AFCP

ATALA

JEP-TALN-RECITAL 2016

Journées d'Études sur la Parole
Traitement Automatique des Langues Naturelles
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

PARIS Inalco du 4 au 8 juillet 2016

Organisé par les laboratoires franciliens

https://jep-taln2016.limsi.fr

Conférenciers invités:

Christian Chiarcos (Goethe-Universität, Frankfurt.)
Mark Liberman (University of Pennsylvania, Philadelphia)

Coordinateurs comités d'organisation

Nicolas Audibert et Sophie Rosset (JEP)
Laurence Danlos & Thierry Hamon (TALN)
Damien Nouvel & Ilaine Wang (RECITAL)
Philippe Boula de Mareuil, Sarra El Ayari & Cyril Grouin (Ateliers)

Préface

Thierry Poibeau[1] Teresa Lynn[2] Delyth Prys[3] John Judge[2]
(1) LATTICE, CNRS-ENS-U. Sorbonne Nouvelle, Paris, France
(2) ADAPT Centre, Dublin City University, Dublin, Ireland
(3) Language Technologies Unit, Bangor University, Bangor, Wales, UK
tlynn@computing.dcu.ie
d.prys@bangor.ac.uk
jjudge@computing.dcu.ie
thierry.poibeau@ens.fr

Le traitement automatique des langues a permis le développement d'un grand nombre d'outils et de ressources pour des langues variées. Des outils comme des correcteurs orthographiques, des interfaces vocales embarquées et bien encore des corpus de milliards de mots ont vu le jour et ont permis le développements d'applications utiles dans la vie quotidienne de millions d'utilisateurs.

Jusqu'à récemment, les langues avec un nombre moindre de locuteurs n'ont pas bénéficié des mêmes avancées. Cependant, les techniques permettent aujourd'hui de mettre au point des outils et des ressources efficaces à partir de moins de données et en un temps limité. De fait, les langues sous dotées disposent à leur tour de plus en plus souvent d'outils et de ressources de qualité.

Les ateliers sur le Traitement automatique des langues celtiques (CLTW) visent à rassembler les chercheurs intéressés par le développement d'outils et de ressources pour cette famille de langues. Comme celles-ci sont largement sous-dotées, le but est aussi d'encourager le dialogue et les collaborations entre chercheurs.

L'édition 2016 de l'atelier sera le deuxième de la série, après la première édition qui avait eu lieu lors de COLING 2014 à Dublin. Nous sommes heureux d'organiser cette deuxième édition en France, conjointement avec la conférence JEP-TALN. Dix articles ont été retenus pour présentation et montrent la richesse et la variété des recherches effectuées par les équipes impliquées.

Nous remercions Robin Owain, pour avoir accepté d'intervenir comme conférencier invité. Nous remercions également tous les auteurs pour la qualité et l'intérêt des travaux soumis, les participants et évidemment les membres du comité de programme pour la qualité de leurs remarques et de leur travail de relecture.

Preface

Thierry Poibeau[1] Teresa Lynn[2] Delyth Prys[3] John Judge[2]
(1) LATTICE, CNRS-ENS-U. Sorbonne Nouvelle, Paris, France
(2) ADAPT Centre, Dublin City University, Dublin, Ireland
(3) Language Technologies Unit, Bangor University, Bangor, Wales, UK
tlynn@computing.dcu.ie
d.prys@bangor.ac.uk
jjudge@computing.dcu.ie
thierry.poibeau@ens.fr

Language Technology and Computational Linguistics research innovations in recent years have given us a great deal of modern language processing tools and resources for many languages. Basic language tools like spell and grammar checkers through to interactive systems like Siri, as well as resources like the Trillion Word Corpus, all fit together to produce products and services which enhance our daily lives.

Until relatively recently, languages with smaller numbers of speakers have largely not benefited from attention in this field. However, modern techniques in the field are making it easier to create language tools and resources from fewer resources in a faster time. In this light, many lesser spoken languages are making their way into the digital age through the provision of language technologies and resources.

The Celtic Language Technology Workshop (CLTW) series of workshops provides a forum for researchers interested in developing NLP (Natural Language Processing) resources and technologies for Celtic languages. As Celtic languages are under-resourced, our goal is to encourage collaboration and communication between researchers working on language technologies and resources for Celtic languages.

This will be the second Celtic Language Technology Workshop (CLTW). The first event was held in Dublin during COLING 2014 and was hugely successful in bringing this fledgling community together. We are pleased to organise this second workshop in France, during the JEP-TALN conference. This year the workshop will present 10 selected papers covering a variety of topics of relevance to the Celtic languages and their associated technologies.

We thank Robin Owain for the invited conference. We also want to thank all our authors and presenters for their hard work and workshop attendees for their participation, and of course we are very grateful to our programme committee for reviewing and providing invaluable feedback on the work published here.

Comité de programme

— Colin Batchelor, Royal Society of Chemistry, Angleterre
— Aoife Cahill, ETS, USA
— Andrew Carnie, University of Arizona, USA
— Brian Davis, INSIGHT/ National University of Ireland, Irlande
— Jeremy Evas, University of Cardiff, Pays de Galles
— Mikel Forcada, Universitat d'Alacant, Espagne
— Annie Foret, Université Rennes 1, Bretagne
— William Lamb, University of Edinburgh, Ecosse
— Montse Maritxalar, University of the Basque Country, Espagne
— John McCrae, INSIGHT/ National University of Ireland, Irlande
— Neasa Ní Chiarán, TCD, Irlande
— Brian O Raghallaigh, Fiontar/ Dublin City University, Irlande
— Kevin Scannell, Saint Louis University, USA
— Mark Steedman, University of Edinburgh, Ecosse
— Nancy Stenson, UCD, Irlande
— Fran Tyers, Prompsit, Russie
— Elaine Uí Dhonnchadha, Trinity College Dublin, Irlande
— Pauline Welby, CNRS/UCD, France

Organisateurs de l'atelier

— Thierry Poibeau, CNRS, France
— Teresa Lynn, DCU, Irlande
— Delyth Prys, Bangor University, Pays de Galles
— John Judge, DCU, Irlande

Table des matières

Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in Scottish Gaelic
Colin Batchelor 1

Building a Dictionary-Based Lemmatizer for Old Irish
Oksana Dereza 12

CALLIPSO – CALL for Irish for Parents Students and Others
Monica Ward 18

Developing Word Embedding Models for Scottish Gaelic
William Lamb, Mark Sinclair 31

English to Irish Machine Translation with Automatic Post-Editing
Meghan Dowling, Teresa Lynn, Yvette Graham, John Judge 42

Enrichissement de données en breton avec Wordnet
Annie Foret 55

Insular Celtic Language Mark-up in WordPress
Mícheál Mac Lochlainn 62

Towards a lexicon of Irish-language idioms
Katie Ní Loingsigh 69

Universal Dependencies for Irish
Teresa Lynn, Jennifer Foster 79

Vocab : a dictionary plugin for web sites
Dewi Bryn Jones, Gruffudd Prys, Delyth Prys 93

Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in Scottish Gaelic

Colin Batchelor
Royal Society of Chemistry, Cambridge, UK CB4 0WF
colin.r.batchelor@gmail.com

RÉSUMÉ

Grammaire catégoriale dérivé automatiquement d'un corpus des textes en gaélique écossais avec annotations syntaxiques

Nous presentons une grammaire catégoriale preliminaire pour le gaélique écossais qui nous avons derivé automatiquement du corpus de texte ARCOSG (*Annotated Reference Corpus of Scottish Gaelic*) de l'Université d'Édimbourg, qui contient plus que 80 000 des entités lexicales en plusieurs genres avec annotations syntaxiques. Nous discutons nos méthodes pour la dérivation de cette grammaire, les traits distinctifs du gaélique écossais et du corpus, l'analyse lexicale categoriale, et dont on a besoin pour une évaluation rigoureouse et systématique d'une telle grammaire.

ABSTRACT

We present a preliminary categorial grammar for Scottish Gaelic derived automatically from the University of Edinburgh's Annotated Reference Corpus of Scottish Gaelic (ARCOSG), which contains over 80 000 tokens of part-of-speech-tagged text in multiple genres. We discuss our methods for deriving this grammar, the distinctive features of Scottish Gaelic and of the corpus, parsing CCG, and set out what is needed for a rigorous and systematic evaluation of the work presented here.

MOTS-CLÉS : gaélique écossais, grammaire categoriale, CCG.

KEYWORDS: Scottish Gaelic, categorial grammar, CCG.

1 Introduction

Scottish Gaelic, like the other Celtic languages, is marked by VSO word order, fused preposition–pronouns, word-initial mutation and extensive use of periphrastic constructions (Lamb, 2003). As in Irish the copula and verb "to be" are separate, and psychological states are typically expressed with a combination of either of those, prepositional phrases and nouns. As such it is a challenging language for automatic processing, a situation which is not helped by its having historically been an under-resourced language for natural language processing, but this started to change at the first Celtic Language Technology Workshop in Dublin in 2014 with the publication of three papers by Lamb & Danso (2014), Scannell (2014) and Batchelor (2014). Subsequently the University of Glasgow has launched the *Corpas na Gàidhlig* 'Corpus of Gaelic' as part of the Digital Archive of Scottish Gaelic (DASG) (University of Glasgow, 2016). The potential for developing resources for Scottish Gaelic has been strengthened by a recent flurry of activity in Irish, which is very closely related, the two having shared a common literary form until the 18th century. Irish now boasts a dependency treebank (Lynn, 2016), a mapping of this Irish Treebank annotation scheme to the scheme in the

Universal Dependencies Project (Nivre *et al.*, 2015), [1] and tools for POS-tagging tweets (Lynn *et al.*, 2015). In this paper we present a Scottish Gaelic categorial grammar bank derived, in contrast to our small hand-built grammar presented in Batchelor (2014), wholly automatically from a part-of-speech tagged corpus, the Annotated Reference Corpus of Scottish Gaelic (ARCOSG) (Lamb *et al.*, 2016), the longer-term background to which is described in Lamb (2008).

2 Methods

2.1 Categorial grammar

Combinatory categorial grammar (CCG) (Steedman & Baldridge, 2003) is a fully-lexicalized theory. This means that all of the grammar resides in the lexicon and that parsing involves applying those rules stored within the lexical entries. Each lexical entry, or word, has a type which may either be atomic or composite. As is standard we work with a small set of atomic types, which in this exercise are the clause (S), the noun phrase (N) and the prepositional phrase (PP). The composite types are functions and are written with slashes indicating whether their arguments are to their right or to their left. To take a simple example, intransitive verbs in Scottish Gaelic have type S/N, indicating that they expect a noun phrase to their right, and attributive adjectives have type N\N, indicating that they expect a noun phrase to their left. Parsing in its simplest form then involves function **application** using the rules :

$$\mathtt{A/B} \quad \mathtt{B} \quad \rightarrow_{>} \quad \mathtt{A} \tag{1}$$

$$\mathtt{B} \quad \mathtt{A\backslash B} \quad \rightarrow_{<} \quad \mathtt{A} \tag{2}$$

To give a concrete example, the phrase *Thàinig corra-ghridheach ghiùigeach* 'A demure heron came' parses as follows :

$$\dfrac{\dfrac{\textit{Thàinig}}{\mathtt{S/N}} \quad \dfrac{\dfrac{\textit{corra-ghridheach}}{\mathtt{N}} \quad \dfrac{\textit{ghiùigeach}}{\mathtt{N\backslash N}}}{\mathtt{N}}<}{\mathtt{S}}> \tag{3}$$

the N\N of *ghiùigeach* combines backwards with the N of *corra-ghridheach* to yield an N, which is then consumed by the S/N of the verb *thàinig* to yield a complete clause.

In addition to application, there are also **harmonic composition** operations.

$$\mathtt{X/Y} \quad \mathtt{Y/Z} \quad \rightarrow_{>\mathbf{B}} \quad \mathtt{X/Z} \tag{4}$$

$$\mathtt{Y\backslash Z} \quad \mathtt{X\backslash Y} \quad \rightarrow_{<\mathbf{B}} \quad \mathtt{X\backslash Z} \tag{5}$$

Operation (4) enables us to use types such as N/S[gu] for "propositional" nouns such as *dùil* 'expectation' or *dòchas* 'hope' so that they can combine with clauses that begin with the word *gu* 'that'.

1. http://universaldependencies.org/

2.2 Assigning types

The usual process for generating a categorial grammar bank, as exemplified for English (Hockenmaier & Steedman, 2007), and Chinese (Tse & Curran, 2010), is to take a pre-existing set of context-free grammar parse trees, to convert any non-binary nodes to binary node, and to assign a category to every node. For German, Hockenmaier Hockenmaier (2006) describes an analogous process based on the TIGER dependency treebank.

However, there being no treebanks for Scottish Gaelic, we need to take a different approach. The main resource for Scottish Gaelic is ARCOSG, which is a corpus of 76 texts from a variety of genres. These have been part-of-speech tagged by hand according to a tagging scheme described in Naismith & Lamb (2014). What we can do, therefore, is to build a categorial grammar in which each lexical entry contains a category that is assigned purely on the token and tag information for a given word in ARCOSG. This is similar to supertagging (Bangalore & Joshi, 1998), an approach which is usually the first step in CCG parsing, in which all of the possible CCG categories are applied to each word in the text and the CCG parser then attempts to find the best overall parse. The difference here is that we are doing this on the level of the original corpus itself, in order to generate a grammar.

The initial version of the mapping was based on the scheme in Batchelor (2014), which is itself largely based on Hockenmaier & Steedman (2007) with adjustments for VSO order in Gaelic. This was refined first by ensuring that there was complete coverage of all of the parts of speech in ARCOSG, and then that it was possible to parse the corpus itself. A summary is given in Table 1.

There are some subtleties which we shall discuss here. The ARCOSG tagset is based closely on the PAROLE tagset used by Uí Dhonnchadha (2009). (Lynn, 2016) describes in detail how the PAROLE tagset is not completely appropriate for her work in dependency grammar. Many of these are familiar topics in Celtic linguistics and are also relevant toe our categorial grammar treatment.

In ARCOSG the prepositional pronouns, for example *orm*, *ort* ("on me", "on you") are treated as pronouns whereas for verbal subcategorization they should be treated in the same way as prepositions. We treat transitive verbal nouns as `S[small]/N/N` and the aspectual particles *a'*, *ag*, *air*, *gu* and *ri*, which precede verbal nouns and are in most cases identical to prepositions, as type-changing particles. [2] *Airson* is tagged as a fossilized noun (*Nf*) in ARCOSG, whereas we treat it here as a preposition (`PP/N`). If a word in ARCOSG is in the "wrong" case according to the accepted grammar of Scottish Gaelic, then it will be tagged with the correct case and the part of speech marked with an asterisk. In these cases we disregard the asterisk and treat the word as a variant.

If we allow dashes and commas to act as noun-coordinators and noun-postmodifiers then we can handle apposition introduced by punctuation. More difficult are plural genitives, which are often identical to either the singular or plural nominative and may be tagged as such.

2. One longer-term reason for doing this is to make the semantics more transparent. First consider the verbal nouns as a whole :

— Intransitive verbs : `S[small]/N`: $f(e) \wedge agent(e, x)$

— Transitive verbs : `S[small]/N/N`: $f(e) \wedge agent(e, x) \wedge patient(e, y)$.

The particles that are unmarked for person, *a'/ag*, *gu*, *ri* and *air*, supply the aspect, hence *a' cluinntinn* ("hearing") gives us

$$progressive(e) \wedge hears'(e) \wedge agent(e, x) \wedge patient(e, y). \quad (6)$$

gam, *gad* and so forth supply not only the aspect but also the patient, hence *gad chluinntinn* ("hearing you") :

$$progressive(e) \wedge hears'(e) \wedge agent(e, x) \wedge patient(e, thu'). \quad (7)$$

ARCOSG	CCG	Comments	Example
Ap	S[adj]/N	predicative adjective	
Aps	(S[adj]/N)/N	second comparative	*feairrde*
Aq	N*N	attributive adjective	
Ar	N/*N	premodifying adjective	*droch*, *seann*
Av	N*N	past participle	
Cc	N*N/N, S*S/S	coordinators	*agus*, *ach*
Cs	S*S/S	subordinators	
Csw	S[gu]/N/N	*gur*	
D	N/*N	determiners	
Fq	S/*S	open quote	
all other *F*	S*S	punctuation	
Mc	N	cardinal numbers	
Mo	N/*N	ordinal numbers	
Nf	N	fossilized noun	
except *airson*	PP[airson]/N	preposition	
Nn-mn	N/*N	forename	
Nv	as verbs	verbal noun	
N...g	N*N	genitive noun	
N...v	S/S	vocative noun	*a Sheumais*
all other *N*	N	nouns	
Pn	N	numerical pronouns	*ceithir*
Pp	N	pronouns	*mi*, *mise*, *i*, *iad*
Pr	PP	personal prepositions	
Q	S[x]/S[y]	clause feature value changers	*cha*, *do*, *gu*
except *Q-s*	(S*S)/S[dep]	"if"	*nam*, *nan*
R	S*S	adverbs	
Sa	S[asp]/N/S[small]/N	aspect	*a'*, *air tighinn*
	S[asp]/N/S[inf]/n		*air a chumail*
Sap	S[asp]/S[small]/N	personal aspect	*gad*, *gam*
Sp	PP/N	prepositions	
T...n, *T...d*	N/*N	articles	
T...g	(N*N)/(N*N)	genitive articles	
Uf	N	fossilized noun	*dòcha*, *urrainn*
Ug	S[inf]\N/S[small]/N/N	agreement particle	
Uv	(S/*S)/(S/*S)	vocative particle	*a Sheumais*
V	varies	verbs	
W	varies	copula	
Xfe	N	foreign words	
Xsc	S/*S	marks a speaker	

TABLE 1 – The most important part-of-speech classes from ARCOSG and the types they map to in our categorial grammar treatment.

ARCOSG POS	Description	Procedure
Nv	verbal noun	see Table 3
all *W*	copula	*is*
*V*s*	past tense	delenite
Vm-1p	1p imperative	remove *-eamaid* or *-amaid*
Vm-2s	singular imperative	preserve
Vm-2p	plural imperative	remove *-ibh* or *-aibh*
V-h, *Vm-3*	conditional, 3p.imp.	delenite, remove *-eadh* or *-adh*
*V.*d*	dependent form	delenite
*V.*f*	future tense	remove *-idh* or *-aidh*
*V.*r*	relative	remove *-eas* or *-as*
V-s0	past impersonal	delenite, remove *-eadh* or *-adh*
V-p0	present impersonal	remove *-ear* or *-ar*

TABLE 2 – Operation of the lemmatizer on verbs. In each case the slenderized form of the suffix is given first.

For determiners, conjunctions and adjectives we use the non-associative, non-permutative slash $/_\star$ from multimodal combinatory categorial grammar (Baldridge & Kruijff, 2003). We ban forward-crossed composition, though this may prove to be unnecessary if we make full use of the multimodal slash repertoire.

2.3 Lemmatization

The ARCOSG tagset marks nouns and articles for number and case, verbs and prepositions and pronouns for person and number, and verbs for tense and whether they are the independent, dependent or relative form of the verb. These are incorporated as features ; for example the verb *thòisich* with the tag *V-p* gets the tense feature `pres`.

However, it does not mark them for transitivity or which prepositional phrases they subcategorize with. This is clearly beyond the scope of a POS tagger, especially one for a corpus of this size, and a full treatment requires a larger dictionary. For this we require a lemmatizer for verbs. We are not aware of any publications about a verb lemmatizer for Scottish Gaelic. Lemmatizers for Irish have previously been presented by Uí Dhonnchadha & Van Genabith (2005) and Měchura (2014). The lemmatizer requires the surface form of the verb and a part-of-speech tag, but Gaelic, while morphologically rich, is largely systematic and it mostly proceeds by delenition [3] where necessary and removing endings. [4] The procedure for this, which covers all of the grammatical categories for verbs found in ARCOSG, is listed in Table 2. The irregular verbs *bi*, *abair*, *beir*, *cluinn*, *dèan*, *faic*, *faigh*, *rach*, *ruig*, *thoir*, *thig* and all verbal nouns are treated separately, the irregular verbs by means of a lookup table and verbal nouns by deleniting where necessary and following the procedure in Table 3.

3. In contrast to the mutations in Welsh, Cornish and Breton, lenition in Irish and Scottish Gaelic is marked orthographically by inserting an *h* after the initial consonant.

4. The endings take different forms according to whether they follow a 'slender' consonant or a 'broad' consonant. These are marked in the orthography as follows : a slender consonant has the vowels *i* or *e* as neighbours ; a broad consonant has the vowels *a*, *o* or *u*. There are occasional exceptions, usually compound words such as *airson* and *rudeigin*, but they do not affect the algorithm.

Input	Output	Example
in list	return dictionary form	*tuiteam* → *tuit*
-sinn, *-s'*, *-tainn*	strip ending	*creidsinn* → *creid*
-eachadh	*-ich*	*cruinneachadh* → *cruinnich*
-achadh	*-aich*	*sgioblachadh* → *sgioblaich*
-gladh	*-gail*	*fosgladh* → *fosgail*
-eadh	strip ending	*tilleadh* → *till*
-adh	strip ending	*glanadh* → *glan*
otherwise	preserve	*ruith* → *ruith*

TABLE 3 – Operation of the lemmatizer on verbal nouns.

#	Rule	Explanation
1	$\texttt{PP} \rightarrow \texttt{N}\backslash_\star\texttt{N}$	PPs modifying noun phrases
2	$\texttt{PP} \rightarrow \texttt{S}\backslash_\star\texttt{S}$	postposed PPs modifying clauses
3	$\texttt{S}\backslash_\star\texttt{S} \rightarrow \texttt{S}/_\star\texttt{S}$	preposed adverbials
4	$\texttt{N[place]} \rightarrow \texttt{N}\backslash_\star\texttt{N}$	placenames used attributively
5	$\texttt{S[small vowel]}/\texttt{N} \rightarrow \texttt{S[inf]}$	intransitive verbal nouns without agreement particle
6	$\texttt{S[small vowel]}/\texttt{N} \rightarrow \texttt{S[inf]}\backslash\texttt{N}$	transitive verbal nouns without agreement particle
7	$\texttt{PP}/\texttt{N} \rightarrow \texttt{PP}/\texttt{S[int]}/\texttt{PP}$	relative clauses in PPs

TABLE 4 – Type-changing rules.

2.4 Unary rules

The binary combinators listed above are insufficient for practical CCG parsing. We have a set of systematic **type-changing rules**, listed in Table 4, which serve to reduce the size of the lexicon by minimizing the number of types. Rules 1 and 2 convert the atomic category PP into a modifier of both NPs and clauses. Rule 3 allows clausal modifiers, be they PPs or adverbial phrases, to go before the clause. Rule 4 deals with placenames, which are not marked for case, when they are used attributively.

Passive-type constructions using the verb *rach* 'to go' are common in the news genre. For example, *Chaidh barrachd dhaoine a mharbhadh* 'more people were killed', where *chaidh* is the past tense of *rach* 'go', *marbhadh* is the verbal noun 'killing', and *a* is an "agreement particle". To handle the case where the verbal noun begins with a consonant, we use **type-raising rule** 1 in Table 5 and give the agreement particle the type $\texttt{S[inf]}\backslash\texttt{N}/\texttt{S[small]}/\texttt{N}/\texttt{N}$.

$$\dfrac{\dfrac{\textit{chaidh}}{\texttt{S[dcl]}/\texttt{S[inf]}} \quad \dfrac{\dfrac{\dfrac{\textit{barrachd dhaoine}}{\texttt{N}}}{\texttt{S}/\texttt{S}\backslash\texttt{N}}\,{>}\mathbf{T} \quad \dfrac{\dfrac{\textit{a}}{\texttt{S[int]}\backslash\texttt{N}/\texttt{S[small]}/\texttt{N}/\texttt{N}} \quad \dfrac{\textit{mharbhadh}}{\texttt{S[small]}/\texttt{N}/\texttt{N}}}{\texttt{S[inf]}\backslash\texttt{N}}\,{>}}{\texttt{S[inf]}}\,{>}}{\texttt{S[dcl]}}\,{>} \tag{8}$$

Rules 5 and 6 in Table 4 handle the analogous case where the verbal noun begins with a vowel, so there is no agreement particle. Rule 7 deals with prepositional relative clauses, for example *leis a bheil an taigh* 'who owned the house'. Unlike conventional relative clauses, which take a declarative

#	Rule	Explanation
1	$\mathtt{N} \rightarrow_{>\mathbf{T}} \mathtt{S/S\backslash N}$	For the *rach* passive
2	$\mathtt{PP} \rightarrow_{<\mathbf{T}} \mathtt{S\backslash S/N}$	For relative clauses
3	$\mathtt{S[adj]/N} \rightarrow_{<\mathbf{T}} \mathtt{S\backslash S/S[adj]/N}$	For relative clauses

TABLE 5 – Type-raising rules

or relative future form of the verb after the relativiser *a*, these take the interrogative form of the verb, for example *a bheil* 'is ?'. We then use forward composition (eqn. 4)

The other type-raising rules in Table 5 enable us to form relative clauses with *a*. To take the example NP *an gille a tha bochd* 'the boy who is ill' :

$$\dfrac{\dfrac{\textit{an gille}}{\mathtt{N}} \quad \dfrac{\dfrac{\textit{a}}{\mathtt{N\backslash N/S/N}} \quad \dfrac{\dfrac{\textit{tha}}{\mathtt{S[dcl]/(S[adj]/N)/N}} \quad \dfrac{\dfrac{\textit{bochd}}{\mathtt{S[adj]/N}}}{\mathtt{S\backslash S/S[adj]/N}}<\mathbf{T}}{\mathtt{S[dcl]/N}}<\mathbf{B}_{\times}}{\mathtt{N\backslash N}}>}{\mathtt{N}}> \tag{9}$$

we use the additional backward crossed composition operation

$$\mathtt{Y\backslash Z} \quad \mathtt{X\backslash Y} \rightarrow_{<\mathbf{B}_{\times}} \mathtt{X/Z}. \tag{10}$$

in addition to type-raising rule 3.

3 In practice

3.1 Pre-processing

The POS-tagged text in ARCOSG treats multiword expressions such as toponyms *e.g. Beinn na Faoghla* 'Benbecula', multiword prepositions such as *an aghaidh* 'against' and fixed phrases such as *Gu sealladh ort !* 'Heaven preserve you !' as single tokens. For simplicity we apply a preprocessing step to ARCOSG where lexical entries containing spaces have them replaced with underscores in place of spaces, thus `ann_an` instead of `ann an`.

3.2 Parsing

Out of the available CCG parsers, we chose OpenCCG, a categorial grammar parsing and realization toolkit, [5] to parse Gaelic text taken from ARCOSG. The key strengths of OpenCCG for rapid prototyping and development of categorial grammars are that it has an interactive mode and a transparent syntax (dotccg format (Baldridge *et al.*, 2007)) for specifying grammars, and an efficient chart parser. One weakness is that by default it doesn't handle out-of-vocabulary text. We also considered the CCG parser in the NLTK ; however the version in NLTK 3.1 (October 2015) doesn't

5. `http://openccg.sourceforge.net/`

support features, such as the type of clause, gender or tense, and as such it is not usable for our purposes. Otherwise the excellent and well-established C&C parser (Curran *et al.*, 2007) is too closely entangled with the underlying CCGbank to be used for this sort of development work.

For the word *ann* 'in it', 'there', 'in him', the OpenCCG parser produces seven parses for which we list here the final result without the full derivations :

```
Parse 1: pp/n
Parse 2: pp
Parse 3: pp<1>/(s{clause=int}/pp<1>)
Parse 4: n<2>\n<2>
Parse 5: s<3>\s<3>
Parse 6: s<6>\@i(s<6>/@ipp)
Parse 7: s<11>/s<11>
```

The first parse comes from the phrase *ann a bhith* 'in which... is', which appears several times in the corpus, and the others are from the type-raising and type-changing rules we have discussed before. Clearly there is no one correct parse for a single word. The correct full derivation (out of six found by OpenCCG for our grammar) for *tha i fliuch* 'it is wet' (used usually of the weather) is :

```
(lex)  tha :- s{clause=dcl, phon=cons, tense=pres}/(s{clause=bi_arg}/n)/n
(lex)  i :- n{ont=pron}
(>)    tha i :- s{clause=dcl, phon=cons, tense=pres}/(s{clause=bi_arg}/n)
(lex)  fliuch :- s{clause=adj}/n
(>)    tha i fliuch :- s{clause=dcl, phon=cons, tense=pres}
```

In the grammar `bi_arg` stands for a `clause` feature value of either `asp` or `adj`, indicating which sorts of clause can be an argument for the verb *bi*.

For development purposes we use the interactive parser `tccg`.

3.3 Towards evaluation

Clark and Hockenmaier (Clark & Hockenmaier, 2002), in the context of CCGbank, compare methods for evaluating the performance of a CCG system. These involve the CCG system being able to output dependencies, whether they be the Universal Dependencies mentioned earlier or ones obtained directly from the steps in a CCG derivation, and comparing those dependencies to a gold standard. This allows for a systematic check of not only whether the correct parts of speech have been assigned, but also, for example, subjects, objects and PP attachment. In contrast, the default testing framework for OpenCCG involves counting the number of parses for a given sentence and comparing it with the expected number. This is useful for pedagogical reasons, but knowing that the correct number of parses has been returned for a sentence is less helpful than knowing how much of it was assigned correctly. A further difficulty is that parsing a sentence in CCG is equivalent to deriving a proof, and if that proof fails for whatever reason, then there is no way of recovering the partial parses to award partial credit to the parser. Hence the program both flatters successful parses and unduly penalizes unsuccessful ones, and so we have not been able to provide a sensible evaluation of the parsing performance. Lastly, because the CCG parser doesn't handle out-of-vocabulary text, we cannot have separate training and testing data.

We can, however, give a qualitative account of the situations where more work is needed. Our examination has focussed on the section of ARCOSG consisting of news scripts from Radio nan Gàidheal, a genre which has been described in detail by Lamb (1999). This section has 11354 tokens and is about 13% of the total 87038. It is amenable to automatic sentence-splitting and does not contain interjections or direct speech, which make parsing harder. The grammar works accurately on simple clauses based on transitive and intransitive verbs, relative clauses and passives formed with the verb *rach*.

Apposition, despite the measures above to deal with punctuation, is still not fully handled. *Rùnaire Èirinn a Tuath Mo Mowlam* 'Northern Ireland Secretary Mo Mowlam', for example, doesn't parse. Similarly if there is a sequence of words tagged as 'foreign', which are treated as nouns for simplicity, then the whole parse will fail. Sequences of nominative nouns also occur in temporal and spatial expressions and chains of possession where only the last noun is grammatically marked as genitive.

Cosubordination, a sort of coordination where the coordinated clause can express, among other things, reason, *dh'fhalbh Alasdair agus i 'na suain*—"Alasdair left because she was fast asleep" or time, is, contrary to initial suspicions, attested in the news subcorpus. *Chaidh bratach Bhreatainn a thoirt a-nuas ann an seirbheis taobh muigh an taighe, 's an Last Post ga chluiche* 'The British flag was taken down in a service outside the house as the Last Post was played' exemplifies this. The conjunction *'s* 'and' joins a *rach* passive to a non-constituent. We anticipate that it should be possible to handle this elegantly in CCG using type-raising rules such as we have seen previously, but this is future work.

