a С n

National Sponsors





CAMBRIDGE UNIVERSITY PRESS The Second Annual

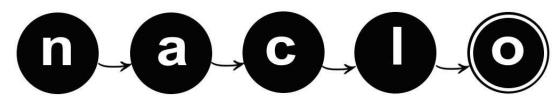
North American Computational Linguistics Olympiad

2008

www.naclo.cs.cmu.edu

INVITATIONAL ROUND

March 11, 2008



The North American Computational Linguistics Olympiad www.naclo.cs.cmu.edu

Contest Booklet

Your Name:	
Registration Number:	
Your School:	
City, State, Zip:	
Your Grade:	
Start Time:	
End Time:	
Your Teacher's Name: _	

Please also make sure to write your registration number on each page that you turn in. Each problem will be graded by a different judge and pages with no registration numbers will not be graded.

Rules for participants

Welcome to the **invitational** round of the second annual North American Computational Linguistics Olympiad! Only 1/6 of the participants (128 out of 763) in the first round got this far. Congratulations!

Because this is an international competition, in order to be completely fair to all participants across North America, we need you to read, understand and follow these rules completely.

- 1. The contest is five hours long. Part I should be done in 3.5 hours and Part II in 1.5 hours.
- 2. Follow the facilitators' instructions carefully.
- 3. If you want clarification on any of the problems, talk to a facilitator. The facilitator will consult with the jury before answering.
- 4. You may not discuss the problems with anyone except as described in 3.
- 5. Each problem is worth a specified number of points, with a total of 100 points in the contest. About half of the points are given for "practice," that is, for getting the right answers. The rest of the points are given for "theory," that is, for the quality of your written description of how you solved the problem. It is therefore essential that you show all your work.
- 6. We will grade only work in this booklet. All your answers should be in the spaces provided in this booklet.
- Write your name and registration number on each page: Here is an example: Jessica Sawyer #850
- 8. Each problem has been thoroughly checked by linguists and computer scientists as well as students like you for clarity, accuracy, and solvability. Some are more difficult than others, but all can be solved using ordinary reasoning and analytic skills. You don't need to know anything about linguistics or about these languages in order to solve them. If we have done our job well, almost no one will solve all these problems completely in the time allotted. So don't be discouraged if you don't finish everything.
- 9. After the invitational round on March 11, we will rank all students based on their scores on that round only. In the case of ties in the invitational round, we will use the score in the first round as tie breaker. In the event of a second tie, we will look at the scores on the hardest problems (as determined by the average scores of all students), in descending order.

Oh, and have fun!

Rules for facilitators

Thank you for your invaluable help with running this contest. Without you, this event would not happen!

Please read the rules carefully.

- 1. You need to send us all problem pages (even if a student didn't work on a particular problem).
- 2. We need to receive all submissions by the Monday following the contest. If you mail them on Tuesday or Wednesday, you can use regular mail. After that date, use Priority mail. You will receive an email with the addresses of each grader. ALTERNATIVELY, you can scan each submission into a separate PDF file and email it to

radev@umich.edu using NAMCLOxxx.pdf as the file name (where xxx is the reg. number).

- 3. Make sure to keep copies of all problems in case something gets lost in the mail. Also, include a cover sheet in each envelope indicating the name of the site, an email and phone number for contact, as well as the registration numbers of the students included. We will be using the same registration numbers as in round 1.
- 4. Part I has 5 problems and Part II has 2 problems. Each site will start at 9 AM at which time only Part I will be handed out to the students. At 12:30 PM, Part I should be collected. At that time, you can decide to have a break. As soon as the contest restarts, Part II will be handed out. Students can work on it for another 1.5 hours. Students are not allowed to work on Part I after the first 3.5 hours have passed and they are not allowed to work on Part II before finishing Part I. In the event that all students at your site finish Part I early, you can move on to the break and Part II, however, the time allotted to part II will remain at 1.5 hours.
- 5. Clarifications will again be available through naclo08jury@umich.edu
- 6. Top scorers from this round will be eligible to participate on the North American team at the International Linguistics Olympiad, to be held in Bulgaria this Summer. The North American team or teams will consist of four or eight of the top students, depending on funding. The NACLO organizers will also pick a small number of students as alternates.

