Building Resources for Languages with Scarce Resources
UROP (Spring 2003) Final Report
May 23, 2003

By: Brian Rassier
Faculty Advisor: Dr. Ted Pedersen

The objective of this UROP was to develop a web interface that allows users from around the world to assist in the construction of large amounts of multilingual parallel corpora. This is an important task because parallel text (that is a text and its translation into another language) is a useful source of information for statistical and machine learning techniques that automatically create computerized translation systems from large samples of written text.

The resulting site is online (http://www.d.umn.edu/~tpederse/trans_site/newsite.cgi) and provides users with a selection of news articles from United Press International (http://www.upi.com) that are to be translated into their native language. The site allows for the convenient display and archiving of this material. The site itself primarily advertises itself as a way to create and obtain the news of the day in various languages. However, behind the scenes we are collecting parallel text for use with statistical natural language processing applications such as machine translation that require large samples of parallel text in a variety of languages.

We had planned to carry out some initial experiments using data collected from users. In particular we wanted to carry out preliminary experiments across a range of languages in sentence boundary detection, sentence and word alignment, and automatic lexicon induction. However, the development of the web site raised a number of unexpected challenges that, while overcome, resulted in our being unable to carry out the experimental work during the period of the UROP.

Our initial challenge had to do with the programming language we selected for creating the web site. We decided to explore the use of JavaScript, since it is known to support very dynamic web pages and allows for the incorporation of user friendly graphics. However, it turns out to be ill suited for actually transmitting data from a remote user to a central site. JavaScript is able to create vibrant and exiting web pages that run on a remote site, but it is unable to transmit data from that remote site. This is clearly a problem for our application, since we want users to provide a translation of an article and send it back to us for storage. As a result, we ultimately rejected the use of JavaScript and implemented the web site using Perl/CGI.

Our second challenge was how to deal with the encoding of text that is normally written in different alphabets, such as Hindi or Arabic. We learned quite a bit about this issue, and it’s much more complex than expected. There is a world-wide standard form for representing text in all languages known as Unicode (http://www.unicode.org). However, the Unicode standard is relatively new, and in technically advanced countries such as

India many different encoding schemes have been already developed and are in use. It may not be an exaggeration to say that every online newspaper in India has a unique way of representing and displaying Hindi text. This creates a very difficult challenge for our site, where the idea is to display text from a variety of sources and make it possible for users to provide their own translation. Largely due to the complexities of encoding other alphabets, our site currently only provides English news articles to translate into other languages. The display and encoding of English and other languages that use the Roman alphabets is very standardized and not a problem. There is a danger that users may provide us with text that is translated using some specialized encoding of their alphabet (and not the Unicode standard) but we have provided users with a means of providing us that information.

Dr. Pedersen will have some of his graduate students (several of whom know Hindi and other languages with non-Roman alphabets) translate articles from English into their language, to evaluate how the site can be improved by future UROP students to collect parallel corpora in many languages. Dr. Pedersen has a continuing interest in this area, as demonstrated by the fact that he recently organized a workshop on parallel text that was held at an international conference in Edmonton, Canada.

This was an interesting project, and helped me understand how difficult it is to truly internationalize a web site. We have also seen that there are relatively small amounts of parallel text available to researchers, and that a site like this can have a big impact on improving the availability of such resources and furthering the success of translation and other multi-lingual systems.

This was my second UROP project with Dr. Pedersen, and I have also worked on another project for Dr. Pedersen developing a web site to collect information from researchers in Latin America. I graduated at the end of this semester, and will start working in June at West Publishing in the Twin Cities. My job there will involve web programming, and it was only through these projects with Dr. Pedersen that I gained the skills that qualified me for this position. I think this experience had a big impact on my getting this job, and I would recommend that UMD consider adding classes in web programming since it seems to be an increasingly important area. I also hope that Dr. Pedersen continues to sponsor UROP opportunities, as they significantly extend the range of a student’s experience.