4 Conclusions and future work

We have produced a medium-coverage categorial grammar of Scottish Gaelic using all of the Annotated Reference Corpus of Scottish Gaelic and where every type is assigned based solely on the token value and its POS tag. The key difficulty has been in providing a convincing evaluation of the foregoing. To this end we need firstly a gold standard corpus of dependencies, of the sort we previously presented in Batchelor (2014) which can be used to evaluate successful parses. The other key requirement is to migrate to a statistical approach, ensuring that there are some successful parses to evaluate. A conventional CCG workflow involves a statistical supertagging stage prior to parsing. Supertagging is similar to POS-tagging but typically uses a larger tagset. Whereas the focus in the ARCOSG POS set is on morphological features, supertags can indicate subcategorization, whether a PP modifies a noun or a clause, or whether a comma is appositive or not, among other functions. The C&C supertagger for English uses around 500 supertags as opposed to 50 Penn Treebank POS tags. As such, the problems described in Lamb & Danso (2014) with ordinary POS-tagging in Scottish Gaelic will be harder for supertagging, but it seems plausible that because of different focus, the number of supertags required for Gaelic will be similar to that for English. A working solution to this would also handle the problems of out-of-vocabulary text and foreign words described in the section above. The code, a small set of Python scripts is available at `https://github.com/colinbatchelor/gdbank/`.

Acknowledgements

Many thanks to William Lamb for a preview copy of ARCOSG, to Teresa Lynn for a critical reading of the manuscript, and to the anonymous referees for their very helpful suggestions.

Références

BALDRIDGE J., CHATTERJEE S., PALMER A. & WING B. (2007). DotCCG and VisCCG : Wiki and Programming Paradigms for Improved Grammar Engineering with OpenCCG. In T. H. KING & E. BENDER, Eds., *Proceedings of the GEAF 2007 Workshop* : CSLI Publications, Stanford, CA.

BALDRIDGE J. & KRUIJFF G.-J. M. (2003). Multi-Modal Combinatory Categorial Grammar. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary.

BANGALORE S. & JOSHI A. (1998). Supertagging : an approach to almost parsing. *Computational Linguistics*, **22**, 1–29.

BATCHELOR C. (2014). gdbank : The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, p. 60–65, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

CLARK S. & HOCKENMAIER J. (2002). Evaluating a Wide-Coverage CCG Parser. In *Proceedings of the LREC 2002 Beyond Parseval Workshop*, p. 60–66, Las Palmas, Spain.

CURRAN J., CLARK S. & BOS J. (2007). Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 33–36, Prague, Czech Republic : Association for Computational Linguistics.

HOCKENMAIER J. (2006). Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, p. 505–512, Sydney, Australia.

HOCKENMAIER J. & STEEDMAN M. (2007). CCGBank : A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, **33**, 355–356.

LAMB W. (1999). A diachronic account of Gaelic news-speak : The development and expansion of a register. *Scottish Gaelic Studies*, **XIX**, 141–171.

LAMB W. (2003). *Scottish Gaelic, 2nd edn*. Munich, Germany : Lincom Europa.

LAMB W. (2008). *Scottish Gaelic Speech and Writing : Register Variation in an Endangered Language*. Belfast : Cló Ollscoil na Banríona.

LAMB W., ARBUTHNOT S., NAISMITH S. & DANSO S. (2016). *Annotated Reference Corpus of Scottish Gaelic (ARCOSG), 1997–2016 [dataset]*. Rapport interne, University of Edinburgh ; School of Literatures, Languages and Cultures ; Celtic and Scottish Studies. `http://dx.doi.org/10.7488/ds/1411`.

LAMB W. & DANSO S. (2014). Developing an Automatic Part-of-Speech Tagger for Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, p. 1–5, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

LYNN T. (2016). *Irish Dependency Treebanking and Parsing*. PhD thesis, Dublin City University and Macquarie University.

LYNN T., SCANNELL K. & MAGUIRE E. (2015). Minority Language Twitter : Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 1–8, Beijing, China : Association for Computational Linguistics.

MĚCHURA M. B. (2014). Irish National Morphology Database : a high-accuracy open-source dataset of Irish words. In *Proceedings of the First Celtic Language Technology Workshop*, p. 50–54, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

NAISMITH S. & LAMB W. (2014). Scottish Gaelic Part-of-Speech Annotation Guidelines. Celtic and Scottish Studies, University of Edinburgh.

NIVRE J. *et al.* (2015). Universal Dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

SCANNELL K. (2014). Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, p. 33–40, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

STEEDMAN M. & BALDRIDGE J. (2003). *Combinatory Categorial Grammar*. Rapport interne, University of Edinburgh. `http://homepages.inf.ed.ac.uk/steedman/papers/ccg/SteedmanBaldridgeNTSyntax.pdf`.

TSE D. & CURRAN J. R. (2010). Chinese CCGbank : extracting CCG derivations from the Penn Chinese Treebank. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 1083–1091, Beijing, China.

Uí DHONNCHADHA E. (2009). *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. PhD thesis, Dublin City University.

Uí DHONNCHADHA E. & VAN GENABITH J. (2005). Scaling an Irish FST morphology engine for use on unrestricted text. In *Fifth International Workshop on Finite-State Methods in Natural Language Processing*, Helsinki.

UNIVERSITY OF GLASGOW (2016). Corpas na Gàidhlig, Digital Archive of Scottish Gaelic (DASG), `http://www.dasg.ac.uk/corpus/`. accessed 15 April 2016.

Building a Dictionary-Based Lemmatizer for Old Irish

Oksana Dereza
School of Linguistics, National Research University
«Higher School of Economics», Moscow, Russia
oksana.dereza@gmail.com

ABSTRACT

This paper explores the problem of developing NLP tools for morphologically rich and orthographically inconsistent classical languages. It is a case study of building a lemmatizer for Old Irish using only a dictionary and an unlabeled corpus as sources of data. At the current stage, the lemmatizer shows 76.31% average recall score on a corpus of ca. 100,000 tokens and is able to predict lemmas for out-of-vocabulary words. However, as it is the work in progress, the lemmatizer lacks some functionality such as disambiguation. There is no gold standard to measure accuracy yet either.

RÉSUMÉ

Le développement d'un programme de lemmatisation pour le vieil irlandais

Cet article vise à présenter le développement d'un logiciel de traitement automatique des langues anciennes qui sont caractérisées par une morphologie riche et par une orthographie irrégulière. Dans ce cas, il s'agit d'un outil de lemmatisation des textes en vieil irlandais créé uniquement à partir du dictionnaire et du corpus de textes non annotés. A ce stade, le rappel moyen du programme de lemmatisation est 76.31% sur un corpus d'environ 100,000 de jetons. Le programme peut prédire les lemmes pour des mots qui n'apparaissent pas dans son vocabulaire. Néanmoins, comme c'est un travail en cours, il manque encore de certaines fonctions comme la désambiguïsation sémantique. En plus, il est encore impossible de mesurer l'exactitide, parce qu'il n'y a pas de corpus annoté qui puisse servir de référence.

MOTS-CLÉS : lemmatisation, lemme, vieil irlandais, moyen irlandais, distance de Damerau–Levenshtein, données non annotées, analyse automatique de la morphologie.

KEYWORDS: lemmatisation, lemma, Old Irish, Middle Irish, Damerau-Levenshtein distance, unlabelled data, automatic morphological analysis.

1 Introduction

The interest to automatic morphological analysis of classical languages arose at the very start of computational linguistics, but still this field is underrepresented in comparison to other NLP tasks. The majority of the related works are quite old and cover only the most popular classical languages, such as Latin (Marinone, 1990), (Passarotti, 2004), ancient Greek (Packard, 1973) and Sanskrit (Verboom, 1988), (Huet, 2003). Most of the functionality of modern NLP tools for classical languages, such as CLTK,[1] is also confined to Latin and ancient Greek. However, there are many other well-documented

[1] http://docs.cltk.org/en/latest/

classical languages where a statistical-based approach to linguistic analysis may prove useful. Such languages are usually morphologically rich and orthographically inconsistent, which complicates automatic processing and requires NLP instruments to be language-specific; the lack of annotated corpora is an even bigger problem. This paper is a case study of building a lemmatizer for Old Irish using only a dictionary and an unlabeled corpus as sources of data.

2 Approach and Data

In Celtic languages, there are two ways to encode morphological information in a word form, which often occur together: initial mutations that come in the beginning of a word, and flections that come in the end. Moreover, in Old Irish some words can be incorporated into a verb between the preverb and the root: cf. *caraid* 'he / she / it loves' and *rob-car-si* 'she has loved you', where *ro-* is a perfective particle, *-b-* is an infixed pronoun for 2nd person plural object, and *-si* is an emphatic suffixed pronoun 3rd person singular feminine. The presence of a preverb with dependent forms triggers a shift in stress, which causes complex morphophonological changes and often produces a number of very differently looking forms in a verbal paradigm, particularly in the case of compound verbs, cf. *do-beir* 'gives, brings' and *ní tab(a)ir* 'does not give, bring'. This morphophonological complexity compounded by the many non-transparent features of Old Irish orthography makes the traditional dictionary approach to lemmatization with hard-coded lists of possible pseudo-suffixes and rules of their treatment less suitable for Old Irish than for other languages. A more reliable way for a start is building a full form dictionary where every word form corresponds to a lemma; an electronic edition of the Dictionary of the Irish Language[2] proves an excellent source of data for this purpose. DIL is a comprehensive historical dictionary of Irish, which covers Old and Middle Irish periods. The latest version of its electronic edition is organized in the following way: each of the 43,345 webpages is a single entry, which contains a headword, a list of possible forms and a 'main body' with translations and examples of use.

Given the aforementioned features of Old Irish, the task of building a dictionary for a lemmatizer reduces to parsing the DIL and extracting all possible forms for each lemma. However, it is not as simple as it seems. First, the list of forms cited in DIL is incomplete; it covers only about 36% of unique words in the working corpus. Second, some of the forms in DIL are contracted; for example, the list of forms for *carpat* 'chariot' looks like *cairpthiu, -thib, -tiu, -tib*. Words can be abbreviated in many different ways, which is a consequence of the fact that there were many scholars who contributed to the DIL throughout 1913-1976, and each of them used his own notation, as preserved in the digital edition.[3] Thus, one either has to drop contracted forms altogether or derive a number of rules to restore them. Third, the markup and punctuation are also inconsistent, which causes various technical problems.

The working corpus of ca. 100,000 tokens[4] was compiled from Ulster cycle sagas published on UCC CELT website.[5] It includes 24 thematically related pieces of narrative that differ in length and orthography. In the future, the corpus will be extended with texts of other forms and genres for better

[2] http://dil.ie

[3] See (Toner *et al.*, 2007) and http://dil.ie/about for details.

[4] I used a pre-trained Punkt tokenizer for English provided in NLTK, a Python library for natural language processing, which is the easiest, but obviously not the best solution for Old Irish. Building an Old Irish tokenizer is a separate important task to be solved in the future research.

[5] www.ucc.ie/celt/publishd.html

representation.

3 Algorithm and Implementation

The current version of the program consists of a dictionary compiler, a lemmatizer, and a lemma predictor for out-of-vocabulary words. The source code in Python 3 is available on GitHub.[6]

3.1 Dictionary Compiler

The dictionary compiler is a separate script that parses the DIL, extracts the list of forms for each lemma, restores contracted forms and builds a 'form : lemma' dictionary which is then dumped in JSON format for future use. It copes well with various contracted, syncopated and bracketed forms, e.g. *carat(r)as* for *caratas* and *caratras*; *carthain, -ana* for *carthana*; *cairpthiu, -tib* for *cairptib*; *caibidil, -lech* for *caibidlech*. However, the end user does not have to compile a dictionary from scratch, as its latest version always goes together with the lemmatizer.

3.2 Lemmatizer

The lemmatizer takes a file in plain text as input, cleans out punctuation and other non-word characters, and then analyses words one by one. Every word is first demutated (i.e. the changes at the beginning of the word are eliminated) and then looked up in the dictionary. The lemmatizer returns a lemma for each known word and a demutated form for each unknown word. In addition to that, the unlemmatized forms are stored in a special list. There is no word sense disambiguation for the moment, which means that if two or more different lemmas have identical forms, we cannot say for sure which lemma should be chosen for a particular instance of a homonymous form. There are two options for such cases in the current version of the lemmatizer: either return a list of all possible lemmas or choose the lemma with the highest probability. Lemma probability here equals the sum of probabilities of forms belonging to a lemma, and word form probability is a frequency count computed for each word in the corpus.

The lemmatizer has several methods, the major ones being the following:

- lemmatize a text;
- show unlemmatized words;
- evaluate performance;
- update a dictionary with a preformatted file containing new lemmas and forms.

3.3 Lemma predictor

The last module predicts lemmas for unknown words with the help of Damerau-Levenshtein distance. For every unknown word, the program generates all possible strings on edit distance 1 and 2, checks

[6] https://github.com/ancatmara/old_irish_lemmatizer

them up in the dictionary and adds those that prove to be real words to the candidate list. Then the candidates are filtered by the first character: if the unknown word starts with a vowel, the candidate should also start with a vowel, and if the unknown word starts with a consonant, the candidate should start with the same consonant. Those parameters were chosen empirically as they yield the best results, i.e. the highest percentage of correctly predicted lemmas. Finally, the lemma of the candidate that has the highest probability is taken as a lemma for the unknown word. Although this algorithm gives very promising results, it is not a default option for out-of-vocabulary words in the lemmatizer yet. There are two major reasons for this: first, the dictionary is still rather small (there are 26,160 unique tokens in the working corpus and 16,742 of them are non-dictionary forms), and second, there is no gold standard to evaluate accuracy. At the current stage, the triplets of unknown words, best candidates and their lemmas are written into an output file that requires manual revision, after which it can be uploaded as an update to the default dictionary.

The rule-based approach to lemma prediction was chosen over machine learning due to the scarcity of available data. For the moment, there are only 79,140 different forms in the 'form : lemma' dictionary compiled from the DIL, and ca. 100,000 tokens in the unlabeled Ulster cycle corpus, which is not enough for training a classifier that would be able to predict tens of different lemma classes.

4 Evaluation

As long as out-of-vocabulary words are left unlemmatized and homonymy is not taken into account, recall seems to be the most important metric for evaluating the lemmatizer's performance as it indicates the percent of forms that the program is able to process. When the recall score exceeds at least the 85-90% threshold, it will be reasonable to make a gold standard and to switch to accuracy, which is more suitable for evaluating disambiguation and unknown word treatment algorithms, because it shows the ratio of correctly predicted instances to the total number of instances.

I conducted three minor experiments to evaluate the lemmatizer's performance. First, I ran it on the whole working corpus with a default dictionary that consisted only of forms and lemmas retrieved from the DIL. This gave the average recall score of 74.7%, with the worst result of 62.5% for *Síaburcharpat Con Culainn* and the best result of 84.8% for *De Gabáil in t-Sída*.

Then I chose three random texts of different length (1,930 tokens in total, where 1,051 are unique), ran the lemmatizer with the default dictionary, manually analysed proposed lemmas for unknown words and added correctly guessed ones to the dictionary. The lemma predictor found candidates for 269 of 368 unique unlemmatized words, and 163 of them, or 61%, were correct. After that, I re-ran the lemmatizer with an updated dictionary, and the average recall score increased by ca. 10%. The results of the experiment are shown in Table 1.

Text	Tokens	Recall before update	Recall after update
Aided Óenfir Aífe	1,093	79.69%	89.75%
Aided Conrói maic Dáiri	738	78.35%	89.04%
Compert Conchobuir	99	78.79%	87.88%
Overall	**1,930**	**78.94%**	**88.89%**

Table 1: Updating the dictionary with predicted lemmas

Finally, I ran the lemmatizer on the whole corpus (99,717 tokens, 26,160 unique) again, but with an updated dictionary. Although I added only 163 forms derived from only 3 texts, the recall score increased for almost every text in the corpus, the average now being 76.3%, which is 1.6% higher than before. The results also show that the score does not correlate with a text's length, but depends on the period when it was created. It is not surprising that later texts are lemmatized worse than texts written in more or less classical orthography, because the DIL contains a lot more Old Irish forms and spellings than Middle and Early Modern Irish ones. The overall results are given in Table 2.

Text	Tokens	Recall before update	Recall after update
Aided Óenfir Aífe	1,093	79.69%	89.75%
Aided Conrói maic Dáiri	738	78.35%	89.04%
Aislinge Óenguso	1,267	78.61%	79.64%
Compert Conchobuir	99	78.79%	87.88%
Compert Con Culainn	1,048	83.30%	84.73%
De Chopur in dá Muccida	868	72.58%	73.27%
Do Faillsigud Tána Bó Cúailnge	326	78.53%	78.53%
Fled Bricrenn	9,006	65.33%	65.75%
Do Fogluim Chonculainn	5,486	65.33%	65.35%
De Gabáil in t-Sída	231	84.85%	84.85%
Immacallam in Dá Thúarad	637	80.69%	80.69%
Fochond loingse Fergusa meic Roig	314	75.48%	75.48%
Longes mac n-Uislenn	2,352	67.01%	67.09%
Scéla Mucce Meic Dathó	2,716	75.22%	75.85%
Mesca Ulad	7,678	76.93%	77.53%
Noínden Ulad	152	65.13%	65.13%
Serglige Con Culainn	5,943	80.67%	81.24%
Síaburcharpat Con Culainn	1,505	62.52%	62.72%
Táin Bó Fráich	3,623	80.13%	81.07%
Talland Etair	2,817	79.91%	80.90%
Táin Bó Cúailnge (Recension I)	35,744	78.78%	79.51%
Tochmarc Emire	9,576	64.00%	64.55%
Tochmarch Ferbe	6,424	71.16%	77.76%
Togail tSitha Truim	84	63.10%	63.10%
Overall	**99,717**	**74.67%**	**76.31%**

Table 2: Updating the dictionary with predicted lemmas: the whole corpus

5 Conclusion

As this is a work in progress, there are still many tasks to tackle and problems to solve. At the last run, the lemmatizer predicted lemmas for 12,156 unknown words, which is too many to filter manually. Therefore, the first priority is developing a dictionary updater that would be able to extend the dictionary with little or no human supervision. The next important task is to compile a gold standard to be able to measure accuracy and evaluate unknown word treatment and disambiguation.

The third biggest problem is disambiguation itself, which most probably requires a statistical approach. All in all, the lemmatizer is ready-to-use and shows promising results even at the current stage.

References

HUET G. (2003). Towards computational processing of sanskrit. In *International Conference on Natural Language Processing (ICON)*.

MARINONE N. (1990). A project for latin lexicography: 1. automatic lemmatization and word-list. *Computers and the Humanities*, **24**(5-6), 417–420.

PACKARD D. W. (1973). Computer-assisted morphological analysis of ancient greek.

PASSAROTTI M. C. (2004). Development and perspectives of the latin morphological analyser lemlat. *Linguistica computazionale*, **20**(A), 397–414.

TONER G., FOMIN M., BONDARENKO G. & TORMA T. (2007). : Royal Irish Academy.

VERBOOM A. (1988). Towards a sanskrit wordparser. *Literary and Linguistic Computing*, **3**(1), 40–44.

CALLIPSO – CALL for Irish for Parents Students and Others

Monica Ward
Dublin City University, Dublin, Ireland
monica.ward@dcu.ie

RESUME

L'irlandais est une langue complexe et opaque qui présente des difficultés pour les apprenants car une compréhension approfondie du système orthographique est nécessaire pour pouvoir lire et prononcer les mots correctement. Cet apprentissage de la langue peut se faire en autonomie grâce à des livres ou à des ressources d'apprentissage assisté par ordinateur (EAO) ou bien avec un enseignant. Toutefois, les règles de prononciation de l'irlandais sont très difficiles à comprendre pour les non-linguistes, y compris pour les enseignants. De plus, il existe très peu de ressources pédagogiques pour l'EAO de l'irlandais qui expliquent clairement les règles de prononciation de la langue. Cet article présente le système CALLIPSO 1 (CALL for Irish for Parents Students and Others) pour l'enseignement et l'apprentissage de la logique du système orthographique irlandais. CALLIPSO est un système modulable et évolutif qui s'adapte facilement à d'autres langues.

ABSTRACT

CALLIPSO – CALL for Irish for Parents Students and Others

Irish is an orthographically deep (opaque) language and presents difficulties for learners. There is a need for learners to understand the logic of the orthographical system in order to help them to read and pronounce words correctly. In order for learners to get this knowledge they either have to learn it via a teacher, a book or Computer Assisted Language Learning (CALL) resource. However, there are problems in this regard as teachers may not know the rules and information on pronunciation are often hard for the non-linguist to understand. There are very few CALL resources for Irish pronunciation that focus on explaining the rules in an accessible manner. This paper provides an overview of the CALLIPSO[1] (CALL for Irish for Parents Students and Others) system for teaching and learning the logic of the Irish orthographical system. CALLIPSO is modular and could be adapted for other languages.

MOTS-CLÉS: Enseignement Assisté par Ordinateur (EAO), irlandais, prononciation
KEYWORDS: CALL, Irish, pronunciation

[1] Available at : callipso.computing.dcu.ie

1 Introduction

Irish is an orthographically deep (opaque) language and learners need to be able to understand the pronunciation rules in order to be able to read and speak correctly. The fact that it has a deep orthography means that it presents difficulties for learners. They cannot make educated guesses as to how a word should be pronounced. There is a need for learners to understand the logic of the orthographical system in a way that they can easily understand. In order for learners to get this knowledge they either have to learn it via a teacher in a classroom setting, a book (perhaps for adult learners) or some Computer Assisted Language Learning (CALL) resource. However, there are problems in this regard. Many teachers themselves do not know the rules of Irish orthography and therefore cannot teach it to their students. They may not have been made aware of the rules themselves. This might be the case for primary school teachers who cover many subjects and are not usually Irish language specialists. Unless learners have a good knowledge of linguistics, it is often difficult to understand the correct pronunciation of Irish words just by reading the International Phonetic Alphabetic (IPA, 1999) phonetic translation of the words. There are some books that aim to show the pronunciation of words using English approximations, but these books tend to focus on explaining how to pronounce certain words and not the overall logic of the orthographical system. In terms of CALL resources for Irish pronunciation that focus on explaining the rules in an accessible manner, there are very few resources available. This paper provides an overview of the CALLIPSO (CALL for Irish for Parents Students and Others) system for teaching and learning the logic of the Irish orthographical system. CALLIPSO is designed in a modular fashion and to be language independent so that it could be adapted for other languages. The system is aimed at a broad spectrum of learners including parents, teachers and Irish language learners themselves.

1.1 Background

The term Computer Assisted Language Learning (CALL) covers all aspects of the use of computers in the language learning process. The degree of difficulty in developing different types of resources varies greatly. Using generic software to develop static resources is quite easy, while developing sophisticated resources that use Natural Language Processing (NLP) and Intelligent Tutor Systems (ITS) techniques is very complex and complicated. The fact that the effort involved in developing one hour of instruction (~ 50 – 100 hours of development time, with the use of authoring tools Aleven et al., 2009) is obviously an important factor. Pirolli and Kairam (2013) note that this may be worthwhile for widely deployed course (e.g. mathematics related material), but may not be feasible for other domains and contexts. The resources (financial, technical and time) required to design and develop such systems and the technical challenges that need to be addressed mean that it is very difficult to build an ICALL system that would be suitable for a range of learners in the real-world. Holland et al., (2013) provide a good recent overview of the field. Irish is a Minority Language (ML) and like other minority languages, there are limited good-quality, accurate CALL resources available for students. There are some Natural Language Processing (NLP) resources available for Irish, but they were not specifically designed for language learners (Uí Dhonnchadha, 2002; Scannell, 2014; Abair, 2016).

Pedagogical design decision should be a key part of the design process of CALL resources. This includes what to teach and how to teach the material. The field of pedagogy and language pedagogy in particular, is well researched and there are a wide variety of pedagogical approaches that can be incorporated into CALL resources. CALL designers may focus on a particular language skill (e.g. listening or writing) or a particular language level. Consideration is usually given into how to teach a

particular language construct, and perhaps, mechanisms to test and evaluate a learner's progress. These are important considerations.

1.2 Motivation

Most CALL resources for students concentrate on the needs of the immediate learners. However, many younger learners often receive help and guidance from their parents as part of the learning process. This is particularly true at primary school level and specifically when learners are doing their homework. Parents may be asked to check that their children know how to spell words or to check their reading of a piece of text. This is usually straightforward for parents – their literacy levels are usually higher than primary school children. However, problems can arise when parents are asked to check their children's ability in a language that the parents may not be competent and/or confident in or may not even speak the language.

This is a situation which arises in Ireland. Irish is a compulsory subject (with some exceptions) in schools in Ireland, yet it is the spoken as a first language by only a tiny minority of the population. Motivation is one of the key factors in determining learning success. Gardner and Lalonde (1985) noted that intrinsic and instrumental motivation are important factors in the learning process, while Dörnyei and Csizér (1998) state that learner level and learning situation level are also important components in L2 motivation. In the case of Irish, there is no real need to be able to speak Irish. Apart from the fact that only about 10,000 people speak it as their main language of communication on a daily basis, all native Irish speakers are fluent in English, and there are very few situations where it would be absolutely essential to be able to speak Irish to communicate with someone. Thus, for Irish, there is no real communicative need to learn the language (Watson, 2008). There are several socio-cultural reasons for learning the language e.g. heritage reasons such as fearing the loss of cultural identity without the language (Darmody and Daly, 2015), but these have less of an impact on learners than if a real communicative need existed. There are instrumental reasons for learning the language – principally, the need to do well in state exams in Irish – but many learners aim to learn just enough, without any high aspirations to master the language. The situation for parents is complex. On the one hand, they would like their children to learning the language, but on the other, they bemoan the time spent learning the language, the difficulty of the language, and their own lack of ability in the language. They tend to have a Machiavellian attitude towards the language. They just want their children to learn enough to not struggle too much with the language and to learn enough pass the state examinations (Darmody and Daly, 2015). Also, Hickey and Stenson (2011) note that the teaching methods can aslo have an influence on motivation to learn Irish.

1.3 Different Learner Groups

There are several personas or learner groups to consider in the context of Irish. The most common persona (false beginners) are parents who has learnt Irish in school, but may have forgotten it, never really mastered it or lack confidence in the language. The next persona are (novices) are parents that have never studied the language, as they immigrated to Ireland as an adult or did not study in either primary or secondary school in Ireland. The third persona (intermediates) consists of parents who have studied Irishand generally have some ability in the language, but may not sufficient competence to help their children. CALL resources for these learners must be tailored to their differing needs. A CALL resource should provide basic information for novices, refresher information for false beginners and access to more detailed information for intermediates. An interesting feature about

parents and their Irish language capability is that parents may often underestimate their ability. MacIntyre et al. (1997) showed that anxious students underestimated their language ability and this has implications for the language learning process. It is useful feature for a CALL resource to help users to identify their own ability and perhaps show them that they are not as 'bad at Irish' as they think they are. An adaptable CALL resource e.g. the learner can decide on the interface language and audio speed, gives learners more control over the learning process and this is good. As more data is gathered about learners, a CALL resource can be modified to adopt a more adaptive approach e.g. determining an appropriate path through the resources based on the learners' achievements and knowledge to date.

2 Irish Pronunciation and Orthography

Irish pronunciation is challenging for learners, including learners whose first language is English. Spelling and pronunciation in Irish is very regular – but there are problems for learners. There are about basic rules that learners may need to know, but in general, they are never taught them. This is in part because language teachers themselves are generally unaware of some of the more common rules. Furthermore, there are dialectal differences (there are three main dialects) and the rules have not been fully defined (Hickey and Stenson, 2011).

2.1 Irish – an Orthographically Deep Language

An orthographically shallow (transparent) language is one in which the letter and phoneme relationships is maintained, whereas an orthographicallay deep (opaque) language is one in which the letter/phoneme correspondance is not as consistent and complete. While not as deep as English, Irish is an orthographically deep (opaque) language (DAI, n.d.). This means that the logic of the orthography is not as transparent as an orthographically shallow language like Spanish. For example, in Spanish, the word for house is 'casa' and most readers will be able to pronounce this word correctly as 'kasa'. The word for house in Irish is '*teach'*, which many English native speakers would pronounce as the English word 'teach' (as in teacher). However, the word is actually pronounced 'chock' and this lack of immediate correspondence between the written form and the spoken form of the word can be disconcerting for many learners. Frost and Katz (1992) outline the Orthographic Depth Hypothesis (ODH) which indicates that the reading process is different for deep orthographies compared to shallow orthographies.

Consonants

The Irish alphabet uses the following consonants for Irish words: b, c, d, f, g, h, l, m, n, p, r, s, t. It also uses the letters j, k, p, v, w, x, y and z in loanwords. In many cases, the pronunciation of a consonant is predictable, but there are differences in pronunciation depending on whether a broad vowel (a, o, u) or a slender vowel (e, i) follow the consonant. For example, 's' with 'ú' (a broad vowel) is pronounced as /sᵛ/, whereas 's' with 'i' (a slender vowel) is pronounced as /ʃ/. Thus, the word *súil* (expected) is pronounced as suul, whereas the word *siúl* (walk) is pronounced as shuul.

Irish, in common with other Celtic languages, uses lenition and eclipsis. Lenition occurs when a stop becomes a fricative. In Irish orthography, lenition is denoted by a h after the consonant being lenited. For example, *peann* (pen, /pʲaːn̪ᵛ/) becomes *pheann* (/fʲaːn̪ᵛ/). Eclipsis, sometimes know as

nasalisation, causes the letter of the new sound to be placed in front of the original letter. For example, *peann* (pen; /pʲaːn̪ˠ/) becomes *bpeann* (/bʲaːn̪ˠ/).

Vowels, Di-graphs and Tri-graphs

Depending on classification, Irish could be considered to have 10 vowels but there are many more digraphs and trigraphs. The most basic vowels are a, e, i, o and u and they can be either stressed or unstressed. Each vowel also has an accented form: á, é, í, ó and ú. The accents denote a long form of the vowel. For example, *a* is pronounced like the 'a' in bat, while *á* is pronounced like the 'aw' in raw. However, there are di-graphs and tri-graphs in Irish. For example, ai is pronounced as 'a' (*baile*, /ˈbˠalʲə/, home) and iai is pronouced / iə/ (bliain, /bʲlʲiənʲ/, year). The list of di-graphs and tri-graphs can be initially daunting, but there are rules that can help the learner understand the system.

3 Overview of CALLIPSO

CALLIPSO (CALL for Irish for Parents Students and Others) is a CALL resource for learning about Irish pronunciation. The aim of the CALLIPSO system is to provide a learning resource that explains the basics of Irish pronunciation in a layperson's terms, rather than in a more technical (i.e. linguistically-oriented) manner. The learners can listen to letters, both individual, digraphs and trigraphs. There are a series of language exercises for learners to try out their knowledge of Irish pronunciation. The learner can also choose the interface language (e.g. English or Polish).

3.1 CALLIPSO Design Approach

The main design approach of the CALLIPSO system is modular and language-independent. A modular system leverages the benefits of good software engineering design. A modular system means that components are organised in a logical manner and that the code has good cohesion and is lightly coupled. This means that it is easy to change a part of the system without impacting on other parts of the system.

Sustainability was one of the key principles behind the CALLIPSO design. Very often resources are developed for a particular learning domain but cannot be adapted to another domain or different type of end-user. This is obviously wasteful in terms of money, time, effort and general resources. In recent years, the concept of sustainability has emerged as a theme in software development, including CALL (e.g. Sanz, 2015). The CALLIPSO system is designed with sustainability in mind. It has a modular design to facilitate extendibility and reuse. This is especially important when developing resources for Minority Languages, particularly if the resources could be used for other languages that currently lack CALL materials. User Modelling has been discussed in the literature for many years and has been used in Intelligent Tutoring Systems, but it has not been used extensively by CALL researchers when designing and developing CALL resources. There may be some consideration given to different user groups and their needs, but the use of user models per se is either very limited or under-reported. One of the future goals of the CALLIPSO system is to use user models in the system to try to enhance the learner's experience (along Fischer's (2001) lines). The target end-users are time-poor and Machiavellian and want the maximum gain with minimal effort.

