Code-switching in Irish tweets: A preliminary analysis

Teresa Lynn
ADAPT Centre, School of Computing
Dublin City University
Ireland
teresa.lynn@adaptcentre.ie

Kevin Scannell
Department of Computer Science
Saint Louis University
Missouri, USA
kscanne@gmail.com

Abstract

As is the case with many languages, research into code-switching in Modern Irish has, until recently, mainly been focused on the spoken language. Online user-generated content (UGC) is less restrictive than traditional written text, allowing for code-switching, and as such, provides a new platform for text-based research in this field of study. This paper reports on the annotation of (English) code-switching in a corpus of 1496 Irish tweets and provides a computational analysis of the nature of code-switching amongst Irish-speaking Twitter users, with a view to providing a basis for future linguistic and socio-linguistic studies.

1 Introduction

User-generated content (UGC) provides an insight into the use of language in an informal setting in a way that previously was not possible. That is to say that in the pre-internet era (where most published content was curated and edited), text that was available for analysis was not necessarily reflective of everyday language use. User-generated content, on the other hand, provides a clearer snapshot of a living language in natural, everyday use.

Analysis of minority language UGC in particular provides much insight into the evolution of these languages in the digital age. In some bilingual environments, the overwhelming dominance of a majority language can sometimes restrict and discourage the natural use of a minority language. Online environments, on the other hand, can offer a kind of 'safe space' in which these languages can co-exist and the minority language can thrive. Additionally, various interesting linguistic phenomena occur online that may be frowned upon in more formal settings. The present paper aims to investigate one such phenomenon among Irish-speaking users of the micro-blogging platform Twitter.

Code-switching occurs whenever a speaker switches between two (or more) languages in a multilingual environment. Negative attitudes towards code-switching have been documented widely in this field – in particular earlier beliefs that code-switching indicated a communicative deficiency or lack of mastery of either language. In fact, the phenomenon is now understood to be indicative of bilingual proficiency (Grosjean, 2010).

Solorio and Liu (2008) note that "when the country has more than one official language, we can find instances of code-switching". Given that Irish is the first official language of the Republic of Ireland, with English as the second,[1] and given the well-known existence of code-switching in the spoken Irish of the Gaeltacht regions (Hickey, 2009), it is unsurprising that Lynn et al. (2015) and Caulfield (2013, p. 208ff) report that code-switching is a common feature in Irish UGC. Our earlier work (Lynn et al., 2015), however, focused only on a part-of-speech (POS) tagging analysis of an Irish Twitter data set, without further exploration of the code-switching phenomenon that was observed. In fact, the English (code-switched) segments of tweets were given a general tag that

[1] Note that English is the more dominant language, with only 17.4% of the population reporting use of Irish outside the education system http://www.cso.ie/en/media/csoie/releasespublications/documents/population/2017/7._The_Irish_language.pdf

was also used to label abbreviations, items, out-of-vocabulary (OOV) tokens and other (non-English) foreign words.

Our current study is a continuation of our earlier work. We annotate, document, and analyse the specific nature of code-switching between English and Irish in our corpus of Irish language tweets (Lynn et al., 2015). With this we provide a basis for linguistic research into the way in which Ireland's official languages interact in an online social context. Our contributions are as follows: (i) an enhancement of the POS-tagged Twitter corpus in which English code-switched segments are annotated, (ii) a categorisation of the types of code-switching present in Irish tweets, and (iii) a quantitative report as to the relative frequency and use of English within Irish tweets.

2 Background and Related Work

2.1 Code-switching

Code-switching has been a focus of study for many years, particularly in the area of bilingualism (e.g. Espinosa (1917) and Muysken (1995)). Despite being an early topic of study in the field of computational linguistics (e.g. Joshi (1982)), interest in the computational study of code-switching has grown substantially in recent years with the increased availability of online UGC (Solorio et al., 2014; Molina et al., 2016). This area of study is applicable to many facets of natural language processing (NLP), including automatic language identification (Rosner and Farrugia, 2007) and POS tagging (Solorio and Liu, 2008), for example. In fact, Minocha and Tyers (2014) carried out some preliminary analysis on English-Irish code-switching in the context of automatic language identification. It is worth noting that the advances in the area of NLP for UGC represent a valuable contribution to the field of sociolinguistics, as NLP allows for easier and more efficient processing of large data sets than traditional manual methods (e.g. Nguyen et al. (2016)).