PART I (3.5 hours)

YOU ARE ONLY ALLOWED TO WORK ON PART I **BEFORE** THE BREAK. AT THAT TIME, RETURN IT TO YOUR FACILI-TATOR.

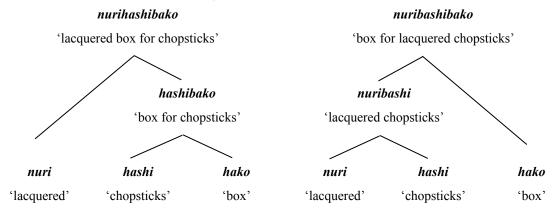
(10 points)

(F) Fakepapershelfmaker (1/2)

In English, we can combine two nouns to get a compound noun, such as in 'mailbox' or 'sandcastle'. We can do this in Japanese as well, but just sticking the two words together isn't enough. Instead, the words themselves undergo predictable changes:



Compound words can then be compounded again, creating compounds with three or more members. Study the diagrams below carefully. You'll notice that the order in which the compound is built affects both the meaning and the final form of the word.



F1. The following is a list of several Japanese words with their English meanings. Use this word bank to write definitions of the Japanese compounds (a)-(f). Be very specific with how you phrase your definition. If your definition is ambiguous (has two meanings), it will not be counted.

sakura	cherry blossom	kami	paper	nise	fake
shiru	soup	tana	shelf	tsukuri	maker
iro	color(ed)	tanuki	raccoon	hako	box
(a) <mark>niseta</mark> n	nukijiru				
b) nisedanukijiru					
c) irogamibako					
d) irokamibako					
e) nisezakuradana					
(f) nisesakuradana					

(F) Fakepapershelfmaker (2/2)

F2. Match the following four-member Japanese compound words on the left with their English meanings on the right. (Some will require you to stretch your imagination a bit!) One of the Japanese words will correspond to two possible English meanings.

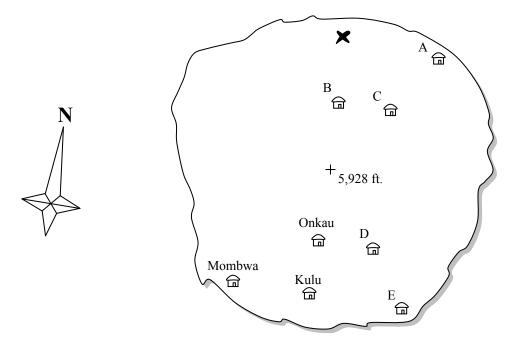
(1) a fake (fraudulent) shelf-maker made of paper	(A) nisegamidanadzukuri
(2) a maker of fake shelves for paper	(B) nisekamitanadzukuri
(3) a fake (fraudulent) maker of shelves for paper	(C) nisegamitanadzukuri
(4) a shelf-maker made of fake paper	(D) nisekamidanadzukuri
(5) a maker of shelves for fake paper	

F3. Explain your answers to F1 and F2 in the space provided below.



(G) Manam, I'm Anam (1/2)

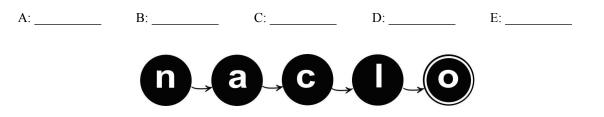
Manam Pile ("Manam Talk") is a Malayo-Polynesian language spoken on Manam Island off the coast of Papua New Guinea. Manam is one of the most active volcanoes in the world, and during violent eruptions the population must be evacuated to the mainland.



Below, a Manam islander describes the relative locations of the houses above.