CALLIPSO is designed using an agile approach (Beck et al., 2001), with a focus on designing and developing components that use useful and usable, rather than a more traditional software

engineering approach which places more emphasis on a structured, sequential model. The aim of using an agile approach is to be able to deliver working pieces of software in small increments so that users can avail of some functionality without having to wait until the entire system is complete. In recent years, the use of the agile paradigm has become more common in the area of software development. CALLIPSO developed its functionality incrementally and sought feedback from target users after each component was developed.

CALLIPSO also aims to reuse existing resources where possible. This is especially important for Minority Languages were resources are limited. It would be challenging and time consuming to provide the audio files required by CALLIPSO from scratch. Native speakers would be required to articulate the words used by CALLIPSO and given the agile approach adopted by CALLIPSO it would not be feasible to do this in a piece-meal manner. In order to overcome this problem, CALLIPSO uses resources from Abair (Abair, 2016), the Irish Text-To-Speech (TTS) tool. Abair is a high-quality system that provides TTS audio for three different dialects of Irish: Donegal (Gweedore), Connemara and most recently Munster (from the Dingle Peninsula). This is important as learners may wish to hear a word spoken in the particular dialect or even compare two different dialects. Abair also provides five speed settings: very slow, slow, normal, fast and very fast. The slower speed is particularly useful for learners as they often find it difficult to understand an L2 spoken at normal pace. As learners progress with their understanding of Irish pronunciation, they can progress towards understanding a word, phrase or text spoken at normal pace. The required audio files are generated by Abair and these are then incorporated into CALLIPSO.

CALLIPSO also builds on the research of Hickey and Stenson (2011) who outline an approach for teaching Irish pronunciation to learners. They combine knowledge of language pedagogy with Irish linguistics to propose a mechanism for teaching the sound system of the language to beginners. They present a mechanism for teaching Irish pronunciation in a logical and coherent fashion and this approach is used in CALLIPSO. For example, they suggest that learners should be explicitly taught the basic values of simple (orthographic) vowels and the length difference indicated by an acute accent), that in a vowel sequence with an accent, the vowel with the accent is the one to pronounce and that word-final vowels are never silent. They also suggest teaching that c in Irish is always pronounced a /k/ and g as /g/, that s is pronouced as s or sh depending on the adjacent vowels and that the use of h in lenition changes the pronunciation (learners often ignore the h). The system provides a brief summary of these rules to learners.

3.2 CALLIPSO Design

CALLIPSO uses the LAMP (Linux Apache MySQL PHP) stack. It uses an Apache HTTP server with MySQL relational database management system and the PHP programming language. The benefits of using the LAMP stack include ease of developing applications, easy to deploy, flexibility, security and there is a large support community. One other important benefit in the context of CALLIPSO and Minority Languages is the fact that it is open sourced and non-commercial. The CALLIPSO files are stored in Git, which is a free and open source version control system. Figure 1 shows an overview of the CALLIPSO LAMP stack.

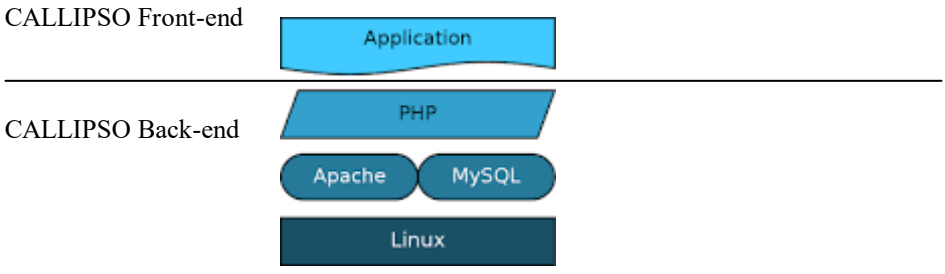

Figure 1 : Overview of the CALLIPSO LAMP Stack

3.3 CALLIPSO Database

CALLIPSO contains information on different entities. Table 1 shows the Business Rules for some of the main entities in the CALLIPSO system – these include letter, language, word, learning item and syllabus. The Business Rules provide a non-technical explanation of what each component means in the CALLIPSO system and the information that the system should store about that component. There are also Business Rules for user related components (e.g. user, login), quiz related components (e.g. quiz, mix quiz, match quiz, cloze quiz) and learner analytics (e.g. user progress and badge).

A **letter** is written character or combination of characters. It has associated audio, an International Phonetic Alphabet (IPA) phonetic representation and an example. A letter can have different examples depending on the language of instruction.
A **language** is a language that is used as the language of instruction for the CALLIPSO system. A language will have a language id, and the name of the language in English and the language itself.
A **word** is a combination of letters. A word has a word id, the word itself and a translation. There will be at least one translation for each language of instruction. A word may also have an associated image.
A **learning item** is a component that has information about some part of the orthography. It has an id, a type, an associated letter (or letter combination) or word. It also has textual information for the user. The textual information can be in different languages.
A **syllabus** is a collection of learning items and quizzes. A syllabus has an id, a type and a description.

Table 1 : Main Business Rules in the CALLIPSO system

CALLIPSO is built using a relational database (MySQL). The main linguists-related tables are Letter and WordDetail. Each letter has a unique id (letter_id), the character of the letter, a classification (vowel, consonant), a class type (simple alphabetic, limited, eclipsed, double, simple accent, acute accent, digraph, diphthong), a sample word with the letter, the IPA representation of the letter, a dialect_id, an audio of the letter, and a speed. The WordDetail table contains information on how to pronounce a letter. It contains a unique word_id, the example word, the meaning of the word, the letter(s) being explained, if the word is with a broad or slender consonant, or stressed/unstressed for a vowel, what it sounds like (for a lay learner) and a word with a similar sound in the learner's L1 (initially English). Table 2 shows the Letter table with an example of a consonant and a vowel. Table 3 shows the WordDetail table with examples. (Note that there is a LetterInfoCode table used to join the Letter table with the WordDetail table to cater for consonants combined with broad or slender vowels, but this is omitted here for simplification purposes). O'Siadhail (1988) provides further examples of broad/slender pairs.

There are also tables for quiz related data, including Mix, Match and Cloze which contain information on mix quizzes (i.e. multiple choice quizzes), match quizzes (match items on the left hand side with their corresponding match on the right-hand side) and cloze quizzes (where the learner has to fill in the blank). There are also tables to keep track on learners and their progress through the system. There are plans to incorporate learner analytics in CALLIPSO to help improve the learner experience and to improve the resource over time.

Letter Field	**Function**	**Example: 'b'**		**Example: 'á'**
		Broad vowel	**Slender vowel**	
letter_id	Unique id	700		800
characters	Letter(s) used	*b*	*b*	*á*
classification	Type of letter	simple_alpha	simple_alpha	simple accent
word_id	Link to sample word	131	132	110
IPA	IPA transcription	/bˠ/	/bʲ/	/aː/
dialect_id	Dialect identifier			
audio	Audio file			
speed	Very slow, slow, normal			

Table 2 : Letter table with examples

WordDetail Field	Function	Example: 'b'		Example: á
		Broad vowel	Slender vowel	
wd_id	unique id	131	132	110
wd_word	word in Irish	*bord*	*béal*	*bán*
wd_meaning	meaning of the word	table	mouth	white
wd_letter	the letter(s)	b	b	á
wd_classification	Broad or slender	s	b	stressed
wd_like	Similar sound	b	b	aw
wd_in	Similar sound in L1 word	boot	beautiful	raw
wd_lang	Language of instruction	English	English	English

Table 3: WordDetail table with examples

3.4 CALLIPSO User Interface

The design of the User Interface was a simple, consistent look and feel. There is an overview of the alphabet, as well as information on consonants and vowels. Within consonants, there are the simple consonants (b, c, d, f, g, h, l, m, n, p, r, s and t), the consonants with lenition (bh, ch, dh, fh, gh, mh, ph, sh and th), the consonants with eclipsis (bhf, bp, dt, mh, nd, and ng) and the special consonants (ll, nc, nn, rr and ts). Within the vowels, there are the simple vowels (a, e, i, o and u), the accented vowels (á, é, í, ó and ú) as well as the many digraphs (e.g. ai, ui) and trigraphs (e.g. aei, uai) that exist in Irish.

Figure 2 shows the CALLIPSO information for the letter b. It explains how the letter b is pronounced with a slender vowel (e or i) and a broad vowel (a, o, or u). In this example, the information is provided in English. A word for b with a slender vowel (*béal* – mouth) and with a board vowel (*bord* – table) are provided. These words can be clicked and the learner can hear the word being pronounced. All this information comes from the database so it can be modified as required. For example, if a different word was preferred as the example word, that could easily be changed. Also, the standard text explaining something (e.g. 'The letter … sounds') can easily be shown in a different language based on learner preference.

Figure 3 shows the CALLIPSO information for vowels. It provides a simple explanation of the vowels in Irish. The learner can click on the panel on the left-hand side to see more details on the vowels.

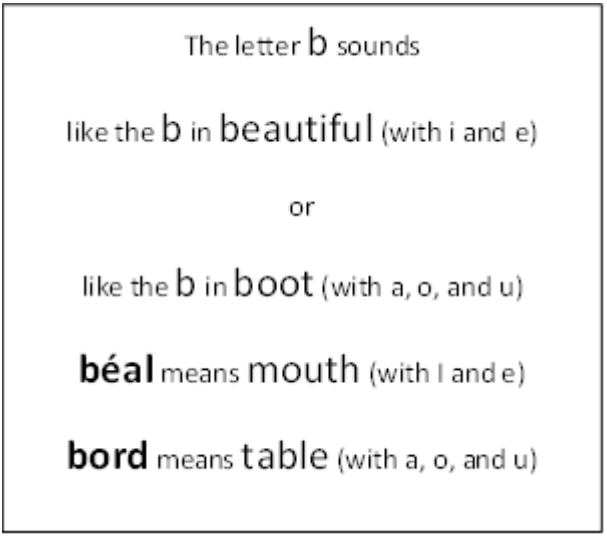

Figure 2 : CALLIPSO information for the letter b

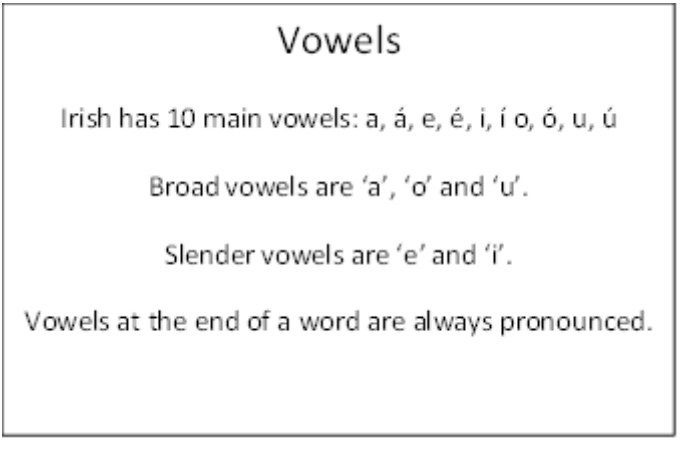

Figure 3 : CALLIPSO main vowel information

4 CALLIPSO – Future Developments

4.1 CALLIPSO for Other Celtic Languages

CALLIPSO was designed to be modular and language independent. It would be possible to adapt CALLIPSO for another Celtic language by populating the database tables with the relevant information for that language. All the information displayed on the CALLIPSO pages is generated by data in the database tables so it is fully flexible. The remaining infrastructure elements of the LAMP stack would not need to be changed. This is the theory – in reality there may be some changes that would be required, but these should not be too substantial.

4.2 CALLIPSO – Animation Module, Gamification, Learner Analytics

Further modules are being developed to enhance CALLIPSO. There is an animated visualisation module under development that will show the steps involved in pronouncing a word. Figure 5 shows an example of how to pronounce the word *Seán* (a popular male name in Ireland). It shows that the *á* means that the *á* is pronounced as 'aw', while the 's + e' means that the 's' is pronounced as 'sh', which gives a final pronunciation of 'shawn'.

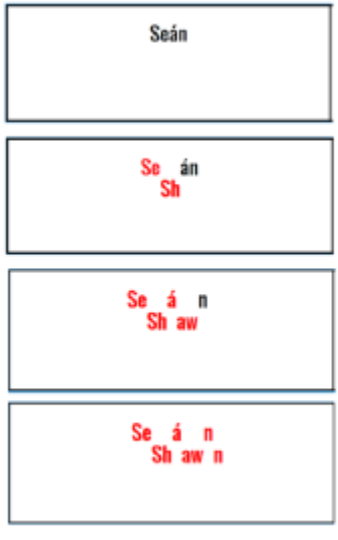

Figure 5 : Diagrams of animated visualisation of how a word is pronounced

Gamification is an increasingly popular area in CALL and in Computer Assisted Learning (CAL) in general (Kapp, 2012). Gamification provides learners with a chance to experiment and try things out with a safe set of boundaries (e.g. test their understanding of Irish pronunciation in private rather than in public). Deterding et al., (2011) define gamificaiton as the use of game design elements in non-game contexts, including education. CALLIPSO aims to provide exercises (or mini-games) where learners can test their knowledge and receive immediate feedback. There will be different levels as learners increate their knowledge and understanding of Irish. CALLIPSO is designed to be

able to award badges to learners as they progress through the system to encourage them to continue their learning. This feature has not been fully implemented. Learner Analytics is an area of active research interest (Ferguson, 2012), particularly in the area of educational data mining and there are plans to incorporate an element of learner analytics in CALLIPSO. Extra information about how learners interact with the system (e.g. do they look at the basic vowels or the more complex combinations ?) can be used to enhance CALLIPSO and provide an improved learning experience.

5 Conclusion

There is a need for a CALL resource to help learners understand Irish pronunciation. At first glance, Irish pronunciation looks difficult, as many learners are not explicitly made aware of the rules and they try to map their understanding of English pronunciation to Irish, which results in incorrect pronunciation. Furthermore, while there are some books that explain pronunciation, they often explain things in linguistic terms and use the IPA which may not be comprehensible for the average learner. A CALL resource has the advantage of allowing learners to be able to hear sounds and words being spoken and this can facilitate their understanding. CALLIPSO is a CALL system designed to explain the rules of Irish pronunciation to a non-technical learner. It is aimed at parents who want to help their children with their Irish homework, children themselves and even teachers who may wish to revise their knowledge of Irish pronunciation. CALLIPSO is module in design and could be adapted for other Celtic languages.

References

Abair. Abair.ie Available at : http://www.abair.tcd.ie/

Aleven, V., Mclaren, B. M., Sewall, J., Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, *19*(2), 105-154.

Beck, K., Beedle, M., Van Bennekum, A, Cockburn, A., Cunningham, W., Fowler, M., Kern, J. (2001). Manifesto for agile software development.

DAI. (n.d.). Irish/Language Learning. Dyslexia Association of Ireland. Available at: http://www.dyslexia.ie/information/information-for-parents/irishlanguage-learning/

Darmody, M., Daly, T. (2015). Attitudes towards the Irish Language on the Island of Ireland. *The Economic and Social Research Institute*.

Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* (pp. 9-15). ACM.CALL

Dörnyei, Z., Csizér, K. (1998). Ten commandments for motivating language learners: Results of an empirical study. *Language teaching research*, *2*(3), 203-229.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Fischer, G. (2001). User modeling in human–computer interaction. *User modeling and user-adapted interaction*, *11*(1-2), 65-86.

Frost, R., Katz, L. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. *Orthography, phonology, morphology and meaning*, *94*, 67.

Gardner, R. C., Lalonde, R. N. (1985). Second Language Acquisition: A Social Psychological Perspective.

Hickey, T., Stenson, N. (2011). Irish orthography: what do teachers and learners need to know about it, and why?. *Language, Culture and Curriculum*, *24*(1), 23-46.

Holland, V. M., Sams, M. R., Kaplan, J. D. (2013). *Intelligent language tutors: Theory shaping technology*. Routledge.

International Phonetic Alphabet. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Kapp, K. M. (2012). *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.

MacIntyre, P. D., Noels, K. A., Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language learning*, *47*(2), 265-287.

Siadhail O', M. (1988) Learning Irish: An Introductory Self-Tutor.

Pirolli, P., Kairam, S. (2013). A knowledge-tracing model of learning from a social tagging system. *User Modeling and User-Adapted Interaction*, *23*(2-3), 139-168.

Sanz, A. M. G. (2015). *WorldCALL: Sustainability and Computer-Assisted Language Learning*. Bloomsbury Publishing.

Scannell, K. (2014). An Gramadoir. Available at: http://borel.slu.edu/gramadoir/

Uí Dhonnchadha, E. (2002). A Two-level Morphological Analyser and Generator for Irish using Finite-State Transducers. In *LREC*.

Watson, I. (2008). Irish language and identity. *Nic Pháidín, C. agus Ó Cearnaigh, S.(eds.). A New View of the Irish Langauge*.

Developing Word Embedding Models for Scottish Gaelic

William Lamb[1] Mark Sinclair[2]
(1) Celtic and Scottish Studies, University of Edinburgh, School of Literatures, Languages and Cultures, 50 George Square, Edinburgh EH8 9LH, United Kingdom
(2) The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
w.lamb@ed.ac.uk, mark.sinclair@ed.ac.uk

RÉSUMÉ

Développement de modèles vectoriels continus de mots pour le gaélique écossais

Nous présentons ici un projet préliminaire pour la construction et l'évaluation de représentations vectorielles continues des mots appliquées au Gaélique écossais. Les méthodes de représentation vectorielles continues des mots ont déjà été appliquées avec succès à de nombreuses tâches en traitement automatique de la langue (TAL) et ont pour avantage de pouvoir être construites à partir de texte brut et non structuré. Ces méthodes sont ainsi particulièrement adaptées aux langues faiblement dotées en ressources linguistiques telles que le Gaélique. Nous avons construit trois différents modèles vectoriels continus des mots à partir de deux versions d'un corpus de 5.8 millions d'occurrences de mots (tokens). La première version consiste contient la simple segmentation en occurrences alors que la deuxième version comprend les occurrences et les formes lemmatisées. La représentation syntaxique des modèles est évaluée à partir d'un étiqueteur syntaxique en Part-of-Speech (POS). Par ailleurs, diverses requêtes sémantiques effectuées sur les modèles permettent de mesurer et caractériser leur richesse sémantique. Les modèles construits à partir du corpus d'occurrences seules s'avèrent peu robustes aux requêtes sémantiques en raison de la parcimonie des données. En revanche la lemmatisation améliore la robustesse des modèles pour les requêtes sémantiques mais au prix d'une sensibilité flexionnelle accrue. Nous illustrons les différences entre les modèles ainsi que l'apparent compromis entre leurs capacités sémantiques et syntaxiques. Finalement, nous soulignons le potentiel des représentations vectorielles continues des mots pour toute une série d'applications futures.

ABSTRACT

Developing Word Embedding Models for Scottish Gaelic

We detail a preliminary project on encoding and evaluating word embeddings for Scottish Gaelic. Word embedding methodologies show promise for diverse natural language processing (NLP) tasks and can be built from raw, unstructured text. Accordingly, they are attractive for under-resourced languages like Gaelic. We instantiated three embedding models on two versions of a 5.8 million token corpus : 1) tokenised and 2) tokenised / lemmatised. Using a simple POS tagger, we quantitatively measured the syntactic similarity between nearest neighbours for each model's vector-space representations of words. We also queried the models to assess their semantic specificity and breadth. Models built from the tokenised corpus exhibited the effects of data sparsity for semantically constrained queries. The lemmatised versions had more semantic robustness, but at the expense of inflectional sensitivity. We note divergences between the models and an apparent inverse relationship between their semantic and syntactic capacities. Finally, we highlight the promise of word embeddings for a range of future work and downstream applications.

MOTS-CLÉS : gaélique écossais, modèles vectoriels continus de mots.

KEYWORDS: Scottish Gaelic, word embeddings, neural networks, natural language processing, word2vec, part-of-speech tagging.

1 Introduction

When reflecting on the position of our language technologies, it is common for those working on minority languages to express some degree of English envy. State-of-the-art natural language processing (NLP) technologies typically require large quantities of labelled training data. These are readily available for English and other majority languages, but not normally for under-resourced languages. Yet, as in other data-driven fields, NLP has recently been dominated by approaches leveraging artificial neural networks. While these approaches do not necessarily mitigate requirements for labelled data directly, they are attractive for their language-independence and the fact that they can be generated unsupervised from relatively raw data (Lin *et al.*, 2015; Chen *et al.*, 2013). Large annotated corpora are unlikely to exist for under-resourced languages, but copious amounts of on-line text are often available. After light processing, this text can be made suitable for approaches based on neural networks. As we demonstrate in this paper, useful models can result from modestly-sized datasets.

A key difference between a neural network and conventional NLP approach is that the former typically requires words to be represented as numerical vectors. The process of mapping atomic word units (tokens, lemmas, etc.) to vectors is known as 'word embedding'. Neural network word embeddings, or vector space models (VSMs), use high-dimensional geometry to map associations between words. Embedding algorithms exploit the iconic relationship between semantics and linguistic context, typically mapping similar words to nearby vector points. The underlying principal recalls Firth's observation that '[y]ou shall know a word by the company it keeps' (1957 : 11). Although vectors are difficult to interpret — each dimension represents multitudinous concepts and concepts are spread multi-dimensionally (Al-Rfou *et al.*, 2013) — word embedding models have proven effective as input to a variety of standard NLP tasks, such as part-of-speech (POS) tagging (Fonseca *et al.*, 2015).

Given the above characteristics and possibilities, word embedding models could be useful for work involving Scottish Gaelic (Al-Rfou *et al.*, 2013). Although improvements have been made to Gaelic language technology in recent years (see Batchelor, this volume), it still lags behind that of most larger languages and even some minority languages (e.g. Irish Gaelic). We present this paper as proof of concept in the interest of using word embeddings methodologies to expedite the development of Scottish Gaelic NLP resources. In the sections below, we overview our methodology, provide an initial assessment of the models' strengths and weaknesses and comment on potential downstream applications and future possibilities.

2 Background and Methodology

Scottish Gaelic is a Celtic language that is closely related to Irish and Manx Gaelic, and more distantly related to Cornish, Welsh and Breton. Working with Gaelic in an NLP context presents several challenges. As aforementioned, one is the low availability of high quality data, such as tagged

corpora. [1] Additionally, word forms in Gaelic are remarkably protean due to its complex morphology (see Lamb, 2008). For example, its nominal system features initial and terminal mutations that are sensitive to grammatical categories such as case, number and definiteness. A word like *cailleach* 'old woman' may appear as *chailleach, caillich, cailliche, chailliche, cailleachan, chailleachan, cailleachaibh* or *chailleachaibh*, depending on grammatical context. (From the lexicon described below, we calculate an average surface-form to lemma ratio of 6.84 to 1.) In addition, although orthographic standards exist (SQA, 2009), few writers adhere to them exclusively ; spelling can be idiosyncratic. Given this variability, data sparsity is a common problem as we discuss further below.

Our data came from a 21 million word web-crawl of Gaelic text, available as part of the Crúbadán [2] project (Scannell, 2007). The source texts are diverse in register and quality, ranging from biblical prose to chat room dialogue. Much of the text stems from the Gaelic version of Wikipedia (gd.wikipedia.org). Scannell took a random sample of sentences from larger sources to lessen any bias towards them, and extirpated the data of much ambient English. He provided us with a file of 5.8 million tokens (263,858 lines ; 133,287 unique tokens) and, from this, we generated two training files : 1) tokenised and 2) tokenised and lemmatised. We built the lemmatiser [3] using a large, manually constructed lexicon of Scottish Gaelic (Am Faclair Beag : [4] see Patton, 2016). It is capable of handling common orthographic variations, such as 'Uidhist' for 'Uibhist' (Eng : 'Uist'), although not out-of-dictionary items ; for this study, we replaced the latter with "#IGNORE". Light standardisation was applied in the form of automatically de-capitalising wrongly capitalised tokens.

2.1 Word Embedding Method

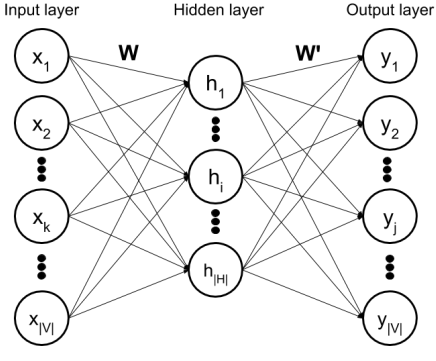

FIGURE 1 – A typical multi-layer perceptron used for learning vector-space word embeddings.

Figure 1 shows a typical multi-layer perceptron (MLP) with input, hidden and output layers. The input and output layers have a node for every word in a given vocabulary V, of size $|V|$. Each of these layers is fully connected to a single hidden layer of size $|H|$. The connections are represented by input and output weight matrices, W and W', respectively.

1. But see (Maolalaigh, 2013; Lamb *et al.*, 2016)
2. *Crúbadán* is Irish for 'crawler' : see http ://crubadan.org
3. We hope to develop the lemmatiser further and make it available in the future
4. www.faclair.com

The nodes in the hidden and output layers perform simple functions that aggregate their inputs and produce a single output. These functions are generally fixed to be common across all nodes in a given layer (specifically they are often softmax or sigmoid functions). It is the weight matrices that describe how to emphasize or de-emphasize a given input to a node. By learning the weights that maximise the likelihood of input/output example pairs, the network can learn inherent structures within a given dataset. Once the learned weights are fixed, we can present an arbitrary input vector to the network and compute a corresponding output vector.

As an example, consider the following dataset of word pairs (bi-grams) that describe Scottish geographical features — lochs (lakes), rivers and bens (mountains) — along with specific names.

$$D = \{loch|lomond, river|ness, ben|lomond, loch|ness, river|tay, ben|vorlich, loch|tay,$$
$$river|clyde, ben|more, loch|more, river|forth, ben|nevis\}$$

We may wish to have the network learn to associate a feature with a name or vice-versa. The vocabulary required to describe this complete set would be :

$$V = \{ben, clyde, forth, loch, lomond, more, ness, nevis, river, tay, vorlich\}$$

Each pair of words can then be described as an input and output vector that is 1 at the position of the word and 0 otherwise , e.g. for the pair $loch|lomond$:

$$loch = \{0,0,0,1,0,0,0,0,0,0\}, lomond = \{0,0,0,0,1,0,0,0,0,0\}$$

By providing the network with these "one-hot" vectors during training, it can learn the weight matrices (typically by means of the back-propagation algorithm) that are best able to make a correct mapping from input to output. As the number of hidden units $|H|$ is typically much smaller than $|V|$, it compresses the input through the hidden layer and decompresses it at the output. It is these compressed vectors that allow us to encode words into a more condensed vector space than that of the input or output layers. We can then measure the distance between vectors in order to find out how 'close' words are in a given model. In this example, we may expect to find $clyde$ close to $river$ and far from ben and $loch$ because there are no such places. However, $lomond$ may be close to both ben and $loch$ as both places exist, so it can co-occur with either of these words. This is a very simple example but serves to show how such neural network architectures can be used to perform NLP tasks such as answering the question "which word comes next ?". Derivatives of similar architectures can be used to model more complex relationships betweens words.

2.2 Tools and Model Types

In order to train the word embedding models, we used the tool [5] developed by (Ling *et al.*, 2015), which itself is a modified version of the popular word2vec [6] algorithm (Mikolov *et al.*, 2013a,b,c). This tool allows word vector representations to be learned from any raw text corpora. Several different learning schemata have been made available so we chose three of the most popular, as described briefly below.

Constrained Bag-of-Words (CBOW)

This training schema works by learning to predict a word given a context. For example, if we consider that we have 5-gram training examples such as “quick brown fox jumps over”, we can set our input vector to a one-hot representation of “quick brown jumps over” and the output vector to “fox”. The model will then learn weight matrices that can predict this

5. https ://github.com/wlin12/wang2vec
6. https ://code.google.com/archive/p/word2vec

relationship.

Skipngram
Skip-gram works as a kind of 'inverse' of *CBOW* – given an input word, we want to predict the context. The step size for the N-grams used for context can also be altered to provide a wider context.

Structured skip-gram
Structured skip-gram works in a similar way to the standard skip-gram model, except that we also provide the relative positions of each context word with respect to the target word. This allows the model to learn more about the inherent linguistic structure such as the proximity of adverbs to verbs or adjectives to nouns. As a consequence, the model can potentially better learn syntactic relationships between words.

The Polyglot project [7] (Al-Rfou *et al.*, 2013) offers a selection of word embedding models that have been automatically generated for a number of languages, including Scottish Gaelic. While this model was trained on different data, we were still able to use it for comparisons.

Unless otherwise stated, all of our models were trained with a 5-gram window over 3 iterations of training with a hidden layer size of 64 (to match Polyglot) or 200. Any words with a frequency of less than 5 were ignored.

3 Evaluation

3.1 Overview

The evaluation of word embedding models is typically performed in one of the following three ways :

The use of a lexical relationship database
For some well resourced languages such as English, the availability of a well-curated lexical database of word relationships can be exploited to analyse word embedding models. WordNet (Miller, 1995) is an example of such a database. In WordNet, words are divided into high level syntactic classes (noun, verb, adjective, adverb, etc.) which are then grouped into sets of cognitive synonyms (synsets). Within each synset, words are interlinked according to conceptual-semantic and lexical relations. Given such a resource, it is possible to evaluate if a word embedding model is in agreement with the database across many different criteria. We can, for example, query the database for relations of a given word and then check the cosine distance between the word and each relation in the embedding model — whereby lower distance would indicate better agreement with the database (Handler, 2014). Lexical databases such as WordNet are the product of a substantial cumulative effort involving thousands of work-hours from expert annotators and, unfortunately, a similar resource does not yet exist for Scottish Gaelic.

Subjective experiments
The concept of word similarity inherently contains a subjective component, particularly concerning semantic relationships. Therefore, it is pertinent to consider designing subjective experiments to evaluate word embedding models. Typical forms of such experiments may include asking subjects to : rate or rank word groups in order of similarity ; suggest a similar

7. https ://sites.google.com/site/rmyeid/projects/polyglot

word y given word x ; choose a best match or selection preference, e.g. which noun typically goes with this verb ? Such queries can also be posed to the word embedding models so that answers can be compared with human subjects.

There are well-documented issues associated with subjective evaluations such as the above, one being inadequate sample size. The recent trend of using crowdsourcing technologies such as Amazon Mechanical Turk (AMT) provides the opportunity to alleviate some of these issues (Schnabel *et al.*, 2015). AMT allows users to set up web-based experiments whereby volunteers can participate for a small financial reward. However, this requires a large pool of potential subjects in order to yield enough respondents. We feel it would be unlikely to garner a large enough response from Scottish Gaelic speakers by means of AMT due to its relatively low number of native speakers. However, in the future we may consider smaller scale subjective experiments that do not utilise crowdsourcing.

Downstream task performance

Instead of evaluating the performance of word embedding models directly, we can also evaluate how they affect other downstream NLP tasks. For example, word embeddings from a given model could be used to train a language model or a document classifier. We can then use evaluation metrics and/or datasets for those tasks. This can often be a more informative evaluation if the word embedding model is designed for a specific task.