There is much debate in the literature about whether the correct umbrella term is *code-switching* or *code-mixing*, or in fact whether both refer to specific types of switching depending on where in the sentence it occurs. In our work, we use the term *code-switching* to cover all instances of the linguistic phenomenon that results in mixed-language text. We divide the instances of code-switching in Irish tweets into four main types:

Inter-sentential: where the switch occurs at a sentence or clause boundary:

(1) *Má tá AON Gaeilge agat, úsáid í! It's Irish Language Week.*
'If you speak ANY Irish, use it! It's Irish Language Week.'

Intra-sentential: where the switch occurs within a sentence or clause:

(2) *Ceol álainn ar @johncreedon on @RTERadio1 now.*
'Lovely music on @johncreedon on @RTERadio1 now.'

Word-level alternation: where the switch occurs within a word:

(3) *Bhfuil do kid ag **mixáil** Gaeilge agus English?*
'Is your kid **mixing** Irish and English?'

Bilingual text is, strictly-speaking, a special case of inter-sentential code-switching, in which the same content is provided in both languages in a single tweet. This is typical on Twitter for users whose followers can be divided into two groups – Irish speakers and non-Irish speakers.[2] Bilingual tweeting aims to be inclusive of a wider audience along with assisting learners in reading the Irish content. Due to the prevalence and importance of such examples, they are given special annotations in the resources described below.

(4) *Happy St Patrick's Day! La Fhéile Pádraig sona daoibh!*

2.2 Code-switching in Irish

Until recently, investigation into the use of code-switching in Irish has focused mainly on transcribed speech. In recent work, Ní Laoire (2016) noted that "[code-switching] has been underrepresented in Irish language corpora and in linguistic and dialectological description and analysis of Irish". In fact, much of the existing literature in this domain focuses on the impact of English as a dominant language in a bilingual environment (e.g. Stenson (1993)), in the context of raising concerns for the survival of the Irish language. In

[2] Stenson (1993) refers to the ability of all Irish speakers to speak English as "Universal Bilingualism"

the same vein, Hickey (2009) looks at the contrast between code-switching and borrowing, and its potential prevalence amongst the next generation of native Irish speakers. Her study focuses on such occurrences in unscripted speech of leaders of Irish-language pre-schools in Irish-speaking communities. Atkinson and Kelly-Holmes (2011) also investigate the nature of code-switching in spoken Irish and took a slightly different angle by looking at the use of English-Irish code-switching in comedy – with respect to the relationship between identity and language.

In terms of written text, Bannett Kastor (2008) provides a summary of the few examples of code-switching in Irish literature from the 17th-20th century, which in many cases incidentally can also be noted as being a conduit for comedy.[3] She notes that "multiliterate texts are constructed deliberately so that switch points or other points of linguistic contact within the text often signal additional, metaphorical levels of meaning which are coherent with the theme and/or other aspects of the work." However, such deliberate and planned code-switching differs from the nature of the switching behaviour we are concerned with here.

Interestingly, while code-switching is sometimes regarded in Ireland today (often negatively) as a 'modern' feature of the language, a number of studies have reported on the prevalence of Latin code-switching in Medieval Irish manuscripts, reflecting the multilingual environment in which medieval Irish monks and scribes worked (Dumville, 1990; Müller, 1999; Stam, 2017). These studies highlight code-switching as a natural feature of language use and as a linguistic activity that has continued across generations and across radically different linguistic environments. In the most recent study, Stam (2017) remarks, a propos of the current study, that "it appears that code-switching in writing and in speech are in some ways comparable, especially in informal textual genres". In our current study, we bridge several centuries from the analysis of medieval Irish-Latin code-switching to analyse and process Irish-English code-switching as it is used by today's Irish language community online.

