileng mworob mwisib.*	**The water runs down.**
3.	*Mwelebte bwet mwesak.*	**He brings the taro up.**

Problem 3:COUNTING IN *ETRUSCAN*

Etruscans flourished as a separate people inhabiting parts of northern Italy centred on the region now known as Tuscany for several centuries until the 1st century B.C. when they were effectively absorbed into the expanding Roman Empire. They traded throughout the Mediterranean and acquired their alphabetic writing system from the Greeks with whom they traded extensively. They left many written texts which we can easily read, as the Greek alphabet was used. However, their spoken language became extinct and because Etruscan bears little resemblance to any Indo-European language, we cannot understand the meaning of many Etruscan words.

Generally, identification of Etruscan numbers remains difficult, but the first six numbers form a group apart. They are found in epitaphs, in which age of the deceased and the number of their children is given, and in the *Book of the Mummy* which specifies dates of the periodical religious ceremonies and the size of various offerings.

On a pair of Etruscan dice, known as the Tuscan dice, there are inscribed the following six words listed here in alphabetic order: *ci, huth, max, sa, thu, zal.* Each of these words corresponds to one of the numbers between 1 and 6 (compare English "one"-1; "two"-2; *etc.*). You can see how these number words are arranged on the two-dimensional representation of a die (cube) below:

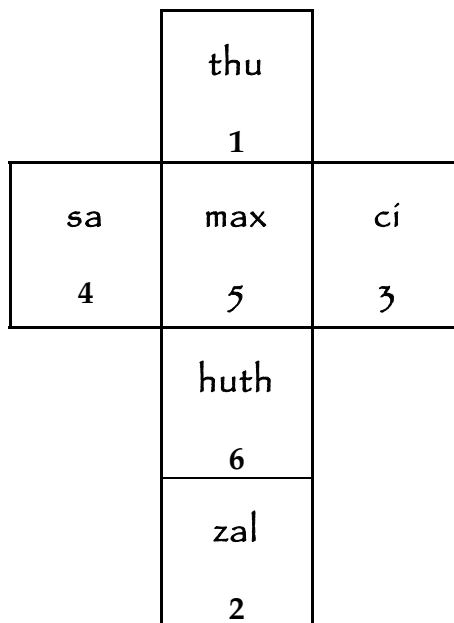|		
	thu	
sa	max	cí
	huth	
	zal	

Question 1:

Which word corresponds to which number?

At the time of the decipherment, linguists had the following clues:

1) each die in a pair of dice has four pairs of opposite faces; the sum of the number on each pair equals 7;

2) *thu*, *ci* and *zal*, in a certain order, represent *1, 2, 3*

3) *ci*, but not *thu* and *zal*, occurs very frequently in the Book of the Mummy;

4) the following pairs of words were found in epitaphs:

   thu clan; thu at; thu mezu; thu vinac; thu thuscu;

   ci clenar; zal clenar; ci atr; zal atr; ci mesur; zal mesur; ci vinacr; zal vinacr; ci thuscur; zal thuscur

5) in several ancient Mediterranean cultures the number '3' had special magic-like significance.

Write the correct number under its corresponding written version on the graphic of the die below.

1 POINT FOR EACH CORRECT NUMBER   TOTAL: _____ /6

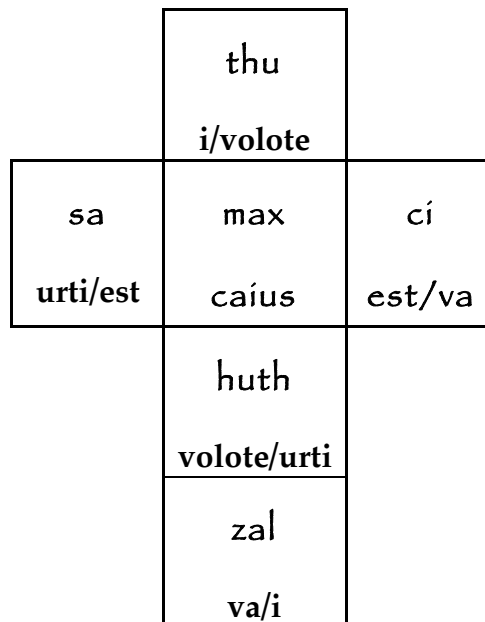|		
	thu 1	
sa 4	max 5	ci 3
	huth 6	
	zal 2	

Now here's another twist.