3.2 Syntactic (quantitative)

We were able to derive a quantitative measure of syntactic representation for each model by exploiting an existing, manually composed lexicon (see Patton, 2016). This lexicon provides highly detailed information with respect to potential POS tags for a given token, including all valid combinations of case, gender, number, etc. In total, the lexicon offers 198 possible POS tags, reflecting the relatively rich morphology of Scottish Gaelic as compared with languages such as English (the Penn Treebank Project, for example, considers only 36 POS tags).

We do not expect our models to able to distinguish accurately between POS tags at a high level of granularity, so we considered only the first-order tags of the lexicon : verb, noun and adjective. We then removed all tokens from each model that did not have a corresponding entry in the lexicon. This meant that we were able to look at the n-nearest neighbours of a given token within the vector space of each model and observe any homogeneity among the associated POS tags. By counting how often the POS tags of a token and its neighbours are in agreement, we are able to quantify the tendency of each model towards a more syntactically informed clustering.

The syntactic score, S, is formulated as follows :-

$$S = \frac{\sum_{i=1}^{|V|} \sum_{j=1}^{n} f(POS_i, POS_{i,j})}{|V|n}, f(A,B) = \begin{cases} 1 & \text{if } a = b \text{ for any } a \in A, b \in B \\ 0 & \text{otherwise} \end{cases}$$

Here, POS_i represents the set of part-of-speech tags for word i, and $POS_{i,j}$ represents the set of tags for the j-th nearest neighbour to word i according to cosine distance.

Table 1 shows the results for several model types. In order to compare with the Polyglot model, we used the 2000 most frequent tokens that were in common across all models to calculate the score. We present results with a hidden-layer size of $|H| = 64$, which matches the Polyglot model, and also with a larger value of $|H| = 200$. In all cases the smaller hidden-layer performs better. This may be an effect of inherent regularization offered by having fewer parameters to model our dataset which is

TABLE 1 – Agreement of tokens with N-nearest vector neighbours in POS category (Syntactic Score)

	Syntactic Score, S	
Model	$\|H\| = 64$	$\|H\| = 200$
Polyglot	64.87%	-
CBOW	82.11%	82.08%
skipgram	75.06%	73.59%
structskipgram	85.06%	84.32%

still relatively small for this task. If we had more data, then a larger hidden-layer may perform better.

We find that *CBOW* performs better than *skipgram*, which is consistent with other findings in the literature (Mikolov *et al.*, 2013a; Qiu *et al.*, 2014). As expected, adding structural information to the *structskipgram* model significantly increases the syntactic score. This is likely due to the extra information helping the model to learn local grammatical structures that can push words into certain clusters based on their relative positions. The Polyglot model has the lowest performance. This may be due to the differences in training data or it could be simply that the model was designed to capture another type of linguistic relationship. Due to the near-fully automatic method used for training the minority Polyglot languages the training data may only have gone through generic — rather than language specific — tokenisation.

It is worth noting, however, that improvement in syntactic modelling may be at the expense of semantic modelling and a suitable choice of model may depend on the target application. A strong syntactic model may, in future work, be able to provide supplementary information to tasks such as part-of-speech tagging for Scottish Gaelic.

3.3 Semantic (qualitative)

As expected from a base of 5.8 million words, the models capture robust semantic and syntactic relationships between common words. However, they lack the sophisticated nuances reported for languages with more available text. We begin by discussing the differences between the models, followed by how well the models encode semantic information. It is worth noting that the empirical evaluation of word embeddings semantics is at an early stage. Although paradigms exist, as detailed above (Schnabel *et al.*, 2015), they require significant human resources. Our comments, perforce, are impressionistic at this point.

We queried the models with terms selected to test their semantic granularity and breadth. Although the models provide similar returns for very common words, they diverge notably with more semantically constrained ones. For example, in Table 2, we report the top five returns ranked by cosine similarity for *Uibhist* 'Uist', a well-known Hebridean island. (NB : Unless noted with '*', the models were trained on the lemmatised data.)

From this result and others, it would appear that the models are sensitive to different semantic domains. *CBOW* is effective at locating the general semantic category ; it groups 'Uist' with a variety of other place-names. *Skipgram* returns nearby island place-names only, indicating greater specificity. *Structskipgram* is similar to *skipgram*, but includes 'America'. When considering returns for *eaglais* 'church', we see similar tendencies : *CBOW* returns other buildings (e.g. hotel, palace, abbey) while *skipgram* returns other ecclesiastic nouns (parish, Catholic, abbey, graveyard). Again, *structskipgram*

TABLE 2 – Nearest neighbours per model for input query Uibhist 'Uist' (lemmatised). English translations are provided for convenience.

CBOW	**skipgram**	**structskipgram**
Aimeireaga 'America'	*Èirisgeigh* 'Eriskay'	*Tiriodh*'Tiree'
Èirisgeigh 'Eriskay'	*Barraigh* 'Barra'	*Ìle* 'Islay'
Afraga 'Africa'	*Tiriodh* 'Tiree'	*Slèite* 'Sleat'
Barraigh 'Barra'	*Slèite* 'Sleat'	*Leòdhas* 'Lewis'
Àisia 'Asia'	*Leòdhas* 'Lewis'	*Aimeireaga* 'America'

TABLE 3 – Nearest five neighbours for common semantic domains. English translations are provided for convenience.

TOKEN	**TRANS.**	**TOKEN**	**TRANS.**	**TOKEN**	**TRANS.**
mara*	**sea (gen)***	***obh***	**oh (dear)**	**faicinn**	***seeing***
beinne	hill (gen)	*Obh*	Oh (dear)	*cluinntinn*	hearing
coille	forrest (gen)	*siuthad*	go on	*tuigsinn*	understanding
gaoithe	wind (gen)	*och*	oh	*faireachdain*	feeling
mòintich	moor (gen)	*ist*	listen	*saoilsinn*	thinking
creige	rock (gen)	*siuthadaibh*	go on (pl)	*smuaintinn*	thinking
dearg	***red***	***beagan***	***a bit***	***bus***	***bus***
geal	white	*cus*	too many	*bàta*	boat
gorm	blue	*tòrr*	many	*trama*	tram
glas	gray	*mòran*	many	*trèana*	train
uaine	green	*barrachd*	more	*trèan*	train
donn	brown	*moran* (sic)	many	*plèan*	plane
(An) Eadailt	**Italy**	***dithis***	**two people**	***craobhan****	**trees***
(An) Ruis	Russia	*triùir*	three people	*creagan*	rocks
(An) Ostair	Austria	*ceathrar*	four people	*glinn*	glens
(An) Fhraing	France	*dithist*	two people	*lusan*	plants
(An) Òlaind	Holland	*còignear*	five people	*cnuic*	hills
(An) Spàinn	Spain	*sianar*	six people	*rathaidean*	roads

is intermediate in focus. Unless otherwise stated below, we report results from *CBOW*, which seemed to be the most semantically coherent model overall.

As reported in Table 3, the model discriminates common semantic domains such as colours, countries and modes of transport. [8] Interestingly, it also groups returns based upon the case, number and grammatical category of the query word. For example, in the case of *obh* 'oh (dear)', the model returns other interjections. The genitive of the feminine noun *muir* 'sea' prompts other feminine, genitive nouns associated with physical geography. Quantifiers and quantitative pronouns are also grouped together, as are psychological verbal-nouns (e.g. *creidsinn* 'believing'). When queried, *craobhan* 'trees' produces other landscape-oriented plural nouns.

These results are promising given the relative paucity of data and the sparsity associated with Gaelic morphology (see Danso & Lamb, 2014). However, these issues come into sharp relief with other

8. Note : *trèana* is a variant of *trèan* as *dithist* is a variant of *dithis*

TABLE 4 – Effects of data sparsity : tokenised vs lemmatised models for clàrsach 'harp' (structskipn-gram reported)

TOKEN	**TRANS.**	**LEMMA**	**TRANS.**
clàrsach	**harp**	***clàrsach***	**harp**
teip	tape	*pìob*	bagpipe
coimpiutairean	computers	*druma*	drum
dubhan	hook	*fonn*	tune
fònaichean-làimhe	mobile phone	*giotàr*	guitar
tulach	hill	*seinn*	singing

queries. For example, the noun *clàrsach* 'harp' – when run through the models trained on the tokenised corpus – produces seemingly unrelated nouns (see Table 4). However, when queried on the models instantiated from the lemmatised corpus, other musical instruments and terms are returned as expected. As is well known, lemmatisation is an effective way to handle data sparsity. On the other hand, it restricts potential searches to root forms and this can be disadvantageous when grammatically sensitive models are required. Additionally, lemmatisers can introduce their own errors, and exclude a significant proportion of tokens as 'out of vocabulary' if the data is orthographically inconsistent. Therefore, one must balance potential gains and losses when considering whether to use them.

4 Conclusions

Although the word embedding techniques employed here are already well-represented in NLP literature, this is the first example of their application and evaluation for Scottish Gaelic. We adapted existing resources for our task, and trained and evaluated a variety of models on two versions of the textual data : 1) tokenised and 2) tokenised and lemmatised. Although the resources required for conventional evaluation approaches were not available, we were able to derive an objective measure of syntactic modeling capacity and an initial, qualitative assessment of semantic modeling capacity. We find the performance in each category to be dependent upon the choice of the model, potentially with an inverse relationship obtaining between the two. In other words, improving syntactic performance may be at the expense of semantic performance, although additional work is required.

The relative lack of resources for Scottish Gaelic compounded with the language's morphological complexity presents significant data sparsity issues. We have shown how these issues can be partially mitigated by lemmatising the training data *a priori*. However, the lemmatisation process introduces its own errors into the overall end-to-end system and may not be suitable for applications requiring sensitivity to grammatical inflection.

A motivating factor for this study, as aforementioned, is that approaches based upon word embeddings facilitate the exploitation of raw textual data without the need for manual annotation or intervention. Therefore, they provide a gateway for dealing with the resource constraints that commonly face minority languages.

We anticipate applying what we have learned from this study in two ways : 1) improving the model training by bootstrapping from better-resourced, related languages such as Irish and Welsh (e.g. by pre-training the model on those languages before fine-training on Scottish Gaelic) and 2) substituting conventional atomic representations with vector-space representations for a variety of potential

downstream NLP tasks (e.g. POS tagging, language modelling and machine translation). Vector space representations are a prerequisite for accessing artificial neural network solutions, which are increasingly driving state-of-the-art language technology. Therefore, this initial work is promising for the future of Gaelic NLP.

Acknowledgements

We wish to thank Prof Kevin Scannell (Saint Louis University) for providing the data that we used to train the models. We are also indebted to Michael Bauer and Will Robertson of 'Am Faclair Beag' [9] for allowing us to use their lexical database.

Références

AL-RFOU R., PEROZZI B. & SKIENA S. (2013). Polyglot : Distributed word representations for multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 183–192.

ANDREAS J. & KLEIN D. (2014). How much do word embeddings encode about syntax ? *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 1–9.

BATCHELOR C. (2016). Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in Scottish Gaelic. In *Proceedings of the Celtic Technology Workshop (CLTW 2016)*, volume 2.

CHEN Y., PEROZZI B., AL-RFOU R. & SKIENA S. (2013). The expressive power of word embeddings. In *Proceedings of the 30th International Conference on Machine Learning*, p. 1–9.

DANSO S. & LAMB W. (2014). Developing an automatic part-of-speech tagger for Scottish Gaelic. In J. JOHN, L. THERESA, W. MONICA & B. Ó RAGHALLAIGH , Eds., *Proceedings of the Celtic Technology Workshop (CLTW 2014)*, volume 1, p. 1–5.

FIRTH J. R. (1957). *A Synopsis of Linguistic Theory, 1930-1955*. Blackwell.

FONSECA E. R., ROSA J. L. G. & ALUÍSIO S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, **21**(1), 1–14.

HANDLER A. (2014). *An empirical study of semantic similarity in WordNet and Word2Vec*. PhD thesis, Columbia University.

LAMB W. (2008). *Scottish Gaelic Speech and Writing : Register Variation in an Endangered Language*, volume 16 of *Belfast Studies in Language, Culture and Politics*. Cló Ollscoil na Banríona.

LAMB W., ARBUTHNOT S., NAISMITH S. & DANSO S. (2016). Annotated reference corpus of Scottish Gaelic (ARCOSG). http ://dx.doi.org/10.7488/ds/1411.

LIN C.-C., AMMAR W., DYER C. & LEVIN L. (2015). Unsupervised pos induction with word embeddings. *arXiv preprint arXiv :1503.06760*.

LING W., DYER C., BLACK A. & TRANCOSO I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1299–1304.

9. http ://www.faclair.info/

MAOLALAIGH R. Ó. (2013). Corpas na Gàidhlig and singular nouns with the numerals 'three' to 'ten' in Scottish Gaelic. *Scottish Cultural Review of Language and Literature*, **19**, 113–142.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, p. 3111–3119.

MIKOLOV T., YIH W. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT-2013)*, p. 746–751.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

PATTON C. (2016). Review of Am Faclair Beag online Gaelic-English dictionary. *Language Documentation and Conservation*, **10**, 155–163.

QIU L., CAO Y. & NIE Z. (2014). Learning word representation considering proximity and ambiguity.

RONG X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv :1411.2738*.

SANTOS C. D. & ZADROZNY B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 1818–1826.

SCANNELL K. P. (2007). The crúbadán project : Corpus building for under-resourced languages. In *Building and Exploring Web Corpora : Proceedings of the 3rd Web as Corpus Workshop*, volume 4, p. 5–15.

SCHNABEL T., LABUTOV I., MIMNO D. & JOACHIMS T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 298–307.

SQA (2009). *Gaelic Orthographic Conventions*. Scottish Qualifications Authority.

English to Irish Machine Translation with Automatic Post-Editing

Meghan Dowling Teresa Lynn Yvette Graham John Judge
ADAPT Centre, Dublin City University, Dublin, Ireland
meghan.dowling@dcu.ie, tlynn@computing.dcu.ie, graham.yvette@gmail.com, jjudge@computing.dcu.ie

RÉSUMÉ

Traduction Automatique de l'Anglais vers l'Irlandais Incluant un Module de Post-Édition Automatique

Cet article présente l'adaptation d'un système de traduction automatique statistique, anglais→irlandais, à un nouveau domaine d'utilisation. Ce système est actuellement utilisé par une équipe de traducteurs du gouvernement irlandais. Nous décrivons également le nouveau module de post-édition automatique qui a été développé pour améliorer le système actuel et faciliter le travail de post-édition des traducteurs.

ABSTRACT

This paper reports on the continued development of a domain-tailored English→Irish Statistical Machine Translation system currently in use by an in-house translation team of an Irish government department. We describe the new automatic post-editing module that has been developed to enhance the current system and reduce the post-editing required of translators.

MOTS-CLÉS : Traduction automatique statistique, Post-édition automatique, langue morphologiquement riche, langue irlandaise..

KEYWORDS: statistical machine translation, automatic post-editing, morphologically rich language, Irish language.

1 Introduction

The Irish language holds the status of national and first official language in the Republic of Ireland. This status has led to a government requirement for all official documents and public services to be made accessible in both Irish and English, with the official status of English being a second official language. The demand for Irish-translated content exceeds the productivity capabilities of current translation services in Irish government departments.

In particular, the in-house translation team of the Department of Arts, Heritage and the Gaeltacht (DAHG), the government department responsible for Irish language affairs, has a significant workload and considerable amount of backlog of documents required to be translated into Irish, due to high demand from within their own and across other government departments.

In the past, translators relied solely on translation memory (TM) tools for translation into Irish. While

TM goes some way towards speeding up manual translation and increasing productivity, its benefits are of course limited only to working with previously translated text that are similar to or fully matching source language input text. When MT is not available and source language text has not been previously encountered, this results in translators translating text entirely from scratch. To this end, DAHG has provided funding for development of an English to Irish Statistical Machine Translation (SMT) system to bridge the wide gap between supply and demand of Irish language translations. The resulting system provides translators with the choice of an MT output translation in addition to any matching TM.

The specific requirement of the system was to achieve high-quality translation of domain specific data – that is, the system was required to produce high-quality translations specifically for public administration text. A feasibility study was carried out to determine the appropriate use-case, which amounted mostly to translation of documents, such as annual reports, staff announcements and public notices, for example. The feasibility study ensured the opportunity to ascertain the most appropriate data to train the required SMT system. As there was not a corpus of suitable quality within the required domain readily available, the priority of the project became the collection, cleaning and curation of parallel data. Dowling *et al.* (2015) provide a summary of this corpus development along with a report on preliminary translation scores for an English to Irish Phrase-based SMT system based on Moses (Koehn *et al.*, 2007), often referred to within the Irish-speaking community as the Tapadóir project.

This paper reports on recent enhancements to Tapadóir. In Section 3, we describe the development of an Automated Post-Editing (APE) module, which addresses morphological challenges encountered by the SMT system and results in modest BLEU score improvements. We then report on the evaluation of the APE module in Section 4. Finally, in Section 5, we show the success of the integration of this MT software into the translator's work-flow by reporting on positive user-engagement with the newly introduced technology.

2 Related Work

There have been various approaches to addressing the problem of translation into morphologically-rich languages. For example, the approach taken by Avramidis & Koehn (2008) involves adding per-word linguistic information to the source language, while Virpioja *et al.* (2007) use unsupervised morphology learning. El Kholy & Habash (2012) report success in this area through the use of a discriminative lexicon model applied to the SMT system. The method suggested by Chahuneau *et al.* (2013) involves a two-tiered approach: building a discriminative model which can predict target-side inflections, and then using this model to generate additional translations which can be included in the standard translation model as "synthetic" phrases. More recently, the Dagstuhl seminar on Statistical Techniques for Translating to Morphologically Rich Languages (Fraser *et al.*, 2014), has brought together researchers from a number of NLP (natural language processing) disciplines to identify new techniques to translating into morphologically rich languages.

Automatic Post-Editing (APE) of MT aims to improve MT output quality in order to reduce post-editing effort required of professional translators (Knight & Chander, 1994). The most widely applied method of APE for MT currently in use is statistical phrase-based post-editing, proposed by Simard *et al.* (2007), where the APE uses the MT output and its corresponding human post-edited data as a parallel corpus. Béchara *et al.* (2011) propose a significant variant that includes the source information

along with the MT output on the source side of the parallel corpus. Chatterjee *et al.* (2015) compare these two approaches for English to Spanish MT, the approach of Simard *et al.* (2007) achieving lower TER scores. Pal *et al.* (2015) apply hybrid word alignment techniques, while Wisniewski *et al.* (2015) take a rule-based approach in addition to Statistical APE. In this paper, we apply a simple rule-based approach to APE for English to Irish MT.

3 Automated Post Editing for Irish

Usability and user experience are extremely important factors in the Tapadóir project. As the primary aim of Tapadóir is to improve the speed and productivity of translators, it is crucial to produce a tool that does not hinder the user in any way. As part of our translator-developer feedback loop, translators reported some repetitive errors in the MT output that were causing frustration. On closer examination, most of the errors were grammatical problems arising from Irish language morphology that Tapadóir was not yet equipped to deal with. In comparison to English, Irish has a richer morphology, such as inflected prepositions and the initial consonant mutations, and causing challenges for SMT due to data sparsity. This problem is compounded in the case of lesser-resourced languages where there are low instances of various inflected forms in the training data.

This gap in knowledge could be bridged through a number of methods such as increasing the volume of training data (where the system becomes familiar with various inflected forms of a word), factored models (where the system uses part-of-speech and lemma information to improve its knowledge) or through the introduction of post-processing module that could address simple grammatical issues on a word level basis.

To this end, we designed an Automated Post Editing (APE) module that could address trivial spelling issues or contraction issues that challenged the SMT system. By automatically post-editing these errors, translators can dedicate more time to more important issues such as language style. The addition of APE is intended to improve the translator user-experience and avoid any negative impact of repetitive grammatical or orthographic errors, thus creating a more enjoyable user experience.

3.1 Designing the APE module

To develop the APE module, our translator-developer feedback loop enabled us to acquire information on frequently occurring errors, and occurrences of mistranslations. On inspection, translations contained a high number of errors related to Irish language prepositions, eclipsis, lenition and contractions. This motivated the development of a set of manually written rules to correct regularly occurring errors in Irish MT output. Rule sets were developed for individual prepositions and contractions and are triggered by the presence of lexical items in MT output. The APE module is split into two parts: one part which deals solely with orthographic rules, and another which addresses errors caused by grammatical case. In total there are 167 hand-written rules, which have been divided into 55 rule groups (according to preposition and error type).

3.1.1 MT Errors related to orthographic rules in Irish

16 of the most common Irish simple prepositions can be in inflected to mark pronominal objects (Christian-Brothers, 1962), (Christian-Brothers, 1960), known as prepositional pronouns or pronominal prepositions. For example, it is ungrammatical in Irish for a pronominal object to occur separated from the preposition (Ó Múrchú, 2013). Such occurrences on occasion arise in the translation output, however, possibly due to a specific phrase being unseen by the MT system and subsequently translating the phrase on the individual word level. An example of an APE rule now implemented in the systems produces correctly inflected forms of these prepositions when the system incorrectly generates word for word translations (see examples 1 and 2).

Examples of rules:

(1) le mé* → liom
'with me'

(2) ag sinn* → againn
'with us'

Irish includes orthographical rules that aid pronunciation and reduce ambiguity from sentences, such as the rule driven by the pronunciation of neighbouring vowels. For example, if a word ending in a vowel is followed by a vowel-initial word, morphophonemic rewrite rules are applied to change the spelling to aid pronunciation (Ó Siadhail, 1989). Examples 3 and 4 show eclipsis and h-prefixing respectively being applied to prevent vowel elision.

(3) **Eclipsis**

(i + vowel) → (in + vowel)
i Éirinn → i**n** Éirinn
'in Ireland'

(4) **h-prefix**

(le + vowel) → (le + h+vowel)
le úll → le **h**úll
'with an apple'

3.1.2 MT Errors with Grammatical Case in Irish

The second type of error the APE module is designed to correct arise due to the system's occasional incorrect choice of grammatical case. Modern Irish includes three main grammatical cases: nominative, genitive and vocative. In Irish, nouns are marked with case through various morphological changes such as lenition (e.g. *an buidéal* 'the bottle' → *dath an b**h**uidéil* 'colour of the bottle'), eclipsis (e.g. *na fir* 'the men' → *foirgneamh na **bh**fear* 'the men's building), and slenderisation or broadening of consonants (e.g. *an dochtú**ir*** 'the doctor' → *ainm an dochtúr**a*** 'the doctor's name'). The nominative form is sometimes regarded as the 'common case' (Christian-Brothers (1962), Christian-Brothers

(1960)) as it also replaces the dative and accusative cases. While the dative case is not expressly marked in Modern Irish, definite nouns that are objects of prepositions still undergo an inflection process. This morphological change may also vary depending on dialect.

The Irish language has three main dialects – the Ulster dialect, Connacht dialect and Munster dialect. Inflection of definite prepositional objects (in the form of initial mutation) is realised through either lenition (Ulster dialect) or eclipsis (Connacht and Munster dialects) (Ó Siadhail, 1989). From a spelling standards perspective, the translators in the DAHG follow the standard orthography for Irish (An Caighdeán Oifigiúil (Rannóg an Aistriucháin, 1962)), which means they should be consistent within a document, given their chosen type of initial mutation. This means that, while MT output of a lenited form of prepositional object may in fact be grammatically correct, it often requires correction to ensure consistency. Through observation of the data at hand, we chose to consistently use eclipsis as the default for the APE. If the translator wishes to instead apply lenition in a given document, they have the option to then post-edit the text manually.

In some instances, the nominal prepositional object is directly translated as a unigram (i.e. without taking into context the other elements of the prepositional phrase such as preposition and determiner) resulting in the use of an incorrectly inflected form. This is likely to be the result of the MT system backing off to translate on a unigram basis due to data sparsity in the training data. Example 5[1] shows the editing step required in such cases. Our APE module, removes the need for this correction and ensures consistency by applying rewrite rules to capture the mapping between the two dialectal forms.

(5) **MT output:** *leis an p**h**róiseas pleanála teanga*
Post-APE output: *leis an **b**próiseas pleanála teanga*
'with the language planning policy'

In example 6, we show two rewrite rules, which inflect definite nouns following the prepositions *as* 'from' and *ar* 'on' to conform to the official standard spelling.

(6) **(PREP + DEF. ART + NOUN) → (PREP + DEF. ART + eclipsed NOUN)**
*as an baile → as an **m**baile*
'from the town'

*ar an geata → ar an **n**geata*
'on the gate'

Rule precedence The order in which the APE rules are applied are important. We apply the orthographic rules described in Section 3.1.1 ahead of the grammatical case rules described in Section 3.1.2. Example 2 shows the steps (1 & 2) of the APE module working together on the phrase *faoin gcathaoir* 'under the chair'.

(7) **(vowel-final-PREP + DEF.ART + NOUN) → (contracted-PREP/DEF.ART + eclipsed NOUN)**

1. **Contraction**
faoi an cathaoir → faoi**n** cathaoir

[1] Taken from actual system output.

2. **Eclipsis**
   faoi**n** cathaoir → faoin **g**cathaoir
   'under the chair'

The combination of vowels in *'faoi'* and *'an'* contract to form *'faoin'* (see example 7.1). The presence of *faoin* before an ecplipsable consonant in turn triggers an initial mutation (*'**g**cathaoir'* instead of *'cathaoir'* in example 7.2). Rule precedence is clearly important here so that the orthography component of the APE module is run before the case component, resulting in the output of the first set of rules triggering the need for the second set of rules.

As with any language, there are exceptions to these rules. For example, in some instances, the combination of both rules can produce non-grammatical character strings (e.g. *ngC*, *mbhF*). Therefore, a small number of 'clean-up' rules were introduced to prevent the module introducing such errors. See Example 8 for a list of these rules.

(8)
1. $ngc \rightarrow gc$
2. $ngC \rightarrow gC$
3. $mbp \rightarrow bp$
4. $mbP \rightarrow bP$
5. $mbhf \rightarrow bhf$
6. $mbhF \rightarrow bhF$

Currently this post-editing module alters 13% of sentences on average, with 4% of these sentences having both sets of APE rules applied.

4 Evaluation

In this section, we describe experiments carried out to evaluate the addition of our APE module. We summarise the training data used to train and test the MT system. We then highlight the BLEU score changes following the introduction of the APE module. In addition, we discuss our observation that improvements introduced by the APE from a post-editing perspective may not always be reflected in an increase in BLEU scores.

4.1 Experiment Set-up

Training Data Our training data comprises mainly data received from the DAHG. The Tapadóir project represents a specific use case for professional translators working in the Department of Arts, Heritage and Gaeltacht (DAHG). As the system is tailored to their specific translation demands, it is important that the MT output is of a certain domain and register. The type of text generally translated by this team comprises of annual reports, staff notices, public announcements, and so on. To achieve accurate domain-specific translation, we have worked closely with the translation team to ensure that we can retrain the system at regular intervals on text they have translate in the interim. This text is provided to us in the form of translation memory (TMX) files. Such a data format is easily

fed back into the MT system as it is well-structured, aligned, and does not require much cleaning or pre-processing. This data set is the most crucial component of the training corpus as it helps to tune the system to the text genre of the DAHG use case. Currently the Tapadóir training set benefits from 42,500 sentence-pairs of DAHG data.

To add to the domain-specific data, we also make use of two additional translation memories, (Digital Corpus of the European Parliament)[2] and DGT-TM[3] (Directorate General for Translation, Translation Memories). Together they provide us with 29,000 sentence-pairs of good quality data of a similar domain.

While parallel data from the DAHG, DCEP and DGT is extremely beneficial to the Tapadóir project, it also requires some support from general-domain data. To achieve this, we used the ILSP web-crawler (Papavassiliou *et al.*, 2013) [4] to gather parallel English-Irish data from websites. Websites containing public reference material were crawled in order to ensure (i) a high level of quality and (ii) close alignment to our domain as possible. Currently 10,000 sentence-pairs of this parallel data crawl are included in the training set.

In addition to this, we made use of some previously publicly available datasets: Corpas Comhthreomhar Gaeilge-Béarla (CCGB), a bilingual corpus crawled from the web[5] and 'Paradocs', a parallel English-Irish corpus of legal texts[6]. While this data did not reflect our domain accurately enough, it was, however, useful in the language model. CCGB and Paradocs contain 6,000 and 89,000 sentence-pairs respectively.

Test data A random sample of 1,500 sentence pairs received from DAHG were held out from the training set to form the test set. The test set is therefore domain-specific, and representative of the type of texts the system will be used to translate (letters, reports, press releases, etc.).

4.2 APE Results

In Table 1, we present BLEU scores for various data combinations before and after the APE module has been included in the Tapadóir pipeline evaluated on our held-out test set. The results show a modest increase in BLEU across the board when the APE module is applied to correction of errors. The maximum increase in BLEU scores occurs when the system is trained on the translation memory and crawled data combined of +0.1 BLEU. Although the increase is small, we believe the impact on translation quality to be more substantial than is apparent from the BLEU scores alone, as approximately 200 of the 1500 test set translations are changed by the APE. Therefore, a small-scale human evaluation of the sentences was carried out for translations of the best-performing model to investigate the precision of our rule application.

Sentence-Level Analysis To further analyse the performance of the APE, we conducted a sentence-level BLEU analysis, which brought to light several instances where the inclusion of the APE module triggered a decrease in BLEU, even though the sentence was in fact improved from a post-editing

[2] https://ec.europa.eu/jrc/en/language-technologies/dcep
[3] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
[4] Maligna Jassem & Lipski (2008) was used to align segments
[5] http://borel.slu.edu/corpas/index.html
[6] http://gaois.ie/crp/en/

System Training Data	No APE (BLEU)	APE (BLEU)
TM	42.21	42.28
TM + Crawled	42.24	42.33
TM + Paradocs	42.91	42.96
TM + Paradocs + Crawled	42.79	42.83
TM + (Paradocs)	43.11	**43.19**
TM + Crawled + (Paradocs)	**43.13**	43.18
TM + (Crawled)	42.89	42.99

Table 1: BLEU evaluations for the Tapadóir system trained on various combinations of the data available, with and without the APE module. Brackets indicate that the data was used to train the language model, but not the translation model.

perspective. In order to understand this conflict, the nuances of Irish grammar need to be understood first.

For example, where the translation from English included some words in French, and lenition was applied to the French words in the sentence. In Irish, however, foreign words should not be lenited. For example, *sa* 'in the' normally triggers lenition on words beginning with *b, c, d, f, g, m, p, s, t*. However, this rule cannot apply to non-Irish words (e.g. sa C**h**ôte d'Azur*). This type of incorrect use of lenition results in an error output in the APE.

An additional example occurs when the APE module is applied to the phrase given in Example 9, there is a decrease in BLEU from 25.93 to 25.68, yet the overall grammaticality of the sentence has been improved [7]. In this example, the reference translation for the phrase 'with my department's officials' is *le mo chuid oifigigh* 'with my own officials' (*chuid* does not trigger a h-prefix on *oifigigh*). However, the MT output is actually more exact than the reference translation: *le oifigigh* mo Roinne* 'with my department's officials', although it does still contain a grammatical error *oifigigh**. This machine translation, while matching the orthography of the reference translation (thus contributing to a higher BLEU score), is missing a h-prefix that should be triggered by the preposition *le* 'with'. The APE accurately corrects this error, resulting in an accurate and grammatical translation of the source text and removing the need for post-editing. However, the application of the APE rule lowers the BLEU score because of the increased edit distance from the reference translation. This is a clear example of how the BLEU metric can miss grammatical improvements in translation output. These differing analyses of automated translation are therefore worth considering in the case of MT evaluation.