2.3 Irish language on social media

Despite a relatively small population of speakers, the Irish language has a strong online presence on social media platforms such as Facebook, YouTube, Instagram, Snapchat, and Twitter (Lackaff and Moner, 2016). In fact, according to the Indigenous Tweets website, which curates tweets from indigenous and minority languages worldwide, there have been over 3 million tweets sent in Irish to date.[4] With the increased availability of user-generated Irish language content, it is unsurprising that there has been an increased interest in the application of technology to analyse Irish language use online, in order to gain insights into how the language is used (e.g. POS-tagging (Lynn et al., 2015), machine translation (Dowling et al., 2017) and sentiment analysis (Afli et al., 2017)).

3 Code-switching annotation

Of course, in order to carry out an in-depth analysis of the nature of code-switching amongst Irish speakers, sufficient data – in terms of quantity and richness of annotation – must be made available. Such data generally comes in two main forms. One is text that is based on recorded and transcribed speech. In terms of recorded speech, one is more likely to find instances of code-switching in spoken content that is spontaneous and non-scripted (such as the data that was used for an earlier Irish code-switching analysis by Ní Laoire (2016)). Another similarly unedited source is uncurated text such as that found in UGC, which is more likely to contain natural examples of code-switching than standard, well-curated text. In the following section, we describe the creation of our data set of tweets, which we have annotated for code-switching.

3.1 Data Set

Our starting point is the gold standard POS-tagged corpus of 1537 Irish language tweets from our earlier study (Lynn et al., 2015), which sought to provide a basis for NLP analysis of the use of the Irish language online.[5]

For this current study, we review the tags previously assigned to English tokens. In the initial corpus development, English tokens were assigned

[3] The data studied came from Irish language poetry, drama, fiction and nonfiction.

[4] `http://indigenoustweets.com`, figures as of July 2019.

[5] Please refer to cited paper for details concerning the POS tagset for Irish tweets.

the catch-all tag ‘G’ (‘general’), which is also used to label other foreign words (of which there are a few, e.g. Japanese), items, and abbreviations. We refine this annotation by now annotating English tokens separately.

Annotation takes place at three levels: (i) Irish part-of-speech tag level: The POS-tag is changed from ‘G’ to ‘EN’ for all English tokens. (ii) Code-switching tag level: The types of code-switching that are present (if any). As per our description in Section 2.1, these labels include INTRA (intra-sentential), INTER (inter-sentential), INTER-BI (bilingualism for the purposes of providing one message in both languages) and WORD (where the word contains code-switched morphemes). (iii) English part-of-speech tag level: The INTRA tags have been extended to identify the English POS (e.g. INTRA-V, INTRA-O, etc). These tags are explained in more detail in Table 1.

During annotation, we identified 41 tweets in the Lynn et al. (2015) corpus that contain both English and Irish words, but for which English was the matrix language. This type of issue often can arise within the context of language identification of tweets that contain instances of code-switching. As our interest is in English code-switching within otherwise-Irish language tweets, we have taken the viewpoint that these are errors in language identification, and have removed these examples from the corpus.

(5) *I added a video to a @YouTube playlist <URL>Sharon Shannon - Geantraí na Nollag 2008 - 25-12-08 <URL>*

(6) *Nuacht is déanaí - Twitter Competition - Help us Reach 20K!*

This leaves us with a data set of 1496 tweets annotated with POS and the above-mentioned code-switching tags.

4 Analysis

4.1 Tag distributions

We observe that 254 (16%) of the tweets in the data set contain some form of code-switching. Firstly, it is worth looking at the POS tag distribution across all POS tags in our data:

INTER 412 tokens (representing 43% of the English tokens) are used in an inter-sentential manner – that is, as strings of English that form separate phrases or sentences.