- 1. Onkau pera kana auta ieno, Kulu pera kana ilau ieno.
- 2. Mombwa pera kana ata ieno, Kulu pera kana awa ieno.
- 3. Tola pera kana auta ieno, Sala pera kana ilau ieno.
- 4. Sulung pera kana awa ieno, Tola pera kana ata ieno.
- 5. Sala pera kana awa ieno, Mombwa pera kana ata ieno.
- 6. Pita pera kana ilau ieno, Sulung pera kana auta ieno.
- 7. Sala pera kana awa ilau ieno, Onkau pera kana ata auta ieno.
- 8. Butokang pera kana awa auta ieno, Pita pera kana ata ilau ieno.

G1. Onkau's, Mombwa's, and Kulu's houses have already been located on the map above. Who lives in the other five houses?

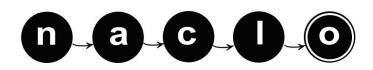


(G) Manam, I'm Anam (2/2)

G2. Arongo is building his new house in the location marked with an X. In three Manam Pile sentences like the ones on the previous page, describe the location of Arongo's house in relation to the three closest houses.

1	 	
2	 	
3		
<i>J</i>	 	

G3. Explain your answers to G1 and G2 in the space provided below.



(10 points) (H) Thorny Stems (1/3)

Wouldn't it be nice if your computer could understand English? In this problem, you will write down a small set of rules encoding one piece of your knowledge about English.

The problem you will approach is called "stemming." You know that "work", "working", "works", and "worked" are all forms of the same verb: "work". Similarly, "guesses" and "guess" are both forms of the same noun: "guess". Below you will find a list of pairs of a word and its stem, both nouns and verbs. Your goal is to write down a list of rules which is as short as possible, but covers all of the example pairs. You must also list exactly **one** exception for every rule which has one.

Example

Rule 1: If a word ends in <u>s</u>, then replace <u>s</u> with <u>s</u> to form the stem. **Rule 2:** If a word ends in <u>s</u>, then replace <u>s</u> with <u>to form the stem</u>. **Rule 3:** Otherwise the word is its own stem.

Let's look at how these rules will apply to a few examples. We always use only the first numbered rule that applies. For the word "work", Rules 1-2 do not apply, so we are left with Rule 3, "work" is its own stem. For the word "works", Rule 1 does not apply, but Rule 2 does, so the stem of "works" is formed by replacing the final "s" with nothing - i.e. deleting it, to form "work". Finally, for the word "grass", Rule 1 does apply, and so we replace "ss" with "ss", i.e. the word is unchanged and then we stop.

Your goal is to write one list of rules which will apply to both the nouns and the verbs listed on the next page.

Exceptions

The rules you write will not always work. Any word for which your rules give the wrong stem is called an "exception". You will write down exceptions for your rules - an exception is written next to the first rule whose "if" part applies to it. For example, "guess" is not an exception to the rules above, since even though Rule 2 does not handle it, Rule 1 (which comes first) does. However, "cries" is an exception (these rules gives its stem as "crie" instead of "cry"), and it should be written next to Rule 2 as follows:

If a word ends in <u>s</u>, then replace <u>s</u> with <u>s</u> to form the stem. Exception: <u>-none-</u> If a word ends in <u>s</u>, then replace <u>s</u> with <u>to form the stem</u>. Exception: <u>cries</u>



(H) Thorny Stems (2/3)

Judging

Your score will be determined according to the following criteria:

You should have rules to cover all the words in the list below.

You should use as few rules as you can.

You should list an exception next to as many rules as you can.

Words and stems

NOUNS			
word	stem		
backs	back		
books	book		
chiefs	chief		
companies	company		
duties	duty		
dwarves	dwarf		
grass	grass		
moss	moss		
potatoes	potato		
presidents	president		
roses	rose		
shelves	shelf		
stores	store		
stapler	stapler		
times	time		
toe	toe		
tomatoes	tomato		
wives	wife		

VE	RBS

word	stem
cried	cry
cries	cry
dished	dish
flies	fly
married	marry
killed	kill
listened	listen
ordered	order
resorts	resort
sailing	sail
tailing	tail
tell	tell

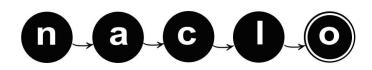


(H) Thorny Stems (3/3)

H1. Your rules

You do not need to use all the blank rules below. Cross out any rules you do not use.

