# The Conference Programme

# LULCL
# Conference on
# **L**esser **U**sed **L**anguages and **C**omputer **L**inguistics
# European Academy of Bolzano, Italy
# 27th -28th October 2005

**Thursday 27th October 2005**

| | |
|---|---|
| 08:30-09:30 | Registration |
| 09:30-10:00 | Opening<br>*Werner Stuflesser*, President of EURAC<br>*Florian Mussner*, Provincial councillor for Ladin Culture |
| 10:00-10:45 | **Spracherneuerung im Rätoromanischen: Linguistische, soziale und politische Aspekte**<br>*Clau Solèr* (University of Geneva) |
| 10:45-11:15 | Coffee break |
| 11:15-11:45 | **Designing a Sardinian Corpus: problems and perspectives**<br>*Nicoletta Puddu* (Università di Pavia) |
| 11:45-12:15 | **Il progetto "Zimbarbort" per il recupero del patrimonio linguistico cimbro**<br>*Luca Panieri* (Istituto Cimbro di Luserna) |
| 12:15-12:45 | **The relevance of lesser used languages for theoretical linguistics: the case of Cimbrian and the support of the TITUS corpus**<br>*Alessandra Tomaselli* (Università di Verona),<br>*Ermenegildo Bidese* (Università di Verona/ Studio Teologico Accademico Bressanone),<br>*Cecilia Poletto* (Padova-CNR) |
| 12:45-14:30 | Lunch break |
| 14:30-15:00 | **Il progetto VERBA, Una rete gli strumenti web-based di comunità linguistiche per permettere alle lingue meno diffuse di accedere a strumenti d'eccellenza nel campo del trattamento automatico della lingua** |

*Carlo Zoli* (Dipartimento di Ingegneria Linguistica di Open Lab),
*Diego Corraine* (Ufitziu pro Sa Limba Sarda)

15:00-15:30 **Speech-to-Speech Translation for Catalan**
*Victoria Arranz* (ELDA - Evaluation and Language resources Distribution Agency),
*Elisabet Comelles* (TALP - Centre de Tecnologies i Aplicacions del Llenguatge i la Parla, Universitat Politècnica de Catalunya),
*David Farwell* (Institució Catalana de Reserca i Estudis Avançats TALP - Centre de Tecnologies i Aplicacions del Llenguatge i la Parla, Universitat Politècnica de Catalunya)

15:30-16:00 **SpeechCluster: a speech database builder's multitool**
*Ivan Uemlianin* (Canolfan Bedwyr, University of Wales, Bangor)

16:00-16:30 Coffee break

16:30-17:00 **XNLRDF, A Framework for the Description of Natural Language Resources. A proposal and first implementation**
*Oliver Streiter* (National University of Kaohsiung),
*Mathias Stuflesser* (Eurac research, Accademia Europea di Bolzano)

17:00-17:30 **Towards Effective and Robust Strategies for Finding Web Resources for Lesser Used Languages**
*Baden Hughes* (Department of Computer Science and Software Engineering, University of Melbourne)

**Friday 28th October 2005**

09:00-09:45 **Implementing NLP-Projects for Small Languages: Instructions for Sponsors, Strategies for Developers**
*Oliver Streiter* (National University of Koahsiung)

09:45-10:15 **Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping**
*Danie Prinsloo* (Department of African Languages, University of Pretoria, South Africa),
*Ulrich Heid* (IMS-CL, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart)

10:15-10:45 **A comparison of approaches towards word class tagging: disjunctively vs conjunctively written bantu languages**
*Elsabé Taljard* (University of Pretoria),
*Sonja E. Bosch* (University of South Africa)

10:45-11:15 Coffee break

11:15-11:45 **Grammar-based language technology for the Sámi languages**
*Trond Trosterud* (Det humanistiske fakultet, Universitetet i Tromsø)

11:45-12:15 **Annotation of documents for electronic edition of Judeo-Spanish texts: problems and solutions**
*Soufiane Rouissi* (Université de Bordeaux 3, Cemic -Gresic),
*Ana Stulic* (Université de Bordeaux 3, Ameriber)

12:15-12:45 **Stealth Learning with an on-line dog**
*Ambrose Choy, Gruffudd Prys*
(Canolfan Bedwyr, University of Wales, Bangor)

12:45-14:30 Lunch break

14:30-15:00 **The Igbo Language and Computer Linguistics: Problems and Prospects**
*Chinedu Uchechukwu* (Otto-Friedrich-University, Bramberg, Germany)

15:00-15:30 **Il ladino fra polinomia e standardizzazione: l'apporto della linguistica computazionale**
*Evelyn Bortolotti, Sabrina Rasom*
(Istitut Cultural Ladin "Majon di Fascegn")

15:30-16:00 **The Welsh National On-line Database**
*Dewi Jones, Delyth Prys*
(Canolfan Bedwyr, University of Wales, Bangor)

16:00-16:30 Coffee break

16:30-17:00 **Lexicelt: an on-line Welsh/Irish Dictionary**
*Delyth Prys, Dewi Evans*
(Canolfan Bedwyr, University of Wales, Bangor)

17:00-17:30 Conclusions

**Invited Keynote speakers**:
*Clau Solèr* (University of Geneva, Swizerland)
*Oliver Streiter* ( University of Kaohsiung, Taiwan)

**Scientific Committee**:
*Dafydd Gibbon* (University of Bielefeld, Germany)
*Christer Laurén* (University of Vasa, Finland)
*Oliver Streiter* (University of Kaohsiung, Taiwan)
*Marcello Soffritti* (University of Bologna, Italy)

**Interpreters:**

English: *Francesco Cappello, Anna Lubin*
Italian: *Leonora Bruno, Sigrid Hechensteiner*

**Organisation:**
Isabella Ties
e-mail:ities@eurac.edu
telephone: +39 0471 055 123
fax: +39 0471 055 199

# Table of Contents

**Designing a Sardinian Corpus: problems and perspectives**

Nicoletta Puddu
(Università di Pavia)

Creating a corpus for minority languages has proved to be important in order to both study and preserve minority languages (see for example the DoBeS project at MPI Nijmegen). Sardinian, as an endangered language, could certainly profit from a well-designed corpus. A first digital collection of Sardinian texts is the *Sardinian Text Database*, which, however, cannot be considered a corpus. In this paper, I discuss the main problems in designing and developing a corpus for Sardinian.

