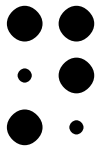
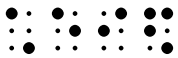
2009 Solutions

(A) Tenji



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 alphabet, such as Japanese, Korean, or Chinese, are often organized in a very different manner!

To the right 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.

karaoke 

1. 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. haiku 

b. sake 

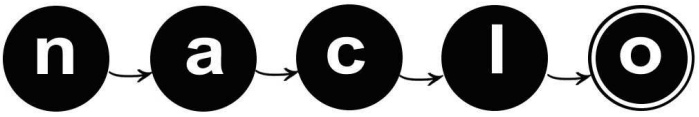
c. katana 

d. kimono 

e. koi 

f. atari 

2. What are the following words?



(A) Tenji

a. karate  b. anime 

3. Write the following words in *tenji* characters:

a. samurai  b. miso 

(B) Spelling Tutor

The spelling tutor computes the EDIT DISTANCE between the given word spelling and the correct spelling. We use the standard definition of operations required for converting one of the two given strings into the other, where each operation is one of the following three:

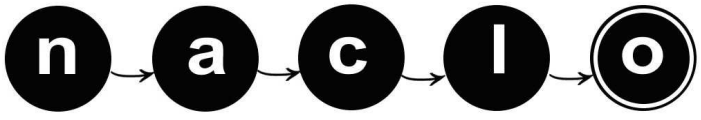
•Removal of a letter.
•Insertion of a letter (anywhere in the string).
•Replacement of a letter with (any other) letter.

The spelling tutor converts the edit distance into a comment using the following scheme:

DIST	COMMENT
0	no comment; correct
1	almost right
2	quite close
3	a bit confusing
4	very confusing

The examples given in the problem do NOT show comments for the edit distance of 5 or more, because Christopher Robin never makes so many mistakes, not even in long and delicate words.

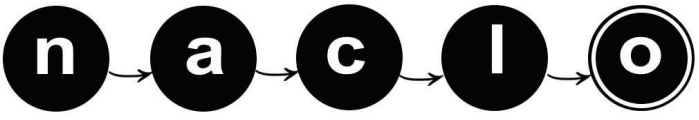
The edit distances and related comments for the given misspellings of "typo" are as follows:



(B) Spelling Tutor

MISSPELL	DIST	COMMENT
oooo	3	a bit confusing
opyt	4	very confusing
pyto	2	quite close
typ	1	almost right
typa	1	almost right
typotypo	4	very confusing

The spelling tutor computes the EDIT DISTANCE between the given word spelling and the correct spelling. We use the standard definition of the EDIT DISTANCE; that is, this distance is the minimal number of operations required for converting one of the two given strings into the other, where each operation is one of the following three:



(C) Orthography Design

Orthography design is the process of developing an alphabet and spelling rules for a language. A good orthography has several features:

> Given a spoken word, there's no question of how to spell it.
> Given a written word, there's no question of how to pronounce it.
> In the modern world, it's increasingly important that it be reasonably easy to type!

Quechua is spoken today by millions of people in Peru, Ecuador, and Bolivia, the descendents of the citizens of the Incan Empire. Quechua speakers are rapidly joining the Information Age, and both Google and Microsoft Windows now come in Quechua!

Like in English, there are more sounds in Quechua than there are letters on a keyboard, but there are ways around that. For example, we can assign one letter to multiple sounds so long as a reader can always predict, from its position in the word or from other letters in the word, which sound is meant. So if the sound [b] only ever occurs right after [m], and [p] never occurs right after [m], we can just write "p" for both, since you'll be able to predict from the previous letter whether "p" means [b] or [p].

This "phonemic principle" is the central principle of most orthographies, not just because it reduces letters but also because our minds categorize sounds in the same way.

Here are 18 words in Cuzco Quechua, as they are pronounced but not necessarily as they are written. [q] and [χ] represent special sounds that don't occur in English.