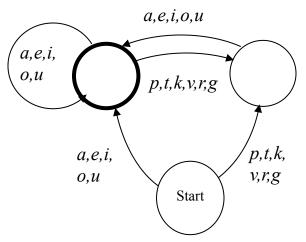
- 1. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 2. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 3. If a word ends in ______, then replace ______ with _____ to form the stem. Exception:
- 4. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 5. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 6. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 7. If a word ends in ______, then replace ______ with _____ to form the stem. Exception:
- 8. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 9. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 10. If a word ends in _____, then replace _____ with _____ to form the stem. Exception:
- 11. Otherwise the word is its own stem.
- H2. Explain your reasoning in the space below.



(I) aw-TOM-uh-tuh (I/2)

Finite-state automata (FSA) are a type of abstract "machine" with many possible uses. One possible use is to guess what language a document (such as a webpage) is in. If we make an automaton that can distinguish between possible English words and impossible ones, and then give it a webpage with a bunch of words that are impossible in English (like "*aioaepa*" or "*ragaiiare*"), we can be pretty sure that the webpage isn't written in English. (Or, at least, isn't *entirely* written in English.)

Here is a finite state automaton that can distinguish between possible and impossible words in Rotokas, a language spoken on the island of Bougainville off the coast of New Guinea. Rotokas has a very simple system of sounds and allows us to create a very small FSA.

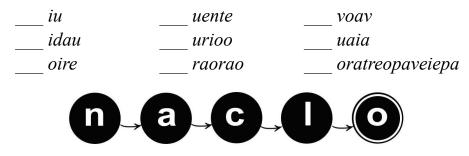


An FSA works like a board game. Choose a word, and place your pencil on the space marked "Start". Going through the letters of the word one at a time, move your pencil along the path marked with that letter. If the word ends and you're at a space marked with a thicker circle, the word succeeds: it's a possible Rotokas word! If the word ends and you're not at a thicker circle, or you're midway through the word and there's no path corresponding to the next letter, the word fails: it's *not* a possible Rotokas word!

Try it out with these possible and impossible words; the automaton should accept all the possible words and reject the impossible ones.

Possible Rotokas words		Impossible Rotokas words	
tauo	kareveiepa	grio	ouag
puraveva	ovokirovuia	ovgi	vonoka
avaopa	ouragaveva	gataap	oappa

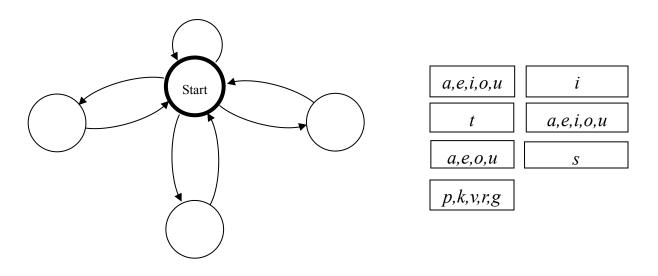
I1. Now, using the automaton above, put a check mark next to each possible Rotokas word:



(I) aw-TOM-uh-tuh (2/2)

I2. Actually, the system that people on Bougainville use to write Rotokas is a little more complicated than the one we've presented here. In addition to the eleven letters above, the real Rotokas alphabet has a twelfth letter, S. This letter represents the sound "s" or "ts", which in Rotokas only occurs in very specific situations.

Below is a skeleton of an FSA for Rotokas with all the path labels removed and set to one side.