(9) *Source:* the Minister said : "I recently met with my department's **officials**.."
Irish reference: dúirt an tAire: "bhí cruinniú agam le déanaí le mo chuid **oifigigh**"
Before APE: dúirt an tAire: "chas mé le déanaí le **oifigigh** mo Roinne.."
After APE: dúirt an tAire: "chas mé le déanaí le **hoifigigh** mo Roinne.."

BLEU decrease: 25.93 to 25.68

(10) *Source:* submissions received about the public advisory **process**...
Irish reference: aighneachtaí a fuarthas mar chuid den **bpróiseas** comhairliúcháin phoiblí...

[7] The words changed as a result of the APE module are highlighted in bold.

Before APE: aighneachtaí a fuarthas **faoi an próiseas** comhairliúcháin phoiblí...
After APE: aighneachtaí a fuarthas **faoin bpróiseas** comhairliúcháin phoiblí...

BLEU increase: 35.43 to 38.60

Example 10 [8], taken from MT output, shows the importance of rule precedence (see also example 7). The contraction of *faoi an* to *faoin* is carried out by the first set of rules in the APE module. The presence of the word *'faoin'* then triggers an eclipsis, mutating *'próiseas'* to ***'b**próiseas'*. Had the rules been applied in reverse order, this eclipsis would not have been triggered. The sentence-level BLEU score for this translation is increased from 35.43 to 38.60. Similar to example 9, the reference translation and automated output differ somewhat in their translation of 'about the public advisory process' (***mar chuid den*** *bpróiseas* vs ***faoin*** *bpróiseas*). Yet, in contrast to Example 9, both of these possible translations of 'about' trigger initial mutation of *próiseas*, and thus the APE results in an the increase the BLEU score.

5 Integration into the User Workbench

The use of technology in translation work-flow has changed considerably over the past two decades. Computer Assisted Translation (CAT) tools such as translation memories (TM) have been widely embraced by the translation community as they help to eliminate repetitive errors and increase consistency in terminology use (Garcìa (2006), Heyn (1998)). In more recent years, there has been a drive towards the integration of MT systems into the translator's work-flow. In general, MT does not aim to replace TM, but instead complement it.

When integrating a SMT system into an existing translation work-flow, it is important to consider translator experience or preconceptions of MT as it is widely acknowledged that there is still some resistance amongst the translation community towards using MT (Lingo *et al.*, 2013).

Fortunately, the in-house translation team were open to trying new types of technology and as a result, integration of MT into translators' daily work-flow has been practically seamless. Figure 1 is a screen-shot of the typical DAHG translator's view within SDL Trados Studio 15[9]. Within the workspace, translators are given a choice to post-edit output which has been found in the translation memory or generated by the Tapadóir MT system. The lower section shows the current segment being translated. The upper section (lines 1 and 2) show the sentence translation options for the current segment as presented by the TM (line 1, indicated by a 71% fuzzy match) and the MT system (line 2, indicated by AT (Automated Translation)).

Figure 2 shows the number of words translated by MT as part of the DAHG translators' work-flow during the period April-August 2015. The steady rise from month to month[10] indicates that the translators have responded positively to the inclusion of MT, and are embracing it in as part of their day-to-day workload.[11]

[8] The sentence was shortened for clarity in this example.
[9] http://www.translationzone.com/products/trados-studio/
[10] The dip in activity in July is a result of the Irish parliament summer break period.
[11] The total number of translated words per month is unfortunately unavailable to us at present.

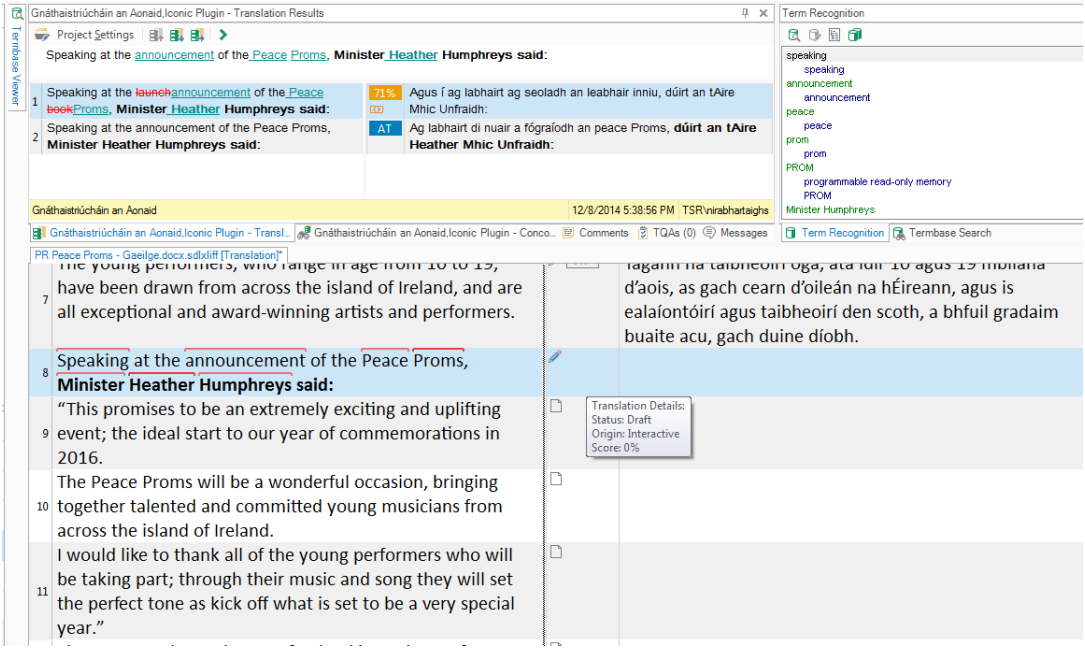

Figure 1: Integration of Tapadóir into SDL Trados Studio 15

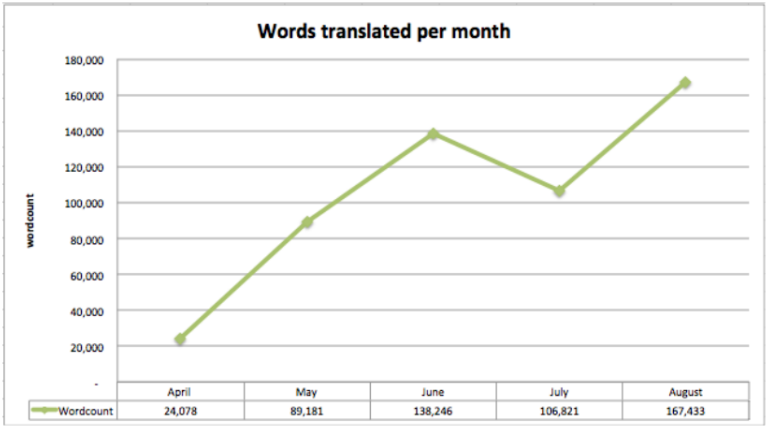

Figure 2: Words translated by Tapadóir in DAHG translation workflow

6 Conclusion

The Irish government Department of Arts Heritage and the Gaeltacht (DAHG) have funded the development of the Tapadóir SMT project to assist their in-house translators meet the growing

demand for English to Irish translation.

While we have evaluated the system using traditional MT evaluation metrics such as BLEU (Papineni *et al.*, 2002) in earlier work (Dowling *et al.*, 2015), we show here that we are also focusing on improving the post-editing user-experience as much as possible. We have described in this paper how, through analysis of examples of MT output inaccuracies (provided by DAHG translators) there is still plenty of room for improvement and we plan to embark on further development and improvement of the system.

We identified grammatical output errors that could easily be addressed by the introduction of an APE module. We also summarised the various nuances of Irish orthography and how to produce the rewrite rules to seamlessly include them in a post-processing step, thus reducing the need for translators to consistently correct simple mistakes.

Thus far the addition of this APE prototype has shown promising results. Therefore the expansion of this module is a natural next step. Future work will also include the adaption of resources such as rules contained in Irish language grammar-checkers (Scannell, 2008) to the domain-specific translation required by the Tapadóir project, as well as the application of factored models (Koehn & Hoang, 2007) to improve translation with respect to Irish language morphology. We also hope to adapt factor templates originally developed for deep-syntax transfer rules (Graham & van Genabith, 2010; Graham & van Genabith, 2008) to factored phrase-based models.

Acknowledgements

This work was part-funded by the Department of Arts, Heritage and the Gaeltacht (DAHG), and by the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21), and is also supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded by the European Regional Development Fund. We would also like to thank the three anonymous reviewers for their useful comments.

References

AVRAMIDIS E. & KOEHN P. (2008). Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*, p. 763–770.

BÉCHARA H., MA Y. & VAN GENABITH J. (2011). Statistical post-editing for a statistical mt system. In *Proceedings of Machine Translation Summit Conference*, volume 13, p. 308–315.

CHAHUNEAU V., SCHLINGER E., SMITH N. A. & DYER C. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1677–1687: Association for Computational Linguistics.

CHATTERJEE R., TURCHI M. & NEGRI M. (2015). The fbk participation in the wmt15 automatic post-editing shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 210–215, Lisbon, Portugal: Association for Computational Linguistics.

CHRISTIAN-BROTHERS (1960). *Graiméar Gaeilge na mBráithre Críostaí*. Dublin: M.H. Mac an Ghoill agus a Mhac, Tta.

CHRISTIAN-BROTHERS (1962). *New Irish Grammar*. Dublin: C J Fallon.

DOWLING M., CASSIDY L., MAGUIRE E., LYNN T., SRIVASTAVA A. & JUDGE J. (2015). Tapadóir: Developing a statistical machine translation engine and associated resources for irish. In *The 4th LRL Workshop: Language Technologies in support of Less-Resourced Languages*.

EL KHOLY A. & HABASH N. (2012). Orthographic and morphological processing for english–arabic statistical machine translation. *Machine Translation*, **26**(1-2), 25–45.

FRASER A. M., KNIGHT K., KOEHN P., SCHMID H. & USZKOREIT H. (2014). Statistical Techniques for Translating to Morphologically Rich Languages (Dagstuhl Seminar 14061). *Dagstuhl Reports*, **4**(2), 1–16.

GARCÌA I. (2006). Translators on translation memories: a blessing or a curse? *Translation Technology and its Teaching*, p. 97–105.

GRAHAM Y. & VAN GENABITH J. (2008). Packed rules for automatic transfer-rule induction.

GRAHAM Y. & VAN GENABITH J. (2010). Factor Templates for Factored Machine Translation Models. In *Proceedings of the seventh International Workshop on Spoken Language Translation*, p. 275–282.

HEYN M. (1998). Translation memories: Insights and prospects. In L. BOWKER, M. CRONIN, D. KENNY & J. PEARSON, Eds., *Unity in Diversity. Current Trends in Translation Studies*, p. 123–136: St Jerome Publishing.

JASSEM K. & LIPSKI J. (2008). A new tool for the bilingual text aligning at the sentence level. *Intelligent Information Systems*, p. 279–286.

KNIGHT K. & CHANDER I. (1994). Automated post-editing of documents. In *Proceedings of 12th National Conference on Artificial Intelligence*, p. 779–784.

KOEHN P. & HOANG H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 868–876.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, p. 177–180, Stroudsburg, PA, USA: Association for Computational Linguistics.

LINGO A. W., GREATER G. & UK M. (2013). Traditional and emerging use-cases for machine translation.

Ó MÚRCHÚ P. (2013). A grammar of modern irish: An annotated guide to graiméar gaeilge na mbráithre críostaí.

Ó SIADHAIL M. (1989). *Modern Irish: Grammatical structure and dialectal variation*. Cambridge University Press.

PAL S., VELA M., NASKAR S. K. & VAN GENABITH J. (2015). Usaar-sape: An english–spanish statistical automatic post-editing system. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 216–221, Lisbon, Portugal: Association for Computational Linguistics.

PAPAVASSILIOU V., PROKOPIDIS P. & THURMAIR G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, p. 43–51, Sofia, Bulgaria: Association for Computational Linguistics.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 311–318: Association for Computational Linguistics.

RANNÓG AN AISTRIUCHÁIN (1962). *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil*. Oifig an tSoláthair.

SCANNELL K. P. (2008). An gramadóir: A grammar-checking framework for the celtic languages and its applications. In *14th annual NAACLT conference*.

SIMARD M., GOUTTE C. & ISABELLE P. (2007). Statistical phrase-based post-editing. p. 508–515.

VIRPIOJA S., VÄYRYNEN J. J., CREUTZ M. & SADENIEMI M. (2007). Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI*, p. 491–498.

WISNIEWSKI G., PÉCHEUX N. & YVON F. (2015). Why predicting post-edition is so hard? failure analysis of limsi submission to the ape shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 222–227, Lisbon, Portugal: Association for Computational Linguistics.

Enrichissement de données en breton avec Wordnet

Annie Foret
IRISA & Université Rennes 1, 35042 Rennes Cedex, France
foret@irisa.fr

RÉSUMÉ

Nous décrivons une expérience d'enrichissement automatique de données en breton. Les données sont des unités de texte en breton. Certaines unités sont enrichies avec des synsets (synonym sets) de Wordnets en exploitant d'une part les ressources d'Apertium pour la paire de langues breton et français et d'autre part des ressources de type Wordnet pour le français et pour l'anglais. Le résultat peut-être visualisé et exploré de diverses manières : notre réalisation est sous forme de système d'information interactif. Notre approche repose d'une part sur des chaînes automatiques de traitements linguistiques en breton et en français et sur un environnement d'exploration de systèmes d'information logiques.

ABSTRACT

Breton data enrichment with Wordnet.

We describe an automatic data enrichment experiment in breton. The data consists in text units in breton. Some units are enriched with synsets (synonym sets) of wordnets exploiting Apertium resources for the language pair breton and french and Wordnet resources for french and english. The result can be viewed and explored in various ways : our proposal is in the form of an interactive information system. Our approach is based on an automatic tool chain for natural language processing in breton and french and a platform for logical information systems.

MOTS-CLÉS : breton, lexique, wordnet, système d'information, recherche d'information, sémantique.

KEYWORDS: breton, lexicon, wordnet, information system, information retrieval, semantics.

1 Introduction

Nous décrivons une expérience d'enrichissement automatique de données en breton et nous présentons la réalisation en cours de cette chaîne de traitement. Les données initiales sont des unités de texte en breton (Foret *et al.*, 2015) saisies manuellement à partir d'un ouvrage d'une série illustrée Les "Mille premiers mots en breton" (Kergoat *et al.*, 2007). Ce lexique, bien que restreint (la chaîne de traitement pourrait s'appliquer à des lexiques plus étendus), offre un vocabulaire de base, de référence pour le breton, et utile aux apprenants. Il a de plus été saisi en conservant l'organisation thématique du livre, et en ajoutant des indications spécifiques au breton : les *mutations* en breton sont les variations de la consonne initiale, (par exemple, l'expression *an daol* pour "une table" est indiqué en figure 1 par : *an d|taol*, le lemme du nom étant *taol*, et la mutation avec cet article *an* étant $d > t$) une description est fournie sur Le site ARBRES (`http://arbres.iker.cnrs.fr`) (Jouitteau, 2005) et une approche pour les gérer dans (Poibeau, 2014).

Certaines unités sont enrichies avec des groupes de synonymes, synsets (synonym sets) de Wordnet en exploitant d'une part les ressources GPL d'Apertium (Tyers, 2010) `apertium.org` pour la paire de langues breton et français et d'autre part des ressources de type Wordnet pour le français et pour l'anglais. Le résultat peut-être visualisé et exploré de diverses manières : notre réalisation est sous forme de système d'information interactif. Notre approche repose d'une part sur des chaînes automatiques de traitements linguistiques en breton et en français et sur un environnement d'exploration de données basé sur les systèmes d'information logiques.

Une chaîne de traitement construisant un système d'information lexical à explorer à partir d'articles en français et en anglais a été proposé par (Cellier *et al.*, 2016). Notre objectif général est assez proche et vise un système d'information interactif, utile et d'emploi sûr et aisé ; mais la chaîne de traitement présentée ici concerne le breton, pour lequel certains problèmes et traitements sont bien sûr spécifiques. Une autre caractéristique de cette réalisation est la possibilité de l'utiliser localement (hors connexion internet) [1].

2 Jeu de donnnées et présentation initiale

Nous considérons pour cette réalisation, un lexique saisi (Foret *et al.*, 2015) à partir du livre les "Mille premiers mots en breton" (Kergoat *et al.*, 2007).

Dans (Foret *et al.*, 2015), le lexique est ensuite chargé comme système d'information avec des facettes logiques, dans l'outil Camelis (version 1, accessible à `http://www.irisa.fr/LIS/ferre/camelis/`) : Camelis (Ferré, 2009; Ferré & Ridoux, 2004; Ferré & Ridoux, 2004) est basé sur une extension de l'analyse de concepts formels (Ganter & Wille, 1999) et peut gérer des hiérarchies de propriétés assez générales, il s'agit en ce sens d'un système de gestion de contextes et de propriétés logiques ; les couples propriétés et objets les vérifiant forment un *treillis de concepts* comme une facette spécifique fermée ou ouverte selon le type et le niveau d'exploration et de filtrage/sélection choisis.

Rôles des fenêtres Camelis par rapport à un contexte. L'outil Camelis, chargé avec un contexte initial, présente trois fenêtres relatives à un contexte courant, qui évolue au fil des sélections dans ces fenêtres. Un tel contexte peut être hétérogène, il peut contenir plus de sortes d'objets et de propriétés, selon les préférences et les usages prévus.

Fenêtre d'objets : la partie droite présente les objets du contexte courant, par leur label.

Fenêtre de propriétés : la partie gauche indique les propriétés, organisées en arbres selon les relations entre les propriétés. Il s'agit d'un index cliquable qui permet de passer d'un contexte à un autre. Les cardinalités des liens/sous-contextes y sont aussi affichées.

Fenêtre de requête : la partie du haut contient une requête caractérisant le contexte courant : c'est une propriété satisfaite par tous les objets du contexte courant ; elle n'a pas besoin d'être saisie puisqu'elle est mise à jour automatiquement selon les sélections dans les deux autres fenêtres. L'utilisation ne nécessite pas de connaissance *a priori*, mais il est aussi possible de rédiger directement les requêtes.

1. y compris les synsets Wordnet, mais hormis les liens-action vers Babelnet via http

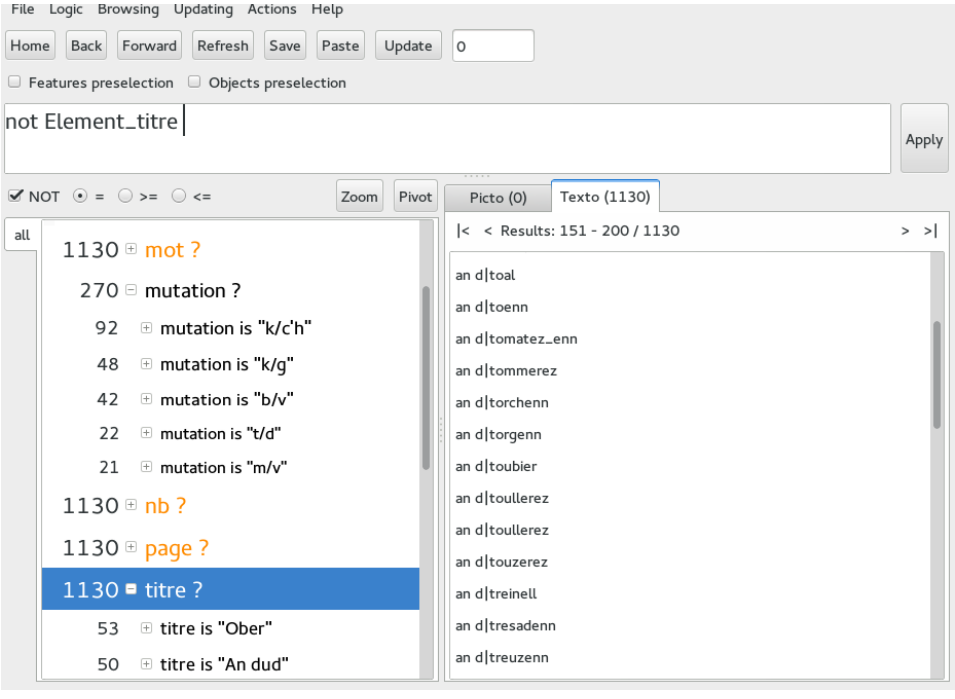

FIGURE 1 – Capture d'écran de Camelis, avec le contexte de départ

Pour ce petit lexique plusieurs sortes d'objets sont distinguées :

— des titres représentant des thèmes comme ("à la maison", "à l'école", etc.) ; ceux-ci pouvent comporter des variantes (c'est le cas de "Ar mezeg" pour le titre is "Ar medisin", désignant "le médecin"), voir la figure 2 ;

— et les expressions rattachées à un titre (correspondant à une page ou un intervalle de pages dans le livre) ; la figure 1 présente un contexte courant, avec à droite l'objet "an dltaol" suivi d'autres mots subissant la même mutation, la fenêtre en haut affiche la propriété sélectionnée (les objets courants ne sont pas des titres), et la fenêtre à gauche est une vue de l'arbre courant des propriétés (cliquable pour affiner la recherche).

Les objets sont étiquetés par leur classe, cette information peut être structurée comme ici en hiérarchie taxonomique et affichée par Camelis dans l'index de navigation à gauche, comme dans la figure 4.

Remarque. En poursuivant cet exemple de terme, on peut noter que Babelnet (utilisé plus loin) propose le breton dans sa liste de langages, mais ne reconnaît pas correctement le mot "mezeg", désignant "médecin", pourtant considéré dans un vocabulaire de base ; c'est ce que montre ce simple test : `http://babelnet.org/search?word=mezeg&lang=BR` qui donne un résultat, mais pour un nom de lieu.

3 Enrichissement avec les synsets de wordnet

Wordnet (wordnet.princeton.edu) est un réseau lexical d'abord développé pour l'anglais, qui sert aussi de référence pour d'autres langues, où des codes *synset* regroupent des ensembles de synonymes. Pour l'anglais, une forme XML est proposée par (Lapalme, 2014). Pour le français, nous avons exploité une autre ressource XML, appelée WoNef, accessible à wonef.fr et qui permet de relier les unités de sens dans les deux langues (par les codes synset).

Réalisation. Le contexte produit présente les unités de sens par leurs codes et ensemble de mots en français. Nous avons privilégié ici les mots en français, mais nous pourrions procéder de même pour associer la liste de mots en anglais, à partir de la version XML de Wordnet.

Cette construction utilise la paire br-fr de Apertium. Pour chaque langue ajoutée, les mots inconnus d'Apertium sont marqués par * (au début). Notons que la traduction d'une forme dictionnaire pour un terme du lexique breton n'est cependant pas toujours un lemme dans la langue cible : tel que le pluriel pour un nom collectif en breton sans suffixe visible [2], par exemple "ar *gwez*" pour "les *arbres*".

Actuellement, pour l'étape d'ajout des synsets comme propriété, nous considérons uniquement les expressions du lexique AvecArticle (cela sélectionne la plupart des termes du lexique : 1000 mots, dont des titres), ce qui amène à relier dans le sous-ensemble des noms (code n) de Wordnet. Cependant au stade actuel, l'association est partielle (pour 416 unités, avec en moyenne 7,5 synsets par unité) ; les termes sans synset en propriété comprennent notamment ceux sans traduction par Apertium br-fr (voir (Foret *et al.*, 2015)).

Remarques. Une première méthode consisterait à appliquer TreeTagger sur le mot français afin de relier ensuite le terme selon le lemme du mot français. Cette piste a été amorcée mais non poursuivie, il faudrait disposer auparavant de la catégorie pour améliorer les résultats. Une alternative est d'exploiter l'ensemble des outils de Apertium pour produire plus de détails dans chacune des deux langues. Une autre difficulté concerne les expressions composées.

4 Actions associées

Lien à Babelnet. Babelnet (https://en.wikipedia.org/wiki/BabelNet) est un réseau sémantique multilingue à très large couverture. Il a été construit automatiquement, notamment avec l'encyclopédie Wikipedia et Wordnet. Comme dans Wordnet les mots sont regroupés en ensembles de synonymes : les *label synsets*. En pratique, les codes synsets de Wordnet peuvent être utilisés en les préfixant par wn:, c'est un procédé que nous pourrions utiliser ici.

Nous avons actuellement réalisé le lien à travers la traduction en français, en deux points :

- dans le fichier d'information principal, où chaque objet est écrit par une ligne (voir figure 4), nous ajoutons *sur chaque ligne*, que le mot en français est un argument possible (pour une commande interactive), par exemple avec : {"fr", "cmdFR", "Couverture"} pour le nom "ar golo"

cet objet appartient dans le contexte au thème du voyage, de titre "beajiñ", voir figure 4 ;

2. voir http://arbres.iker.cnrs.fr/index.php?title=Noms_collectifs

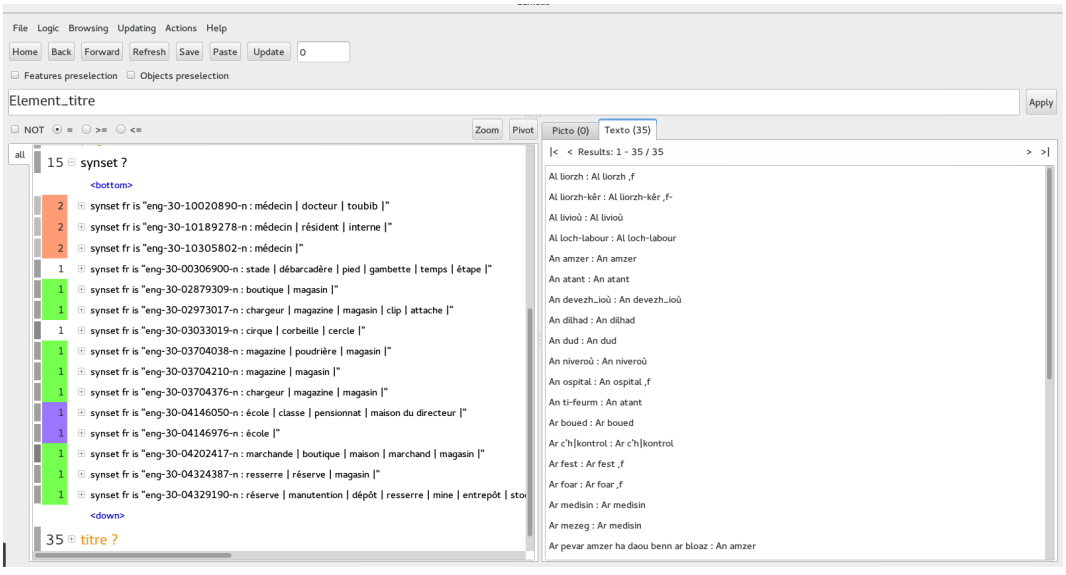

FIGURE 2 – Capture d'écran de Camelis, montrant les synsets obtenus pour des titres

```
mk {"id", "cmdLabel", "ar golo : Beajiñ"} {"titre", "cmdTopic", "Couverture"}
{"ref", "cmdExpr", "ar golo"} {"fr", "cmdFR", "Couverture"}
 "ar golo : Beajiñ" nb Apertium = 2,mot gb is "*Litkovrilo",mot fr is "Couverture",mot eo is "Litkovrilo",
 synset fr is "eng-30-06389398-n : fourchette | couverture |",
 synset fr is "eng-30-04605726-n : papier d'emballage | couverture | emballage | peignoir | papier |",
 synset fr is "eng-30-04605446-n : couverture | peignoir |",
 synset fr is "eng-30-04118021-n : moquette | tapis | plaid | couverture | petit tapis | carpette | descente de lit | postiche | moumoute
 synset fr is "eng-30-02849154-n : nappe | couverture | couvrante |",
 synset fr is "eng-30-01049685-n : masquage | couverture | recouvrement | plaque | enveloppe |",
...
MilleMotsBzh,mot Ref is "ar golo",mot Dico is "golo",AvecArticle
titre is "Beajiñ",Element_mot,page in [20,21]
```

FIGURE 3 – Extrait du fichier de contexte, pour le terme "ar golo" du thème voyage

- dans le fichier d'information, nous indiquons par *une ligne générique* le choix d'action, cette ligne comprend quatre parties, le mot action, le nom de l'action, la commande avec l'argument `$(fr)` et la propriété de filtrage de contexte (ici `all` pour l'ensemble) :

```
action "cmdFR" "firefox \"http://babelnet.org/search?word=$(fr)&lang=FR\"  & " all
```

En suivant le même principe et selon les préférences, d'autres actions peut être prévues, par exemple pour éditer ou interroger localement (par un outil XPATH, comme BaseX) le fichier de ressources XML Wonef (ou pour l'anglais Wordnet en version XML).

Ainsi, à une étape de navigation, en cliquant dans la fenêtre des objets, l'utilisateur verra les actions associées à un objet du contexte courant et pourra en déclencher.

5 Conclusions et perspectives

Nous avons proposé une chaîne de traitements qui enrichit des données en breton, avec des informations d'un réseau sémantique de type Wordnet. Les données enrichies peuvent alors être chargées

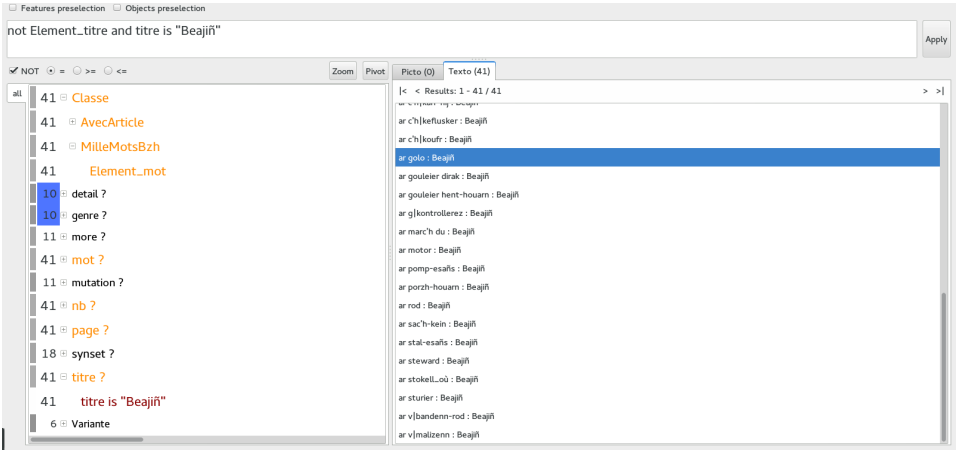

FIGURE 4 – Capture d'écran de Camelis, montrant les mots du thème voyage

comme système d'information (loqique) et être explorées de diverses façons : en multi-facettes par simples sélections successives dans un arbre de propriétés avec des liens vers d'autres ressources. Ce système pourra être proposé à un apprenant du breton pour faciliter sa recherche de l'information. Cela peut aussi servir au spécialiste, par exemple pour une forme d'évaluation d'une ressource par rapport à une autre (Foret *et al.*, 2015).

L'outil Camelis utilisé dans cette étude permet une exploration sûre (jamais de réponse vide, tout ce qui est accessible l'est en suivant l'arbre de navigation) généralisant en particulier les interrogations de type hiérarchique, et bases de données avec des outils de l'analyse de concepts logiques.

Un aspect important de ce travail avec des étapes automatisées est sa réutilisabilité : pour de nouvelles versions ; pour d'autres données à caractéristiques proches.

Il s'agit d'un travail en cours, [3] la couverture étant encore partielle en ce qui concerne l'association directe aux objets des synsets Wordnet (avec leur ensemble de mots en français).