Kennedy (1998: 70) identifies three main stages in compiling a corpus: (1) corpus design; (2) text collection and capture; (3) text encoding or markup. As for the first stage, I propose that a Sardinian corpus should be mixed, monolingual, synchronic, balanced and annotated and I discuss the reasons for these choices throughout the paper. Text collection seems to be a minor problem in the case of Sardinian: both written and spoken texts are available and the number of speakers is still high enough to collect a sufficient amount of data. The major problems arise in connection with the third step. Sardinian is fragmented into different varieties and does not have a standard variety (not even a standard orthography). Recently, several proposals for standardization have been made albeit without success (see Calaresu 2002, Puddu 2003). First of all, I suggest to use a standard orthography which allows us to identify some different macrovarieties. Then, it will be possible to structure the corpus into subcorpora which are representative of each variety. The creation of an adequate morphological tagging system will be fundamental. Thanks to a homogeneous tagging system, it will be possible to operate searches throughout the corpus and to study linguistic phenomena both in each single macrovariety and in the language as a whole.

Finally, I propose a morphological tagging system and present a tagged pilot corpus of Sardinian based on written samples.

**References**

Calaresu, E. (2002) 'Alcune riflessioni sulla LSU (Limba Sarda Unificada)', in V. Orioles, (a cura di) *La legislazione nazionale sulle minoranze linguistiche. Problemi, applicazioni, prospettive*: 247-266.

Kennedy, G. (1998) *An introduction to Corpus Linguistics*, London, Longman.

McEnery, T.; Wilson, A. (1996) *Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Puddu, N. (2003) 'In Search of the real Sardinian', in Brincat, J.; Boeder, W.; Stolz, T. (eds.) *Purism in minor languages, endangered languages, regional languages, mixed languages . Papers from the conference on 'Purism in the Age of Globalization' Bremen, 2001*, Bochum: Universitätsverlag Dr. N. Brockmeyer: 27-42.

www.lingrom.fu-berlin.de/sardu/textos.html

www.mpi.nl/DOBES

## Il progetto “Zimbarbort” per il recupero del patrimonio linguistico cimbro

Luca Panieri
(Istituto Cimbro di Luserna)

L’idea di questo progetto nasce dalla consapevolezza della situazione precaria in cui versano le tre isole linguistiche cimbre sopravvissute nei secoli fino ai giorni nostri: Luserna (TN), Giazza (VR) e Roana-Mezzaselva (VI). La condizione relativamente rosea in cui fortunatamente si trova ancora la varietà cimbra di Luserna impone l’attuazione di ogni possibile strategia di difesa e consolidamento del patrimonio linguistico cimbro, essendo diventata Luserna l’ultima roccaforte di una tradizione etnica un tempo estesa in tutto il territorio prealpino tra l’Adige e il Brenta.
L’intervento che qui si presenta consiste nella realizzazione di una banca dati lessicale globale, comprendente la tradizione linguistica cimbra storicamente attestata in tutte le sue varietà diatopiche e diacroniche. L’idea di fondo è quella di creare un luogo virtuale della memoria linguistica cimbra che fornisca lo spessore storico necessario al consolidamento e al futuro sviluppo della lingua cimbra, affinché essa trovi prima di tutto nella propria stessa tradizione le risorse per rinnovarsi ed estendere il proprio dominio espressivo agli ambiti concettuali tipici della cultura moderna. Dalla conoscenza profonda delle proprie radici linguistiche emana il rispetto della parlata materna e la fiducia nella sua autonomia espressiva, oltre l’ambito familiare e tradizionale.
In questo luogo virtuale della memoria linguistica ogni lemma attestato viene dotato di una scheda informativa in cui, tra l’altro, se ne indica la fonte di provenienza. Nella scheda virtuale figurano inoltre annotazioni di carattere grammaticale, fraseologico, lessicologico, fonologico, etimologico, ecc. Data la struttura aperta e flessibile della banca dati informatica, si rendono possibili continue revisioni e aggiornamenti dei dati e delle informazioni su essi disponibili, in un processo teoricamente all’infinito, poiché l’acquisizione dei dati lessicali oltre che esaurire le fonti storiche in cimbro, scritte a partire dal ‘600 circa, tiene conto di tutte le fonti orali a disposizione e dei neologismi che si vanno a creare nell’uso linguistico attuale. Oltre a ciò l’aggiornamento progressivo del corredo informativo della banca dati deriva dalla sempre maggior capacità analitica della scienza linguistica, che porta alla consapevolezza di nuovi aspetti e considerazioni correlati ai dati lessicali.

**The relevance of lesser used languages for theoretical linguistics: the case of Cimbrian and the support of the TITUS corpus**

Alessandra Tomaselli
(Università di Verona),
Cecilia Poletto
(Padova-CNR),
and
Ermenegildo Bidese
(Università di Verona/Philosophisch-Theologische Hochschule Brixen)

In recent years the Department of Germanic and Slavic Philology at the University of Verona (Italy) has undertaken several research projects devoted to the syntactic exploration of **a unique Germanic language** which is "surviving" in a few linguistic isles in the North-East of Italy: Cimbrian. The syntax of this less commonly used language reveals interesting peculiarities, which make it a surprising "mixture" of both Germanic and Romance features, among all: i) the loss of the V2 restriction, ii) the acquisition of word order patterns largely convergent with the typology of SVO languages; iii) a very structured set of pronominal clitics.

Since the use of this language is nowadays limited to the community of Lusern/Luserna (TN), whereas the Venetian varieties (which are still sporadically spoken in the communities of Mittoballe/Mezzaselva (VI) and Ljetzan/Giazza (VR)) could be considered almost extint, it is very important for any kind of linguistic analyses to make available corpora of texts or, at least, of sentences. In this perspective, it is worth to consider the publication of two relevant Cimbrian texts from the XIX century, i.e., the catechisms of 1813 and of 1842, on the World Wide Web, at the TITUS site (Thesaurus Indogermanischer Text- und Sprachmaterialien: http://titus.uni-frankfurt.de/indexd.htm). This represents, in fact, a good example of just how the employment of online resources can support the linguistic research of less commonly used languages. The TITUS corpus is particularly interesting for syntactic investigations because the two Cimbrian texts have been provided with a first analysis of the clitic elements, whose relevance for the theory of grammar represents the core of our presentation.

In the last decade, the (morpho-)syntax of unstressed pronouns has become the subject of intensive studies, in particular within the theoretical framework of Generative Grammar, primarily because the positions of clitic elements within the clause are regulated by strong syntactic restrictions. Hence the syntax of clitics should be considered as one of the most relevant topics for any theoretical speculation about sentence structure and movement theory.

