Corpus Navigator Status

Jonathan Clark

Monday, November 19, 2007

1 Project Background

1. Automatically collect sentences from a relatively naive bilingual person
2. First, we statically selected sentences to ask based on semantic features
3. Now, we would like to actively select sentences based on what we know (what we've asked) so far

2 Goals

1. Create Urdu Dataset for "Post-NIST Experiment"
2. Provide evidence for the hypothesis that "The right data is better data" instead of "More data is better data"

3 Questions for Monica

1. What interesting things have you found in your thesis work?
2. Might you have any results that support "the right data" hypothesis?
3. Do you know of any other work in this area?

4 Project Questions

1. How much more experimentation is necessary to show this hypothesis is true?
2. How do we go about the creation of the Urdu Dataset? (Perhaps the free word order is a good starting place?)