LOL	LOL	G	
–	–	,	
tell	tell	EN	INTER
ya	ya	EN	INTER
what	what	EN	INTER
–	–	,	
más	má	&	
féidir	féidir	N	
leatsa	le	P	
foclóir	foclóir	N	
Ioruaise	Ioruais	N	
a	a	T	
sheoladh	seol	V	
chúm	chuig	P	

INTER-BI 246 tokens (26% of the English tokens) represent code-switching for the purposes of providing comprehension for two groups of followers (Irish speakers and non-Irish speakers)

Lón	lón	N	
sa	i	P	
Spéír	spéir	N	
/	/	,	
MEN	MEN	EN	INTER-BI
AT	AT	EN	INTER-BI
LUNCH	LUNCH	EN	INTER-BI
FILM	FILM	EN	INTER-BI

WORD Interestingly, our data only contains two instances of word level code-switching. This ran counter to our intuition before examining the corpus data, as examples of this kind are heard in the spoken language quite frequently. The two tagged examples both use the Irish emphatic prefix ‘an-’ with an English root: *an-talent* ‘a lot of talent’; *an-time* ‘a great time’. It is possible, of course, to find other examples of this word level switching on Twitter by focused searching (e.g. examples of verbs with the gerund suffix ‘*-áil*’: *creepáil, buzzáil, textáil, snapchatáil, flirtáil*, etc.). However, our data suggest that the relative frequency of such types may not be as high as our intuition leads us to believe.

tá	bí	V	
an-talent	talent	EN	WORD
go	go	P	
deo	deo	N	

```
agaibh      ag      P
in          i       P
Éirinn      Éire    ^
```

INTRA(+EN-POS) 297 tokens (31% of the English tokens) are used in an intra-sentential manner, that is to say that these English tokens are inserted comfortably within the syntax of Irish phrases. Figure 1 shows the distribution of the EN POS tags for intra-sentential code-switching. This feature is of the most interest to us, and is therefore described in more detail in the next section.

```
Don't    Don't    EN INTRA-V
forget   forget   EN INTRA-V
to       to       EN INTRA-P
use      use      EN INTRA-V
the      the      EN INTRA-D
cúpla    cúpla    D
focal    focal    N
ag       ag       P
obair    obair    N
agus     agus     &
ar       ar       P
scoil    scoil    N
```

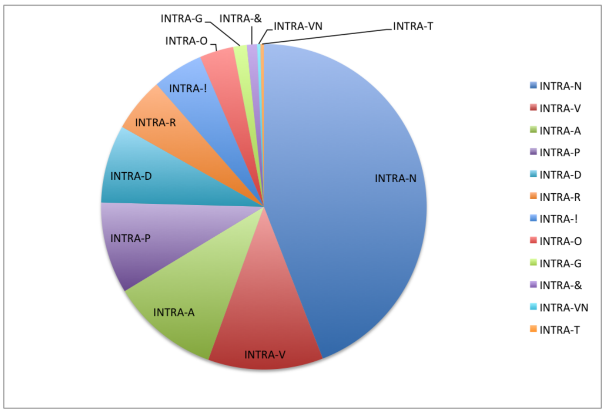

Figure 1: Distribution of INTRA tags, showing the syntactic role of code-switched tokens. Tag descriptions are given in Table 1

INTRA+POS	POS meaning
INTRA-N	Noun
INTRA-V	Verb
INTRA-A	Adjective
INTRA-P	Preposition
INTRA-D	Determiner
INTRA-R	Adverb
INTRA-!	Interjection
INTRA-O	Pronoun
INTRA-G	General
INTRA-&	Conjunction
INTRA-VN	Verbal Noun
INTRA-T	Particle