As we know, in **standard German** a set of clitics does not exist morphologically. Even so, unstressed German pronouns obey peculiar syntactic restrictions: only pronominal elements are allowed to realize (move to) the so-called "Wackernagelposition", i.e., the position immediately to the right of the inflected verb in the main clause/the subordinating conjunction in the dependent one (in other words, the initial portion of the "middle field"/*Mittelfeld*).

On the other hand there are many languages which have two sets of pronouns, stressed (free pronouns) and unstressed (clitics). All **northern Italian dialects**, for example, have a morphologically realized set of clitics, both subject and object

clitics, which differ with respect to their proclitic versus enclitic position according to the structural location of the verb.
The **Cimbrian dialect** has both subject and object clitics, but they still behave *ala German* allowing just enclisis either to the inflected verbal form or to the subordinating conjunction (preservation of the "Wackernagelposition").
As we are going to demonstrate, the comparative analysis of these three different manifestations of cliticization processes, with a particular attention devoted to the Cimbrian configuration, will allow us to shed a new light on the principles which underlie the theory of movement, ultimately the relation between overt morphology and syntactic derivations along the lines put forwards by Chomsky's Minimalist Program.

Thus, our contribution to the conference has a threefold aim: 1. The on-line presentation of integrated Cimbrian texts as part of the TITUS corpus and of their use for a syntactic analysis; 2. The comparison of the syntax of clitic elements in the northern Italian vernaculars with those in Cimbrian. 3. The analysis of the Cimbrian clitics with particular consideration of the object clitics in relation with the morph-syntactic features of the verbal phrase (realization versus non realization of agreement morphology on the past participle).

**VERBA**

**Una rete di strumenti web-based di comunità linguistiche per permettere alle lingue meno diffuse di accedere a strumenti d'eccellenza nel campo del trattamento automatico della lingua**

Diego Corraine
(Ufitziu pro Sa Limba Sarda)
&
Carlo Zoli
(Dipartimento di Ingegneria Linguistica di Open Lab)

Tre sono i limiti che i progetti di Trattamento Automatico del Linguaggio "lingue meno diffuse" scontano tipicamente:

1. insufficiente comunicazione tra le varie minoranze: comunità linguistiche anche vicine sviluppano "in parallelo"
progetti simili anziché unire le forse e sviluppare un unico grande progetto "in serie"
2. scarsa attenzione alla qualità tecnica degli strumenti sviluppati (si crede soddisfare le esigenze funzionali senza valutare,
e spesso senza saper valutare la qualità e l'orizzonte temporale della soluzione implementata)
3. mancanza di uno standard condiviso e universale per lo scambio dati

Si è mirato a creare una rete unificata di comunità linguistiche e di applicativi *web-based* così concepita:

- ogni comunità linguistica contribuisce da un punto di vista teorico, pratico, economico, ecc, all'avanzamento costante
- del progetto
- gli applicativi, nello stato in cui si trovano, sono immediatamente disponibili a tutti i partecipanti alla rete
- gli standard tecnici di sviluppo devono a loro volta soddisfare questi requisiti:
  - assoluta eccellenza nello sviluppo Java2EE; conformità alle indicazioni W3C; rispetto dei più stringenti standard esistenti per la qualità del codice (ex: 21 cfr part 11 della FDA, a cui sono obbligate le grandi farmaceutiche)
  - rifiuto dei formati proprietari e chiusi, massima attenzione ai formati di interscambio XML, all'apertura delle specifiche
  - standard elevatissimi di progettazione software per garantire scalabilità, stratificazione tra dati, *business-logic* e presentazione, ecc (metodologia UML, 100% Object Oriented, documentazione javaDoc)

Il progetto in un anno ha avuto riscontri superiori alle aspettative.
Partecipano:

- i Sardi tramite l'ULS delle province di Nuoro e Oristano
- la rete terminologica LinMiTer tramite l'Unione Latina di Parigi
- il TermCat di Barcellona
- i Ladini con gli istituti di Fassa e Badia-Gardena
- gli Occitani italiana tramite la Chambra D'Òc
- gli Arumeni di Romania-Grecia

Nel corso del 2005 entreranno nella rete: l'Ofis Ar Brezhoneg - TermBret, l'Istituto di Sociolinguistica Catalana
(Barcellona), la minoranza Grika di Puglia. Nel 2006 l'Ufici de la Lenghe Furlane.
Sono in corso contatti molto promettenti
con tutte le minoranze d'Italia e d'Europa, e con il Colegio de México.
Attualmente il sistema conta strumenti per:

1. la creazione on line di dizionari monolingue e multilingue
2. la creazione on line di repertori terminologici multilingue
3. l'interscambio e fusione XML di dizionari esistenti (dizionario di dizionari)
4. analizzare *corpora*
5. gestire di grandi portali e quotidiani on-line
6. la correzione ortografica che, oltre ai classici algoritmi di "distanza ortografica" utilizzi algoritmi che individuino e correggano errori d'ortografia indotti dalla conoscenza, da parte dello scrivente, di una variante "non standard" della lingua (caso assai frequente per lingue di recente normalizzazione)

La divisione ha solo scopo esplicativo: si tratta di moduli tutti legati (e allo stesso tempo indipendenti), che raggiungono e raggiungeranno uno stato di integrazione finora mai realizzato, anche nell'ambito delle grandi lingue internazionali; ad esempio di mira a unificare i dizionari on-line con i repertori ad uso degli *spell-checkers*, l'analizzatore di concordanze con i dizionari, ecc.
Alcune parti del progetto saranno oggetti di altri interventi a questo convegno.

## The Igbo Language and Computer Linguistics: Problems and Prospects

Chinedu Uchechukwu
(Otto-Friedrich-Universität, Bamberg)

*Computer Linguistics* is a wholly undeveloped and an almost unknown area of research in the study of Nigerian languages. Two major reasons can be given for this state of affairs. The first is the lack of training of Nigerian linguists in this area, while the second is the general newness of computer technology in the country as a whole. This situation, however, is most likely to change as a result of the increasing introduction of the technology in the country and in the institutions of higher learning in particular. Such a change is highly promising and most welcome, but also brings up other computer technology related issues, most of which have to be properly addressed one after the other before one can with confidence speak of the onset of computer linguistics in connection with any Nigerian language.

This paper looks at the Igbo language in the light of this state of affairs in the country. Section 1, which serves as the introduction, presents the major problems confronting the language with regard to its realization in the new technology. Section 2 presents the strategies adopted to take care of these problems. Section 3 examines the benefits of such strategies on the development of Igbo corpus and lexicography, as well as the issue of computer linguistic tools (like spell checkers) for the language. Finally, section 4, the conclusion, examines the prospects of full-fledged computer linguistics in the Nigerian setting.

