NLP **Quiz 3** **30.09.2010** **Name:** ____________________

1. Define the term *Language Model.*

A probabilistic estimation for the frequency of words and word sequences.

2. Assume you have a tokenized corpus, *tokens.txt*, with one token per line. Show the sequence of commands (using *sort*, *uniq* and Unix/Linux pipes) that produce unigram frequencies, *tokens.freq*, for the tokens found in the corpus.

```
sort tokens.txt | uniq -c | sort -nr > tokens.freq
```

3. Let $S = w_1, w_2, \ldots, w_n$ be a word sequence and let

$$P(S) = P(w_1) \prod_{i=2}^{n} P(w_i|w_{i-1})$$

be the probability of the sequence using bigrams (and one unigram). Show the formula which approximates the bigram probabilities (the formula uses frequencies from a corpus).

$$P(w_i|w_{i-1}) \approx \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

where C stands for the count of the corresponding unigram or trigram in the corpus.

4. Define the term *morpheme.*

The smallest meaningful unit of a word. A morpheme divide into stems and affixes.

5. What is a *finite-state transducer*?

A finite-state automaton recognizing or generating a **pair of strings**.