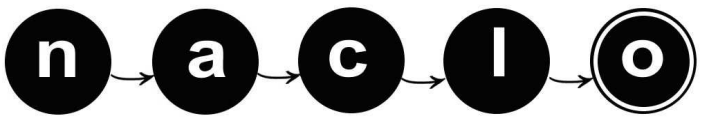
2009 Solutions

(C) Orthography Design

awtu	car	**qasi**	free	**seqay**	to climb
kanka	roasted	**qatoχ**	merchant	**sikasika**	caterpillar
karu	far	**qatuy**	to barter	**sipeχ**	murderer
kiru	teeth	**qatisaχ**	I will follow	**sipiy**	to kill
kisa	nettle	**qelqaχ**	writer	**soχtaral**	sixty cents
kisu	cheese	**qelqay**	to write	**sunka**	beard
kunka	neck	**qolqe**	silver	**toχra**	ball of ash
kusa	great	**qosa**	husband	**uyariy**	to listen
layqa	witch	**qosqo**	Cuzco	**uywaχ**	caretaker
oqe	spotted	**saqey**	to abandon	**waleχ**	a lot
qasa	frost	**saχsa**	striped	**weqaw**	waist

Notes:

- It is quite expected that few if any contestants are going to get all 20 points. There are going to be entirely correct and well-explained answers that don't get quite as many points as another entirely correct and well-explained answer because the latter was more thorough. (In the first version of this rubric, we found that the minimal "correct score" was about 12, give or take.)
- Some solvers will have completely misunderstood what they were supposed to do. This is too bad, but they don't get any points for well-meaning but bizarre answers! This is a contest, rather than a homework assignment, and for some of the puzzles the puzzle *is* to figure out what's being asked.



2009 Solutions

(C) Orthography Design

- Half-points may be awarded.
- It is *not* necessary for a complete solution that the solver to chooses <u> and <i> to be basic rather than <o> and <e>. From a phonemic point of view, the label of a sound is arbitrary – these could be Smiley Face and Labialized Smiley Face for all we care – and from an orthographic point of view, it doesn't matter, they're all just symbols.

I. Show that we don't need separate letters for [q] and [χ]. (3pts)

- 1a. **1pt**. for noticing that they never occur in the same environments
- 1b. **1pt**. for correctly specifying what these environments are.
- 1c. **1pt**. for clearly explaining why this means they can be the same letter. (This explanation doesn't have to be *long*, just clear.)

II. Show that we can't represent [a] and [i] by the same letter. (3pts)

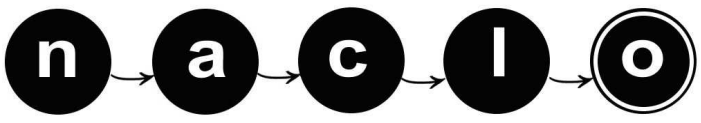
- 2a. **1pt**. for noticing that they do occur in the same environments
- 2b. **1pt**. for finding a minimal pair, like "karu ~ kiru" or "qasa ~ qasi". (If they have this but not 2a., give them the point for 1a anyway, since this subsumes that.)
- 2c. **1pt**. for clearly explaining why this means they have to have different letters.

III. Show that we can't represent [a] and [e] by the same letter. (3pts)

- 3a. **1pt**. for noticing that they do occur in the same environments.
- 3b. **1pt**. for noticing the pair "saqey" ~ "seqay". (If they have this but not 3a, again give them that point anyway.)

3c. **1pt**. for clearly explaining why this means they have to have different letters.

IV. Most modern Quechua orthographies get by with only three of the five vowels [a], [e], [i], [o], and [u]. Show how this is possible. (11pts)



2009 Solutions

(C) Orthography Design

First, they should establish which sounds *can't* be merged into a single letter:

- 4a**. 1pt**. for finding a pair kisa ~ kisu or kanka ~ kunka.
- 4b. **1pt**. for finding the pair qasa ~ qosa.
- 4c. **1pt**. for finding the pair kisa ~ kusa.