**Speech-to-Speech Translation for Catalan**

Victoria Arranz
(ELDA),
Elisabet Comelles
and
David Farwell
(TALP, Universitat Politècnica de Catalunya)

This abstract describes the FAME Interlingual Speech-to-Speech Machine Translation System for Catalan, English and Spanish, which is intended to assist users in the reservation of a hotel room when calling or visiting abroad. This system is part of an initiative to support Catalan within the European Union-funded FAME project (IST-2001-28323).

We will begin by giving some information about Catalan, then provide a general description of the system and show some results from its most recent evaluation.

The Catalan language, with all its variants, is the language spoken in the Països Catalans, that includes the Spanish regions of Catalonia, Valencia and Balearic Islands, the French department of the Pyréneés Orientales and in the Italian area of Alghero. Inside the Spanish territory Catalan is also spoken in some parts of Aragon and Murcia as well. Catalan is a Romance language and shows similarities with other languages belonging to the Romance family, in particular with Spanish, Galician and Portuguese. Nowadays, Catalan is understood by 9.000.000 people and spoken by 7.000.000 people.

Our Speech-to-Speech Translation (SST) System consists of four components: a speech recognizer, an analyzer that uses a CFG analysis grammar to map spoken language transcriptions into interlingua representation (called Interchange Format), a generator that uses a generation grammar to map from interlingua into natural language text, and a speech synthesizer. The main advantage of this interlingua-based architecture is that it only requires developing analysis and generation modules when adding new languages. In fact, this SST system was already used for other languages such as English and Spanish, so we created the analysis and generation components for Catalan. The Catalan analysis grammar was developed adapting and extending an existing Spanish analysis grammar. This was a smooth transition though some points required a considerable effort. As for the generation component, there was no Spanish grammar and thus the Catalan grammar was created from scratch. The whole process lasted 6 months. Nowadays we can translate from Catalan into Spanish, English, German, French and Italian and vice versa.

An evaluation of the translation component has been carried out on text input, both for a Catalan-speaking travel agent and an English-speaking tourist for a hotel reservation task. The data have been evaluated using a subjective methodology based on judgements of the fidelity and naturalness of the translations given the task. The evaluation data used were obtained from 10 dialogs recorded with 10 speakers.

A set of evaluation criteria was defined *a priori* according to the *form* and *content* of the translations. The following categories were considered:

- *Good*: well-formed output (form) or full communication of speaker's information (content).
- *Ok*, *divided into Ok+/Ok/Ok-*: acceptable output, grading from only some minor form error or non-communicated information (Ok+) to more serious form/content problems (Ok-).
- *Bad*: unacceptable output, either essentially unintelligible or semantically inaccurate.

The results obtained were as follows:

Catalan -> English

| | FORM | CONTENT |
|---|---|---|
| GOOD | 85.72% | 73.10% |
| OK+ | 5.89% | 13.45% |
| OK | 2.52% | 4.20% |
| OK- | 4.20% | 6.73% |
| BAD | 1.69% | 2.52% |

English -> Catalan

| | FORM | CONTENT |
|---|---|---|
| GOOD | 89.75% | 88.89% |
| OK+ | 8.55% | 1.70% |
| OK | 1.70% | 0.85% |
| OK- | 0% | 4.28% |
| BAD | 0% | 4.28% |

## SpeechCluster: a speech database builder's multitool

Ivan Uemlianin
(Canolfan Bedwyr, Univeristy of Wales, Bangor)

When collecting and annotating speech data, to build a database for example, speech researchers face a number of obstacles. The most obvious of these is the sparseness of data, at least in a usable form. A less obvious obstacle, but one which is surely familiar to most researchers, is the plethora of available tools with which to record and process the raw data. Example packages include, EMU, Praat, SFS, JSpeechRecorder, Festival, HTK, Sphinx. Although prima facie an embarrassment of riches, each of these tools proves to address a slightly different set of problems, to be slightly (or completely) incompatible with the other tools, and to demand a different area of expertise of the researcher.

At best this is a minor annoyance. At worst, a project must expend significant resources to ensuring that the necessary tools can interoperate. As this work is no doubt repeated in unrelated projects around the world, an apparently minor problem becomes a possibly major - and undocumented - drag on progress in the field. This danger is especially extreme in research on minority and lesser-spoken languages, where a lack of resources or expertise may preclude research completely.

Researchers need some way of abstracting from all these differences, so they can conduct their research. The simplest approach would be to provide an interface which can read and write the existing formats, and provide other facilities as required. On the WISPR project, developing speech processing resources for Welsh and Irish, we have adopted this approach in developing SpeechCluster. The vision behind SpeechCluster is to enable researchers to focus on research rather than file conversion and other low-level but necessary preprocessing. SpeechCluster is a freely available software package, released and maintained under an open-source licence.

In this paper we present SpeechCluster, reviewing the requirements it addresses and its overall design, we demonstrate SpeechCluster in use, and finally we evaluate its impact on our research, and outline some future plans.

**XNLRDF, A Framework for the Description of Natural Language Resources.**
**A proposal and first implementation**

Oliver Streiter
(National University of Kaohsiung, Taiwan),
and
Mathias Stuflesser
(European Academy Bozen Bolzano, Italy)

With the advancement of Unicode, the presentation and processing of many languages, for which previously specific implementations and resources were required, has become possible or simplified. This advancement is due on the one hand to the fact that Unicode assigns a unique 'code point' to a character of a language script. On the other hand, Unicode assigns 'properties' to characters, like 'uppercase', 'lowercase', 'decimal digit', 'punctuation' or 'separator', the writing direction, or the script, that a character belongs to. In addition, operations on the characters like uppercasing, lowercasing and sorting have been defined. Any computer application which has not been endowed with particular linguistic knowledge is thus much better off when processing a text in Unicode than in traditional encoding systems such as 'latin1', 'big5' or 'koi-r.' With Unicode, the recognition of words, numbers and sentences may be performed in many languages without additional knowledge.

The wisdom of Unicode, however, is limited to characters only. A computer application might require, or at least profit from, additional information that Unicode cannot give, e.g. how to transform the Latin number IX into an Arabic number. Much of the required information is available on Web-pages or within linguistic databases. But the databases might not be accessible on-line or the web-pages might have been designed for human reading.

