

Position paper for Stephen Soderland Digital Tool Summit in Linguistics

I am helping to direct the Panlingual project of the University of Washington's Turing Center. One of our main goals is machine translation to and from minor languages. We are exploring how a monolingual author can improve MT accuracy by reducing ambiguity in the source message, and are also exploring corpus-based techniques to learn a translation lexicon.

A variety of tools and resources are interesting to me:

- A source of comparable corpora in an arbitrary pair of languages. These may be parallel Bible translations, multi-lingual versions of the same Web pages, or collections of news stories from the same time period.
- Tools for creating parallel corpora from comparable corpora, such as news stories from the same time period, but not necessarily from the same news feed.
- Tools for developing a part-of-speech tagger for a new language. This could be learned automatically from tags in a major language where there are parallel texts. It could be gathered from field linguistic elicitation.
- Tools for developing a morphological analyzer for a new language. This may be a by-product of a field linguist's toolkit.
- Tools for developing a rudimentary syntactic parser for a new language. This may be induced from parallel texts with a major language. It may be induced from an elicitation corpus as in the AVENUE project.