Table 1: Explanation of fine-grained (INTRA+) intra-sentential tags

4.2 Nature of INTRA code-switching in Irish tweets

One striking outcome of preliminary observations of this work is the distribution of syntactic patterns that arise within intra-sentential code-switching between Irish and English. There is a clear ease with which English nouns are used to replace Irish nouns in a single instance. For example *Figiúirí nua tally do Chonamara* 'New tally figures for Connemara'. In this instance, the English word 'tally' is part of a noun compound.[6] In Irish, the head of noun compound is the first noun (figiúirí tally) – in English it is the last (tally figures). In addition, the position of the adjective *nua* 'new' follows the rules of Irish syntax by following rather than preceding the head noun. Several similar noun-adjective examples exist in the corpus: *keyboards* beag 'small keyboards', *podcast* úr 'new podcast', *an album* nua 'the new album', *an stuff* corcra 'the purple stuff'. It is interesting to note that in these final two cases, there exist Gaelicized spellings of the code-switched words, *albam* and *stuif* respectively. In the first case at least, had the Irish form been intended, one would have *an **t**-albam nua* to satisfy Irish grammatical constraints related to the gender of the noun.

The relative infrequency of intra-sentential verb usage is particularly interesting when we consider the variations across English and Irish with respect to word order (SVO vs VSO). We observe that 5 out of the 34 INTRA-V (English verb) occurrences occur alone in an Irish context. (e.g. *Wish nach raibh aon obair le déanamh agam* 'Wish I didn't have work to do') All other 29 instances are part of an Irish string of two or more tokens (e.g. *am éigin an bhliain seo sounds good* 'some time this year sounds good'). Interestingly, Müller (1999) observed a similarly rare switching of verbs from Irish to Latin in historical texts.

[6] Interestingly, the Gaelicized spelling *teailí* is recorded in dictionaries and is seen occasionally in Irish writing.

4.3 Automatic POS tagging and detection of code-switching

We also reproduced the automatic POS tagging experiments from Lynn et al. (2015) with the addition of the EN tag in order to evaluate the impact of the slightly richer tagset on tagger performance, and to assess this as an approach to detecting code-switched segments. Due to the relative infrequency of code-switching types (INTRA, INTER, INTER-BI, WORD), we do not yet have enough data to train an effective tagger for this level of annotation, so the results below involve only the introduction of the tag EN.

In Table 2, the results given for the models BaseMorf and NormMorf (using the Morfette tagger of Chrupala et al. (2008)) and ArkLemma#URL@ (using the ARK tagger of Gimpel et al. (2011)) are the same as those of Lynn et al. (2015), to which we refer the reader for full experimental details. The results given with the suffix +EN repeat the same experiments but this time based on the training and test data that we have retagged, using the EN tag for code-switched tokens. Given the relative infrequency of the EN tag in the overall corpus, it is not surprising that the results change only slightly. The slight improvements coming from the introduction of the EN tag might be explained in part by the use of the G tag as a kind of "catch-all", making it difficult for the tagger to learn generalizations over examples of G tags. For example, in the original training corpus without the EN tag, there many sequences of two or more consecutive G tags. As a consequence, the taggers of Lynn et al. (2015) sometimes incorrectly assign G tags to one or more Irish words following an English word, but this seems to happen less often after introduction of the separate EN tag.

Training Data	Dev	Test
Baseline		
Rule-Based Tagger	85.07	83.51
Morfette		
BaseMorf	86.77	88.67
BaseMorf+EN	87.16	88.64
NormMorf	87.94	88.74
NormMorf+EN	**88.06**	**89.22**
ARK		
ArkLemma#URL@	**91.46**	91.89
ArkLemma#URL@+EN	91.23	**91.98**

Table 2: Changes in POS-tagging accuracy following separate labelling of English tokens (+EN indicates new experiments).

5 Inter-annotator Agreement

In order to assess consistency, levels of bias, and reliability of the annotated data, we carried out an Inter-Annotator Agreement (IAA) study. There are a number of metrics used widely to calculate IAA in classification tasks (Artstein and Poesio, 2008). In this study, we report IAA between two annotators using Cohen's Kappa (Cohen, 1960):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of observed agreement among annotators, and $P(E)$ is the proportion of expected agreement. By correcting for $P(E)$, this measurement accounts for the fact that the annotators are expected to agree a proportion of times just by chance. Di Eugenio and Glass (2004) present the calculation of Cohen's $P(E)$ as:

$$P(E) = \sum_j p_{j,1} \times p_{j,2}$$

where $p_{j,a}$ is the overall proportion of items assigned to a label j by annotator a.