XNLRDF has been designed to grant computer applications access to linguistic information on written languages, which goes beyond that offered by Unicode. This is especially important for Languages with few electronic language ressources. XNLRDF sets out to answer a computer application's questions like those listed below:

* Where is language X spoken?
* Which languages are spoken in region Y?
* What is the script used for language X in region Y?
* What is the default encoding/are the encodings for language X in region Y?
* How can I identify words/sentences in language X?
* What are the function words of language X?
* How can I perform stemming of language X?
* Which standard abbreviations are used in language X?
* Which non-Arabic numbers are used in language X, and how are they mapped onto Arabic numbers?
* Where can I find dictionaries/corpora related to language X and how are they encoded?
* Where can I find parallel texts to language X in language Z?

XNLRDF does so by storing the relevant information for hundreds of languages in an XML-structure. This seems a straight-forward solution to the problem, as any XML-

aware computer application might extract the required information from the XNLRDF XML-files. However, languages may be spoken in more than one region and a region may use more than one language. A language may have different scripts in different times and different regions, or one language may have more than one standards. To disentangle these facts and to prepresent them in a computer readable form, is the ultimate purpose of XNLRDF.

XNLRDF could have adopted RDF-XML from the beginning. However, RDF does not allow for defaulting and the overwriting of default values, something which is at least handy, e.g. when describing the French character set as 'latin' plus a set of accented characters. XNLRDF allows to define default values for groups of languages, and to overwrite some of these default values for a particular language. We adopt a non-monotonic representation which might be compiled out into RDF at a later stage.

While XNLRDF is constantly developing, the current version can be freely downloaded at: http://140.127.211.213/research/nlrdf.html.

A first implementation of XNLRDF has been integrated into Gymn@zilla, a CALL systems which supports languages like Afrikaans, Catalan, Chinese, Dutch, Faroese, Irish, Khasi, Ladin, Latin, Russian, Sanskrit, Scottish Gaelic, Swahili and Ukrainian.

# Towards Effective and Robust Strategies for Finding Web Resources for Lesser Used Languages

Baden Hughes
(Department of Computer Science and Software Engineering,
University of Melbourne)

Locating resources of interest on the web in the general (ie non-linguistic) case is at best a low precision activity owing to the large number of pages on the web (for example, Google covers more than 8 billion web pages). As language communities (at all points on the spectrum) increasingly self-publish materials on the web, so interested users are beginning to search for them in the same way that they search for general internet resources, using broad coverage search engines with typically simple queries. Given that language resources are in a minority case on the web in general, finding relevant materials for low density or lesser used languages on the web is in general an increasingly inefficient exercise even for experienced searchers. Furthermore, the inconsistent coverage of web content between search engines serves to complicate matters even more.

A number of previous research efforts have focused on using web data to create language corpora, mine linguistic data, building language ontologies, create thesauri etc. The work reported in this paper contrasts with previous research in that it is not specifically oriented towards creation of language resources from web data directly, but rather, increasing the likelihood that end users searching for resources for minority languages will actually find useful results from web searches. Similarly, it differs from earlier work by virtue of its focus on search optimisation directly, rather than as a component of a larger process (other researchers use the seed URLs discovered via the mechanism described in this paper in their own work). Moreover, this work does not use language data itself as a seed for seeking related resources as is a feature of much prior work in the area. The work here can be seen to contribute to a user-centric agenda for locating language resources for lesser-used languages on the web.

In this paper we report the development of effective and robust strategies for finding web resources for lesser used languages. Using empirical evidence, we show how a metasearch approach, combined with principled query permutation and result aggregation, significantly increase the likelihood of locating online resources for lesser used languages, in both qualitiative and quantitative dimensions. These methods are used in a range of research applications involving the curation from web data of various types of corpora for lesser used languages.

We implement a query expansion and permutation strategy around language name and linguistic terms, spawning numerous programmatic queries to web search engines given an initial input. In the first instance, we use language name variants from the Ethnologue which supports expansion from a single language name to an average of 6.7 language names in a given instance (based on 46K language name 2 Hughes variants in the 14th Edition of the Ethnologue). In the second instance, that of linguistic terms, we supplement a language name with a range of linguistic terms such as grammar, dictionary etc. Across the 7K primary language names given in the Ethnologue, on average given a single language name as a starting point, we instantiate and execute more than 100 queries to each of 3 broad coverage web search

engines, collecting the top 100 results provided. In essence, the number of queries is derived from the number of language name variants for a given input.

Having executed these queries, we require a result aggregation ('folding') policy for the large number of results obtained (on average, 1000 URLs per query). In this case our approach is relatively straightforward: we sort the resulting URLs according to ranking scheme based on frequency of occurrence. In effect, this weights the relevance of a URL directly on the number of queries the result URL appeared in the top 100 results.

Validation using of ranked URLs is undertaken using classic information retrieval measures, namely p@1, p@5 and p@10. Precision is taken to be the accuracy of a result URL with relation to language resources for the language in focus.

Hand validation of 10% of the results for the 7K primary language names from the Ethnologue reveals that p@1 and p@5 approach 100% precision, with variance introduced for p10 based on the classification of the language in question as to its number of speakers (for languages with under 2M speakers, p@10 remains in the 95-100% range, while for languages with more than 2M speakers, p@10 drops to around 87%.)

Beyond simply conducting empirical experiments based on this methodology, we also provide user-centric services based on the URLs we discover using our expansion, permutation and aggregation strategy. The source URLs are provided as collections through an Open Language Archives (OLAC) data provider – allowing access to the results using the OLAC Search Engine in general, and through a customised search interface at a higher layer. Furthermore, since OLAC provides a DP9 gateway service to broad coverage web search engines, these collections can be discovered directly by users using widely used web search engines. In essence, this publishing methodology is a 'round-trip' approach to lesser used language resource discovery on the internet, allowing collection-centric, community-grounded and broad coverage search engines to index relevant content.

# Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping

Danie Prinsloo
(Department of African Languages, University of Pretoria)
and
Ulrich Heid
(IMS-CL, Institut für maschinelle Sprachverarbeitung, Univerität Stuttgart)

Sesotho sa Leboa (= Northern Sotho) is one of South Africa's eleven national languages, spoken by about 4.2 million people in north-eastern South Africa. Typologically, it belongs to the Sotho group of the Bantu languages. It is written disjunctively (contrary to the Nguni group). With the installation, in the year 2000, of dictionary units for all South African languages, a need for corpus data for dictionary writing was felt. Since 1991, an unannotated corpus of Northern Sotho has been collected at University of Pretoria (now ca. 6 million tokens, cf. [Prinsloo 1991]). We report on the development of a word class (= POS) tagset, of POS guessers and of resources for a stochastic POS tagger for Northern Sotho. The objective is to provide
word class annotated corpora for lexicography.

