A web-based system for multilingual school reports

David Chan
Canolfan Bedwyr
Bangor University
Safle'r Normal
Bangor, Gwynedd
LL57 2PX UK
d.chan@bangor.ac.uk

Dewi Jones
Canolfan Bedwyr
Bangor University
Safle'r Normal
Bangor, Gwynedd
LL57 2PX   UK
d.b.jones@bangor.ac.uk

Oggy East
Semantise Ltd
Unit 4, Coed y Parc Ind. Est.
Bethesda, Gwynedd
LL57 4YY UK
oggy@semantise.com

Abstract: Schools in Wales have a statutory duty to produce bilingual Welsh/English school reports. Since computerization of the reporting process, this has mainly been achieved with English-only reporting software, using a labour-intensive and error-prone process of disparate add-ons and ad-hoc post-processing to work round the underlying monolingual data model. In the scope of this project a reporting tool, osisADRODD, with a fully multilingual data model was designed from scratch to support the various bilingual scenarios operating in the schools of Wales. A multilingual co-ordinated comment bank system was designed with support for gender-dependent pronoun substitution and agreement in any language. A general library for language-specific text proofing within an HTML form was implemented and incorporated into the system, and a grammar checker and spell checker were included for Welsh and English respectively.

1   School Reporting in Wales

1.1   Welsh Language Act

The Welsh Language Act (United Kingdom 1993) requires all state-run schools in Wales to treat Welsh and English "on a basis of equality [...] so far as is both appropriate in the circumstances and reasonably practicable".

1.2   Language use in schools

Both Welsh and English are used in every school in Wales. The distribution of both the medium of instruction and the linguistic proficiency of the teachers will vary depending on the school.

Irrespective of these factors, the Welsh Language Act requires school reports to be provided in Welsh and/or English according to parental preference.

1.3   Monolingual systems in use

Since the introduction of Student Information Systems (SISs) into education in Wales, schools have generally been using reporting systems designed for a monolingual environment. These typically have a user interface and data model which cannot accept and store parallel bilingual text.

In many cases, the monolingual SISs have been customised with extensions which accept text in an alternate language and store it in a duplicate record or a separate database table. Because the core system is unaware of the extension there is usually a risk of corrupting data or destroying the interlanguage correspondence, which is managed by painstaking care and manual verification.

1.3.1   Comment Banks

Comment banks (ready-made, pre-entered sentences of general utility) reduce the effort required to write reports. However, the monolingual SISs do not generally support the maintenence of parallel bilingual comment banks. This is in some cases addressed by institutions manually creating an alternate language version of the comment database with parallel comments stored using the same internal database record ID as in the main language record.   Naturally,   this

correspondence is fragile and manual care is required when editing either data set.

Furthermore, comment bank systems generally have a language-specific pronoun substitution mechanism; for example, a comment such as "^ does ~ homework" may be rendered as "Adam does his homework" or "Eve does her homework" depending on the student's name and gender. Major international software vendors have generally declined to support Welsh-language pronouns due to the small size of the market, notwithstanding the simplicity of such textual subsitutions. It therefore seems there is little prospect of support for more complex variations, such as the requirement in Welsh to vary the word for "homework" to agree with the gender of the student.

In some cases institutions have worked round these limitations by paraphrasing comment bank entries to avoid the use of pronouns or gender agreement altogether; the disadvantage is a final report containing unnatural, forced sentences which sound excessively formal and are harder to understand.

1.3.2 Post processing

Due to the limitations of the authoring systems, most institutions perform some amount of post-processing, whereby the system outputs word processor documents which are then modified manually. This compensates for the missing flexibility in the main system, but at the expense of losing the convenience, data integrity and uniformity which is the advantage of a reporting system.

1.4 Lack of market power

As indicated above, multinational vendors have proven unwilling to customise their software to support the market in Wales due to its comparatively small size. There are, nevertheless, nearly 2000 schools in Wales (Welsh Assembly Government 2007) and therefore it might be expected that ample funding exists to pay for the required customisations, if implemented for example by small to medium-sized enterprises, which may have lower overheads and greater flexibility than the international software vendors which dominate the market for SIS software.

The feasibility of such a solution depends on an institution's chosen SIS being sufficiently modular or open that a third party extension can interface with it.