> It seems that Etruscans enjoyed gambling as many pairs of dice have been found. On another pair there are inscribed the following six words which we give here in their alphabetic order: *caius, est, i, va, volote, urti.* These were inscribed on the dice rather than the number words found on the "Tuscan dice ".
>
> Moreover, this choice of words is not random. It is claimed that they make up a sentence expressing a popular Etruscan proverb: *volote i va est, caius urti* meaning 'to a docile horse, the ford is pleasant'.

Question 2:

Supposing that these words were arranged on these dice to symbolize the numbers written on the Tuscan dice, inscribe each word of the proverb below its corresponding number word on this two-dimensional figure of a Tuscan die.

1 POINT FOR EACH CORRECT ANSWER - EITHER USING THE FIRST OPTION THROUGHOUT OR THE SECOND ONE - BUT NOT MIXING   TOTAL: ____ /6

	thu	
	i/volote	
sa	max	ci
urti/est	caius	est/va
	huth	
	volote/urti	
	zal	
	va/i	

Question 3:
Briefly explain your reasoning for the answer you gave to Question 2.
2 POINTS FOR <u>ONE</u> OF THESE ANSWER - WILL DEPEND ON WHICH STRATEGY THEY GO WITH   TOTAL: ____ /2

(for first set of options) the number of letters in each word symbolizes the number

(for second set of options) the order of the words in the proverb correspond to the number, e.g., first word = 1, last word = 6

Problem 4: LET'S PLAY AROUND WITH MINANGKABAU

Minangkabau is spoken by about seven million people around the West Sumatran city of Padang in Indonesia. Its speakers generally also speak Indonesian but Minangkabau is a distinct language.

Minangkabau has a number of 'play languages' that people use for fun, like Pig Latin in English. Ordinary language words are changed into play language by following just a few rules. One of these 'play languages' is called *Sorba* while another is called *Solabar*.

Here are some examples of standard Minangkabau words and their Sorba play language equivalents:

Standard Minangkabau	Sorba	English Translation
raso	*sora*	'taste, feeling'
rokok	*koro*	'cigarette'
rayo	*yora*	'celebrate'
susu	*sursu*	'milk'
baso	*sorba*	'language'
lamo	*morla*	'long time'
mati	*tirma*	'dead'
bulan	*larbu*	'month'
minum	*nurmi*	'drink'
lilin	*lirli*	'wax, candle'
mintak	*tarmin*	'request'
cubadak	*darcuba*	'jackfruit'
mangecek	*cermange*	'talk'
bakilek	*lerbaki*	'lightning'
sawah	*warsa*	'rice field'
pitih	*tirpi*	'money'
manangih	*ŋirmana*	'cry'
urang	*raru*	'person'
apa	*para*	'father'
iko	*kori*	'this'
gata-gata	*targa-targa*	'flirtatious'
maha-maha	*harma-harma*	'expensive'
campua	*purcam*	'mix'

Question 1:

Using the same rules that you have discovered from examining the words in the Table above, write the Sorba equivalents of the following standard Minangkabau words in the Table below.

2 POINTS FOR EACH CORRECT WORD: TOTAL: ____ /14

Standard Minangkabau	Sorba	English
rancak	caran	'nice'
jadi	dirja	'happen'
makan	karma	'eat'
marokok	kormaro	'smoking'
ampek	peram	'hundred'
limpik-limpik	pirlim-pirlim	'stuck together'
dapua	purda	'kitchen'

Question 2:
If you know a Sorba word, can you work backwards to standard Minangkabau? Demonstrate with the Sorba word *lore* 'good'.

1 POINT FOR 'NO'; 4 POINTS FOR DEMONSTRATION   TOTAL: ____ /5

NO. CAN ONLY WORK BACK TO A SET OF POSSIBLE WORDS: (1)

'R' PROBLEM: YOU CAN'T KNOW IF 'r' IN *LORE* WAS IN STANDARD WORD OR WHETHER IT WAS INSERTED BY SORBA 'R' RULE, E.G., STANDARD *ELO* OR *RELO* = SORBA *LORE* (2 POINTS)

FINAL SOUND PROBLEM: CAN'T KNOW IF STANDARD WORD ENDS IN CONSONANT OR ONE OR TWO VOWELS OR NOT AS SORBA DELETES FINAL CONSONANT/VOWEL FOLLOWING A VOWEL. *LORE* COULD BE DERIVED FROM *ELO, RELO, ELOA, RELOA* OR *ELOC* OR *RELOC* WHERE 'C' STANDS FOR ANY POSSIBLE FINAL CONSONANT. (2 POINTS)

Question 3:

