Best Practices for Sharing Language Technology Resources in Minority Language Environments

Delyth Prys

Bangor University, Wales

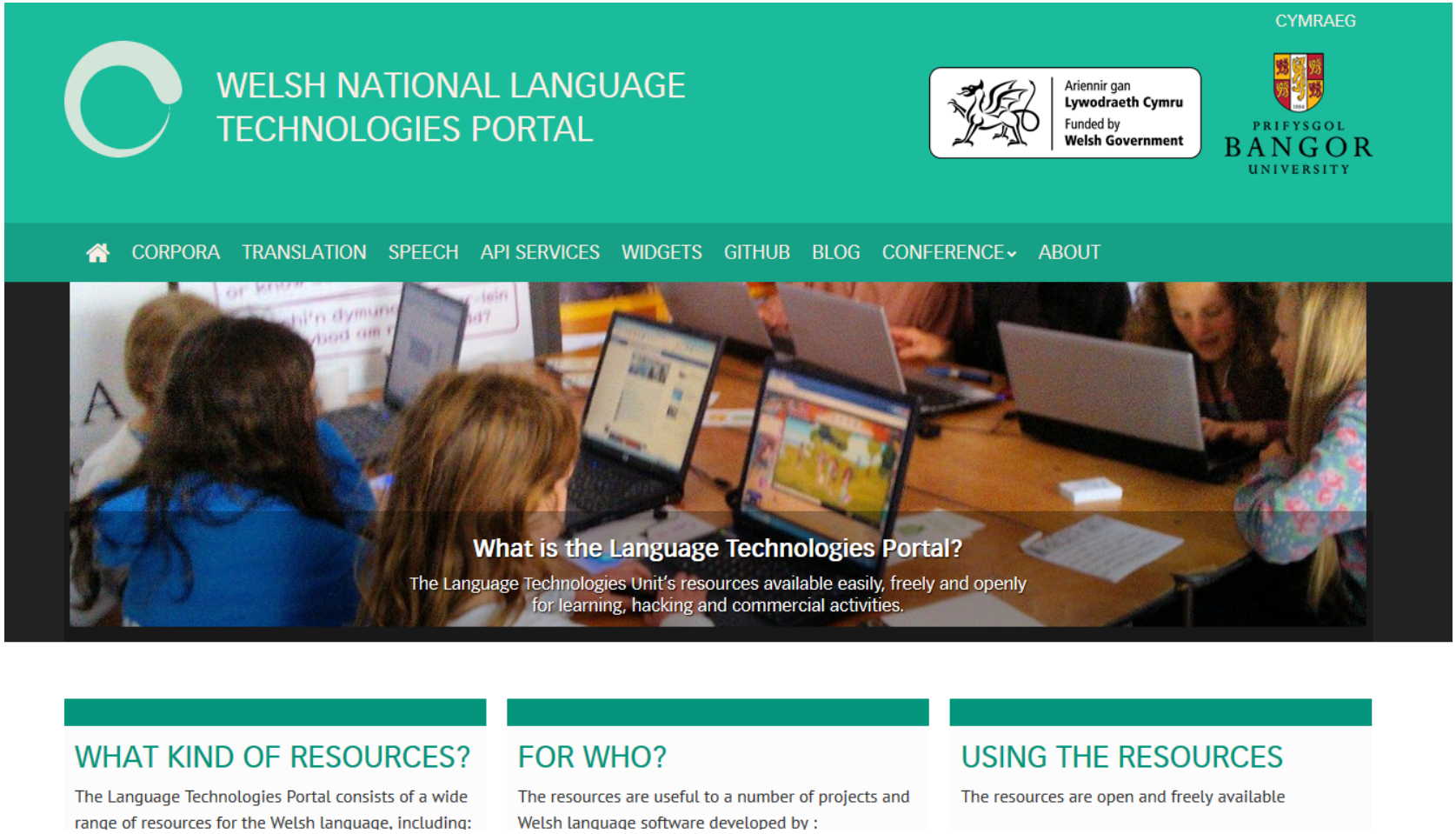
d.prys@bangor.ac.uk

The Language Technologies Unit, Bangor University

- Self-funded research unit
- Established as the Centre for the Standardization of Welsh Terminology in 1993
- Expanded scope from 2001 onwards to include Speech and Language Technologies
- Focus on Welsh and other minoritized/less-resourced languages