Mise à jour. L'outil de gestion de contexte permet non seulement une navigation flexible, il est aussi prévu pour permettre la modification interactivement. Nous avons présenté ici un lien automatisé avec un code Wordnet (code français et anglais compatibles), mais ce type de lien pourrait être aussi repris manuellement, depuis l'outil Camelis interactif puis exporté en nouveau fichier de contexte (c'est aussi un fichier texte facilement modifiable, avec un objet décrit par ligne).

Web sémantique. Une version orientée *web sémantique* (Hitzler *et al.*, 2009) est une variante possible de ce travail. D'une part les ressources Wordnet et Babelnet ont des versions et des liens adaptés à SPARQL (`http://babelnet.org/sparql/`) D'autre part, les ressources dans ce format (RDF, ou un équivalent) peuvent être explorées avec l'outil Sparklis (Ferré, 2014), à la place du système Camelis.

3. une mise à disposition est prévue ici, en licence compatible GPL : `http://www.irisa.fr/LIS/softwares-fr`

Références

CELLIER P., FERRÉ S., FORET A. & RIDOUX O. (2016). Exploration des données du défi EGC 2016 à l'aide d'un système d'information logique. In C. DE RUNZ & B. CRÉMILLEUX, Eds., *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, volume E-30 of *RNTI*, p. 443–448 : Hermann-Éditions.

FERRÉ S. (2009). Camelis : a logical information system to organize and browse a collection of documents. *Int. J. General Systems*, **38**(4).

FERRÉ S. (2014). Expressive and scalable query-based faceted search over SPARQL endpoints. In P. MIKA & T. TUDORACHE, Eds., *Int. Semantic Web Conf.* : Springer.

FERRÉ S. & RIDOUX O. (2004). Introduction to logical information systems. *Inf. Process. Manage.*, **40**(3), 383–419.

FERRÉ S. & RIDOUX O. (2004). An introduction to logical information systems. *Information Processing & Management*, **40**(3), 383–419.

FORET A., BELLYNCK V. & BOITET C. (2015). Akenou-breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton. In *Actes de la Traitement Automatique des Langues Régionales de France et d'Europe*, Caen, France : Association pour le Traitement Automatique des Langues.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis - Mathematical Foundations*. Springer.

HITZLER P., KRÖTZSCH M. & RUDOLPH S. (2009). *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC.

JOUITTEAU M. (2005). *La syntaxe comparée du breton, une enquête sur la périphérie gauche de la phrase bretonne*. PhD thesis, Nantes, France.

KERGOAT L., AMERY H. & CARTWRIGHT S. (2007). *Les 1000 premiers mots en breton*. Skol an Emsav, 8 edition.

LAPALME G. (2014). Wordnet en XML-HTML. In *TALN 2014 - Atelier RLTLN*.

POIBEAU T. (2014). Processing mutations in breton with finite-state transducers. In *Proceedings of the First Celtic Language Technology Workshop*, p. 28–32, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

TYERS F. M. (2010). Rule-based breton to french machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation, EAMT10*, p. 174–181.

Insular Celtic Language Mark-up in WordPress

Mícheál Mac Lochlainn
Acadamh na hOllscolaíochta Gaeilge, Ollscoil na hÉireann Gaillimh,
Roisín na Mainiach, Carna, Condae na Gaillimhe, Éire
micheal.maclochlainn@oegaillimh.ie

RESUME

WordPress est une base populaire pour la création des sites Internet. Selon les statistiques actuelles, 38% des sites construits sur de tels systèmes de gestion de contenu (SGC) l'utilisent (Built With, 2016). Cependant, ses outils d'édition structurent le contenu des documents avec le balisage HTML, qui est sémantiquement compromis parce qu'il préfère le paradigme WYSISWYG (What You See Is What You Get) à l'approche WYSIWYM (…What You Mean), qui est sémantiquement significative. Músgraí WYSIWYM WP est un module d'extension WordPress qui remplace cette fonctionnalité WYSIWYG par sa propre fonctionnalité sémantiquement forte. L'éditeur rudimentaire de son module d'extension central est enrichi par des modules d'extension supplémentaires avec des fonctions spécifiques de balisage sémantique.
Cet exposé traite du développement de deux de ces modules d'extension qui facilitent l'annotation sémantique basée sur la linguistique, le balisage et le style de présentation de textes écrits dans les langues celtiques et leurs dialectes.

ABSTRACT

Insular Celtic Language Mark-up in WordPress
WordPress is a popular website creation framework. Current statistics indicate that 38% of websites built using such content management system (CMS) technologies are based on it (Built With, 2016). However, its editing tools structure document content with HTML mark-up that is semantically compromised, favouring the presentationally focussed WYSIWYG (What You See Is What You Get) paradigm over the semantically meaningful WYSIWYM (What You See Is What You Mean) approach. Músgraí WYSIWYM WP is a WordPress plug-in that replaces this WYSIWYG functionality with semantically sound WYSIWYM functionality of its own. Its plug-in core implements a basic WYSIWYM editing environment and additional plug-in modules extend this with domain-specific tools for rich semantic mark-up.

This paper discusses the development of two such plug-in modules, which facilitate linguistically-based semantic annotation, mark-up, and presentational styling of text written in the Celtic languages and in dialects thereof.

MOTS-CLÉS: WordPress, langues celtiques, WYSIWYM, balisage sémantique.
KEYWORDS: WordPress, Celtic languages, WYSIWYM, semantic mark-up.

1 WordPress, language and dialect

It should be stressed that WordPress and its graphical editor are general-purpose tools; excellent at what they do but not specifically designed for semantically rigorous linguistic mark-up. Nothing in this paper is intended, nor should it be taken, as critical of either. Regarding language, it is of course quite natural for languages and dialects to be in states of chronological and generational flux. So it should also be stressed that the emphasis here is on annotational accuracy, not linguistic purity.

2 'The Insular Celtic territories'

The plug-in extension modules discussed here reflect a design philosophy predicated on support for explicit, granular identification of any established, multi-generational L1 Insular Celtic language communities that exist or have historically existed in any geographic region. This support extends to any non-Insular Celtic languages around which overlaping L1 language communities have coalesced. For convenience, these regions are referred to here as 'the Insular Celtic territories'.

3 WordPress, semantic integrety and Músgraí WYSIWYM WP

Web pages are electronic plaintext documents, delivered to the browser in HyperText Markup Language (HTML). This language logically structures document content by marking-up the text with semantically meaningful plaintext tags: ** **this is a paragraph** **, **<q>**this is a quote**</q>**, **<cite>**this cites a creative work**</cite>** and so on. The browser uses a separate technlogy, Cascading Stylesheets (CSS), to manage the on-screen visual presentation of content elements based on their bounding tags; paragraphs are rendered in body text for instance, while quotes are placed in double-inverted commas and citations are italicised. Although the WordPress graphical editor writes technically valid HTML its semantic scope is quite limited (no means to mark-up quotes or citations, for example). Of greater concern, its **bold** and *italic* buttons work by applying the tags that signify **strong importance** and *emphatic stress* respectively, regardless of the actual semantic meaning intended by the visual styling; consider, for example, the typographical conventions regarding **snag words**, publication titles (such as *Séadna*) and taxonomic designations (such as *felis catus sapiens*). Finally, it allows the user to inappropriatly and inconsistently use visual, CSS-based styling to apply logical pseudo-structure. These things need not necessarily cause problems for sighted human readers but they can spoil the output of Braille readers and speech synthesisers as used by the blind, and can compromise the results of automated data mining. The Músgraí WYSIWYM WP plug-in core strips the WordPress graphical editor of buttons and options that facilitate this semantically compromised WYSIWYG markup and replaces them with semantically sound WYSIWYM-based alternatives. So for example, instead of italicising the publication title *Séadna* by applying inappropriate stressed emphasis (****Séadna****) it can be italicised by explicitly identifying it as a creative work (**<cite>**Séadna**</cite>**) or, with greater and therefore more useful specificity, as a book (**<cite class="leabhar">**Séadna**</cite>**). This mark-up is perfectly standards-compliant, and where the standards do not extend to the desired level of granularity it follows best current practice. Consequently, the semantic metadata are quite parsable and can be made available to the human reader simply by building the requisite functionality into the website back end. A number of plug-in extension modules exist, and more can be developed, to broaden the semantic scope of the editor by providing buttons and options to apply markup using controlled values relevant to specific domains.

4 The Extended Functionality (Celtic Languages) module

This module is a prototype for adding linguistic annotation functionality to WordPress. I created it after searches on keywords such as 'linguistics' in the WordPress plug-in archive returned very few results, none of which provided the desired functionality. Potential real-world applications include language learning and digital archiving of textual artefacts. The module adds options to the WordPress graphical editor, allowing the user to mark-up **<q>**quotes**</q>**, ** **paragraphs** **, **<blockquote>**blockquotes**</blockquote>**, ****spans**** (stretches of text within single paragraphs) and ** **divs** ** (document sections across multiple paragraphs) as being written in a given language. Supported Insular Celtic languages are Irish, Scottish, Manx and Canadian Gaelic; Welsh and Patagonian Welsh; Cornish; and Breton. Supported non-Insular Celtic ones are Irish, British and Canadian English; French and Canadian French; and Argentinian Spanish.

The module works by adding a **lang** attribute, which specifies the primary language of the tagged content, to the HTML tag: ** **This paragraph is in English** **, **<q lang="ga">**Is i nGaelainn athá an athfhriotal so**</q>** [this quote in Irish] et-c. The attribute's value must be a valid BCP 47 language tag, or the empty string (Faulkner, Eicholz, Leithead and Danilo, 2016). BCP 47 language tags, also called IETF tags, comprise one or more defined case-insensitive subtags, separated by hyphens. Subtags have fixed positions within the tag. Each has a maximum length of eight characters and may only include the characters A-Z, a-z and 0-9. (Phillips and Davis, 2016).

BCP 47 tags can be trivially simple, often consisting only of a **language** subtag (derived from an ISO 639 language code), as in the examples given perviously. A slightly longer form, deployed by this module, appends a **region** subtag (derived from ISO 3166-1 alpha-2 codes). For the Insular Celtic languages, the applicable values are **ga-IE** (Irish in Ireland), **gd-GB** (Scottish Gaelic in Britain), **gv-IM** (Manx Gaelic in The Isle of Man), **cy-GB** (Welsh in Britain), **kw-GB** (Cornish in Britain), **br-FR** (Breton in France), **gd-CA** (Scottish Gaelic in Canada), **cy-AR** (Welsh in Argentina). For the non-Insular Celtic ones, the values are **en-IE** (English in Ireland), **en-GB** (English in Britain), **en-IM** (English in The Isle of Man), **fr-FR** (French in France), **en-CA** (English in Canada), **fr-CA** (French in Canada) and **es-AR** (Spanish in Argentina).

The module also includes custom CSS to manage the visual presentation of these marked-up document elements in the browser, automatically highlighting languages with different colours and with shades of the same within language variations. At present, because of the questionable effects of too much conspicuous colour, this styling is only applied in the WordPress graphical editor and not in the site's public view. Uniquely and permanently styling text written in multiple dialects and languages could easily result in a particularly gross form of ransom note typography. With regard to disability and accessibility, there's also the question of how Braille readers might be expected to interpret such styling. However, the publicly presented Web document does remain fully marked-up, and is therefore parsable and linguistically searchable. It even provides speech synthesisers with a potential key for uttering strings in each language in a unique, linguistically appropriate voice.

5 The Extended Functionality (X Language annotation) module

The 'X' here is a place-holder for any of the supported Insular Celtic territory languages; extant or extinct. This module is a development of the preceding one but realised as a set of stackable sibling modules, each customised to support logically-grouped language-territory domains. Although I still consider the modules works in progress, the beta versions are stable and perfectly functional. For the

purposes of this discussion, they can be referred to as a single entity. The Irish instance will be highlighted here as it is currently the most heavily developed but commonalities will be discussed.

As with the previous module, this one works by applying the **lang** attribute, albeit with much greater specificity, and by applying presentational styling based on the attribute's values. The module's potential real-world uses are the same. Again, it was clearly necessary to work with controlled sets of attribute values but in this case some of them had to be composed ad hoc. This made it necessary to create a template for constructing them consistently. Purely for ease of reference, I have called this the **Geata Bán template**. Obviously, the starting point for constructing any such template is the full permitted BCP 47 language tag structure. This can be found in Ishida (2014):

[primary] language-extlang-script-region-variant-extension-privateuse

The **extlang** and **extension** subtags proved unsuitable for use in **Geata Bán** and were excluded. The rest were found useful, albeit to varying degrees.

5.1 Overview of the subtags used in Geata Bán

language : For this subtag to be valid it must specify one of the ISO 639 codes permitted in the Internet Assigned Numbers Authority (IANA) Registry (Ishida 2014). Codes are available for all the Insular Celtic languages, not only in their current forms but in a large number of historical ones.

script : This subtag specifies the script in which textual content is written. For it to be valid it must specify a script supported by the ISO 15924 standard (codes for the representation of names of scripts) (Ishida 2014) and must adhere to the four-character alphabetic code assigned to that script by the standard. The **script** subtag should only be used where it adds distinguishing information however (Phillips and Davis, 2016). For most of the Insular Celtic languages, the Latin script can be assumed, making it un-necessary. But as ISO 15924 also supports the Latin (Gaelic variant) script it was logical to include the subtag in **Geata Bán**, thus enhancing Irish and Scottish Gaelic instances of the module by making it possible to mark text as being set down in either the Gaelic or the Roman script. This might be particularly useful for preparing orthographically faithful digital transcriptions of original artefacts written or printed before the general migration to the Latin script.

region : For this subtag to be valid it must have a value drawn from either the UN M.49 standard, the specification of which is incompatible with the template's requirements, or from ISO 3166-1 alpha-2 (Phillips and Davis, 2016); as used by the previous module. Although ISO 3166-1 alpha-2 was found suitable for purpose, the only Insular Celtic territory recognised by it is the Republic of Ireland. This of course causes a significant loss of specificity (one that attenuated the previous module's accuracy). As specifying any other Insular Celtic territory here would violate BCP 47, it was necessary to devise a workaround using a **privateuse** subtag sequence; discussed presently.

variant : This subtag has great potential but is, for now, rather limited with regard to the Insular Celtic languages. BCP 47 specification states that « Variant subtags are used to indicate additional, well-recognized variations that define a language or its dialects that are not covered by other available subtags » (Phillips and Davis, 2016). Initially, this reference to dialects looked promising but inspection of the sole (Ishida, 2016) authoritative reference for language subtags, the IANA Language Subtag Registry (IANA, 2016) revealed that although values representing Cornish English and no less than four different Cornish orthographic standards are currently available, along with

ones for 'Scottish Standard English' and the 'Ulster dialect of Scots' (and also, interestingly, 'Scouse') there is no mention of any other dialectal variant from within the continua of the Celtic territories.

privateuse : The BCP 47 specification states (Phillips and Davis, 2016) that « Private use subtags are used to indicate distinctions in language that are important in a given context by private agreement ». But it also states that « Private use subtags are simply useless for information exchange without prior arrangement… Private use sequences… are completely opaque to users or implementations outside of the private use agreement… » It does however conceed that « …in some cases… the choice of [whether to use them] sometimes depends on the particular domain in question ». In a similar vein, Ishida (2014) states that « Because [**privateuse**] subtags are only meaningful within private agreements and cannot be used interoperably across the Web, they should be used with great care, and avoided whenever possible ».

While **Geata Bán** was never intended to support public interoperability, it is regretable that it could not be made publically accessible, within formal standards, while still offering the desired functionality. Unfortunately, a **privateuse** subtag sequence, albeit a rigorously defined one, was the only way I could find to work around the current limitations of the **variant** and **region** subtags. I did try to adopt existing standard codes and notations so that the each subtag in the sequence would at least be intuitively recognisable to third parties but even this proved problematic.

5.1.1 The privateuse subtag sequence: data structures and values

The sequence's overall structure is fairly stable, though not yet fixed, and data and values for territories, languages and dialects other than those of present-day Ireland are still rather basic. It is intended to be specific enough to carry accurate and meaningful geographic and linguistic data but flexible enough to avoid imposing inappropriate, one-size-fits-all data structures. The sequence conforms to the rules for **privateuse** subtags as given in BCP 47.
Abstracted, its structure is: **x-geataban-AAA-Bb-Cc-D(D|d)00-1111-2222**.

the x singleton : Required by BCP 47 to mark the start of a **privateuse** sequence.

geataban : A unique identifier subtag. Required by **Geata Bán** to avoid *extremely* unlikely but still not impossible clashes in real-world use. Must always have the lowercase value **geataban**.

AAA : The Celtic territory subtag. I prepared these three-letter uppercase codes to compensate for the lack of specificity in the **region** subtag: **EIR** (Éire), **ALB** (Alba), **EVN** (Ellan Vannin), **CYM** (Cymru), **KNW** (Kernow), **BRZ** (Breizh), **ANU** (Alba Nuadh), **EAP** (Eilean a' Phrionnsa), **TAE** (Talamh an Éisc) and **PTG** (Patagonia). The subtag must always have one of these values.

5.1.1.1 First- to fifth order national subdivision subtags

These provide geographic (and therefore dialectal) specificity currently unavailable in the **variant** subtag. The slightly vague nomenclature is quite deliberate as it allows for locally appropriate linguistic and territorial categorisations. I had hoped to use ISO 3166-2 codes here (representation of names of countries and their subdivisions) but these proved unsuitable, as did other standards such as FIPS 10-4, NUTS (levels 2 and 3) and Chapman.

Bb : First order national subdivisions. In the case of Ireland, these equate to its provinces. I derived two-letter PascalCase codes for these from their names: **Co** (Connachta), **La** (Laighin), **Mu** (An Mhumha) and **Ul** (Ulaidh). The subtag must always have one of these values.

Cc : Second order national subdivisions. In the case of Ireland, these equate to its counties. I took the county abbreviations given in An Brainse Logainmneacha (2007) as codes for these, adhering to the PascalCase convention used in that publication but stripping out diacritics, which would have violated BCP 47. Sample codes : **Ao** (Aontroim), **AC** (Átha Cliath), **TA** (Tiobraid Árann), **TE** (Tír Eoghain) and **UF** (Uíbh Fhailí). The subtag must always have one of these values.

D(D|d)00, **1111** and **2222** : Third- fourth- and fifth order subdivisions. For Ireland, these equate to the historical territorial units of barony, civil parish and townland. A full set of barony codes has been prepared but codes for civil parishes and townlands are still in development at time of writing.

6 Examples

Example markup, all fully interoperable up to the **x singleton**. The **privateuse** sequence then compensates for limitations in the formal standards:

<blockquote lang="br-FR-x-geataban-BRZ">A block quotation in Breton, in the Latin script, originating in France, more specifically in Brittany.**</blockquote>**

** **A paragraph in English, containing ****a span in Welsh, in the Latin script, originating in Argentina, more specifically in Patagonia****.** **

** **A paragraph in Irish Gaelic, in the Gaelic script, originating in the Irish Republic, in Ireland, in the dialect of Munster, more specifically in the dialect of County Waterford, even more specifically in the dialect of the barony of Na Déise.** **

7 Going Forward

It is intended to increase the module's linguistic specificity, to broaden its functionality (most obviously by adding granular search mechanisms) and to explore further styling techniques. It is hoped that the module may ultimately be useful to users deploying WordPress in language learning environments, on digital archiving or transcription projects, or on any online project where there is a desire or requirement to specify the language or dialect in which text is written.

The factors that caused the need for **Geata Bán** and its **privateuse** subtag sequence have less to do with BCP 47 itself than with limitations in the external ISO standards on which it draws. This situation may improve and if it does I shall update the template, and any WordPress plug-ins based on it, to migrate metadata out of the **privateuse** sequence and into the publically interoperable part of the **lang** attribute. Space limitations meant that it was only possible in this paper to give an overview of the module and template. Documentation detailing the full set of data rules and valid data values for **Geata Bán**; the Músgraí WYSIWYM WP WordPress plug-in; and working betas of the annotation (and other WYSIWYM) modules, which are free and open software, can be found online at www.gaoluinn.com/teic/bogearra/músgraí-wysiwym-wp-2/index.en.

References

BRAINSE LOGAINMNEACHA, AN. (2007). Gasaitéar na hÉireann. An Roinn Gnóthaí Pobail, Tuaithe agus Gaeltachta, [Online]. Available at http://www.logainm.ie/eolas/Data/Brainse/gasaitear-na-heireann.pdf [Accessed 14th April 2016].

BUILT WITH. (2016). CMS Usage Statistics. Built With, [Online]. Available at http://trends.builtwith.com/cms [Accessed 14th April 2016].

IANA. (2016). Language SubtagRegistry. IANA, [Online]. Available at http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry [Accessed 14th April 2016].

ISHIDA, I. (2014). Language tags in HTML and XML. W3C, [Online]. Available at https://www.w3.org/International/articles/language-tags/ [Accessed 14th April 2016].

FAULKNER, S., EICHOLZ, A., LEITHEAD, T. AND DANILO, A. (ED). (2016). HTML 5.1 Editor's Draft. W3C, [Online]. Available at http://w3c.github.io/html/dom.html#the-lang-and-xmllang-attributes [Accessed 14th April 2016].

PHILLIPS, A., AND DAVIS, M. (ED). (2016). Htags for Identifying Languages. IETF, [Online]. Available at http://tools.ietf.org/html/bcp47 [Accessed 14th April 2016].

Towards a lexicon of Irish-language idioms

Katie Ní Loingsigh
Fiontar, Dublin City University, Ireland
katie.niloingsigh@dcu.ie

RESUME

Vers un lexique d'idiomes de la langue irlandaise

Le présent exposé fournit un éclairage sur un lexique d'idiomes de la langue irlandaise rassemblés par *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977) et *Foclóir Gaedhilge agus Béarla* (Dinneen, 1927), deux références en matière de dictionnaires de gaélique élaborés au cours du vingtième siècle. Cette lexique d'idiomes est le fruit de recherches effectuées dans le cadre d'études doctorales sur les idiomes de langue irlandaise fondées sur les publications de Peadar Ó Laoghaire, auteur gaélique emblématique du 20ème siècle, et en est un sous-produit utile. La présente compilation offrant une ressource exploitable pour des analyses et recherches à venir en phraséologie, linguistique informatique et lexicographie gaélique.

ABSTRACT

Towards a lexicon of Irish-language idioms

This paper presents the development of a lexicon of Irish-language idioms as collected from *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977) and *Foclóir Gaedhilge agus Béarla* (Dinneen, 1927), the two primary Irish-English dictionaries compiled during the twentieth century. This lexicon of idioms is a beneficial by-product of doctorate research undertaken on Irish-language idioms from the published work of Peadar Ó Laoghaire, one of the foremost Irish-language authors of the twentieth century. The lexicon of idioms presented in this paper can be used as a resource for future analysis and research in the areas of phraseology, computational linguistics and Irish language lexicography.

MOTS-CLÉS: phraséologie, idiome, langue irlandaise, ressources lexicales, lexique.
KEYWORDS: phraseology, idioms, Irish language, lexical resources, lexicon.

1 Introduction

> If natural language had been designed by a logician, idioms would not exist. (Johnson-Laird, 1993 cited in Cacciari and Tabossi, 1993, vii)

There has been a steady increase in idiom-related research in the area of phraseology over the past three decades. Despite this development, there are still many terminological issues which cause difficulties in the classification and description of various linguistic units to the extent that "phraseology is 'bedivilled' (Cowie's description) by the proliferation of different terms for the same category and by conflicting uses of the same terms" (Pawley, 2001). Idioms are classified as a subset of a more general linguistic unit in phraseology which has been described using numerous

terms by various authors, e.g. 'multiword unit' (Grant and Bauer, 2004; Wulff, 2008); 'word-combination' (Zgusta, 1972); 'fixed expression' (Carter, 1987; Alexander, 1984); 'phraseme' (Mel'čuk, 1995); 'composite' (Howarth, 1996; Cowie, 1981); 'phrasal lexeme' (Moon, 1998; Lipka, 1990); 'phraseological unit' (Gläser, 1986); 'multiword expression' (Fernando, 1996) and 'conventional expression' (Pawley, 2001), etc. In this paper an idiom is defined as a type of phraseme or multiword expression (MWE) which has a figurative meaning in terms of its whole, or a unitary meaning that cannot be derived from the meanings of its individual components and whose components can only be varied within restricted definable limits. This description follows the definition of idioms as laid down in the literature (e.g. Abdou, 2012; Hanks, 2004; Howarth, 1998 and Fernando, 1996), for example:

- *Rud a chur ar an méar fhada*[1], 'to put something off indefinitely';
- *Muc i mála a cheannach*[2], 'to accept an offer or deal foolishly without being examined first';
- *Teacht aniar aduaidh ar dhuine*[3], 'to take someone unawares'.

This paper focuses specifically on the development of a linguistic resource for Irish, namely a lexicon of Irish-language idioms for reference and research. This lexicon can be used as a resource in the study of Irish-language idioms in phraseology but also in research related to phrasemes or multiword expressions in the area of natural language processing (NLP). A brief background to the creation of the lexicon is given in Section 2. Section 3 focuses on the areas of phraseology and computational linguistics, while the methodology and compilation of the lexicon is discussed in Section 4. Future work is discussed in Section 5 and the lexicon itself will be made available as an open data set available at http://www.gaois.ie/ under a non-restrictive license.

2 Background

The Irish language belongs to the Celtic branch of the Indo-European family of languages and is one of two official languages of Ireland, the other being English. The study of Irish-language idioms within the field of phraseology is a relatively new and underdeveloped area of research. Even though idioms have been collected and analysed as part of general lexicographic studies from the late nineteenth century onwards, there has been only one major academic study undertaken on Irish-language idioms, i.e. *A concordance of idiomatic expressions in the writings of Séamus Ó Grianna* (Ó Corráin, 1989). Additionally, it is acknowledged that in comparison to other official languages of the European Union, Irish-language technology is under-resourced (Lynn, 2014; Ó Raghallaigh and Měchura, 2014).

The idiom lexicon presented in this paper is a useful by-product of doctorate research which involved the creation of a database of Irish-language idioms from the published work of Peadar Ó Laoghaire (Ní Loingsigh 2016). The idiom database was created in *Léacslann*, an online platform used for building dictionary writing systems and terminology management systems as well as other lexicographic and reference applications (Měchura 2012) and is used in various research projects developed in Fiontar, Dublin City University (Ó Raghallaigh and Měchura, 2014). To facilitate the search and extraction of idioms from Peadar Ó Laoghaire's published work, a lexicon

[1] Literal meaning: 'to put something on the long finger'.
[2] Literal meaning: 'to buy a pig in a sack/bag'.
[3] Literal meaning: 'to come upon someone from the northwest'.

of idioms was manually compiled from the two primary Irish-language dictionaries of the twentieth century and it is this lexicon which is presented in this paper. The most common lemmas from this lexicon were categorized in order of frequency and were used to search a corpus of Ó Laoghaire's published work, which was compiled using *Sketch Engine* tools (Kilgarriff et al., 2004) and a morphological analyser and a part-of-speech tagger (Uí Dhonnchadha, 2009). The search methodology, which used "idiom-prone words" (O'Keefe, McCarthy and Carter, 2007), or the most frequent lemmas from the lexicon of idioms presented here, to search the corpus, will not, however, be examined in this paper.

3 Phraseology and NLP

The central topics of research on idioms in Europe during the past three decades have focused on five main areas: (i) syntax of idioms, (ii) semantics of idioms, (iii) pragmatics of idioms (including text-related modifications), (iv) cognitive approaches to idioms, and (v) contrastive research on idioms (including cultural specifics and cross-cultural comparison of idioms) (Dobrovol'ski and Piirainen, 2005, p.30). Additionally, there has been ongoing research undertaken on idioms as a subtype of MWEs in the field of NLP. Multiword expressions are recognized (Sag et al., 2002) as a key problem for the development of NLP technology and are underappreciated in the field at large. Colson (2015) highlights two serious shortcomings of computational phraseology:

1. There is no universally accepted algorithm for the automatic extraction of phraseology, especially not for ngrams larger than bigrams.
2. There is no consensus as to the proportion of set phrases in relation with the rest of the vocabulary: according to Jackendoff (1995), there are about as many fixed expressions as there are single words in the dictionary, but others (such as Mel'čuk 1995) hold the view that fixed expressions far outnumber single words. (Colson, 2015, p.7)

In a similar vein, the development of tools to recognize and extract Irish-language idioms as a type of MWE from corpora are lacking. This is not a problem specific to Irish-language idioms or Irish-language MWEs but a common stumbling block faced in NLP research on MWEs.

4 Methodology

In this paper, idioms are categorized as a subclass of MWEs following the definition set out by Sag et al. (2002 cited in Baldwin and Kim (2010, p. 269)):

(3) Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity.

While idioms are recognized as a central class of phrasemes or MWEs, other categories such as similes, proverbs, routine formulae and certain restricted collocations, which exhibit similar features to idioms, are not examined here.[4] The majority of idioms collected in the lexicon presented in this paper are primarily lexical items that display semantic and lexical idiomaticity as set out in Baldwin and Kim (2010). However, it is often difficult to distinguish certain idioms from

[4] See Dobrovol'ski and Piirainein (2005) for a more indepth overview of this topic.

other types of MWEs depending on factors such as a speakers age, cultural and linguistic background, etc. Phrasal verbs are not included as "they are such a large group... that they merit separate and thorough research of their own" (Grant 2003, p.19). Additionally, compound words and functional expressions such as proverbs, greetings, blessings, and terms of endearment are not included (Abdou 2012, Moon 1998). The lexicon of idioms presented in this paper provides a base for future research on MWEs and it is accepted that further analysis and scrutiny of the list is needed to ensure its completeness and to also verify the quality of the idioms collected.

According to the definition of idiom as set down in Section 1 of this paper, idioms were manually selected and recorded from the two primary Irish-language dictionaries of the 20th century, *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977) and *Foclóir Gaedhilge agus Béarla* (Dinneen, 1927). Each individual headword in both dictionaries was manually examined and any lexical items falling within the definition of an idiom were manually extracted and recorded. *Foclóir Gaedhilge agus Béarla* (Dinneen, 1927) was the first major Irish-English lexicographic work undertaken during the early twentieth century and is still a valuable resource for students and writers alike (Ó Murchú, 2005). The first edition of Dinneen's dictionary, which was published in 1904, is available as part of the Corpus of Electronic Texts (CELT) project in University College Cork in searchable PDF format[5] and the second more comprehensive edition, which was published in 1927 and used as a basis for analysis in this paper, is also available as a digitized and fully searchable online resource.[6] The foremost Irish-English dictionary available at present is *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977). Ó Dónaill's dictionary is still recognized as the principal orthographical source for the spelling of the language and "provides the most comprehensive coverage of the grammar and other aspects of words in Irish" (Foclóir Gaeilge–Béarla (Ó Dónaill), 2016). It is available as a searchable electronic resource as part of the *Leabharlann Teanga agus Foclóireachta*[7] project. Due to the lack of NLP resources available and the limited research undertaken on Irish-language MWE's, both dictionaries were examined manually and 5,437 idioms were collected. While this approach proved time-consuming, the resultant lexicon of idioms provides a comprehensive base for further research. This lexicon includes some duplication due the inclusion of variant forms of certain idioms used as usage examples in various entries.

> **Bealtaine**, *f.* (*gs.* ~, *pl.* -ní). May, Lá ~, May Day. Oíche Bhealtaine, eve of May Day. Mí na ~, month of May. Idir dhá thine Bhealtaine, in a dilemma. (Ó Dónaill, 1977, s.v. Bealtaine.)

FIGURE 1: Idiom: *idir dhá thine Bhealtaine*[8] ('in a dilemma').

