Homework 4	CMSC 723
Part of Speech Tagging	
Due: October 14, 2013	

1 Tagging and Tag Sets (20 points)

1.1 When taggers go bad (10 points)

Consider the following sentences:

1. British Left Waffles on Falkland Islands
2. Teacher Strikes Idle Kids
3. Clinton Wins Budget; More Lies Ahead
4. Juvenile Court to Try Shooting Defendant

(You're also more than welcome to create or find another sentence that is similarly confusing.) Choose one of these sentences and tag it in two different (but plausible) ways.

1.2 Exploring the tag set (10 points)

There are 265 distinct words in the Brown Corpus having exactly four possible tags (assuming nothing is done to normalize the word forms).

1. Create a table with the integers $1 \ldots 10$ in one column, and the number of distinct words in the corpus having $\{1, \ldots, 10\}$ distinct tags.
2. For the word with the greatest number of distinct tags, print out sentences from the corpus containing the word, one for each possible tag.

2 Viterbi Algorithm (80 Points)

Consider the following sentences written in Klingon. For each sentence, the part of speech of each "word" has been given (for ease of translation, some prefixes/suffixes have been treated as words),

along with a translation. Using these training sentences, we're going to build a hidden Markov model to predict the part of speech of an unknown sentence using the Viterbi algorithm.

N	PRO	V	N	PRO
pa'Daq	ghah	taH	tera'ngan	'e
room (inside)	he	is	human	of

The human is in the room

V	N	V	N
ja'chuqmeH	rojHom	neH	tera'ngan
in order to parley	truce	want	human

The enemy commander wants a truce in order to parley

N	V	N	CONJ	N	V	N
tera'ngan	qIp	puq	'eg	puq	qIp	tera'ngan
human	bit	child	and	child	bit	child

The child bit the human, and the human bit the child

2.1 Emission Probability (25 points)

Compute the frequencies of each part of speech in the table below for nouns and verbs. We'll use a smoothing factor of 0.1 (as discussed in class) to make sure that no event is impossible; add this number to all of your observations. Two parts of speech have already been done for you. After you've done this, compute the emission probabilities in a similar table.

	NOUN	VERB	CONJ	PRO
'e			0.1	1.1
'eg			1.1	0.1
ghaH			0.1	1.1
ja'chuqmeH			0.1	0.1
legh			0.1	0.1
neH			0.1	0.1
pa'Daq			0.1	0.1
puq			0.1	0.1
qIp			0.1	0.1
rojHom			0.1	0.1
taH			0.1	0.1
tera'ngan			0.1	0.1
yaS			0.1	0.1

2.2 Start and Transition Probability (25 points)

Now, for each part of speech, total the number of times it transitioned to each other part of speech. Again, use a smoothing factor of 0.1. After you've done this, compute the start and transition probabilities.

	NOUN	VERB	CONJ	PRO
START				
N			1.1	2.1
V			0.1	0.1
CONJ			0.1	0.1
PRO			0.1	0.1

2.3 Viterbi Decoding (30 points)

Now consider the following sentence: "tera'ngan legh yaS".

1. Suppose that we knew "legh" were a pronoun (it's not, but pronouns often act like its true part of speech in Klingon). What would be the probability of each of the four parts of speech for "yaS"?
2. Create the decoding matrix of this sentence for nouns and verbs (ignore other parts of speech if you want; doing so won't prevent you from finding the right answer). You should have at least six numbers: $\log \delta_n(k)$ for $n = 1 \ldots 3$ and k for both nouns and verbs.
3. What is the most likely sequence of parts of speech?
4. What is the probability of your previous answer?
5. (For fun, not for credit) What do you think this sentence means? What word is the subject of the sentence?