The other 'play language' is called *Solabar*. The rules for converting a standard Minangkabau word to *Solabar* can be worked out from the following examples:

Standard Minangkabau	Solabar	
baso	*solabar*	'language'
campua	*pulacar*	'mix'
makan	*kalamar*	'eat'

What is the Solabar equivalent of the Sorba word *tirpi* 'money'? <u>tilapir</u>

2 POINTS FOR CORRECT ANSWER: TOTAL: ____ /2

Question 4:

In writing Minangkabau does the sequence 'ng' represent **one** sound (as in English *singer*) or **two** sounds (as in English *finger*)? <u>ONE</u>
1 POINT FOR ANSWERING 'ONE'

Provide evidence that supports your answer.

3 POINTS FOR GOOD ANSWER. TO GET FULL POINTS NEEDS TO CITE RELEVANT WORD FROM DATA SET.

'NG' IS ONE SOUND BECAUSE THE SORBA FOR STANDARD M. *MANANGIH* 'CRY' IS *NGIRMANA*. IF 'NG' WERE TWO SOUNDS THE SORBA WORD WOULD BEGIN WITH G AND END IN N I.E., *GIRMANAN*.

TOTAL: ____ /4

Problem 5: HELP THE COMPUTER TO FIND THE END OF A SENTENCE

> A common task that a computer needs to do with text is to identify the words and the sentences. This task is very easy for humans because we can use our understanding of the meaning of the text to identify the sentences, but a computer needs to follow very specific rules that do not require any real understanding of the text. An example of a rule is:
>
> *IF a full stop is followed by blank spaces plus a capital letter THEN this is a sentence boundary.*

Use this rule to find all the sentences in the following text:

> *The Bank of New York ADR Index, which tracks depositary receipts traded on major U.S. stock exchanges, gained 1.3% to 183.32 points in recent session. The index lost 4.63 from the beginning of July. American Depositary Receipts are dollar-denominated securities that are traded in the U.S. but represent ownership of shares in a non-U.S. company.*

Question 1.

Did this rule suffice to find all and every sentence in the above text? YES /NO (Circle your answer.)

NO POINTS FOR "YES" OR "NO". FOR 2 POINTS, NEED <u>ONE</u> OF THE ANSWERS TO THE FOLLOWING RELEVANT QUESTION.

If you answered YES, what implicit assumption did you make?

THAT A CAPITAL LETTER AT THE BEGINNING OF A PARAGRAPH (OR TEXT) MARKS THE BEGINNING OF A SENTENCE AND THAT A FULL STOP AT THE END OF A PARAGRAPH (OR TEXT) MARKS AN END OF SENTENCE.

If you answered NO, indicate any sentence the rule would fail to find, or any non-sentence that the rule would incorrectly take to be a sentence.

IT FINDS ALL THE SENTENCES EXCEPT THE LAST ONE. IT WOULD NOT KNOW IF THIS FULL STOP MARKS THE END OF A SENTENCE BECAUSE IT IS NOT FOLLOWED BY A CAPITAL LETTER.

TOTAL: ____ /2

Question 2:

Give *two* different examples of text that would make the rule fail to split into correct sentences. Your examples should illustrate different types of failure. They should *not* include the type(s) of failure you may have discovered in answering Question 1.

ANY TWO GOOD EXAMPLES GET A POINT EACH, SEE NON-EXHAUSTIVE EXAMPLES BELOW.

TOTAL: ____ /2

(1) A numeral is not a capital letter, so the problem here is that the computer will not split two sentences if the second sentence starts with a numeral. e.g., "10 banks closed today. 10 more will close before the end of the year."

(2) Where full stop does not mark end of sentence as with abbreviations followed by full stop + space + capital letter, e.g. *"Dr. Watson looked puzzled."* would be divided into 2 sentences [Dr.] and [Watson looked puzzled.] OR *"My grandfather was onboard the R.M.S. Titanic."* would be divided before *Titanic*. Similarly in *"C.J. Dennis wrote 'The Sentimental Bloke'."*

(3) Where end of sentence is marked by symbol other than full stop, e.g., *"What a man! He broke every record in the book."* or *"Why did Dr. Watson look puzzled? The solution was elementary."* Each of these would be treated as 1 sentence by the rule.

Question 3:

How would you need to revise the initial rule in order to handle any *two* of the problematic examples that you have identified so far (i.e., in answering Questions 1 and 2)?