Replace the path labels so that exactly half of the words below succeed and the other half fail.

oisio	tiravau	saiuu	kotoe
uasau	utsa	sioparoia	parauos
puapuata	sisigarue	porouativeve	aasiia

I3. Why do T and S get their own paths? What is special about these letters?



REGISTRATION #:

(20 points)

(J) The Curragh of Kildare (1/2)

And straight I will repair To the Curragh of Kildare For it's there I'll find tidings of my dear [Irish Folk Song]

In Ireland, each place name has two versions with equal legal status – an English one and an Irish one. Below are some place-names in their two versions and translations of the Irish ones.

	English	Irish	Translation of Irish name
1	Glenamuckaduff	Gleann na Muice Duibhe	Valley of the Black Pig
2	Clonamully	Cluain an Mhullaigh	Meadow of the Summit
3	Buncurry	Bun an Churraigh	Base of the Marsh
4	Curraghmore	An Currach Mór	The Big Marsh
5	Annaghanoon	Eanach an Uain	Fen of the Lamb
6	Dunard	An Dún Ard	The High Fort
7	Bunagortbaun	Bun an Ghoirt Bháin	Base of the White Field
8	Gortnakilly	Gort na Cille	Field of the Church
9	Binbane	An Bhinn Bhán	The White Peak
10	Ballyknock	Baile an Chnoic	Town of the Hill
11	Ballynaparka	Baile na Páirce	Town of the Park
12	Kilcarn	Cill an Chairn	Church of the Mound
13	Killeshil	An Choill Íseal	The Low Wood
14	Clashbane	An Chlais Bhán	The White Pit
15	Bunbeg	An Bun Beag	The Small Base

Sometimes the English name is no more than a translation of the Irish one:

16	Blackabbey	An Mhainistir Dhubh
17	Bigpark	An Pháirc Mhór
18	Castlepark	Páirc an Chaisleáin
19	Woodland	Talamh na Coille



(J) The Curragh of Kildare (2/2)

J1. What would the Irish names of the following towns and villages be? Provide a translation for each one. If you think more than one Irish name could correspond to a given English name, give all of them:

	English	Irish	Translation of Irish name
20	Mullaghbane		
21	Killananny		
22	Knocknakillardy		
23	Gortnabinna		
24	Clashgortmore		
25	Killbeg		
26	Blackcastle		Black castle

J2. Explain your reasoning and provide any additional observations about this problem.



PART II (1.5 hours)

YOU ARE ONLY ALLOWED TO WORK ON PART II **AFTER** YOU HAVE RETURNED PART I TO YOUR FACILITATOR

REGISTRATION #:

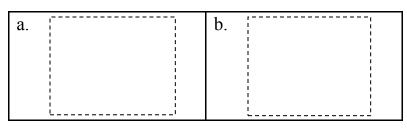
(15 points) (K) Tzolk'in (1/3)

Central to Mayan thought and religion was the concept of time, which was held to be cyclical and without beginning or end. Accurate reckoning of the day was highly important for ceremonial, political, and agricultural reasons, and to this end the Mayas kept several calendars of different length, all cycling simultaneously, while also keeping track of the exact number of days since August 11, 3114 BC.

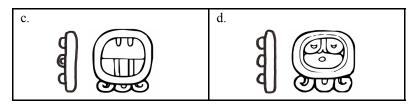
The *tzolk'in* ritual and ceremonial calendar is the oldest and most important of these, and is not exclusively Mayan – it was shared by all of the ancient civilizations of Mexico, and predates Classical Mayan civilization by hundreds of years. It is also the main calendar to have survived, and is still in use today in some Mayan villages.

Given the conception of time as cyclical, it is appropriate that the *tzolk'in* has no identifiable beginning or end, but simply assigns names to days in an infinite loop. Below we've taken a calendar for our August and September, but inserted the day names according to the *tzolk'in*.