4d. **1pt**. for recognizing the relevance of the pairs in II and III to this question.

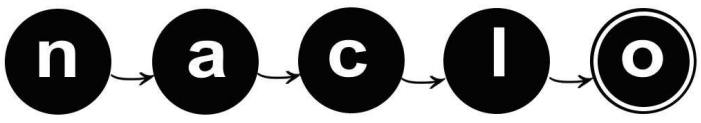
Second, three points for deducing which sounds can be merged:

- **1pt**. for figuring out that either [e]~[i] and [o]~[u], or [e]~[u] and [o]~[i]. (They don't have to get both for this point.)
- **1pt**. for figuring out that both of these are possible.
- **1pt**. for clearly explaining how this follows from the facts above and in parts II and III – that given the minimal pairs, these two are the only solutions that don't cause two different words to be spelled the same.

Four points are available for:

•Up to **2** points for determining the conditioning environment for the difference: [e] and [o] when next to [q] and [χ], [i] and [u] elsewhere. (1 point each for the completeness of the description and the clarity of the explanation.)

•Up to **2** points for determining, based on the alternations *qelqay ~ qelqaχ*, *qatuy ~ qatoχ*, and *sipiy ~ sipeχ*, that [o]~[u] and [e]~[i] is the better or more likely of the possible solutions. (1 point for noticing the pattern and 1 point for correctly deducing the right phonemicization.)



2009 Solutions

(D) Guarani

The Guarani verb consists of:

1. prefix ***n(d)(a)-***, if negation exists;
2. person and number of the subject: ***a-*** 'I', ***o-*** 'he', ***ja-*** 'we', ***pe-*** 'you (pl.)';
3. root;
4. ***-(r)i***, if negation exists;
5. ending ***-ma*** for past tense or ***-ta*** for future tense.

where:

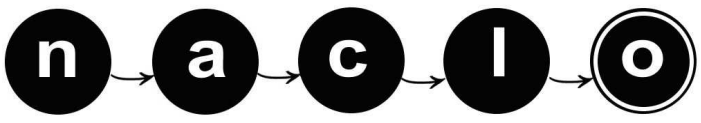
- the negative prefix should start with ***n*** (rather than ***nd)*** in case the root of the verb contains any nasal sound
- the vowel ***a*** is dropped from the negative prefix in case the personal prefix starts with a vowel.
- if a future tense is to be negated, the suffix is ***-mo'ãi,*** rather than ****-(r)i-ta;***
  the negative suffix is ***-ri*** after the vowel ***i; -i*** otherwise.

Part 1

akaruma	I was eating
ojupita	He will be waking up
ndavo'omo'ãi	I will not be taking
napekororõ	you are not crying
ndapyhyima	I wasn't catching

Part 2

you are not shooting	ne-pe-mbokapu-i
he is not singing	ndo-purahei-ri
we will be eating	ja-karu-ta
I will not be singing	nda-purahei-mo'ãi



(E) Summary

An extractive summarizer scores each sentence according to some criteria that are correlated with being a good summary sentence. Then it picks the top 3 sentences from each story based on the sum of its scores on the different criteria.

The first criterion measures primacy. The first sentence gets 3 points, the second 2, and the third 1, to account for the likely increased importance of the initial sentences. Then multiples of .1 are added, starting with .0 for the last sentence, to break the other criteria's ties in favor of earlier sentences.

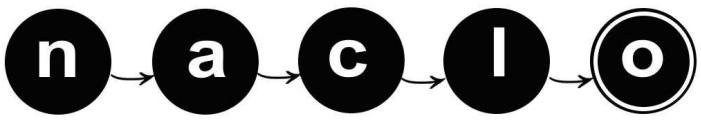
The second criterion measures recency. The last sentence gets 3 points, the second-last 2, and the antepenultimate 1, to account for that they're likely summaryish themselves.

