The risks of spelling variation and reform


Jeff Allen

For those of you who are subscribed to various language and technology e-lists, you might have certainly noticed a flood of computational linguistics and language technology jobs at Microsoft during the periods of August 1998 (for French, German, Spanish and English), April 1999 (for French, German, English, Spanish, Chinese, Japanese), May, June and August 1999 (Natural Language Group), and October 1999 for French. In general, all of the Microsoft jobs tend to be for the major languages. These job offers clearly indicate their focus. Then, Lernout & Hauspie, one of the major players that had made its name from the development of speech-based systems and applications, announced a significant number of new open positions in June, August and December 1999 and in January 2000. These jobs were not only for specialists of Dutch, English, Italian, Spanish, Greek, Swedish, and Danish, but also notably for Polish, Czech, Hungarian, Turkish, Belarussian, Thai, Taiwanese, Vietnamese, Indonesian, Malay, Urdu, Farsi, Tamil, Hindi, and Bahasa. I wonder if the IJLD readership knows where all of these languages are spoken.

These job advertisements demonstrate that there is a significant interest in many languages that have been labelled by different people as being lesser-commonly taught, lesser-used, minority, neglected, sparse-data, low-density, endangered, etc. For lack of space, I do not want to go into the reasons for which there is so much interest in such languages, especially for speech-based applications. Having worked on several minority language projects in previous positions, I have a few hunches on where the interest is.

However, in this column article I do want to focus on the risk of rapidly developing technologies for such languages. While at the Language Technologies Institute of Carnegie Mellon University (CMU) (http://www.lti.cs.cmu.edu/Research/Diplomat), I wrote several conference papers and articles on the topic of developing language technologies for minority languages, primarily for French Creole languages. On the CMU Haitian Creole project, we were confronted with a number of problems pertaining to language standardisation, these being issues that must be resolved in order to process the textual and speech data at an acceptable level. It was obvious that speech- and text-based systems would have difficulty handling issues of spelling variability in Haitian

Creole because of the lack of widespread standardisation and normalisation of the written code. I will not go into details here but simply mention that the list of papers is given at a Web site (http://www.cstr.ed.ac.uk/SALTMIL/jeffall-en.html) ;and I send electronic copies to people upon request.

Spelling variation in words is very common for languages that are in transition toward being written languages. In fact, nearly every human language is spoken (except of course for a few like Braille and Sign Language) and the reality of spoken language is that it is variable. Not only do I pronounce words slightly differently from my neighbour or colleague, but even my pronunciation of the same word varies from moment to moment. Many studies on phonetics, phonemics and dialectology have been conducted for hundreds of the world's languages, so I spare the space to mention them here. However, language variation plagues the work of researchers and developers in the field of text and speech processing. Even major national languages demonstrate spoken variation as shown for German (Burger & Schiel, 1998) and Spanish (Moreno et al., 1998). Such variation is not only quite normal for traditionally unwritten languages, but it also extends to the written forms of these languages where they are transitioning to becoming newly literate cultures. In one of our CMU project papers (Eskenazi et al, 1998), we produced charts that compared the level of spelling variation between English and Haitian Creole, and the results are astonishing. This paper is available on-line at the following site http://www.ling.su.se/Creole/Papers_On-Line.html#DB_Design.

In order for large companies to rapidly develop language technologies for exotic languages, especially with a limited amount of funding and a designated project duration, there is always the temptation to skip over the sociolinguistic factors and to cut corners on language standardisation issues. Some tweaks are advantageous and certainly save development time now, but are these short-term time savers going to serve the language overall for the long run? There is a risk of producing language databases that are only designed for specific systems, and thus resulting in textual databases that do not line up with the grassroots and local literacy and educational efforts and activities for these languages. Limited-term minority language technology projects can risk compiling written data that do not correspond with other efforts for producing consistent written language that native speakers will perceive as beneficial for the long-term.

Spelling is not a trivial issue, although we some-

times believe that it is. The minority languages mentioned above simply highlight a problem that has been captured in the research and development work of Mason Integrated Technologies Ltd in the form of an orthography conversion software package and an optical character recognition product. The number of person years invested into those two functional software programs demonstrates that a significant amount of time and energy must be spent on quality systems in order to adequately provide the type of tool and service that will respond to the customers' needs. This is essential when considering for sociolinguistic factors that greatly challenge the field of language engineering, documentation, and information management.

But now just as you are thinking that minority languages are far from your own documentation and information context and needs, let us remember that there have been multiple spelling reforms (or attempts) for various other languages (German, Dutch, Norwegian, Swedish, Greenlandic, Spanish, French, and Chinese) of which many are major languages (Note: feel free to contact me at postediting@aol.com for details on Web sites that give details on spelling reforms for each of these languages)

As I read through various papers and reports on language spelling reforms, the questions that I ask are: How does one deal with updating Translation Memory databases in the context of spelling reform? How can authoring and translation services adequately implement spelling reform changes when variability is acceptable and when different customers want the texts conforming to the different standards? What are the efforts for spelling localisation within a specific local language or dialect, and not just between different dialects (for example, British versus American English, or Continental vs Canadian French)? These are the kinds of issues that small language technology companies have been addressing for years and are very good at managing. On the other hand, these are also the very issues that will greatly challenge the larger corporations in their new ventures in language and document processing. As I have stated over and over again in conference talks, the technology itself is not the only factor that leads to successful integration and implementation. It must be complemented by 1) usable functionality, 2) training on the application, and 3) a service that allows for it to work optimally within a given context. Otherwise the new technology can build up resistance among the users and can end up just being another software program which gets listed in the technology patent records and finally collects a lot of dust on the company and computer store shelves due to non-use. Will all of the technology companies meet this challenge? Only time will tell.

In the next article for this column, I am considering discussing the issue of ownership of translation memories since this issue keeps coming up over and over again at conferences and in various articles. ■

References:

- Burger, Susanne and Florian Schiel (1998) RVG 1 – A Database for Regional Variants of Contemporary German. In Proceedings of the First International Conference on Language Resources and Evaluation, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 1083-1087.
- Moreno, Asunción, Harald Höge, Joachim Koechler, and José Mariño (1998) SpeechDat Across Latin America: Project SALA. In Proceedings of the First International Conference on Language Resources and Evaluation, 28-30 May 1998, Granada, Spain. Vol. 1, pp. 367-370.
- Eskenazi, Maxine, Hogan, Christopher, Allen, Jeffrey, and Robert Frederking. 1998. Issues in database design: recording and processing speech from new populations (poster session). In Proceedings of the First International Conference on Language Resources and Evaluation, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 1289-1293.