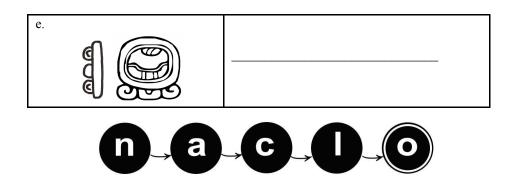
K1. What are the Mayan names of the days labeled *a* and *b* on the calendar? Draw the appropriate glyphs in the boxes below.



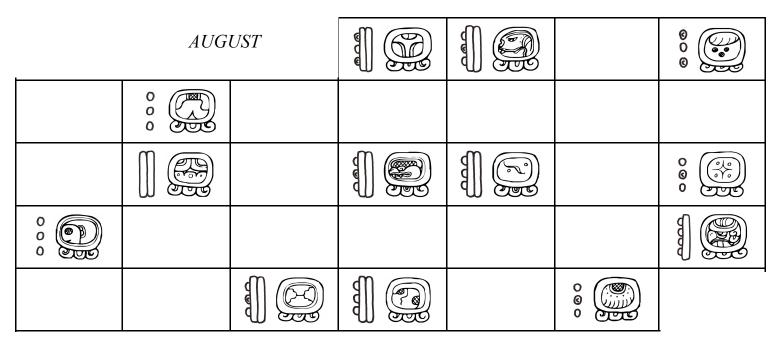
K2. Where on the calendar would the following days fall? Write a *c* and a *d* on the appropriate days on the following page.

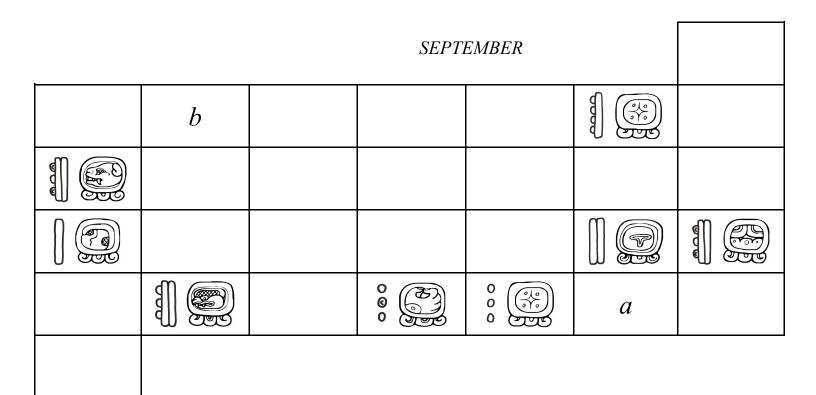


K3. The two months shown are completely representative of the workings of the *tzolk'in*; there are no special sub-cycles, "leap" days, periods with fewer days than other periods, etc. Knowing that, *how often* does the following day occur?



(K) Tzolk'in (2/3)







REGISTRATION #:



K4. Explain your answers in the space provided below.



(20 points) (L) The Whole Spectrum (1/4)

Background

The sounds of human speech (and indeed all sounds) travel through the air in waves, some of which your ear detects as sounds. These waves can be analyzed (by a mathematical technique called "Fourier analysis") into combinations of basic ("sinusoidal") waves whose most important properties are "frequency" and "amplitude." One sound may comprise many such basic waves, and a basic wave of the same frequency or amplitude may appear in many sounds. Linguists sometimes display this analysis on a diagram called a spectrogram: the resulting sine waves' frequencies are plotted vertically (with greater amplitude indicated by darker points) and the time horizontally.

Data

A sequence of spectrograms produced using the computer software "Praat" is shown in the following pages. The first 12 are each labeled with an English word, which is shown in the spectrogram. The following words are given:

sash, lamb, knee, sheesh, soup, pang, loose, (the letter) e, mice, ice, coo, shine

The last four spectrograms each show one of the following eight words:

louse, lass, lease, lice, pass, ash, sheep, lack

Problems

L1. What words are shown in the last four spectrograms? You may give up to two answers on each; if you give two, you will be given half credit if either of them is correct.