For this study, we presented all 1496 tweets to the annotators. Annotators first received instructions to annotate all English tokens as 'EN'. For each English token, according to the code-switching categories described in Section 2.1, a new tag was to be inserted in the next column (INTER, INTER-BI, INTRA or WORD). We refer to these as *coarse-grained tags*. For each INTRA tag we also asked the annotators to identify the POS-tag for the English token (e.g INTRA-N (noun), INTRA-V (verb), etc). We refer to these as *fine-grained tags*. 943 tokens in the corpus are English tokens, and as such our kappa score is based on the agreement of the labelling of these tokens.

We achieved a kappa agreement rate of 0.69 on coarse-grained tags and 0.74 on fine-grained tags. On closer inspection, there were a couple of clear explanations for the coarse-grained tagging disagreements. Some cases involved confusion between INTER vs INTER-BI, and INTER vs INTRA. As an instance of INTER usually consists of a string of tokens (e.g. 'a rock and a hard place'), a single misinterpretation can lead to multiple instances of tag disagreement.

We use Landis and Koch (1977)'s metric shown in Table 3 for interpretation of our Kappa results.

Kappa value	Strength of Agreement
< 0.00	None
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Table 3: Landis and Koch's interpretation of Cohens's Kappa

While our results are regarded as substantial agreement we will take this as an opportunity to identify the areas of confusion and to revise our annotation guidelines for future labelling work.

6 Conclusion and Future Directions

We have reported on the enhancement of a corpus of POS-tagged Irish tweets with code-switching annotations and provided a categorisation of code-switching types of Irish UGC. We have also provided a quantitative report with respect to the distribution of code-switched tweets and tag types in the corpus.[7]

We have also reported more accurate automatic POS tagging results for these tweets, based on the inclusion of updated EN labels.

Our study has revealed that Irish speaking online users switch effortlessly and effectively between Irish and English. This ease demonstrates the clever mix across the syntax paradigms of both languages and supports the argument that code-switching is indeed a reflection of advanced grammatical ability. The various different types of code-switching employed suggested different motivations for this linguistic behaviour.

In terms of future work, the natural progression for this study would be to increase the size of the dataset so that more instances of code-switching can be observed and analysed. Of course, after having observed the disagreements amongst annotators in our IAA study, we will need to update the annotation guideline to make the instructions much clearer and to avoid ambiguity.

In future work, we would also like to take a more socio-linguistic approach to our analysis. We would like to investigate users' motivation for code-switching and assess whether linguistic patterns provide clues as to why and when English text is inserted into Irish tweets. For example, in some instances, we observe that English noun phrases are used where there is no official Irish term for a concept, and in other instances where there is an official Irish term that may not be known to the speaker (or if known, not preferred). This type of information would be a useful source of data for language planning and terminology development.

Given that Stam (2017) notes that "it appears that code-switching in medieval Irish texts may be both a functional communicative device used to structure a text and an unconscious expression of bilingual identity for a like-minded audience", we believe our corpus will provide an interesting dataset to help identify whether this holds true today.

One likely future application of this corpus is to build a tool to automatically identify code-switching in Irish online content. Despite it being a challenging task, there has been much progress in this area, with notable impact on a number of downstream applications, as outlined by Çetinoğlu et al. (2016). Yet, we note that our own data set is still not large enough to support state-of-the-art data-driven approaches. Further development of this corpus is therefore required.

In addition, we see this data set as a starting point for a treebank of Irish user-generated content. Parsing code-switched text is an area of research attracting much attention, and for this reason we have labelled the POS-tag of the switched tokens. Again, this is simply a starting point and much larger data set will be required before a data-driven system can be developed.

7 Acknowledgements

The first author's work is funded by the Irish Government Department of Culture, Heritage and the Gaeltacht under the GaelTech Project, and is also supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

8 Bibliographical References

References

Afli, Haithem, Sorcha McGuire, and Andy Way. 2017. Sentiment translation for low-resourced languages: Experiments on Irish General Election tweets. In *Proceedings of the 8th International Conference on Intelligent Text Processing and Computational Linguistics*, Budapest, Hungary.