> **Bealtaine**, *g.id.*, *f.*, (oft. pron. Beallthaine), the Irish May Festival, the month of May... idir dhá theine (uisce) lae Bealtaine, in a dilemma, from the practise of driving cattle between two fires with a view to their preservation. (Dinneen, 1927, s.v. Bealtaine.)

FIGURE 2: Idiom: *idir dhá theine* (*uisce*) *lae Bealtaine*[9] ('in a dilemma').

The canonical form of the idiom in Figure 1 and Figure 2 is recorded as *idir dhá thine Bhealtaine* ('in a dilemma') and the variant morphological form can be seen in Figure 2 which still retains the same idiomatic meaning, *idir dhá theine (uisce) lae Bealtaine* ('in a dilemma'). The idiom in

5 Available: http://www.ucc.ie/celt/Dinneen1sted.pdf.
6 Available: http://glg.csisdmz.ul.ie/index.php.
7 Available: http://www.teanglann.ie/ga/fgb/
8 Literal meaning: 'between two May fires'. Emphasis added.
9 Literal meaning: 'between two May Day (water) fires'. Emphasis added.

Figure 2 is displayed in non-standardized orthography. As *Foclóir Gaedhilge agus Béarla* (Dinneen, 1927) was compiled prior to the publication of an official standard of Irish (*An Caighdeán Oifigiúil* (Rannóg an Aistriúchán, 1958)), these idioms are recorded in non-standardized orthography. However, these idioms can be standardized using *An Caighdeánaitheoir* (Scannell, 2009), an application which annotates pre-standard words with standardized forms. A more detailed description of *An Caighdeánaitheoir* can be found in Uí Dhonnchadha et al. (2014).

The idioms in the lexicon presented in this paper are listed according to the headword under which they are recorded in the dictionary and have not been classified or analysed. As the idioms are recorded as usage examples in both dictionaries, they are often recorded as part of a longer sentence and not systematically by canonical form. However, certain syntactic structures are more common than others. For example, a number of idioms in the lexicon contain a verbal noun at the beginning of the idiom. Figure 3 and Figure 4 below show an example of this structure through the following example, *ag imeacht le haer an tsaoil* ('pleasure-seeking, leading a gay life'). Although part of this idiom, *aer an tsaoil* ('the pleasures of the world'), is given under the headword *aer* in *Foclóir Gaeilge-Béarla* (Ó Dónaill 1977) in Figure 3, the canonical form of the idiom containing a verbal noun is recorded as an individual example under the same headword. Additionally, in Figure 4, a non-standardized canonical form of the idiom, *ag imtheacht le haer an tsaoghail* ('leading a purposeless, improvident life'), is given as a usage example under the headword *aer* in *Foclóir Gaedhilge agus Béarla* (Dinneen 1927).

> aer, *m*. (*gs*. aeir)… 4. Gaiety, pleasure. ~ an tsaoil, the pleasures of the world. *Ag imeacht le h~ an tsaoil*, pleasure-seeking, leading a gay life. *Chaith sé a chuid airgid le h~ an tsaoil*, he spent his money on pleasure. (Ó Dónaill, 1977, s.v. aer.)
> FIGURE 3: Idiom: *ag imeacht le haer an tsaoil*[10] ('pleasure-seeking, leading a gay life').

> aer, *g*. aeir, *m*., the air, the sky, climate... ag imtheacht le haer an tsaoghail, wandering aimlessly about, leading a purposeless, improvident life. (Dinneen, 1927, s.v. aer.)
> FIGURE 4: Idiom: *ag imeacht le haer an tsaoil*[11] ('wandering aimlessly about, leading a purposeless, improvident life').

Mulhall (2010, p.1358) refers to a limited number of idioms that contain non-words or idioms which are also referred to as 'unique sublexical items' (Gouws, 1991) and gives an example of the use of 'amok' in the idiom 'to run amok' in English. A number of idioms that contain non-words are recorded in the lexicon presented in this paper and can be seen in Figure 5 and Figure 6 below. The idiom *húm ná hám* ('a sound, a move') is recorded as a usage example in both dictionaries under the headword *húm* ('a jot, a word').

> húm, s. (In phrase) Ní raibh ~ ná hám as, there wasn't a sound, a move, out of him. (Ó Dónaill, 1977, s.v. húm.)
> FIGURE 5: Idiom: *húm ná hám* ('a sound, a move').

[10] Literal meaning: 'Going with the pleasure of the world/of life'. Emphasis added.
[11] Literal meaning: 'Going with the pleasure of the world/of life'. Emphasis added.

húm, *m*., a jot, a word; ní fhéadfadh sé húm ná hám do bhaint as an gcloich, he could not get a move out of the stone; ní dubhairt sé húm ná hám, he remained neutral. (Dinneen, 1927, s.v. húm.)
FIGURE 6: Idiom: *húm ná hám* ('a sound, a move').

While Figure 5 includes the prompt 'In phrase' which signifies the phrasal or idiomatic use of the headword, this prompt is not used systematically throughout the dictionary and cannot be relied on to identify all idioms recorded in the dictionary. In addition, it can be seen from Figure 5 and Figure 6 that the idiom *húm ná hám* is used in its negative sense only with the negative verbal particle, *ní*, prefixing any use of the idiom.

As the lexicon of idioms presented in this paper was primarily compiled as part of doctorate research to identify idiom-prone words which were used to facilitate a corpus search, it has not yet been subject to indepth linguistic analysis. This paper does not provide information in relation to the primary linguistic features of idioms listed in the lexicon and further analysis would ensure its quality and its potential use as a gold standard lexicon of Irish-language idioms.

5 Future work

This lexicon is the only current comprehensive representation of Irish-language idioms collected from both written and oral sources during the twentieth century. It provides a general representation of Irish-language idioms and can be used as a foundation for any future development of a comprehensive dictionary of Irish-language idioms. Additionally, it can be used as a base for analysis of the syntactic structure of idioms as a subset of MWEs. This will further the research and analysis on Irish-language syntax generally and Irish-language idiom syntax specifically. For example, the area of syntactic parsing benefits greatly from research in multiword expressions. As Baldwin et al. (2004) note, "a lack of MWE lexical items in a precision grammar is a significant source of parse errors". It follows that statistical parsers, which are trained on syntactically annotated treebanks, perform better if a treebank has multiword expressions identified and annotated. To date, the Irish Dependency Treebank (Lynn, 2012; 2016) does not contain these annotations, primarily due to the lack of sufficient linguistic resources available such as MWE corpora, MWE linguistic analysis or sufficient identification and categorization of MWEs in digital format. This work therefore constitutes a step towards bridging that gap in knowledge between Irish linguistics and NLP tools.

The lexicon of idioms can also be used as a starting point for research on semantics and pragmatics of Irish-language idioms, modification in Irish-language idioms, frequency of various idioms, language change over time and cognitive approaches to idioms. While the lexicon presented in this paper focuses on idioms as a subset of MWEs, it can be used as a stepping stone towards further analysis of various MWEs, e.g. collocations, similes, proverbs, etc. Multiword expressions present obstacles to the development of NLP tools and this lexicon provides a resource which can be developed and utilised to advance the development of other resources for Irish in this area.

Acknowledgements

This research was undertaken with support from Fiontar, Dublin City University. I would also like to thank the useful comments from the three anonymous reviewers.

References

ABDOU A. (2012). *Arabic idioms: a corpus based study.* London: Routledge.

ALEXANDER R.J. (1984). Fixed expressions in English: reference books and the teacher. *English Language Teaching Journal* 38(2): 127-132.

BALDWIN T. and KIM S.N. (2010). Multiword Expressions. N. INDURKHYA and F.J. DAMERAU (eds.) *Handbook of Natural Language Processing*. 2nd ed. Chapman & Hall: United States of America, 267-293.

BALDWIN T., BENDER E.M., FLICKINGER D., KIM, A. and OEPEN, S. (2004). Road-testing the English resource grammar over the British National Corpus. *In Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC 2004*), Lisbon, Portugal, 2047-2050.

CACCIARI C. and TABOSSI P. (1993). *Idioms: processing, structure and interpretation.* Hillsdale: Erlbaum.

CARTER R. (1987). *Vocabulary: applied linguistic perspectives.* London: Allen & Unwin.

COLSON J-P. (2015). The contribution of corpus-based phraseology to translation studies: from experiments to theory. G.C PASTOR (eds.) *Computerised and corpus-based approaches to phraseology: monolingual and multilingual perspectives*, 6-9.
Online at: http://www.europhras2015.eu/europhras2015_bookoffullpapers/! [Retrieved 9 March 2016].

COWIE A.P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2(3): 223–235.

DINNEEN P.S. (1927). *Foclóir Gaedhilge agus Béarla, being a thesaraus of the words, phrases and idioms of the modern language*. 2nd ed. Dublin: Irish Texts Society.

DOBROVOL'SKIJ D.O and PIIRAINEN E. (2005). *Figurative language: cross-cultural and cross-linguistic perspectives.* Amsterdam: Elsevier.

FERNANDO C. (1996). *Idioms and idiomaticity.* Oxford: Oxford University Press.

FOCLÓIR GAEILGE–BÉARLA (Ó DÓNAILL). (2016). Online at: http://www.teanglann.ie/ga/fgb/ [Retrieved 2 April 2016].

GRANT L. and BAUER L. (2004). Criteria for re-defining idioms: are we barking up the wrong tree? *Applied Linguistics* 25(1): 38-61.

GRANT L.E. (2003). *A corpus-based inverstigation of idiomatic multiword units.* PhD Thesis. Victoria: University of Wellington.

GLÄSER R. (1986). *Phraseologie der englischen Sprache.* Tübingen: Walter de Gruyter.

GOUWS R.H. (1991). Toward a lexicon-based lexicography. *Dictionaries: Journal of the Dictionary Society of North America* 13: 75-90.

HANKS P. (2004). The syntagmatics of metaphor and idiom. *International Journal of Lexicography* 17(3): 245-274.

HOWARTH P. (1998). Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24-44.

HOWARTH P. (1996). *Phraseology in English academic writing: some implications for language learning and dictionary making.* Tübingen: Niemeyer.

KILGARRIFF A., RYCHLY R., SMRZ, P. and TUGWELL D. (2004). The Sketch Engine. G. WILLIAMS and S. VESSIER (eds.) *In Proceedings of the 11th Euralex International Congress*, Lorient, France, 6-10 July 2004, 105-115. Online at: http://www.euralex.org/elx_proceedings/Euralex2004/011_2004_V1_Adam KILGARRIFF, Pavel RYCHLY, Pavel SMRZ, David TUGWELL_The_Sketch Engine.pdf [Retrieved 3 November 2012].

LIPKA L. (1990). *An outline of English lexicology: lexical structure, word semantics, and word-formation.* Tübingen: Niemeyer.

LYNN T. (2016). Irish Dependency Treebanking and Parsing. PhD Thesis. Dublin City University.

LYNN T., FOSTER J., DRAS M. and TOUNSI L. (2014). Cross-lingual transfer parsing for low-resourced languages: an Irish case study. *In Proceedings of the First Celtic Language Technology Workshop* (*CLTW '14*), Dublin, Ireland, 41-49. Online at: http://www.aclweb.org/anthology/W/W14/W14-4606.pdf [Retrieved 8 April 2016].

LYNN T., ÇETINOĞLU O., FOSTER J., UÍ DHONNCHADHA E., DRAS M. and VAN GENABITH J. (2012), Irish treebanking and parsing: a preliminary evaluation. *In Proceedings of the 8th International Conference on Language Resources and Evaluation* (*LREC '12*), Istanbul, Turkey, 1939-1946. Online at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/378_Paper.pdf [Retrieved 20 May 2016].

MĚCHURA M.B. (2012). Léacslann: A platform for building dictionary writing systems. R.V. FJELD and J.M. TORJUSEN (eds.) *In Proceedings of the 15th Euralex International Congress*, Oslo, Norway, 7-11 August 2012, 855–861. Online at: http://www.euralex.org/elx_proceedings/Euralex2012/pp855-861 Mechura.pdf [Retrieved 23 November 2012].

MEL'ČUK I.A. (1995). Phrasemes in language and phraseology in linguistics. M. EVERAERT, E-J. VAN DER LINDEN, A. SCHENK, AND R. SCHREUDER (eds.) *In Idioms: structural and psychological perspective.* Hillsdale: Erlbaum, 167-232.

MOON R. (1998). *Fixed expressions and idioms in English: a corpus-based approach.* Oxford: Oxford University Press.

MULHALL C. (2010). A semantic and lexical-based approach to the lemmatisation of idioms in bilingual Italian-English dictionaries. A. DYKSTRA, and T. SCHOONHEIM, (eds.) *In Proceedings of the 14th Euralex International Congress*, Leeuwarden/Ljouwert, The Netherlands, 6-10 July 2010, 1355-1377. Online at: http://www.euralex.org/elx_proceedings/Euralex2010/129_Euralex_2010_9_MULHALL_A Semantic and Lexical-Based Approach to the Lemmatisation of Idioms in Bilingual Italian-Eng.pdf [Retrieved 18 October 2015]

NÍ LOINGSIGH K. (Forthcoming 2016). *Tiomsú agus rangú i mbunachar sonraí ar chnuasach nathanna Gaeilge as saothar Pheadair Uí Laoghaire*. PhD Thesis. Dublin City University.

O'KEEFE A., MCCARTHY M. and CARTER R. (2007) *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Ó CORRÁIN A. (1989). *A concordance of idiomatic expressions in the writings of Séamus Ó Grianna.* Belfast: The Institute of Irish Studies, the Queen's University Belfast.

Ó DÓNAILL N. (1977). *Foclóir Gaeilge–Béarla.* Baile Átha Cliath: An Gúm.

Ó MURCHÚ M. (2005). Dineen and Ó Dónaill. P. RIGGS (ed.) *Dinneen and the Dictionary 1904-2004*. Irish Texts Society: Dublin, 78-101.

Ó RAGHALLAIGH B. and MĚCHURA M. B. (2014). Developing high-end reusable tools and resources for Irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies. *In Proceedings of the First Celtic Language Technology Workshop*, Dublin, Ireland, August 2014, 66-70. Online at: http://www.aclweb.org/anthology/W/W14/W14-4610.pdf [Retrieved 24 March 2016].

PAWLEY A. (2001). Review of *Phraseology, linguistics and the dictionary*, by Anthony Paul Cowie. *International Journal of Lexicography* 14(2): 122-134.

RANNÓG AN AISTRIÚCHÁIN (1958). *Gramadach na Gaeilge agus litriú na Gaeilge: an caighdeán oifigiúil.* Baile Átha Cliath: Oifig an tSoláthair.

SAG I.A., BALDWIN T., BOND F., COPESTAKE A, and FLICKINGER D. (2002). Multiword expressions: a pain in the neck for NLP. *In Computational linguistics and intelligent text processing. Third International Conference*, (*CICLing 2002*), Mexico City, Mexico, 17-23 February 2002. Berlin: Springer, 1-15.

SCANNELL K. (2009). *Standardization of corpus texts for the NEID.* Presentation given in Saint Louis University, 22 May 2009. Online at: http://borel.slu.edu/pub/naaclt09.pdf [Retrieved 3 December 2012].

Uí DHONNCHADHA E. (2009). *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar.* PhD Thesis. Dublin City University.

Uí DHONNCHADHA E., SCANNELL K.,Ó hUIGINN R., Ní MHEARRAÍ E., NIC MHAOLÁIN M., Ó RAGHALLAIGH B., TONER G., MAC MATHÚNA S., D'AURIA D., Ní GHALLCHOBHAIR E, and O'LEARY N. (2014). Corpas na Gaeilge (1882–1926): integrating historical and modern Irish texts. N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK AND S. PIPERIDIS (eds.) *In Proceedings of the Workshop: Language resources and technologies for processing and linking historical documents and archives,* (*LREC 2014*), Reykjavík, Iceland, 26 May 2014. Online at: http://borel.slu.edu/pub/ria.pdf [Retrieved 9 June 2014].

ZGUSTA L. (1972). *Manual of lexicography.* The Hague: Mouton.

WULFF S. (2008). *Rethinking idiomaticity: a usage-based approach.* London: Continuum.

Universal Dependencies for Irish

Teresa Lynn [1, 2] Jennifer Foster[1]
(1) ADAPT Centre, School of Computing, Dublin City University, Ireland
(2) Department of Computing, Macquarie University, Australia
tlynn@computing.dcu.ie, jfoster@computing.dcu.ie

RÉSUMÉ

Dépendances universelles de l'irlandais

Les ressources linguistiques permettant aux études cross-langues de se développer sont très importantes pour les langues minoritaires telles que l'irlandais, car elles favorisent le partage des ressources pour palier au problème du manque de données. Le projet «Universal Dependencies » (UD) a pour but de faciliter les études cross-langues des arbres syntaxiques, des structures linguistiques et de l'analyse syntaxique. L'objectif principal de ce projet est de former un ensemble harmonieux d'arbres syntaxiques en utilisant un schéma d'annotations universelles. Dans cet article, nous présentons la transformation de l'arbre de dépendance syntaxique irlandais (IDT) (Lynn, 2016) au schéma d'annotations universelles du projet UD, suivie d'une description claire des changements structurels nécessaires à cette conversion. Le nouvel arbre est ainsi appelé « Irish Universal Dependency Treebank » (IUDT).

ABSTRACT

Language resources that enable cross-lingual studies have become increasingly valuable for lesser-resourced languages such as Irish, as they allow for easier sharing of resources, thus overcoming the problem of data scarcity. The Universal Dependencies (UD) Project[1] is an initiative aimed at cross-lingual studies of treebanks, linguistic structures and parsing. Its goal is to create a set of multilingual harmonised treebanks that are designed according to a universal annotation scheme. In this paper, we report on the conversion of the Irish Dependency Treebank (IDT) (Lynn, 2016) to a UD version of the treebank which we term the Irish Universal Dependency Treebank (IUDT). We report on the mapping of the IDT labelling scheme to the UD scheme, along with a clear description of the structural changes required in this conversion.

MOTS-CLÉS : Analyse syntaxique, irlandais, langue irlandaise, arbre de dépendance syntaxique, dépendances syntaxiques universelles, conversion, étiquettes.

KEYWORDS: parsing, Irish, dependency treebank, universal dependencies, mapping, labels.

1 Introduction

Dependency treebanks exist for many languages (e.g. Turkish (Oflazer *et al.*, 2003), Czech (Hajič, 1998), Danish (Kromann, 2003), Slovene (Džeroski *et al.*, 2006) and Finnish (Haverinen *et al.*, 2010)). However, these treebanks vary significantly, with labelling notations and linguistic analyses that are usually specific to that language, and often influenced by linguistic theories to which the developers

[1] http://universaldependencies.org/

subscribe. As a result, cross-lingual research is often hampered by variations that exist across the annotation schemes of treebanks. From a statistical parsing perspective, if the labelled training data for both languages is based on different annotation schemes, parser output in one language cannot be easily compared or transferred to another (Søgaard, 2011; McDonald *et al.*, 2011). McDonald *et al.* (2013) reported improved results on cross-lingual transfer parsing using 10 uniformly annotated treebanks. Lynn *et al.* (2014) also reported on similar experiments using the same treebanks to bootstrap parsing for Irish.

In October 2014, the Universal Dependency (UD) Project released guidelines to assist with the creation of new UD treebanks, or mappings and conversions of existing treebanks to a *new* universal scheme. This new annotation scheme is based on (universal) Stanford dependencies (de Marneffe *et al.*, 2006; de Marneffe & Manning, 2008; de Marneffe *et al.*, 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The UD scheme accounts for varying linguistic differences across languages by providing the option of defining language-specific label sub-types when the prescribed list of labels do not adequately cover all linguistic features of a given language. Nivre (2015) clearly explains the motivation behind the project. Ten treebanks were released in January 2015 including Czech, English, Finnish, French, German, Hungarian, *Irish*, Italian, Spanish and Swedish. Since then a large number of additional treebanks have been either (i) built from scratch or (ii) converted from existing treebanks to form new UD treebanks. To date[2], there are 54 treebanks representing 40 languages listed in the UD project.

We have mapped the Irish Dependency Treebank (IDT) (Lynn, 2016) to the UD scheme (v1) for purposes of cross-lingual studies and parser improvement. The IDT is a corpus[3] of Irish sentences that have been annotated with information on deep syntactic structure. This paper summarises the conversion and mapping of the IDT to the Irish Universal Dependency Treebank (IUDT), as part of the Universal Dependencies (UD) Project[4].

2 Mapping the Irish POS tagset to the Universal POS tagset

The UD part-of-speech (POS) tagset is an extension of the The Google Universal POS tagset (Petrov *et al.*, 2012) and contains 17 POS tags. The IDT was built upon a gold-standard POS-tagged corpus developed by Uí Dhonnchadha (2009), and is based on the PAROLE Morphosyntactic Tagset (ITÉ, 2002). The IDT's tagset contains both coarse- and fine-grained POS tags, both of which we map to the Universal POS tags (e.g. Prop Noun → NOUN). Note, however, that we only map to 16 of the UD tags as we do not identify auxiliary verbs in Irish to require the inclusion of `AUX`. We provide a mapping from the Irish POS tagset to the UD tagset in Table 1.

3 Universal Dependency Scheme

The IDT to UD treebank conversion required extensive work on dependency relation renaming, mapping and structural changes. We provide a mapping in Table 2 and describe the changes below.

[2]May 2016

[3]Current treebank size is 1020 trees with 23,684 tokens. See Appendix C of Lynn (2016) for additional statistics.

[4]`http://universaldependencies.org`

Part-of-speech (POS) mappings			
UD	**IDT**	**UD**	**IDT**
NOUN	Noun Noun, Pron Ref, Subst Subst, Verbal Noun,	ADP	Prep Deg, Prep Det, Prep Pron, Prep Simp, Prep Poss, Prep CmpdNoGen, Prep Cmpd, Prep Art, Pron Prep
PROPN	Prop Noun	ADV	Adv Temp, Adv Loc, Adv Dir, Adv Q, Adv Its, Adv Gn
PRON	Pron Pers, Pron Idf, Pron Q, Pron Dem	PART	Part Vb, Part Sup, Part Inf, Part Pat, Part Voc, Part Ad, Part Deg, Part Comp, Part Rel, Part Num, Part Cp,
VERB	Cop Cop, Verb PastInd, Verb PresInd, Verb PresImp, Verb VI, Verb VT, Verb VTI, Verb PastImp, Verb Cond, Verb FutInd, Verb VD, Verb Imper	NUM	Num Num
DET	Art Art, Det Det	X	Item Item, Abr Abr, CM CM, CU CU, CC CC, Unknown Unknown, Guess Abr, Foreign Foreign
ADJ	Prop Adj, Verbal Adj, Adj Adj	PUNCT	. . .. ? ? ! ! : : ? . Punct Punct
CONJ	Conj Coord	INTJ	Itj Itj
SCONJ	Conj Subord	SYM	(Abr)

Table 1: Mapping of the IDT's POS pairs (coarse fine) to the Universal Dependency POS tagset.

3.1 UD labels not used in the Irish UD Treebank

The following is a list of labels in the UD annotation scheme that do not apply to the Irish language:

- `aux`: This label is used for non-main verbs in a clause, i.e. auxiliary verbs. Examples in English are '*has* opened', '*will* be', '*should* say'. There are no equivalent auxiliary verbs in Irish.[5]

- `auxpass, nsubjpass, csubjpass`: These labels are used in passive constructions, respectively as: passive auxiliary verbs, passive nominal subjects and clausal passive subjects. There is no equivalent passive form in Irish (see The Christian Brothers (1988, p.120) and Stenson (1981, p.145)).

- `iobj`: In English, an example is 'Mary gave *John* the book'. There are no indirect objects in Irish, and constructions like these must follow the normal ditransitive verb structure using a preposition (i.e. 'Mary gave the book to John').

Some UD labels are not used in IUDT due to lack of instances observed in the data[6]:

- `reparandum`: This label is used to indicate disfluencies in text. The IDT data does not currently contain any disfluencies.

[5]Stenson (1981, p.86) notes that modal verbs such as *caithfidh* inflect as per regular verbs and are considered the main verb.
[6]This may be related to the well-structured, grammatical nature of the text in the IDT corpus (e.g. newswire, literature).

UD Dependency Label Mappings			
Universal	**Irish**	**Universal**	**Irish**
root	top	*foreign*	for
acl:relcl	relmod	*list*	quant †
advcl	comp †	*mark*	subadjunct, toinfinitive
advmod	adjunct †, advadjunct, advadjunct_q, quant †	*mark:prt*	advparticle, cleftparticle, particle, qparticle, vparticle
amod	adjadjunct	*name* ±	nparticle, nadjunct †
appos	app	*neg*	vparticle
case ±	padjunct †, obl_ag	*nmod*	aug, pobj †±, relparticle †
case:voc	vocparticle	*nmod:poss*	poss
cc ±	–	*nmod:prep*±	obl, obl2
ccomp	comp †	*nmod:tmod*	advadjunct, padjunct †, pobj †±, relparticle †
compound	nadjunct †	*nsubj*	relparticle †, subj, subj_q
compound:prt	particlehead	*nummod*	quant †
conj ±	coord	*parataxis*	comp †
cop ±	NEW	*punct*	punctuation
csubj:cop	csubj	*vocative*	addr
det	det, det2, dem	*xcomp*	xcomp
discourse	adjunct †	*xcomp:pred*	adjpred, advpred, npred, ppred ±
dobj	obj, vnobj, obj_q, relparticle †		

Table 2: Mapping of Irish Dependency Annotation Scheme to UD Annotation Scheme. † marks one-to-many mappings, and ± marks structural changes. The IUDT uses 26 of the 40 UD labels (and 9 Irish-specific sub-labels).

- `goeswith`: This label links to parts of a word that has been split, due to poor editing. There are no instances of this in the Irish data.
- `dep`: This catch-all label is used for unknown relations. We do not require this in the Irish data.

In addition, there are some UD labels that we have not included in the first release version of this treebank, but which we expect will be included in future releases:

- `expl`: There is no existential 'there' in Irish. However, we have not yet fully researched uses of other types of expletives in the IDT data (e.g. *tá sé soiléir go..* 'it is clear that ..').
- `mwe`: Multiword expressions are not marked in the IDT. There is not sufficient linguistic literature on this topic for Irish on which we could base a complete analysis of idioms or multiword units in the treebank. This analysis therefore remains as a future enhancement to the treebanks when such resources are available.
- `remnant`: This label is used for remnants in ellipsis, where a predicate or verb is dropped (e.g. 'Marie went to Paris and Miriam **[]** to Prague'). Instances of remnants in Irish are not easily identified. Further study is required to identify cases, if any, including a possible analysis of crossing dependencies.
- `dislocated`: This label is used for fronted or postposed elements that are not core grammatical elements of a sentence. Example, 'he must not eat it, *the playdough*'. We have not yet identified such cases in the IDT data.

3.2 Manual label updates

Some of the treebank conversion was automated with straightforward mappings. However, there were a number of one-to-many label mappings that required manual mapping. These instances are marked with † in Table 2 and discussed here.

relative particles: In the IDT, the relative particle *a* is attached to a relative modifier verb with the label `relparticle`. In the UD scheme, this particle is labelled with the syntactic role it plays in the relative clause.[7] The *a* can therefore fulfil the role of `nsubj`, `dobj`, `nmod` or `nmod:tmod`[8]. For example, *an rud deireanach a chonaic sé* 'the last thing that he saw' is shown in Figure 1. In this case *a* refers to *rud* 'thing', and therefore is labelled as a `dobj` of *chonaic* 'saw'.

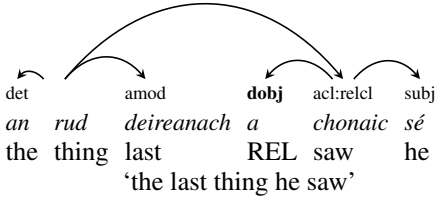

Figure 1: UD `dobj` relative particle analysis

quant → nummod, list, advmod Numerals and quantifiers are given more fine-grained descriptions in UD than the single IDT `quant` label. In addition, list numbering is represented by `list`.

comp → advcl, ccomp, parataxis The tokens labelled in the IDT with the closed complement label `comp` have been divided among three new labels. The UD labels are: `advcl` adverbial clause (normally connected with a subordinator such as *nuair* 'when', *má* 'if' etc); `ccomp` complement clauses that are normally introduced by the complementiser *go*, *nach*, *gur*, or quoting direct speech; `parataxis` labels two phrases or sentences set side-by-side without explicit linking through coordination or subordination, for example. Sometimes punctuation such as colons or semicolons connects the pairs. *Bhí an cál an-ghann; b'fheidir nach mbeadh i ngach baile ach aon gharraí amháin.* 'Kale was very scarce**;** maybe there would only be one garden in every town'.

nadjunct → compound, name The compound label is used for nominal modifiers. In Irish this could take the form of compounding (one noun modifying another) such as *deireadh seachtaine* 'weekend', or ownership *teach Mhichil* 'Michael's house'. Compounding can occur with a string of nouns as per the example in Figure 2.

The new label `name` is explained below in more detail in Section 3.3.

3.3 Structural Changes

Other labels required a manual annotation because they related to structural changes required in the treebank that were not easily automated. The following structural changes were made manually before the dependency labels were mapped to the universal scheme.

[7] This type of annotation that cannot be automated in the absence of additional data on the semantic properties of the element to which the relativiser refers.

[8] Irish language-specific label for temporal modifiers in nominal form.

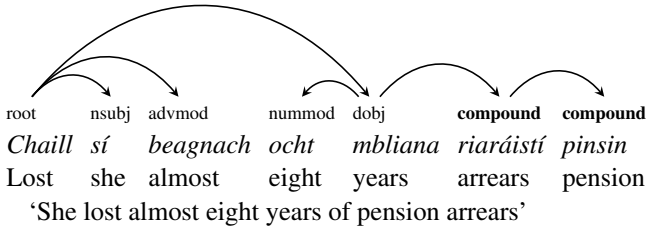

Figure 2: UD compounding analysis

coordination Significant changes were required to the analysis of coordination while mapping to IUDT. The IDT follows the Lexical Functional Grammar (LFG) (Bresnan, 2001) coordination analysis, where the coordinating conjunction (e.g. *agus* 'and') is the head, with each coordinate as its dependents, labelled as `coord` (see Figure 3). The UD annotation scheme, on the other hand, uses right-adjunction, where the first coordinate is the head of the coordination, and the rest of the phrase is adjoined to the right, labelling coordinating conjunctions as `cc` and subsequent coordinates as `conj` (Figure 4).

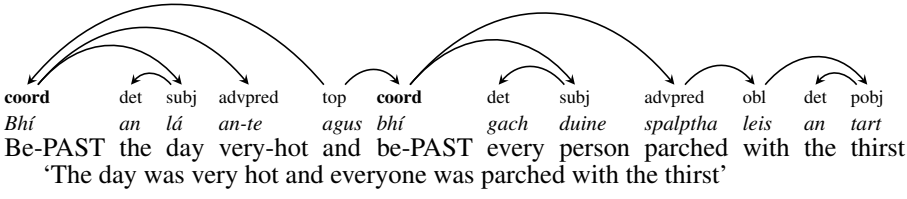

Figure 3: LFG-style coordination of the IDT

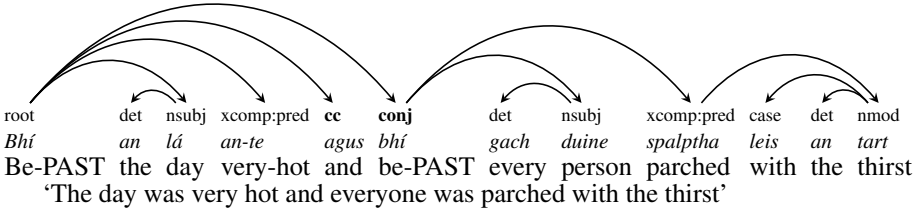

Figure 4: Coordination structure in the IUDT

subordinate clauses In the IDT, the analysis of the relationship between the matrix clause and a subordinate clause is similar to that of LFG: the subordinating conjunction (e.g. *mar* 'because', *nuair* 'when') is a `subadjunct` dependent of the matrix verb, and the head of the subordinate clause is a `comp` dependent of the subordinating conjunction (Figure 5). In contrast, the UD scheme marks the head of the subordinate clause as a dependent of the matrix verb, and the subordinating conjunction is a dependent of the subordinate clause (Figure 6).