(13) _____ (14) _____ (15) _____ (16) _____

L2. For each of the three most significant sounds in ``sash," mark an interval corresponding to the duration of that sound. You should mark each interval by a horizontal line drawn above the spectrogram over all parts of the spectrogram affected by that sound. If you make other markings on the spectrogram, clearly indicate where each of the three requested intervals starts and ends.

L3. Do the same for ``lamb."



REGISTRATION #:

(L) The Whole Spectrum (2/4)

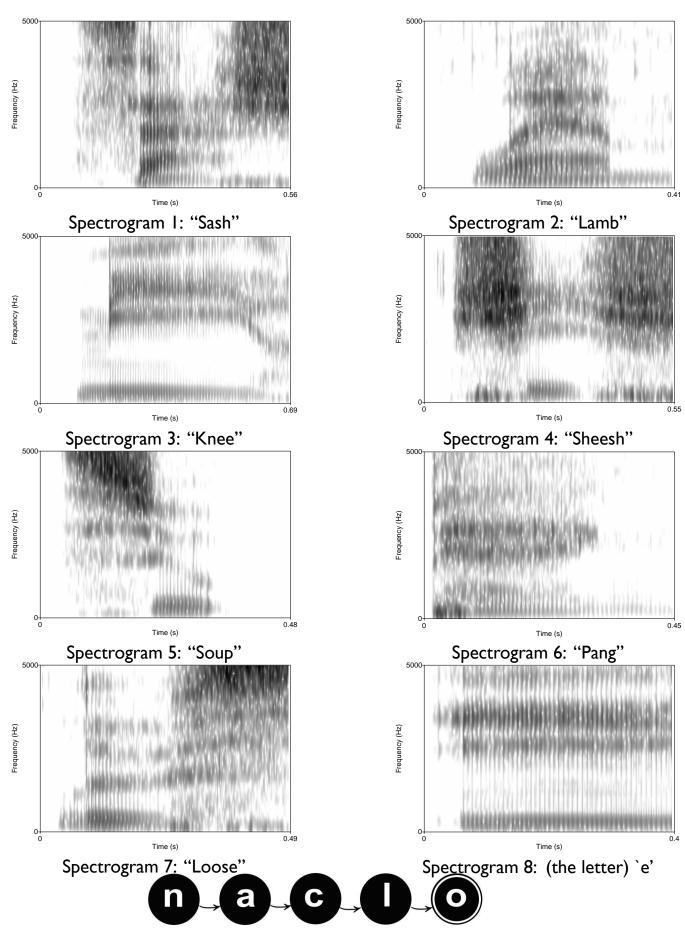
L4. Explain your reasoning for G1.

L5. Discuss the correspondence between the English spellings of the given 12 words and their spectrograms. In particular, some letters' sounds affect longer intervals in the spectrograms than others do. Which letters and sounds affect the longest intervals? Do adjacent sounds' affected intervals overlap, and if so, how? Are there any letters that correspond to no interval at all, and which letters? Conversely, are there any distinct sounds on the spectrograms that correspond to no English letters, and where are those sounds?