The Welsh National Terminology Portal

http://techiaith.org/



The Problem

- Lack of easily accessible tools for developing LT for Welsh
- Expense of developing tools for languages which are not commercially interesting
- Many useful LT tools/resources developed over the years for internal use not available to others
- Lack of knowledge by software/web developers etc on how to use these resources

The Solution

- Hold an audit of all tools/resources developed by the LTU that are available for reuse
- Select the most important ones for repackaging and dissemination
- Create a free, accessible on-line portal
- Include documentation, tutorials and ideas for use

Initial resources selected

1. Social web corpora (Twitter and Facebook)
2. Welsh<>English Machine Translation
3. Welsh<>English Aligner for preparing parallel texts
4. Welsh Text-to-Speech API service
5. Welsh Festival Speech Synthesis resources
6. Spelling and grammar checker API and widget
7. PoS Tagger API
8. Language Detection API
9. 'Vocab' bilingual dictionaries widget

TRANSLATION

Language technologies are currently revolutionizing the world of translation. Here at the Language Technologies Unit, we believe that Welsh translators need be able to understand and use these disruptive new technologies effectively.

We believe that the industry should use translation technologies on the following basis:

- Domain-specific translation engines
- Data from open and closed sources
- Ownership and sharing
- Standards, research and training
- Prioritizing the needs of Welsh translation customers

The Language Technologies Portal contains many free and open resources that can contribute towards achieving these aims. We hope that the Portal will lead to the growth and development of a community of developers and practitioners within the Welsh translation industry.

WHAT RESOURCES ARE AVAILABLE?

These resources for facilitating translation are available from the Language Technologies Portal:

- Welsh <> English Machine Translation with Moses-SMT
- An aligner for preparing data for training Moses-SMT.

Moses-SMT Welsh <> English Machine Translation

This project contains the code, scripts and related documentation which facilitates the use of Moses-SMT for Welsh.

The Project Structure

- **docs** - contains documentation on how to use the scripts. Specifically, how to:
    - install Moses-SMT from docker.com and download the Language Technologies Unit's translation packages click here...
    - install Moses-SMT on a normal computer - click here...
    - run Moses-SMT on a normal computer - click here...
    - train Moses-SMT and create your own engines - click here...
- **get** - contains scripts that are necessary to prepare and install Moses-SMT on a Linux computer
- **mtdk** - containts the scripts that support Moses-SMT's training features

The Project Files

- **Dockerfile** - the file that is used to create docker images

Porth Technolegau Iaith Cenedlaethol Cymru

Casgliad o adnoddau agored ac am ddim technolegau iaith Cymraeg ar gyfer clybiau codio, hacwyr, a chwmnïau i greu cynnyrch digidol Cymraeg

Uned Technolegau Iaith, ... http://www.techiaith.org techiaith@bangor.ac.uk

People 0 >

This organization has no public members. You must be a member to see who's a part of this organization.

Filters Find a repository...

moses-smt

Python ★ 1 1

Adnoddau Cyfieithu Peirianyddol Cymraeg / Machine Translation for Welsh

Updated 9 days ago

alinio

Python ★ 0 0

Cod hwyluso alinio gyda hunalign a dogfennaeth ar sut i ddefnyddio LFAligner / Code for simplifying using hunalign and documentation for LFAligner

Updated 28 days ago

festival

HTML ★ 0 0

Dogfennaeth ac enghreifftiau ar sut i ddefnyddio API testun i leferydd Cymraeg trwy Festival ~~ Documentation and examples on how to use the Welsh text-to-speech API

Updated 28 days ago

llais_festival

Scheme ★ 1 0

Llais Deuffon Festival Cymraeg - Welsh Diphone Festival Voice

Updated 28 days ago

PorthTechnolegaulaith/turing-test-lesso...
GitHub, Inc. (US) https://github.com/PorthTechnolegaulaith/turing-test-lessons
Google

PorthTechnolegaulaith / turing-test-lessons

forked from raspberrypilearning/turing-test-lessons

Watch 3 Star 0 Fork 10

Mae'r cynllun gwaith hwn wedi'i gynllunio fel cyflwyniad i raglennu seiliedig ar destun gan ddefnyddio Python ar y Raspberry Pi. Erbyn diwedd yr holl wersi, bydd myfyrwyr wedi creu robot sgwrsio rhyngweithiol ac wedi'u cyflwyno at technolegau iaith Cymraeg yn ogystal a'r cysyniad cyfrifiadurol o ddeallusrwydd artiffisial / This scheme of work ha...

