NLP

Introduction to NLP

NACLO

NACLO

- Competition in Linguistics (including Computational Linguistics)
  - Since 2007
  - http://www.naclo.cs.cmu.edu
- Best individual US performers so far:
  - Adam Hesterberg (2007)
  - Hanzhi Zhu (2008)
  - Rebecca Jacobs (2007–2009) – 3 team golds + 2 individual medals
  - Ben Sklaroff (2010)
  - Morris Alper (2011)
  - Alex Wade (2012, 2013) – 2 team golds + 2 individual golds + 1 individual silver
  - Darryl Wu (2012, 2014)
- Other strong countries:
  - Russia, UK, Netherlands, Poland, Bulgaria, South Korea, Canada, China
- IOL – the International contest
  - Since 2003
  - IOL 2013 in Manchester, IOL 2014 in Beijing, IOL 2015 in Bulgaria
  - http://www.ioling.org
- Other high school competitions, e.g., IMO, IOI, IPhO, IChO, IBO, IOAA, etc.

Consider these phrases in Ancient Greek (in a Roman-based transcription) and their unordered English translations:

(A) *ho tōn hyiōn dulos*
(B) *hoi tōn dulōn cyrioi*
(C) *hoi tu emporu adelphoi*
(D) *hoi tōn onōn emporoi*
(E) *ho tu cyriu onos*
(F) *ho tu oicu cyrios*
(G) *ho tōn adelphōn oicos*
(H) *hoi tōn cyriōn hyioi*

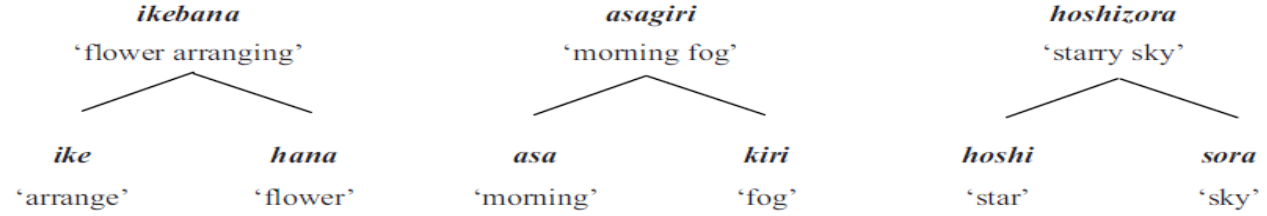
(1) the donkey of the master
(2) the brothers of the merchant
(3) the merchants of the donkeys
(4) the sons of the masters
(5) the slave of the sons
(6) the masters of the slaves
(7) the house of the brothers
(8) the master of the house

C1. Place the number of the correct English translation in the space following each Greek sentence. Explain your answers!

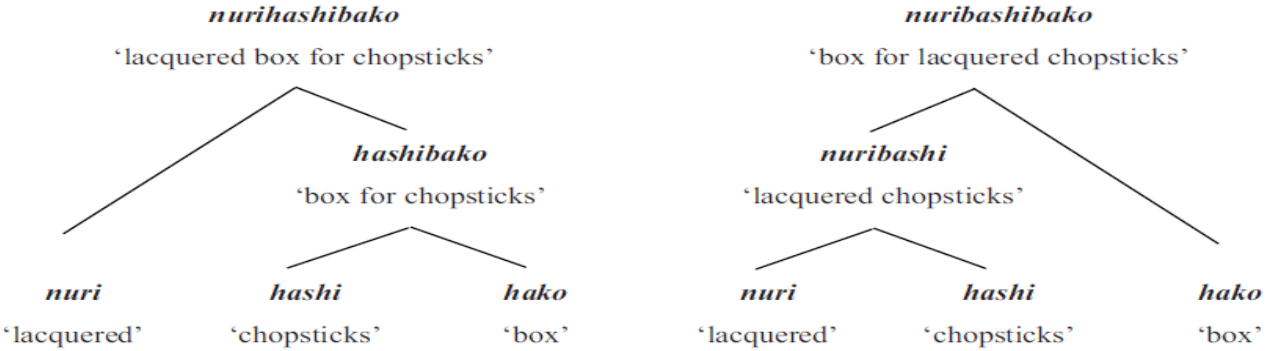
C2. Translate into Ancient Greek:
the houses of the merchants;
the donkeys of the slave
Explain your answers!

A Donkey in Every House, by Ivan Derzhanski

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.