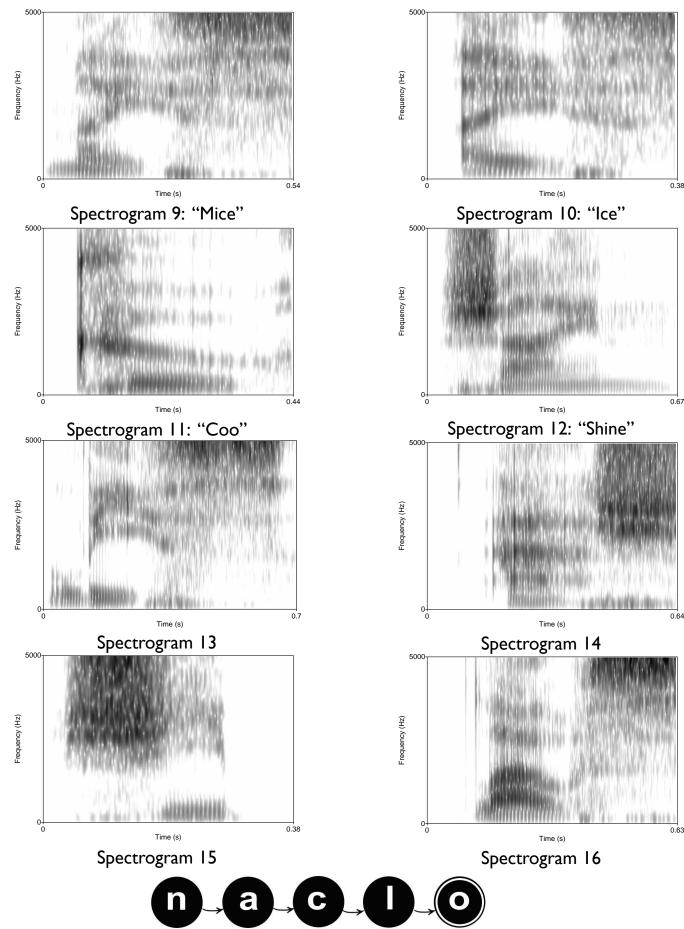
REGISTRATION #:

(L) The Whole Spectrum (3/4)



REGISTRATION #:

(L) The Whole Spectrum (4/4)



NACLO 2008 organizers

General co-chairs: Lori Levin, Carnegie Mellon University Thomas Payne, University of Oregon

Program committee chair and booklet editor:

Dragomir Radev, University of Michigan

Program committee:

Emily Bender, U. Washington Eric Breck, Cornell U. Lauren Collister, U. Pittsburgh Eugene Fink, Carnegie Mellon U. Adam Hesterberg, Princeton U. Joshua Katz, Princeton U. Stacy Kurnikova, Carnegie Mellon U. Lori Levin, Carnegie Mellon U. Will Lewis, U. Washington Patrick Littell, U. Pittsburgh David Mortensen, U. Pittsburgh Barbara Partee, U. Massachusetts Thomas Payne, U. Oregon James Pustejovsky, Brandeis U. Richard Sproat, U. Illinois Todor Tchervenkov, U. Lyon 2

Sponsorship chair:

James Pustejovsky, Brandeis U.

Administrative assistant: Mary Jo Bensasi, Carnegie Mellon University

School liaison: Amy Troyani, Taylor Allderdice High School

Web site and registration:

Justin Brown, Carnegie Mellon University

Problem credits (invitational round only):

Problem F: Willie Costello Problem G: Patrick Littell Problem H: Eric Breck Problem I: Patrick Littell Problem J: Todor Tchervenkov Problem K: Patrick Littell and Erin Donnelly Problem L: Adam Hesterberg

Additional thanks to:

Tanya Korelsky, NSF Paula Chesley, U. Minnesota Alina Johnson, U. Michigan Patti Kardia, U. Michigfan Jennifer Wofford, Cornell U. Cindy Robinson, Cornell U. More than 65 high school teachers from 28 states, provinces, and the District of Columbia

Site coordinators:

Brandeis University: James Pustejovsky Carnegie Mellon University: Lori Levin Columbia University: Kathy McKeown and Barry Schiffman Cornell University: Claire Cardie and Lillian Lee Middle Tennessee State University: Aleka Blackwell San José State University: Roula Svorou University of Illinois: Roxana Girju and Richard Sproat University of Michigan: Steve Abney and Sally Thomason University of Oregon: Thomas Payne University of Ottawa: Diana Inkpen and Stan Szpakowicz University of Pennsylvania: Mitch Marcus University of Toronto: Elaine Gold University of Wisconsin: Catherine Arnott Smith "Online" (high school sites): Dragomir Radev

All material in this booklet Copyright © 2008, North American Computational Linguistics Olympiad, unless otherwise indicated. Please do not copy or distribute without permission.

NACLO 2008 sites



as well as more than 65 high schools throughout the USA and Canada