2 Linguistic use cases for a reporting system

In order to be useful to schools with varying linguistic profiles, a reporting system must support a variety of usage scenarios. Teachers may author in Welsh and English simultaneously, or they may write content in a primary language which is later translated by themselves or by clerical staff. Indeed, a teacher's proficiency in one of the languages may be non-existent or insufficient to author confidently. Content may be composed from scratch, selected from comment banks or a combination of the two processes. Schools may have differing workflows for creation, editing and approval of content. Staff should be able to depend on convenient access to language proofing tools at any stage in the workflow.

3 osisADRODD: A Multilingual Reporting System

In order to maximise compatibility with existing systems and minimise data exchange issues, the data flow from the school's SIS to the reporting system is one way; that is, the reporting system acts as a consumer of data only. Existing monolingual reporting functionality in the SIS is disregarded, due to insufficient support for a multilingual data model.

The system's user interface is implemented in HTML and Javascript using the Google Web Toolkit (Google 2009), with a SQL database backend. This minimises dependence on a school's particular choice of SIS.

3.1 Comment banks

The data model stores parallel comment bank entries as separate records sharing a database ID, with the comment's language code forming part of the record key. This enables multiple language storage (not just two languages).

Comment bank entries contain metalinguistic strings to enable the specification of name- and gender-dependent text. For example, an entry such as "%studentName% thinks hard about %(his|her)% work" specifies two sets of sentences (male and female). This is more powerful than the substitution mechanism described above, since any gender-

dependent variation can be specified, not just a few language-specific cases hardcoded into the software.

3.2   The content authoring process

The content authoring process follows three stages, as follows.

Firstly, the teacher constructs a sequence of sentences from a comment bank; these are displayed in the teacher's preferred language but the equivalent text can also be constructed in any other language which the comment bank supports. The teacher's selections are saved.

Secondly, the sentences are concatenated into paragraphs of free text, one for each language version, which the teacher can then modify. The modifications are highlighted for clarity. Note that the teacher may have no proficiency in one or more of the languages; in this case, the paragraph versions in those languages are not displayed to the teacher, in accordance with stored preferences.

The modified free text is then saved in each language, including any versions which were not displayed.

Thirdly, the free text with highlighted modifications is recalled, and displayed in each language, in order that an editor or translator can proofread the translation correspondence and make any neccessary changes.

3.3   Language proofing subsystem

Where editable free text fields appear, the user can invoke language proofing tools to check grammar and spelling. A common application programming interface has been defined for passing text in a specified language via an HTTP POST request to a separate proofing server, which then returns a list of spelling and/ or grammar errors with requisite details and suggestions. We have implemented a proofing server conforming to this API which wraps Bangor University's Cysill grammar checker (Prys *et al.* 2002) to proof Welsh language text, in addition to the standard Hunspell library (Németh 2009) for spellchecking English and other languages.

The entire proofing subsystem is independent of the other components of osisADRODD in general, and could in principle be used to provide language proofing tools for virtually any HTML forms. An exciting possibility is to invoke the proofing subsystem via a Javascript bookmarklet (a script which can

be run on the content of another webpage at the click of a button). A user could thereby proof text entered into a third party content provider's web form without any action from the content provider to integrate the tools. This may provide an alternative to persuading providers of online services to spend time and resources providing proofing tools for lesser-used languages.

3.4   Security

School reports are inherently confidential personal data and require protection as such. In the original security model, osisADRODD is served from the same security context as the school's SIS, typically running from servers physically located within a school, firewalled and accessible only from a staff intranet. The same authentication credentials used in the SIS are used to restrict users' data access by relevance.

However this model is not generally appropriate for primary schools (for students aged 4-11 years), as these seldom have the budget or expertise to manage application servers internally. Further development is now in progress to provide a hosted solution for such schools; here the system will be accessed across a wide area network, or even directly via the public Internet, and secured in the usual way by a combination of an encrypted HTTP session and digital certificates.

References

United Kingdom. *Welsh Language Act 1993 (c. 38)*, London.

Schools Management Division, Welsh Assembly Government (2007). *Defining schools according to Welsh medium provision*, Cardiff.

Prys D., Hicks B., Jones D. B., Morgan M. (2002), *Cysgliad*, University of Wales, Bangor. http://hdl.handle.net/10242/13190

Németh L. (2009). *Hunspell: open-source spell checking, stemming, morphological analysis and generation*, Hungary, http://hunspell.sourceforge.net .

Google Inc (2009). *Google Web Toolkit*. http://code.google.com/webtoolkit .