Language Research Discussion

Note: Highlighting my suggestions
Legend
[ ] Things to consider
[ ] Done

1) Harvest monolingual data written in that language. (1 point) -
We can crawl the following websites and extract the text in the correct unicode range(http://en.wikipedia.org/wiki/Devanagari_%28Unicode_block%29).
http://tatkakhabar.com/
http://bhojpurimedia.com/
http://www.bhojpuria.com/v2/
http://khabarlahariya.org/
http://bhojpurika.com/
http://www.anjoria.com/
http://www.omniglot.com/
http://www.thebhojpuri.com/
Universal Declaration of Rights - [Downloaded]

2) Try to find bilingual data for the language. (1 point). [DONE]

Not sure if the data is useful!
Bhojpuri proverbs : (~70 proverbs)
http://anjoria.com/v1/bhasha/bhojpuri-proverbs1.htm
http://anjoria.com/v1/bhasha/bhojpuri-proverbs2.htm
http://anjoria.com/v1/bhasha/bhojpuri-proverbs3.htm

Has very trivial sentences with their english translations(~30)
http://tatoeba.org/eng/sentences/show_all_in/bho/none/none/

http://www.udhri.be/BHOJPURI%20Universal%20Declaration%20Of%20Human%20Rights.html

3) Collect ~100 interlinear glosses from one or more grammar books about the language. (2 points).
?? Isn't this same as translation. Instead of sentences we deal with paras
Gloss is word to word translation. Translation is per sentence. This seems easier.

4) If the language is written in a non-Roman script, build a transliteration system for it. (2 points).
Can we use/build a hindi transliteration system ?
http://en.wikipedia.org/wiki/Devanagari_transliteration
http://www.omniglot.com/writing/bhojpuri.htm

5) Build a language identification system that is able to predict whether a sample of text is written in the language or not. (2 points).
We can use the data that we get in the first section to train a LM. Unicode ranges and LM scores can be used to detect the language

6) **Does the language have a presence on Twitter? (2 points). Use the –location-query of this Twitter stream scraper to harvest tweets from the regions that speak the language.**
(2 point) [DONE]
Api already given.

7) Build a named entity tagger for the language. (3 points).
??

8) Collect or find a bilingual dictionary for the language. (1 point).
our dictionary should be in a CSV file with the following fields:
The foreign word
Its English translation
the source of the translation
(optional) an example sentence that uses the word
(optional) its part of speech
(optional) a definition

I came across couple of bhojpuri dictionary links. But they do not have all the information about a word.
http://anjoria.com/v1/bhasha/bhojpuri-verbs.htm
http://anjoria.com/v1/bhasha/wordbank.htm
http://anjoria.com/v1/bhasha/word-for-relations.htm
http://www.bhojpuria.com/v2/dictionary

9) **Find a language informant at Penn who knows both English and the language that you are researching. Have the language informant assemble bilingual ~50 sentence pairs and create manual word alignments for the language.** (2 points).
Let's post on different facebook groups and see if we get anything. I think this should be easy 2 points
Pranav Sahay. Ironically he was our ML TA :P
???

10) Try to locate language informants on a crowdsourcing platform like Mechanical Turk or CrowdFlower. (2 points). How can you ensure that they actually speak the language? Design a test to ensure that the crowd workers know the language:
We should build a basic language proficiency test and put it on Amazon Turk.

11) Have you language informant label part of speech (POS) information for the language. (2 points).
We should check how Universal POS tags work for Bhojpuri. Seems feasible.

12) Create tables of inflectional paradigms for some of the words in your language. (1 point).
Seems feasible