Northern Sotho (as all Bantu languages) shows a particularly uneven distribution of both ambiguity and frequency across word classes: there are several hundred function words which are very frequent (the top 1000 words by frequency cover 77.5 per cent of the occurrences in the 6 million word corpus) and highly ambiguous with respect to PUS; alongside, most noun and verb forms are unambiguous. Moreover, noun and verb morphology are marked by mostly unambiguous affixes. These facts are accounted for in the tagset, which is much more fine grained in the field of function words than for lexical words (details in the presentation).

For POS tagging, we opted for the stochastic TreeTagger [Schmid 1994], as the (manually corrected) reference corpora it requires are smaller than with other tools, i.e. ca. 50.000 word forms.

For the creation of the reference corpus and of a tagger lexicon (word + POS), we opted for a lingustically informed bootstrapping approach, which again takes the above properties of the language into account. We identify noun and verb forms by means of their affixes (pattern based search, automatic classification proposal, manual correction). In addition we project a manually constructed list of 753 function words with their alternative POS tags onto the corpus (ambiguous annotation) and use contextual disambiguation rules to identify the most plausible POS of these function words. The rules are handcrafted (as in rule-based tagging) and implemented as query-and-annotate rules in the format of the corpus query processor QP (URL: http: //www.ims. uni-stuttgart.de/projekte/corpusWorkbench). The following is a sample rule:

• If the particle "le" is preceded by a NOUN and followed by a NOUN, then tag "le" as a conjunction (and);
• Else if "le" is followed by "a", followed by an OBJECT CONCORD, then tag "le" as SUBJECT CONCORD of noun class 5;

• Else if “le” is followed by “tlo/tla” then tag “le” as SUBJECT CONCORD C5/2PP;

• Else if “le” is preceded by a SUBJECT CONCORD with or without “a” and followed by a VERB then tag “le” as OBJECT CONCORD C5/2PP.

In the paper, we will present details about the noun and verb morphology analyzer and about the query-and-annotate rules. We will quantitatively and qualitatively assess the bootstrapping approach we use:

• proportions of automatically assignable, semi-automatically assignable and only manually assignable POS labels;

• kinds of word forms not amenable to semi-automatic annotation;

• results of the use of the ‘fleeTagger, including a first evaluation.

From a theoretical point of view, we are interested in the interplay between linguistic analysis, corpus-based methods and semi-automatic bootstrapping of linguistic resources, also because there are three more Sotho languages in South Africa, for which similar procedures may work.

References:

• [SCHMID 1994] Schmid, Helmut: “Probabilistic Part-of-Speech Tagging Using Decision Trees”. In Proc. International Conference on New Methods in Language Processing (NeMLaP). Manchester, UK.

• [PRINSLOO 1991] Prinsloo, D.J.: “Towards computer-assisted word frequency studies in Northern Sotho”. In SA Journal of African Languages, 11(2) 1991.

• [DESCHRYVER 2000] DeSchryver, G-M: “Electronic corpora as a basis for the compilation of African-language dictionaries”, Part 2: The inicrostructure. South African Journal of African Languages 20/4: 310-330.
Ulrich Held 2 printed on February 28, 2005

# A COMPARISON OF APPROACHES TOWARDS WORD CLASS TAGGING: DISJUNCTIVELY VS CONJUNCTIVELY WRITTEN BANTU LANGUAGES

Elsabé Taljard
(University of Pretoria)
and
Sonja E. Bosch
(University of South Africa)

The disjunctive versus conjunctive writing systems in the South African Bantu languages have direct implications for word class tagging. For purposes of this discussion we selected Northern Sotho, representing the disjunctive writing system, and Zulu as an example of a conjunctively written language. These two languages belong to the Southeastern zone of Bantu languages. The following example illustrates the difference in writing systems:

Northern Sotho: Ke a ba rata 'I like them'

| | | | |
|---|---|---|---|
| Ke<br>I<br>Subject conc. 1p.sg | a<br>PRES<br>Present tense morpheme | ba<br>them<br>Object concord cl 2 | rat-a<br>like<br>Verb root +<br>ending |

Zulu: Ngiyabathanda 'I like them'

| | | | |
|---|---|---|---|
| Ngi<br>I<br>Subject conc. 1p.sg | ya<br>PRES<br>Present tense morpheme | ba<br>them<br>Object concord cl 2 | thand-a<br>like<br>Verb root +<br>ending |

In this paper a two pronged approach is followed. Firstly, the available linguistic and computational resources for the two languages are compared; secondly, a comparison is drawn between the approaches towards word class tagging for Northern Sotho and Zulu.

Both languages have unannotated electronic corpora at their disposal – 6 million tokens for Northern Sotho, and 5 million tokens for Zulu. These corpora are utilized among others for the generation of frequency lists, which are of specific importance for the development and testing of word class tagging, especially in disjunctively written languages. In Northern Sotho, for instance, the top 10 000 types in the corpus represent approximately 90% of the tokens, whereas in Zulu the top 10 000 types represent only 62% of the tokens. This implies that the correct tagging of the top 10 000 tokens in Northern Sotho, be it manual, automatic or a combination, results in a 90% correctly tagged corpus. The low relation between types vs tokens in Zulu, however, results in a much smaller percentage, i.e. 62% of the corpus being tagged.

An additional resource such as a morphological analyser as described in Pretorius & Bosch (2003), would therefore be a useful tool to facilitate a higher percentage in the automatic tagging of the Zulu corpus.

With regard to the tagsets of the two languages respectively, important differences come to the fore. The tagset for Northern Sotho is a hybrid system, containing both morphological and syntactic elements, although biased towards morphology. In the case of Zulu, morphological aspects need not be included in the word class tagging since these are already accounted for in the morphological analysis. This difference in approach to the tagsets can be mainly ascribed to the different writing systems.

In both languages, cases of ambiguous annotation require the application of disambiguation rules based mainly on surrounding contexts. A typical example of ambiguity is that of class membership, due to the agreement system prevalent in these languages. For instance, in orthern Sotho as well as Zulu, the class prefix of class 1 nouns is morphologically similar to that of class 3 nouns, i.e. mo- (N.S) and umu- (Z). This similarity makes it impossible to correctly assign class membership of words such as adjectives, which are in concordial agreement with nouns, without taking the context into account.

References:

De Schryver, G-M & D.J. Prinsloo. 2000. The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies 18/1-4*: 89–106.