The third criterion counts named entities in the sentence, since they're likely to be the most important actors.

The fourth criterion counts words from the title that appear in the sentence (after reducing each word to a stem; e.g. struck = strike). These sentences are likely to pertain most immediately to the topic of the story.

The fifth criterion counts named entities introduced for the first time in this sentence. The first mention of a named entity is probably important for understanding its role.

The sixth criterion counts past-tense verbs. Current information is probably more important, and past-tense verbs are slightly less likely to give new information.



(E) Summary

Solutions

Story 1, sentence 2, criterion 3 - change to 1
Story 1, sentence 2, criterion 5 - change to 1
Story 1, sentence 2, total - change to 2.3

Story 1, sentence 5, criterion 3 - change to 0
Story 1, sentence 5, total - change to 3.0

Story 2, sentence 1, criterion 1 - change to 3.9
Story 2, sentence 1, total - change to 8.9

Story 2, sentence 2, criterion 1 - change to 2.8
Story 2, sentence 2, total - change to 4.8

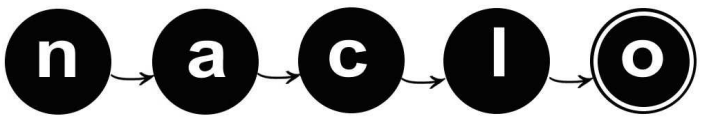
Story 2, sentence 3, criterion 1 - change to 0.6 (similarly change the next 6 numbers down from this one to be 0.1 less than before)
Story 2, sentence 2, total - change to 2.6 (this affects the totals for the next 6 sentences)

Story 2, sentence 5, criterion 5 - change to 2
Story 2, sentence 5, total - change to 4.4

Story 2, sentence 8, criterion 5 - change to 1
Story 2, sentence 8, total - change to 4.1

Part 2:

1.7 0 2 3 2 -4 4.7



2009 Solutions

(F) Anaphora

In the following dialogues, the computer is misunderstanding the human because it is misinterpreting certain words:

Example 1

Human: There's this restaurant on Bancroft that's supposed to be really good that I heard about from my mother. Can you help me find it?

Computer: Where did you last see your mother?

Example 2

Human: My friend is looking for information about on-line movie databases. What should I tell him?

Computer: You can tell movie databases reviews of movies you've seen.

Example 3

Human: My parents are having a hard time choosing a new digital camera. How can I help them?

Computer: I'm afraid I don't understand your question. People don't help cameras.

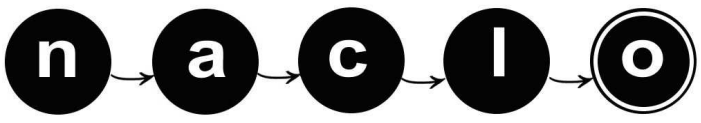
A. What are the words that the computer is misinterpreting?

Ex1: ____him_____ Ex2: ____it_____ Ex3: ____them_____

*** 1 correct: 0 points
2 correct: 1 points
3 correct: 2 points

B) What rule is the computer using to interpret those words? Interpret pronouns as referring to the last noun phrase in the previous sentence.

*** 2 points for any answer that states a generalization that applies to the final NP in each first sentence and nothing else.



(F) Anaphora

*** 1 point for any answer that applies only to the final noun in eachsentence, or to some intermediate category that doesn't fit the data.

C) Give a better rule that would make the computer interpret the words correctly in these examples.

***1 point for just about anything that either works on all three given sentences or is distinct better than the computer's rule,
e.g.:

-Interpret pronouns as referring to the previous sentence's first noun.

-Interpret pronouns as referring to a noun in the previous sentence with the same number/gender properties.

-Interpret pronouns as referring to the previous sentence's subject.

-Check for sentences of parallel syntactic structure first, and refer to a noun (phrase) in the same place if there is one.