[7] Available to download from `https://github.com/tlynn747/IrishTwitterPOS/tree/master/Data/morfette-CS`

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Atkinson, David and Helen Kelly-Holmes. 2011. Codeswitching, identity and ownership in Irish radio comedy. *Journal of Pragmatics*, 43(1):251 – 260.

Bannett Kastor, Tina. 2008. Code-mixing in biliterate and multiliterate Irish literary texts. *Asociación Española de Estudios Irlandeses (AEDEI)*, 3:29–41.

Caulfield, John. 2013. *A social network analysis of Irish language use in social media*. Ph.D. thesis, Cardiff University.

Çetinoğlu, Özlem, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas, November. Association for Computational Linguistics.

Chrupala, Grzegorz, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Di Eugenio, Barbara and Michael Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101, March.

Dowling, Meghan, Teresa Lynn, and Andy Way. 2017. A crowd-sourcing approach for translations of minority language user-generated content (UGC). In *Proceedings of The First Workshop on Social Media and User Generated Content Machine Translation*, Prague, Czech Republic.

Dumville, David. 1990. Latin and Irish in the Annals of ulster, a.d. 431- 1050. *Histories and Pseudo-Histories of the Insular Middle Ages*, pages 320–341.

Espinosa, Aurelio, 1917. *The Pacific Ocean in History*, chapter Speech Mixture in New Mexico: the influence of English language on New Mexican Spanish. New York MacMillan.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Grosjean, François. 2010. *Bilingual: Life and Reality*. Harvard University Press.

Hickey, Tina. 2009. Code-switching and borrowing in Irish. *Journal of Sociolinguistics*, 13(5):670–688.

Joshi, Aravind K. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th International Conference on Computational Linguistics, COLING '82, Prague, Czechoslovakia, July 5-10, 1982*, pages 145–150.

Lackaff, Derek and William J. Moner. 2016. Local languages, global networks: Mobile design for minority language users. In *Proceedings of the 34th ACM International Conference on the Design of Communication*, SIGDOC '16, pages 14:1–14:9, New York, NY, USA. ACM.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. In *Biometrics*, volume 33, pages 159–174. International Biometric Society.

Ní Laoire, Siobhán, 2016. *Irish-English Code-switching: A Sociolinguistic Perspective*, pages 81–106. Palgrave Macmillan UK, London.

Lynn, Teresa, Kevin Scannell, and Eimear Maguire. 2015. Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the 1st Workshop on Noisy User-generated Text (W-NUT 2015)*, Beijing, China.

Minocha, Akshay and Francis M. Tyers. 2014. Subsegmental language detection in Celtic language text. In *Proceedings of the Celtic Language Technology Workshop (CLTW), co-located with COLING 2014*.

Molina, Giovanni, Nicolas Rey-Villamizar, Thamar Solorio, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, and Mona Diab. 2016. Overview for the Second shared task on language identification in code-switched data. EMNLP 2016.

Muysken, Pieter. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press, Cambridge, England.

Müller, Nicole. 1999. Kodewechsel in der irischen Übersetzungsliteratur: exempla et desiderata. *Übersetzung, Adaptation und Akkulturation im insularen Mittelalter*, pages 73–86.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Comput. Linguist.*, 42(3):537–593, September.

Rosner, Mike and Paulseph-John Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *INTERSPEECH*, pages 190–193. ISCA.

Solorio, Thamar and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1051–1060, Stroudsburg, PA, USA. Association for Computational Linguistics.

Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the First shared task on language identification in code-switched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching, held in conjunction with EMNLP 2014.*, pages 62–72, Doha, Qatar. ACL.

Stam, Nike. 2017. *A typology of code-switching in the Commentary to the Félire Óengusso*. Ph.D. thesis, Utrecht University.

Stenson, Nancy. 1993. English influence on Irish: The last 100 years. *Journal of Celtic linguistics*, 2:107–128.