Pretorius, Laurette & Sonja E Bosch. 2003. Computational aids for Zulu natural language processing. *Southern African Linguistics and Applied Language Studies 21/4*: 267-282.

Van Rooy, Bertus & Rigardt Pretorius. 2003. A word-class tagset for Setswana. *Southern African Linguistics and Applied Language Studies 21/4*: 203-222.

## Grammar-based language technology for the Sámi languages

Trond Trosterud
(Det humanistiske fakultet, Universitetet i Tromsø)

Working with language technology for minority languages differs from working with majority languages. In the latter case, the projects often have a long prehistory, and the source code thus involves several generations of technology, it may be restricted to 8-bit or even 7-bit, and due to possible commercial interest, it may even be unavailable to inspection. Contemporary projects for minority languages face a different situation: As new projects, they are not hampered by the legacy of old solutions, but may build on state-of-the-art solutions from the start. 7-bit ascii is never an option, and in most cases, it is in principle desireable to use Unicode UTF-8. Since minority language projects are not financed via income from product sales, sharing source code with other projects are usually not a problem. Seen in this light, the perspectives for portability between different language technology solutions seem promising.

Our project, involving 3 different Sámi languages, is run on Mac/Linux platforms, and uses UTF-8 as its native encoding set. With the latest versions of the respective OS-s and shells, we have access to tools that in most cases are UTF-8 aware, and although it takes some extra effort to tune the development tools to multi-byte input, the advantage is a more readable source code (with correct letters instead of digraphs) and an easier input/output interface, as UTF-8 now is the de facto standard for digital publishing.

We build morphological transducers with two-level morphology and the Xerox fst development tools. Disambiguation is done with constraint grammar. With these basic tools as a starting point, we offer online analysis and generation, at giellatekno.uit.no. Ongoing work includes morphologically annotated corpora for linguistic analysis, a spell checker, and interactive pedagogical grammar learning programs (in cooperation with visl.sdu.dk). Future plans include information retrieval, intelligent bilingual dictionaries and term bases, and basic text-to-speech solutions).

Within language technology, there is a long-term controversy between statistical and grammatical methods. The former often present themselves as language independent, and thus easily portable to new languages. Our experiences with Sámi bring us to the opposite conclusion.

First, good achievements with a statistical approach requires both large corpora, and a relative simple morphological structure (low wordform / lemma ratio). Sámi and many other languages have a rich morphological structure and a paucity of corpus resources, whereas the basic grammatical structure of the languages is reasonably good understood.

Work on minority languages will typically be carried out as cooperation projects between research institutions and ingroup individuals or organisations devoted to the strengthening of the languages in question. Whereas private companies will look at the ratio of income to development cost, and care less about the developmental philosophy, it is important for research institutions to work with systems that are not “black boxes”, but that are able to give insight into the language beyond merely producing a tagger or a synthetic voice.

# ANNOTATION OF DOCUMENTS FOR ELECTRONIC EDITION OF JUDEO-SPANISH TEXTS : PROBLEMS AND SOLUTIONS

Soufiane Rouissi, Ana Stulic
(University of Bordeaux 3)

Issued from the interdisciplinary point of view that comprises Linguistics, Information and Computer Sciences, this contribution consist of modelling the annotated electronic edition of Judeo-Spanish (language spoken by the Sephardic Jews expelled from Spain at the end of 15th century and settled in the large Mediterranean area) texts written in Hebrew characters following the principle of generation of a document in a collaborative work environment.

The Judeo-Spanish texts in Hebrew characters use an adaptation of Hebrew script, but the conventions of its use present many variations due to different way the adaptation of a script can be realized and as well as to the phonological changes in Judeo-Spanish. The main difficulty concerning the edition of Judeo-Spanish texts is to make a transcription easy to read (where the vowels are specified, for example, which is not the case in the original texts where no difference is made between similar vowels /e/ and /i/, and /o/ and /u/) and, at the same time, preserve the original writing system which can be subject to discussion/further interpretation.

Our approach is based on the concept of annotation of a document that places mark up at word/group of words level on the result of transcription. We adopt the point of view by which the annotations of 'translated/interpreted' document can have two different purposes, to interpret (to add new mark up in order to propose different interpretation from the one formulated at the starting point), at one hand, and at the other, to comment (place a comment on the interpretation done by another author). The aim is to make possible for the reader/user to act over the document by adding his own interpretation (translation) and/or comments over the interpretation done by another author. In the environment that facilitates active reading, the reader/user becomes the author of the document generated on the basis of the version proposed as a starting point.

This general model can be presented in a schematic way :

(1) Original document (in the forme of a scanned image, for example) ->(2) « Starting point version » : first interpretation done by an author that can be the basis for « discussion / construction » (the transcription of a document, where the vowels are specified with a unique criterion, in our case) ->(3) Document generated by various authors who can add their interpretation over the existing mark up, introduce new mark up at word/group of words level that can equally be available for discussion. At this point the conditions of storage must be put into question, like the format of annotations and of document itself. ->(4) Starting from (3),it would be possible to construct new documents using a particular criterion (choosing an interpretation of a particular author, or generating new document on the basis of active reading using successive choices). This electronic document can be presented in various formats, ASCII, HTML, XML, TEI or a specific format conditioned by its storage in the data base that would facilitate the ulterior usage.

## Stealth Learning with an On-line Dog

Ambrose Choy, Gruffudd Prys
(Canolfan Bedwyr, Univeristy of Wales, Bangor)

This paper describes an innovative new project designed to improve the language skills of fluent Welsh speakers. Commissioned by BBC Wales, it will appear on the BBC's Welsh language web-site. It comprises six different types of word games, a self marking set of language improvement exercises, and an on-line answering service dealing with grammatical and other language problems.

The word games are a mixture of popular formulas already adapted to a web-based environment: conundrums, hangman and crossword puzzles, together with games to discover Welsh proverbs, geographical entities and word definitions. They are targeted at young professionals, used to working in a computing environment, who are interested in language matters. The Welsh title of the project: *Ieithgi* (literally 'language dog') refers to someone who is interested in language matters. The dog in this instance is Cumberland, who features in the BBC's cartoons for learning Welsh, *Colin and Cumberland* (also available for Irish and Scots Gaelic). In this series, Colin is the slightly ignorant human trying to learn Welsh, with Cumberland as the allknowing fluent Welsh speaker. Cumberland was therefore deemed suitable to be the language teacher, featuring in the games, but also as the language expert in "Ask the dog".