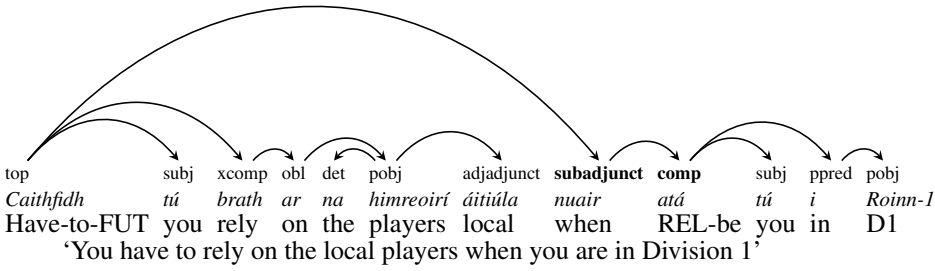

Figure 5: IDT subordinate clause analysis

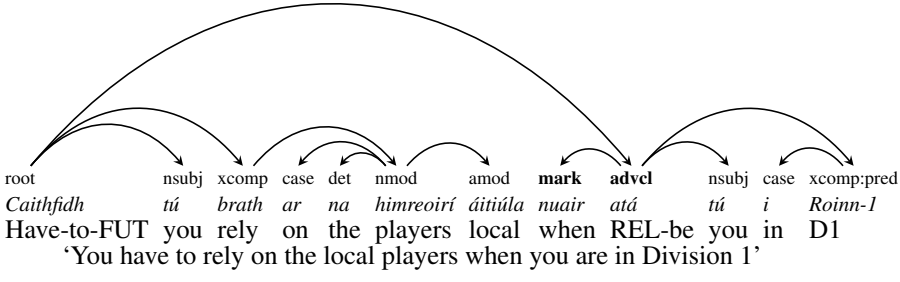

Figure 6: IUDT subordinate (adverbial) clause analysis

cop[9] In the IDT, the copula is treated similarly to a verb, and can function as the root of a sentence, or as the head of a dependency clause. However, the UD scheme analyses copula constructions differently. Instead, the predicate is regarded as the head of the phrase, and the copula is its dependent, as indicated by the `cop` label. This also applies to copula use in fronting or cleft structures. See Figure 7 and Figure 8 for comparison.[10]

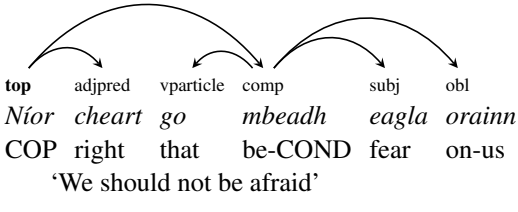

Figure 7: IDT copula analysis

name: The UD relation `name` is used with compounding proper nouns, typically for names of people,

[9]Note that Irish has two forms of the verb 'to be' – the copula and the substantive verb *bí*. Constructions using the substantive verb are not analysed using the UD `cop` label and are treated like regular verbs instead. For example, *tá sé fuar* 'it is cold'

[10]The labels have also been mapped between examples, but the structural change is of interest here.

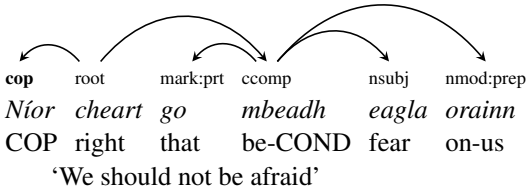

Figure 8: UD copula analysis

places, organisations and so on. In Irish, this not only includes surnames, but also surname particles such as *Mac, Mc, Ó, de, Uí* and *Ní*. In the IDT, the surname is the head noun, and its dependents can either be first names (`nadjunct`) or nominal particles (`nparticle`). See Figures 9 for example. However in the UD analysis, the first word is the head, modified by the rest of the words as `name`. See Figure 10 for comparison.

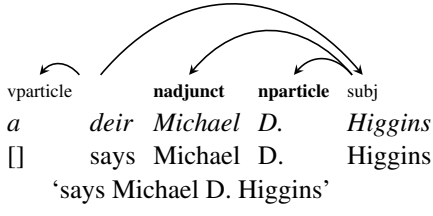

Figure 9: IDT name analysis

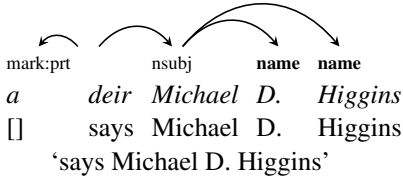

Figure 10: UD name analysis

nmod, case, xcomp:pred In the IDT, the preposition is the head of a prepositional phrase (PP). UD recognises the head noun of the object NP as the PP head. This affects the Irish treebank in a number of ways:

In the UD analysis, the head of regular preposition phrases (object of the preposition) is attached to the verb as `nmod` (formerly `pobj` in IDT). The preposition is a dependent of the object, and this relation is labelled as `case`. Compare Figures 11 and 12 to observe the difference in analyses.

Irish progressive aspectual phrases are constructed with the preposition *ag* followed by a verbal noun. The IDT regards *ag* as the head of the prepositional phrase, and thus the open complement label (`xcomp`) marks the relation between the matrix verb and the preposition. In the UD scheme however, the verbal noun is regarded as the head of the prepositional phrase. Compare Figures 13 and 14.

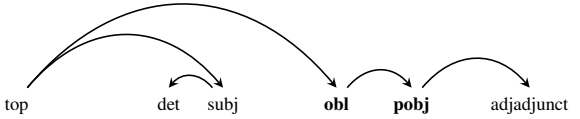

Figure 11: IDT prepositional phrase analysis

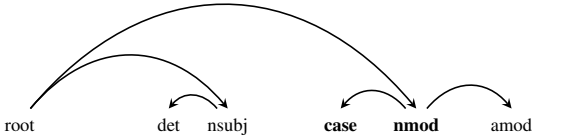

Figure 12: UD prepositional phrase analysis

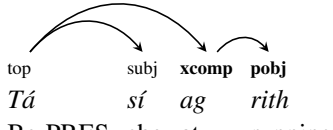

Figure 13: IDT progressive aspectual phrase analysis

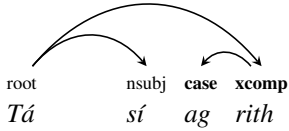

Figure 14: UD progressive aspectual phrase analysis

Prepositional predicates are labelled as `ppred` in the Irish Dependency Treebank. In keeping with the other PP analyses, the preposition is the head of the prepositional phrase. The IDT label `ppred` maps to `xcomp:pred` in the UD scheme.[11] In addition, the object of the preposition is now regarded as the head of the phrase. See Figures 15 and 16 for comparison of prepositional predicate analyses.

[11] The label `xcomp:pred` is an Irish-specific label, these language specific labels are discussed in Section 3.4.

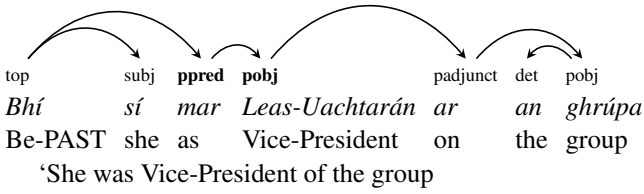

Figure 15: IDT prepositional predicate analysis

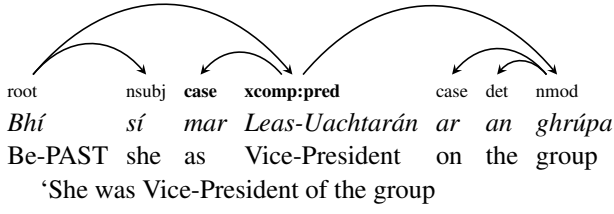

Figure 16: UD prepositional predicate analysis

3.4 Irish-specific relations

The UD scheme provides scope to include language-specific subtype labels. The label naming format is *universal:extension*, which ensures that the core UD relation remains identifiable, making it possible to revert to this coarse label for cross-lingual analysis. During the conversion of the IDT, we defined some labels required to represent Irish syntax more concisely. These labels are discussed below.

acl:relcl: This label is used for relative clause modifiers. We use this subtype label `acl:relcl` in cases where the head of the relative clause is a predicate (usually a verb), and is dependent on a noun in a preceding clause. It is also used in the English, Finnish and Swedish schemes. An example of this subtype used in the converted IUDT is in Figure 17.

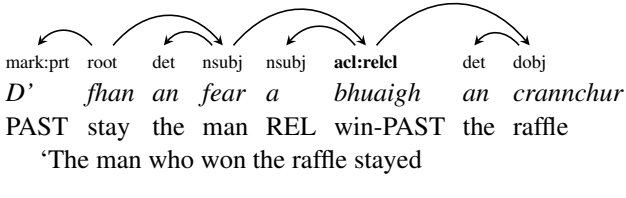

Figure 17: UD relative clause analysis

case:voc: The vocative particle *a* is a case marker in Irish and precedes an addressee. We therefore use the `case:voc` label for vocative particles. For example, *Slán **a** chara* 'Goodbye, friend'.

compound:prt We use `compound:prt` for verbal particle-heads, in order to distinguish them as particles as opposed to nominal compounds (e.g. *leagtha amach* 'laid out').

csubj:cop: The supertype label `csubj` indicates a clausal subject (a clause whose role is the subject of another). In English '[what she `said`] makes sense'. However, Finnish uses an additional specific

subtype label `csubj:cop` to indicate clausal subjects that act as a subject of a copular clause. We observed in the IDT data that clausal subjects in Irish are only ever subjects of copula clauses. For this reason we use only the subtype label `csubj:cop` for clausal subjects (see Figure 18).

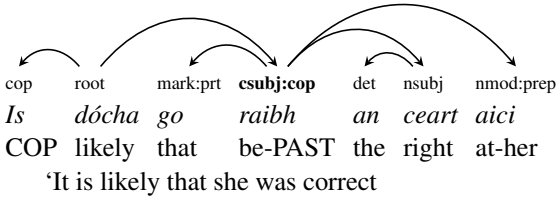

Figure 18: UD copular clausal subject analysis

mark:prt: We introduce a new subtype label `mark:prt` for adverbial particles, cleft particles, quantifier particles, comparative/ superlative particles, verb particles and days of the week particles.

nmod:poss: In Irish, possession is denoted by possessive pronouns (*mo, do, a, ár, bhur*). English, Finnish and Swedish use the subtype label `nmod:poss` to indicate possession, and we also adopt it for Irish. The pronoun is a dependent of the noun to which it denotes ownership. For example, *Chuir mé ceist ar **mo** mhúinteoir* 'I asked **my** teacher a question'.

nmod:prep: 16 of the most common Irish simple prepositions can be inflected to mark pronominal objects (e.g. *le* 'with' inflects as *liom* 'with-me') and are referred to as pronominal prepositions or prepositional pronouns.[12] In the UD scheme, where the object is the head of a PP, these inflected prepositions play nominal roles instead of prepositional roles.[13] We introduce the language-specific label `nmod:prep`, thus retaining information on the presence of the preposition within this synthetic form. An example is given in Figure 19.[14]

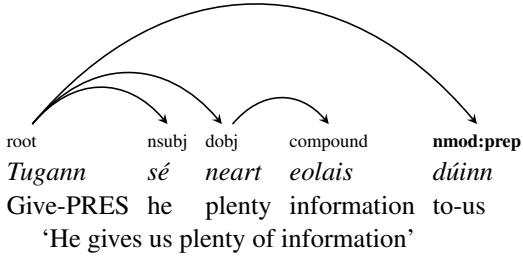

Figure 19: UD prepositional pronoun analysis

nmod:tmod: Temporal modifiers specifying time, in nominal form, are labelled as `nmod`. English also uses this subtype label. An example in Irish is *daoine a mhair na milliúin **bliain** ó shin* 'people who lived a million **years** ago'.

[12] Inflected prepositions were most frequently marked as either `obl` or `obl2` in the IDT.

[13] Their POS-tag remains `ADP`, however.

[14] Note that in some cases, prepositional pronouns behave like nominal modifiers of noun phrases. E.g. *an bheirt **acu*** 'the two **of them**'. These cases take the label `compound`.

xcomp:pred: The IDT uses the following fine-grained labels for predicates: `npred` (nominal), `adjpred` (adjectival), `advpred` (adverbial) and `ppred` (prepositional). These were typically used in copular constructions but are now no longer relevant in the UD, where the predicate heads the copular phrase. However, adjective, adverbial and prepositional predicates can also be arguments of the substantive verb *bí*. Therefore, we extend the open complement label to include the subtype `xcomp:pred`.[15] See Figure 20 for an example of an adjectival predicate.

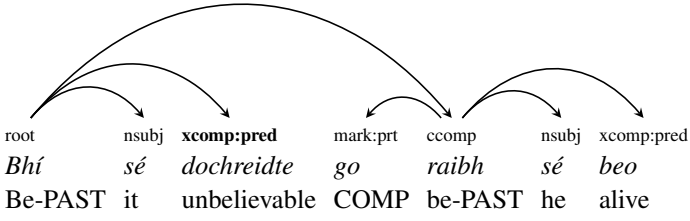

Figure 20: UD adjectival predicate analysis

4 Summary and Future Work

In this paper, we have summarised the conversion of the Irish Dependency Treebank (IDT) to a UD format (IUDT). We have described in detail the mapping and conversion process, including structural changes required, for the release of the IUDT as part of the Universal Dependencies project. We have also discussed linguistic analyses and motivations for choice of Irish language-specific label types. The Irish UD treebank (IUDT) is available to download under an open-source licence from The Universal Dependencies Project repository[16].

We have not discussed here the inclusion of morphological information in the IUDT as this still requires extensive documentation within the UD project. We plan to report on this at a later stage. In addition, as the IDT grows in size (a work in progress), we plan to extend the IUDT in parallel.

Acknowledgements

This work was funded by Science Foundation Ireland (SFI) through the CNGL Programme (Grant 12/CE/I2267) at Dublin City University and supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded by the European Regional Development Fund.

We are extremely thankful for input from the various UD contributors and in particular to Joakim Nivre for his advice on the Irish conversion effort. We would also like to thank the three anonymous reviewers for their useful comments and feedback.

[15] This follows the LFG use of xcomp (open complement) to represent predicates.

[16] v1.3 `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1699`

References

BRESNAN J. (2001). *Lexical Functional Syntax*. Oxford: Blackwell.

DE MARNEFFE M.-C., DOZAT T., SILVEIRA N., HAVERINEN K., GINTER F., NIVRE J. & D. MANNING C. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, p. 4585–4592, Reykjavik, Iceland.

DE MARNEFFE M.-C., MACCARTNEY B. & MANNING C. D. (2006). Generating typed dependency parses from phrase structure trees. In *Proccedings of the 5th international conference on Language Resources and Evaluation (LREC 2006)*, p. 449–454, Genoa, Italy.

DE MARNEFFE M.-C. & MANNING C. D. (2008). The Stanford typed dependencies representation. In *Workshop on Crossframework and Cross-domain Parser Evaluation (COLING2008)*, Manchester, U.K.

DŽEROSKI S., ERJAVEC T., LEDINEK N., PAJAS P., ŽABOKRTSKY Z. & ŽELE A. (2006). Towards a Slovene dependency treebank. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC2006)*, p. 1388–1391, Genoa, Italy.

HAJIČ J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. HAJIČOVÁ, Ed., *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, p. 12–19. Prague Karolinum, Charles University Press.

HAVERINEN K., VILJANEN T., LAIPPALA V., KOHONEN S., GINTER F. & SALAKOSKI T. (2010). Treebanking Finnish. In *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT-9)*, p. 79–90, Tartu, Estonia.

ITÉ (2002). PAROLE Morphosyntactic Tagset for Irish. Institiúid Teangeolaíochta Éireann.

KROMANN M. (2003). The Danish Dependency Treebank and the DTAG Treebank Tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT2003)*, p. 217–220, Växjö, Sweden.

LYNN T. (2016). *Irish Dependency Treebanking and Parsing*. PhD thesis, Dublin City University.

LYNN T., FOSTER J., DRAS M. & TOUNSI L. (2014). Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, p. 41–49, Dublin, Ireland.

MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TÄCKSTRÖM O., BEDINI C., BERTOMEU N. & LEE C. J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL '13*, p. 92–97, Sofia, Bulgaria.

MCDONALD R., PETROV S. & HALL K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 62–72, Stroudsburg, PA, USA.

NIVRE J. (2015). Towards a Universal Grammar for Natural Language Processing. In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, p. 3–16. Springer International Publishing.

OFLAZER K., SAY B., HAKKANI-TÜR D. Z. & TÜR G. (2003). Building a Turkish treebank. In A. ABEILLE, Ed., *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers.

PETROV S., DAS D. & MCDONALD R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 2089–2096.

SØGAARD A. (2011). *Data point selection for cross-language adaptation of dependency parsers*, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. Association for Computational Linguistics.

STENSON N. (1981). *Studies in Irish Syntax*. Tübingen: Gunter Narr Verlag.

THE CHRISTIAN BROTHERS (1988). *New Irish Grammar*. Dublin: C J Fallon.

UÍ DHONNCHADHA E. (2009). *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. PhD thesis, Dublin City University.

ZEMAN D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08).*, p. 213–218, Marrakech, Morocco.

Vocab : a dictionary plugin for web sites

Dewi Bryn Jones, Gruffudd Prys, Delyth Prys
Language Technologies Unit, Bangor University, Bangor, Wales, UK

{d.b.jones, g.prys, d.prys}@bangor.ac.uk

RÉSUMÉ

Vocab : un plugin dictionnaire pour les sites web
Ce document décrit un plugin dictionnaire, Vocab, qui peut être installé sur les sites Web plutôt que dans un navigateur. Le plugin permettre aux utilisateurs de passer le curseur de la souris au-dessus les mots, les entités à mots multiples et les phrases, et voir les entrées de dictionnaire pertinentes dans un pop-up sur la page web elle-même. Les filtres de Bloom et lemmatisation sont utilisés pour identifier les mots-vedette qui se trouvent dans une page. Vocab est disponible comme une ressource gratuite via un portail en ligne d'outils linguistiques. Les instructions sont faciles à suivre afin que les concepteurs de sites Web peuvent l'intégrer dans leurs propres sites Web. Vocab est utile comme aide à l'apprentissage pour les apprenants avancés et pour aider les utilisateurs couramment avec des mots techniques ou inconnus. Il a été principalement développé pour la traduction des mots et des phrases entre gallois et en anglais.

ABSTRACT

This paper describes a dictionary plugin tool, Vocab, that enhances websites by providing a rapid, integrated facility for users to hover the mouse cursor over or touch words, multi word entities and phrases and see relevant dictionary entries aggregated from a number of federated dictionaries as pop-up windows within the website itself. Bloom filters and lemmatization are used to identify dictionary entry headwords within a webpage's text. Vocab is made available as a free resource via an online portal of language tools, with easy to follow instructions on its deployment so that web designers can integrate and customize into their own websites. Vocab is useful both as a learning aid for advanced language learners and as an aid to vocabulary improvement. While primarily developed for word and phrase translation between Welsh and English it could be adapted for use with other language pairs through opportunities for collaboration

MOTS-CLÉS : gallois, dictionnaires en ligne, Les filtres de Bloom
KEYWORDS: Welsh, Online dictionaries, Bloom filters

1 Introduction

Vocab is an easy to install server-side tool that enables users to read the text in a website that they may not completely understand without having to resort to a translated version or to an external reference resource such as a dictionary

When activated, the plugin is able to highlight all words, multi word entities (such as technical terms) and phrases where they occur as entries from a number of dictionaries associated with the plugin. Users are able to simply hover over (with a mouse) or touch (using a touchscreen) any highlighted text in order to view the associated dictionary information in full. A user can also click through to search for related or similar words on the Welsh National Terminology Portal website[1].

It is available as a free resource via the Welsh National Language Technologies Portal[2] (Prys D., Jones., 2016). The Vocab plugin can currently be seen in use on popular Welsh language websites such as Golwg360[3] and the BBC CymruFyw[4] service. Vocab supports all modern desktop and mobile based web browsers.

Figure 1 shows a screenshot of the Vocab widget in action on the Golwg360 Welsh language news website. Recognized dictionary entities have been highlighted with subtle blue underlining. A popup with the dictionary definition for the multiple word phrase 'Cefn Gwlad' is displayed as a consequence of hovering over with the mouse.

Figure 1- Vocab in action the Welsh language newswebsite Golwg360

Other reading assisting plugins and products exist for a wide variety of languages (Shuttleworth, 2014) but only one or two support Welsh as well such as ReadLang[5], Geriaoeg[6]. Vocab is distinguishable from these offerings in that it is integrated into websites where users are more likely to use it and that it can also recognize multi word entities such as terms, placenames and phrases rather than only single words.

[1] http://termau.cymru
[2] http://techiaith.cymru/widgets/vocab/?lang=en
[3] http://www.golwg360.com
[4] http://www.bbc.co.uk/cymrufyw
[5] http ://readlang.com/cy/dashboard
[6] http://wiki.apertium.org/wiki/Geriaoueg

2 Vocab Architecture

Vocab's client is a Javascript library which operates within the containing web page. The Vocab server hosts dictionary data and provides RESTful APIs for dictionary search and lookup services. This means that Vocab exists in two decoupled parts, following the classic client-server model, so that processing is partitioned and distributed and communication can be completed rapidly resulting in a disruption-free user experience. The Vocab client is responsible for collecting all eligible texts and making multiple calls to the server. The server is responsible for recognizing the words, terms and phrases to be highlighted as well as providing the content of the pop-ups in the form of detailed dictionary and lexicographical information.

2.1 Vocab Client

This section provides a brief overview of the Vocab client's internal construction and operation.

A text nodes selector component is responsible for discovering all valid text nodes underneath any given HTML element. Valid text nodes are those considered not to be included within 'iframe', 'script', 'noscript', 'style', 'object', 'input', 'textarea' and meta HTML elements. In the case of mobile based browsers, text nodes with 'a' HTML elements for hyperlinks are also considered invalid so that their touch still activates navigating to another webpage or site.

A server communication component receives all gathered texts. First all texts are split using a simple regular expression into segments and grouped into suitably sized payloads for requesting the services of the Vocab server's REST API. Payloads are packaged as HTTP GET requests with an optimal maximum size of 2048 characters. Larger sized requests are possible with current browser/server expectations, but a reasonable maximum payload was defined so as to limit latency in progressing through the server communication component's queue of requests.

An HTML injection component receives responses containing markup which is able to update original text node locations with new markup that provide highlighting and further Vocab client functionality. A pop-ups handler component attaches event handlers to each recognized dictionary entity's mouseover or touch triggers. When triggered, the Vocab client makes another call to the Vocab server API for the corresponding dictionary entries to be displayed with the pop-up. The amount of event handler attachments typical for a reasonably sized webpage can frustrate the user with its lack of responsiveness, whatever the qualities of the linguistic resource. Vocab client thus attaches delegated event handlers[7] in order to avoid such issues.

2.2 Vocab Server

The Vocab Server is a component of the wider dictionary and terminologies infrastructure developed by the LTU to support its activities in terminology standardization and lexicographical resource building and dissemination. Its Welsh National Terminology Portal allows users to search

[7] http://javascript.info/tutorial/event-delegation

over 20 terminology dictionaries and connect easily to search on other similar resources and services on the web.

Vocab server uses at present two of the largest dictionaries, namely *Y Termiadur Addysg*[8] - a technical dictionary of approximately 45700 standardized terms for the National Curriculum in Wales, and *Geiriadur Cyffredinol Cysgair*[9] – a general language dictionary containing approximately 30,000 entries.

The Vocab server provides two REST API endpoints to the Vocab client. The first provides a means by which recognized dictionary entities are noted as such in any given string of text. The second provides a simple and efficient dictionary entries lookup for a given word or term.

Welsh, in common with other Celtic languages, is a moderately inflected language, where the first letter of a word can change according to certain grammar rules. This together with internal vowel changes and conjugated verbs using different word endings cause complications for dictionary lookups where a root or lemma form is required. The only significant use of a natural language processing component by Vocab is therefore that of a lemmatizer. The lemmatizer used was originally developed for the Cysill spelling and grammar checker (Hicks, 2004) which can recognise over half a million mutated, verb and plural forms to return lemma forms of all words. For example, it has the ability to recognise 'ellir' as the mutated impersonal present tense of the verb 'gallu'.

Once all lemma forms have been derived, the next step is for fast and efficient identification of headwords and terms from the dictionaries associated with the Vocab service. This involves iterating through the given text and looking up sub-sequences of words in one or more dictionaries. Such an algorithm is feasible for such a service if querying the database, where dictionary data resides on disk or over a network, is avoided since it introduces latency and unnecessary iterations and lookups are eliminated. These requirements were addressed by deciding to use two caches implemented with Bloom filters for each dictionary as in-memory caches.

Bloom filters (Bloom, 1970) are highly efficient data structures that are ideal for determining membership queries of a given set. False positives are possible but their use is still beneficial if given sufficient size and tolerable error rate parameters. Bloom filters have traditionally been used in the implementation of spell checkers (Broder, Mitzenmacher, 2004) and more recently in the efficient utilization of massive language models (Talbot, Osborne, 2007) where memory resources are restrictive. Bloom filters were seen as a sensible approach given the LTU's limited server capacity along a need for future proofing for any possible expansion of the service that would include 20 or more dictionaries from its National Welsh Terminology Portal.

The first cache is a Bloom filter of dictionary headwords and multiple word entities split into their sequences of words. For example a standardized education term prescribed in the Termiadur Addysg such as 'gallu i ddatrys problemau' (translation: *'ability to solve problems'*) would be cached into 4 separate lemmatized entries: "gallu", "gallu i", "gallu i datrys", "gallu i datrys problem"

[8] http://www.termiaduraddysg.org
[9] http://geiriadur.bangor.ac.uk

The recognition algorithm iterates through the text a word at a time and is able to look ahead with the first Bloom filter as to whether subsequent sequences of lemmas constitute a possible a dictionary term or phrase. When a multiple word term or phrase has been identified, the algorithm is able to skip by the last value of the look ahead counter. A second Bloom filter contains all dictionary headwords and the lemmatized versions of terms and phrases in their entirety. For example, only one entry exists for 'gallu i ddatrys problemau' i.e. 'gallu i datrys problem'. This filter serves a double check against any false positives that may have arisen from the first Bloom filter.

The Vocab server's second REST API endpoint is called upon the Vocab client when a user has hovered or touched over a highlighted range of text and replies with the result of a normal query on dictionary data residing in databases.

The Vocab Server keeps logs of all of its API usage. User privacy and anonymity is respected as described in the Vocab service's terms and conditions[10] so that no information can be used to individually identify the user. Vocab server logs consist of the webpage URL that a user has used with Vocab; each source text submitted for headword, term or phrase recognition along with the consequent result of recognized (or not) headwords, terms and phrases as well as each word or term the user has hovered or touched on for triggering popups that display further dictionary and lexicographical information.

3 Performance and Uptake

Section 2 described how Vocab's architecture was designed so as to ensure viable performance and usability despite its operation involving a substantial amount of communication and computation. The figures in Table 1 demonstrate that Vocab performs with sufficient performance that user's only experience a 'small perceptible delay' (Grigorik, 2013) when Vocab is initialized on a typical news webpage.

Webpage URL	**Word Count**	**Sentences**	**Total time**	**No. Of Requests**	**Average Request Time**
http://golwg360.cymru/newyddion/cymru/221283-siarad-cymraeg-gyda-chyfrifiaduron	2848	583	286 ms	9	31.7ms
http://www.bbc.co.uk/cymrufyw/36092710	6769	1443	323 ms	9	35.8ms
http://golwg360.cymru/blog	13932	1320	7.22 s	113	63.89ms

Table 1- Performance of Vocab with variously sized webpages

[10] http://techiaith.cymru/api/terms-and-conditions/?lang=en

In its first year of general availability, Vocab has been used on over 6300 distinct URL webpages. Also to date, users have hovered over or touched 209,000 recognised dictionary entries. This compares quite favourably with the usage statistics recorded for other websites and services provided by the LTU. (Prys D., Prys G., Jones D.B., 2015)

4 Future Work

A substantial amount of work has already been done on developing Vocab as a means of applying and disseminating terminological and lexicographical resources maintained by the LTU. Due to its success and uptake by significant and popular Welsh language websites a number of ideas have been suggested and opportunities identified for expanding its use to other languages and media that users consume. However all further work would be dependent on successfully obtaining further funding.

That said, the number of dictionaries that Vocab supports can be easily extended if the requirement ever arose and Vocab as such could be utilised on webpages to push technical terminologies only. The number of recognised dictionary entries could be made to be more focused by adding controls and expanding the Vocab server API to filter all but difficult or unusual words.

This idea has been recently considered for a version of Vocab that would operate on subtitles with browser-based catch up services or news video clips. In such a use case, where the viewer does not want translated subtitles and is not able to hover or touch a word, term or phrase he/she doesn't understand, a Vocab for Video would choose on behalf of the user and display in real-time any difficult word or term used in the source language subtitles.[11] Further research is required to identify the words that are perceived by users as being difficult or unfamiliar, and which are not. The content of the search logs may provide a useful indication of the words that are generally found challenging by users, and this may enable the Vocab server in future to suggest only those words that exceed a general threshold of unfamiliarity.

Acknowledgements

Whilst Vocab itself has received no external funding, we wish to acknowledge grant aid from the Welsh Government towards the establishment of the Welsh National Language Technologies Portal from which the Vocab is available for free to all website developers. We wish to thank Golwg360 and BBC Cymru Wales for their help and support in developing and testing Vocab.

[11] https://vimeo.com/160714756 (Vocab for Video)

References

BLOOM B. (1970) *"Space/Time Tradeoffs in Hash Coding with Allowable Errors."* Communications of the ACM 13:7 (1970), 422-426

BRODER A, MITZENMACHER M (2004) "*Network applications of Bloom filters: A survey"*, Internet Mathematics, vol 1 no. 4, pp. 485-509, 2004.

GREGORIK, I. (2013) *High Performance Browser Networking*. O'Reilly Media.

HICKS W.J. (2004) "*Welsh Proofing Tools: Making a Little NLP go a Long Way*." Proceeding of the1st Workshop on International Proofing Tools and Language Technologies. Greece: University of Patras

PRYS D., JONES D. B. (2016) "*National Language Technology Portals for LRLs: A Case Study"* Language Technologies in Support of Less-Resourced Languages, (LRL 2015) 28 November 2015, Poznan, Poland

PRYS D., PRYS G., JONES D. B. (2016) "*Quantifying the Use of Digital Welsh-language Language Resources"*. Language Technologies in Support of Less-Resourced Languages, (LRL 2015) 28 November 2015, Poznan, Poland

SHUTTLEWORTH M. (2014) "*Approaches to language learning: Blending tradition with innovation"* Presented at: Symposium on International Languages and Knowledge, Penang, Malaysia (2014)

TALBOT D., OSBORNE M. (2007) *"Smoothed Bloom filter language models: Tera-Scale LMs on the Cheap"* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 468-476, Prague, June 2007.