O Galego ILG-RAG e as Novas Tecnologias da Língua

Devido à sua projecção internacional, a língua portuguesa dispõe duma grande quantidade de ferramentas e recursos computacionais: glossários on-line, tradutores automáticos, analisadores morfológicos, POS taggers, bases terminológicas, etc. Face a esta diversidade e riqueça, o galego ILG-RAG, por estar sujeito a uma grafia regional e isolacionista, joga na terceira divisão das Novas Tecnologias da Língua.

A linha oficial de actuação na Galiza baseia-se no desenvolvimento de ferramentas linguísticas específicas para o galego RAG, sem tomar em conta todo o que já foi desenvolvido para o português. No pior dos casos, elaboram-se estratégias de desenvolvimento de software galego que usam como modelo ferramentas para o castelhano. Se queremos jogar na primeira divisão das tecnologias da língua, bastaria com substituir a norma RAG por uma outra, por exemplo a AGAL, baseada no português padrão. Porém, dada a situação socio-política actual, o uso oficial desta norma está ainda longe de se consolidar.

Existe, no entanto, a possibilidade de abrir uma nova via de desenvolvimento de software linguístico para o galego RAG. Esta nova estratégia teria como alvo a adaptação das ferramentas que já forom elaboradas para o português. O processo de adaptação pode ser feito mediante a simples transliteração gráfica dos recursos linguísticos portugueses que usam essas ferramentas. Para ajudar a compreender a filosofia desta nova via, vou dar um exemplo. Suponhamos que temos de construir uma base terminológica galega no eido da economia. Tomando em conta que já existem muitos glossários terminológicos de economia disponíveis na web para o português, o que se pode fazer é transliterar esses glossários por forma a construir uma nova base de termos que respeitem a grafia da RAG. Nos dous links seguintes, podedes acceder a dous pequenos glossários multilíngues com entradas em galego RAG que foram transliteradas automaticamente de glossários com entradas em português:
http://gramatica.usc.es/~gamallo/xml/nortes.html
http://gramatica.usc.es/~gamallo/xml/panlatino.html

Esta estratégia tem dous aspectos positivos: por um lado, aproveita os ingentes recursos duma língua internacional, e por outro, permite criar recursos “mais” galeguizados e, portanto, deturpados de castelhanismos. Em resumo, conseguimos mais e melhores recursos.

Um outro exemplo ainda mais esclarecedor é a possibilidade de elaborar etiquetadores morfossintáticos (“taggers”) para o galego RAG a partir de etiquetadores portugueses. A estratégia é mui simples: translitera-se o corpus de treino português e treina-se o sistema com o novo corpus transliterado. Isto funciona porque o português e o galego partilham a mesma morfo-sintaxe (é dizer, são a mesma língua). Podedes testar um etiquetador para o galego RAG construido desta maneira em:
http://gramatica.usc.es/~gamallo/tagger.htm

A ferramenta essencial para realizar estes trabalhos é o transliterador port2gal:
http://gramatica.usc.es/~gamallo/port2gal.htm
que é uma versão revisada e actualizada do script de Alberto Garcia. Podedes instalá-lo nas vossas máquinas se tendes o sistema Linux. Também podedes fazer consultas on-line.

Acho que deveriamos convencer às administrações, universidades, centros de investigação e empresas da língua para que desenvolvam, a curto prazo, projectos de i+d+i dentro desta linha estratégica. Isto não só permitiria criar mais e melhor software galego senão também contribuiria para a nossa integração dentro do universo das novas tecnologias da língua portuguesa. Para já, o primeiro passo nesta direcção tem de ser a criação urgente duma Associação Galego-Luso-Brasileira para o Processamento da Língua Natural. A Direcção Geral de I+D da Junta vai elaborar as linhas estratégicas prioritárias para 2006-2010. Aqui temos uma que já é viável.

Pablo Gamalho