It is hoped that the games will prove popular for their entertainment value. However, there is a hidden agenda in their design. Their intention is to improve the vocabulary and spelling skills of players, together with their knowledge of Welsh culture and geography. The historically inferior position of Welsh as a minority language means that speakers have less confidence when using it in formal or professional contexts. This project hopes to raise users confidence through pleasurable experiences, through stealth learning, where players who visit the web-site will not necessarily understand that they are learning new language skills, but will enjoy playing with words and language.

## Il ladino fra polinomia e standardizzazione: l'apporto della linguistica computazionale

Evelyn Bortolotti, Sabrina Rasom
(Istitut Cultural Ladin "majon di fascegn")

Il ladino delle Dolomiti (Italia) è caratterizzato da una grande varietà interna, che ha reso necessario un intervento di normazione e standardizzazione, nel rispetto del carattere polinomico della lingua stessa.
Nelle valli ladine dolomitiche si vanno formando lingue di scrittura, o standard di valle. Alcuni idiomi di valle sono piuttosto unitari ed è stato sufficiente codificare questi, ma in Val Badia (con Marebbe) e in Val di Fassa la varietà degli idiomi ha portato alla proposta di una normazione che andasse sopra gli idiomi di paese: il "*badiot unitar*", basato principalmente sull'idioma centrale (*San Martin*), ma aperto anche a elementi provenienti da idiomi di altri paesi, e similmente il "*fascian standard*", orientato verso l'idioma *cazet*, la cui scelta come variante standard è giustificata anche dal fatto che il fassano standard si propone come elemento di congiunzione rispetto agli idiomi delle altre vallate.
Infine si è sentito il bisogno di elaborare un livello ancora più alto di standardizzazione valido per l'intera Ladinia, sulle orme del Rumantsch Grischun, dando il via alla elaborazione del *Ladin Dolomitan*, o *Ladin Standard*.
Dal punto di vista della polinomia quindi, da una situazione linguistica molto differenziata, si è passati prima ad un livello più alto di normazione dove sul piano di valle si raccolgono più varietà in una norma unica, per poi raggiungere un terzo livello, una terza possibilità, che permette di avere a disposizione un unico idioma di riferimento, una norma o lingua standard per tutte e cinque le vallate.
I vari progetti relativi all'informatizzazione delle risorse lessicali e allo sviluppo di strumenti per il trattamento automatico della lingua ladina sono dunque stati portati avanti attenendosi al principio di conservazione e valorizzazione della ricchezza e della varietà in una visione unitaria. Questo principio deriva dalla riflessione teorica del linguista còrso Jean-Baptiste Marcellesi, dove per la prima volta compare il concetto di "lingue polinomiche" (*Langues Polynomiques*).
I principali obiettivi perseguiti in campo linguistico computazionale, che verranno presentati più ampiamente nella relazione sono:
• l'informatizzazione del patrimonio lessicale ladino con la creazione di una banca dati generale lessicale ladina, di banche dati strutturate delle varietà locali e di una banca dati centrale dello standard;
• l'elaborazione di dizionari degli standard di valle e dello standard dolomitano anche in versione elettronica o consultabili online;
• una raccolta di glossari terminologici, parzialmente consultabili online:
• la creazione di corpora elettronici analizzabili tramite un'apposita interfaccia, il *webconcordancer*;
• la realizzazione di strumenti informatici per facilitare l'uso e l'apprendimento delle varianti standard: dizionario elettronico, e-learning, correttori ortografici e adattatori per il fassano standard e per il *Ladin Standard*.

## The Welsh National On-line Terminology Database

Delyth Prys, Dewi Evans
(Canolfan Bedwyr, Univeristy of Wales, Bangor)

Terminology standardization work has been ongoing for the Welsh language for many years. At an early date the decision was taken to adopt international standards such as the ISO 704 and 860 ones for this work. It was also decided to store the terminologies in a standard format in electronic databases, even though the demand in the early years was for traditional paper-based dictionaries.

Welsh is now reaping the benefits of those far-seeing early decisions. In 2004 work began on compiling a national database of bilingual (Welsh/English) standardized terminology. Funded by the Welsh Language Board, it will be made freely available on the world-wide web. Electronic databases already in existence have been revisited and reused for this project, with a view to updating them to conform to an ISO terminology markup framework (TMF) standard.

An additional requirement of this project is that the term lists should be packaged and made available in a compatible format for downloading into popular termbase systems found in translation tool suites such as Trados, Déjà Vu and Wordfast. As far as we know, this is the first time that a terminology database has been developed to provide a freely available termbase download utility at the same time as providing an on-line searchable facility.

Parallel work of utilizing an ISO lexical markup framework (LMF) compliant standard for another project, namely the LEXICELT Welsh/Irish dictionary, has provided the opportunity to research similarities and differences between a terminological concept-based approach and a lexicographical lexeme-based one.
Direct comparison between TMF and LMF have been made, and both projects have gained from new insights into their strengths and weaknesses.

This paper will present an overview of the on-line database, and attempt to show how frugal reuse of existing resources and adherence to international standards both help to maximize sparse resources in a minority language situation.

## Lexicelt: An On-line Welsh/Irish Dictionary

Delyth Prys, Dewi Evans
(Canolfan Bedwyr, Univeristy of Wales, Bangor)

Bilingual dictionaries between two minority languages are comparatively rare. This paper describes an Interreg IIIa (Wales/Ireland) project to create such a dictionary. Welsh and Irish are related Celtic languages, but until now dictionary users have had to use English as an intermediate language in order to translate between the two. As well as being of use to students learning Irish through the medium of Welsh, the dictionary is aimed at the general public, especially as cultural tourism between Wales and Ireland is a growing phenomenon. It is also important for two small business sectors in north and west Wales and the east of Ireland: the publishing industry with its programme of translating literature between Welsh and Irish, and the television industry with its translations of television programmes between the same two languages.

Designing the dictionary as an on-line interactive one has enabled it to use many new features not available to traditional, paper-based dictionaries. The lexicography currently being developed adheres to an ISO Lexical mark-up Framework (LMF) compliant standard, thereby enabling two minority languages to be part of the international mainstream. It also includes a lemmatizer for both Welsh and Irish. This is a vital feature for languages where initial mutations and conjugations make finding words based on traditional alphabetical look-up difficult for those not fluent in those languages. It also includes sound files to aid correct pronunciation. This takes advantage of another Interreg IIIa funded project, namely the WISPR project, which is developing text to speech technology for Welsh and Irish.

Strong interest has been shown in extending the Lexicelt platform to include other Celtic languages. This paper will also explore avenues for future cooperation and the ways using LMF compliant formats facilitate such cooperation.

# Author Index