2 commits 2 branches 0 releases 2 contributors

branch: master turing-test-lessons / +

This branch is 2 commits ahead, 73 commits behind raspberrypilearning:master Pull Request Compare

Ychwanegu 'technolegau iaith Cymraeg' i'r README

techiaith authored on Mar 26 latest commit c83074df38

Lesson-1	Cyfieithiad llawn / Completed translation	a month ago
Lesson-2	Cyfieithiad llawn / Completed translation	a month ago
Lesson-3	Cyfieithiad llawn / Completed translation	a month ago
Lesson-Garn	Cyfieithiad llawn / Completed translation	a month ago

Code
Pull requests 0
Pulse
Graphs

HTTPS clone URL
https://github.com/Pc

You can clone with HTTPS or Subversion.

Clone in Desktop
Download ZIP

‘Vocab’ widget installed on newspaper website

Cynyddu maint testun

RSS VOCAB

golwg360

Swyddfa Ewrop Greadigol y Deyrnas Unedig Cymru | Creative Europe Desk UK Wales
www.creativeeuropeuk.eu
EWROP GREADIGOL
Cyllid a chyfleoedd i'r sectorau diwylliannol, creadigol a chlywedol

Newyddion | Chwaraeon | Celfyddydau | Calendr | Blog | Bwyd | Lle Pawb | Lle Busnes | Adran Swyddi

Cymru | Prydain | Rhyngwladol

Addysg | Arian a Busnes | Gwleidyddiaeth | Iechyd | Rhyfeddodau | Gwasanaethau Brys | Cefn Gwlad | Gwyddoniaeth a Thechnoleg | Iaith | Clonc360

Cwmni cyfieithu yn agor yn Aberystwyth

Ebrill 24, 2015 09:50

Mae un o brif gwmnïau cyfieithu'r gorllewin yn cynnal lansiad i ddathlu agor swyddfa newydd a chreu swyddi newydd yn Aberystwyth heddiw.

Dros y tair blynedd ddiwetha' mae cwmni Trywydd wedi gweld “cynnydd yn y galw” am ei wasanaethau yn ôl y rheolwyr, ac mae hynny wedi arwain at greu pedair swydd newydd.

Fe fydd yr Aelod Cynulliad lleol Elin Jones yn bresennol yn y lansiad sy'n cychwyn am 11.

Mae wyth o bobol yn gweithio yn swyddfeydd y cwmni yng Nghaerfyrddin a Llandeilo.

Cyflogaeth leol

Ffrwyth prosiect Ewropeaidd o'r enw Dewis, a sefydlwyd yn 2002, yw Trywydd.

Dywedodd y Prif Weithredwr Owain Gruffydd: “Rydym yn falch iawn ein bod fel cwmni yn ehangu ymhellach, yn dilyn y galw cynyddol am ein gwasanaethau cyfieithu a chynllunio iaith – nid yn unig yn yr ardal hon, ond ledled Cymru.

Straeon Perthnasol

- Cwmni iogwrt yn wynebu colli swyddi
- Aberystwyth yn brifddinas diwylliant Prydain?
- Tridiau o streic yn Ysgol Penweddig
- Trigolion Wrecsam yn yfed mwy o de nag unrhyw le arall ym Mhrydain

Chwilio Golwg 360...

y ganolfan cynllunio iaith
welsh centre for language planning
cyngor
ymchwil
hyfforddiant
digwyddiadau
prosiectau
☎ +44 (0) 1239 711668
post@iaith.eu www.iaith.eu

Coleg Cymraeg Cenedlaethol

‘Vocab’ dictionary selected on text

Conclusions

- Successfully addresses needs of both local and international developers
  - Hence National Portal + use Github and other international repositories
  - APIs and widgets have proved very popular (due to demand from web developers?)
- Increases interest in open source LTs such as Moses and Festival
- Further queries from public organisations show the need to prioritize ease of use to end users