3 POINTS FOR EACH OF THE TWO SUGGESTED REVISIONS OF THE RULE WHICH WOULD GO SOME WAY TO ACTUALLY SOLVING THE PROBLEM.

TOTAL: ____ /6

HERE ARE SAMPLE ANSWERS OF THE TYPE WE ARE LOOKING FOR HERE. (THEY DON'T NEED TO ACTUALLY REVISE THE WORDING OF THE RULE, BUT THEY DO NEED TO MAKE CLEAR WHERE IT WOULD NEED TO BE CHANGED OR REVISED, OR ADDED TO.)

FOR PROBLEM 1 (IDENTIFIED IN ANSWER TO Q2) NEED TO ADD "OR NUMERAL" AFTER "CAPITAL LETTER".

IF a full stop is followed by blank spaces plus a capital letter OR A DIGIT/NUMERAL THEN this is a sentence boundary.

FOR PROBLEM 2, which can be divided into two different but related problems, NEED TO HAVE AN EXHAUSTIVE LIST OF EXCEPTIONS AFTER "FULL STOP", AND ALSO NEED TO EXCLUDE SEQUENCES LIKE R.M.S. OR C.J.

> *IF a full stop NOT PRECEDED BY ANY MEMBER OF THIS LIST*: Dr, Ms, Mrs, Prof, etc, viz, (*etc*) OR A REPEATED SEQUENCE OF A CAPITAL LETTER FOLLOWED BY FULL STOP *is followed by blank spaces plus a capital letter THEN this is a sentence boundary.*

FOR PROBLEM 3, NEED TO ADD ! AND ? TO "FULL STOP"

IF a full stop OR EXCLAMATION MARK OR QUESTION MARK is followed by blank spaces plus a capital letter THEN this is a sentence boundary.

Problem 6: TURKISH DELIGHT

Turkish is a language from the Turkic group of the Altaic language family. It is spoken by 60 million people in Turkey and roughly 10 million other people around the world.

NOTE:

ç sounds like ch in *church*, c like j in *job*, ş like sh in *shoe*.

e, i, o, and u are pronounced approximately like in *red*, *reed*, *rod*, and *rude*, respectively.

ö and ü are like e and i respectively, but pronounced with the lips rounded.

ı (written like an "i" but without a dot on top) is like u, but pronounced with the lips spread (unrounded).

Here are some Turkish words and their English equivalents. Examine the Turkish words closely to see how each is formed, paying close attention to the vowels.

	Turkish	English
A	*güreşçi*	wrestler
B	*ikbalsiz*	unsuccessful
C	*gözcü*	sentry, eye-doctor
D	*isimsiz*	nameless
E	*ormancı*	forester
F	*sonsuz*	endless
G	*içkici*	drunkard
H	*takatsiz*	lacking strength
I	*barutçu*	gunpower-maker
J	*sütsüz*	without milk
K	*balıkçı*	fisherman
L	*parasız*	cashless
M	*mumcu*	candle-maker

Question 1:

Native Turkish words (as opposed to loan words from other languages) are restricted in the combinations of vowels that may co-occur in the same word. (Linguists refer to this phenomenon found in many of the world's languages as "Vowel Harmony".)

Two of the above words are formed in a slightly different way from the others because they are based on loan words. Identify those two words.

Put their corresponding letters (from column 1) in the boxes:

5 POINTS FOR *BOTH* CORRECT ANSWERS; 2 POINTS FOR ONLY *ONE* CORRECT:

TOTAL: ＿＿ /5

B	H

Question 2:

Translate these two words into Turkish (write one letter in each box, starting from the left; it is ok to leave blank boxes after your answer. Use lowercase letters only.)

3 POINTS EACH   TOTAL: ＿＿ /6

1 FOR CORRECT ROOT; 1 FOR CORRECT SUFFIX; 1 FOR CORRECT VOWEL IN SFX

milkman	s	ü	t	c	ü							
blind	g	ö	z	s	ü	z						

Question 3:

Here are two more Turkish words (not loans from another language):

dil	language
kalıp	form

Translate the following *four* words into Turkish:

(Write one letter in each box, starting from the left; it is ok to leave blank boxes after your answer. Use lowercase letters only.)

3 POINTS FOR EACH CORRECT WORD: (ALLOW C OR Ç IN 'AGENT' SUFFIX) 1 FOR CORRECT ROOT; 1 FOR CORRECT SUFFIX; 1 FOR CORRECT VOWEL IN SFX

TOTAL: _____ /12

linguist	d	i	l	c	i						
mute	d	i	l	s	i	z					
mould-maker	k	a	l	ı	p	c	ı				
shapeless	k	a	l	ı	p	s	ı	z			