Fakepapershelfmaker, by Willie Costello

An excerpt from a well known text is shown below. It is in two languages (X and Y) that are closely linguistically related to each other and also to English. However the two versions are not perfect translations of one another.

Text in language X

X1. Rödluvan: Men mormor, varför har du så stora ögon?
X2. "Mormor": Det är bara för att jag skall se dig bättre, mitt barn.
X3. Rödluvan: Men mormor, varför har du så stora öron?
X4. "Mormor": Det är bara för att jag skall höra dig bättre, mitt barn.
X5. Rödluvan: Men mormor, varför har du så stora tänder?
X6. "Mormor": Det är bara för att jag skall kunna äta upp dig!

(almost) the same text in language Y

Y1. - Så store ører du har, bestemor, sa Rødhette.
Y2. - Det er fordi jeg skal kunne høre deg bedre, svarte ulven.
Y3. - Så store øyne du har, bestemor, sa Rødhette.
Y4. - Det er fordi jeg skal kunne se deg bedre, svarte ulven.
Y5. - Så store hender du har, bestemor, sa Rødhette.
Y6. - Det er fordi jeg skal kunne klemme deg bedre, svarte ulven.
Y7. - Så stor munn du har, bestemor, sa Rødhette.
Y8. - Det er fordi jeg skal kunne ete deg bedre, svarte ulven.

Rødhette, by Dragomir Radev

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	

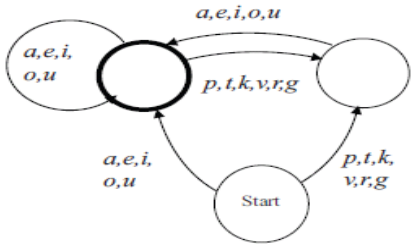
A1. The following *tenji* words represent ***atari***, ***haiku***, ***katana***, ***kimono***, ***koi***, and ***sake***. Which is which? You don't need to know either Japanese or Braille to figure it out; you'll find that the system is highly logical.

a. _____ 	b. _____ 
c. _____ 	d. _____ 
e. _____ 	f. _____ 

Tenji Karaoke, by Patrick Littell

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	
tauo	*kareveiepa*
puraveva	*ovokirovuia*
avaopa	*ouragaveva*

Impossible Rotokas words	
grio	*ouag*
ovgi	*vonoka*
gataap	*oappa*

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

___ *iu*	___ *uente*	___ *voav*
___ *idau*	___ *urioo*	___ *uaia*
___ *oire*	___ *raorao*	___ *oratreopaveiepa*

Rotokas aw-TOM-uh-tuh, by Patrick Littell

On her visit to Armenia, Millie has gotten lost in Yerevan, the nation's capital. She is now at the Metropoliten (subway) station named Shengavit, but her friends are waiting for her at the station named Barekamutyun. Can you help Millie meet up with her friends?

1. Assuming Millie takes a train in the right direction, which will be the first stop after Shengavit?

Note that all names of stations listed below appear on the map.

a. Gortsaranayin
b. Zoravar Andranik
c. Charbakh
d. Garegin Njdehi Hraparak
e. none of the above

2. After boarding at Shengavit, how many stops will it take Millie to get to Barekamutyun (don't include Shengavit itself in the number of stops)?

Lost in Yerevan, by Dragomir Radev

NACLO: Computational Problems

- http://clair.si.umich.edu/naclo/resources/resources.html

- List of computational problems:
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-O.pdf
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-C.pdf
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-J.pdf
  - http://www.naclo.cs.cmu.edu/problems2014/N2014-L.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-C.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-F.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-H.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-L.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-N.pdf
  - http://www.naclo.cs.cmu.edu/problems2013/N2013-O.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-C.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-K.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-O.pdf
  - http://www.naclo.cs.cmu.edu/problems2012/N2012-R.pdf
  - http://www.naclo.cs.cmu.edu/problems2011/F.pdf
  - http://www.naclo.cs.cmu.edu/problems2011/M.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/D.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/E.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/I.pdf
  - http://www.naclo.cs.cmu.edu/problems2010/K.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-E.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-G.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-I.pdf
  - http://www.naclo.cs.cmu.edu/problems2009/N2009-M.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-F.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-H.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-I.pdf
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-L.pdf
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-A.pdf
  - http://www.naclo.cs.cmu.edu/problems2007/N2007-H.pdf

NLP