# Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets

**Teresa Lynn**[1,3], **Kevin Scannell**[2], and **Eimear Maguire**[1]
[1]ADAPT Centre, School of Computing, Dublin City University, Ireland
[2]Department of Mathematics and Computer Science, St. Louis University, USA
[3]Department of Computing, Macquarie University, Sydney, Australia
[1]{tlynn,emaguire}@computing.dcu.ie
[2]{kscanne}@gmail.com
[3]{teresa.lynn}@mq.edu.au,

## Abstract

Noisy user-generated text poses problems for natural language processing. In this paper, we show that this statement also holds true for the Irish language. Irish is regarded as a low-resourced language, with limited annotated corpora available to NLP researchers and linguists to fully analyse the linguistic patterns in language use in social media. We contribute to recent advances in this area of research by reporting on the development of part-of-speech annotation scheme and annotated corpus for Irish language tweets. We also report on state-of-the-art tagging results of training and testing three existing POS-taggers on our new dataset.

## 1 Introduction

The language style variation used on social media platforms, such as Twitter for example, is often referred to as noisy user-generated text. Tweets can contain typographical errors and ungrammatical structures that pose challenges for processing tools that have been designed for and tailored to high quality, well-edited text such as that found in newswire, literature and official documents. Previous studies, Foster et al. (2011) and Petrov and McDonald (2012) for example, have explored the effect that the style of language used in user-generated content has on the performance of standard NLP tools. Other studies by Gimpel et al. (2011), Owoputi et al. (2013), Avontuur et al. (2012), Rehbein (2013) and Derczynski et al. (2013) (POS-tagging), Ritter et al. (2011) (named entity recognition), Kong et al. (2014) and Seddah et al. (2012) (parsing) have shown that NLP tools and resources need to be adapted to cater for the linguistic differences present in such text.

When considering data-driven NLP tasks, a lack of resources can also produce additional challenges. We therefore examine the impact of noisy user-generated text on the existing resources for Irish, a low-resourced language. We also explore options for leveraging from existing resources to produce a new domain-adapted POS-tagger for processing Irish Twitter data. We achieve this by:

- defining a new POS tagset for Irish tweets
- providing a mapping from the PAROLE Irish POS-tagset to this new one
- manually annotating a corpus of 1537 Irish tweets
- training three statistical taggers on our data and reporting results

This paper is divided as follows: Section 2 gives a summary of Twitter and issues specific to the Irish Twitter data. Section 3 discusses the new part-of-speech tagged corpus of Irish tweets. Section 4 discusses our inter-annotator agreement study and the observations we note from annotator disagreements. Section 5 reports our tagging accuracy results on three state-of-the-art statistical taggers.

## 2 Irish Tweets

Irish, the official and national language of Ireland, is a minority language. While it is a second language for most speakers, everyday use outside of academic environments has seen a recent resurgence in social media platforms such as Facebook and Twitter. Twitter is a micro-blogging platform which allows users (*tweeters*) to create a social network through sharing or commenting on items of social interest such as ideas, opinions, events and news. Tweeters can post short messages called *tweets*, of up to 140 characters in length, that can typically be seen by the general public, including the user's *followers*. Tweets can be classified by topic by using *hashtags* (e.g. #categoryname)



*Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pages 1–8,
Beijing, China, July 31, 2015. 

and linked to other tweeters through the use of *at-mentions* (e.g. @username).

The first tweets in Irish appeared not long after the launch of Twitter in 2006, and there have been more than a million tweets in Irish since then, by over 8000 tweeters worldwide[1].

The social nature of tweets can result in the use of informal text, unstructured or ungrammatical phrases, and a variety of typographical errors. The 140 character limit can also lead to truncated ungrammatical sentences, innovative spellings, and word play, such as those discussed by Eisenstein (2013) for English. From our analysis, this phenomenon appears to extend also to Irish tweets.

In Figure 1, we provide an example of an Irish tweet that contains some of these NLP challenges:

*Freezing i dTra Li,Ciarrai chun cinn le cuilin.*
Freezing i dTr**á** Lí, **tá** Ciarra**í** chun cinn le c**ú**il**í**n.
'Freezing in Tralee, Kerry (is) ahead by a point.'

Figure 1: Example of noisy Irish tweet

**Diacritics** Irish, in its standard orthography, marks long vowels with diacritics (á,é,í,ó,ú). Our analysis of Irish tweets revealed that these diacritics are often replaced with non-accented vowels (cúilín => cuilin). There are a number of word pairs that are differentiated only by the presence or absence of these diacritics (for example, *cead* 'permission' : *céad* 'hundred'). There are many possible reasons for omitting diacritics, including shortening the time required to tweet (this tweet is from a spectator at a Gaelic Football match), a lack of knowledge on how to find diacritics on a device's keyboard, carelessness, or uncertainty about the correct spelling.

**Code-switching** Alternating between English and Irish is common in our dataset. This is unsurprising as virtually all Irish speakers are fluent English speakers, and many use English as their first language in their daily lives. In the example given, there is no obvious reason why "Freezing" was used in place of various suitable Irish words (e.g. *Préachta*), other than perhaps seeking a more dramatic effect. Sometimes, however, English is understandably used when there is no suitable Irish term in wide use, for example 'hoodie' or 'rodeo-clown'. Aside from occurring at an intra-sentential level, code-switching at an inter-sentential level is also common in Irish: *an t-am seo an t7ain seo chugainn bei 2 ag partyáil le muintir Ráth Daingin! Hope youre not too scared #upthevillage*. In total, of the 1537 tweets in our gold-standard corpus, 326 (21.2%) contain at least one English word with the tag G[2].

**Verb drop** We can see in this example that the verb *tá* 'is' has been dropped. This is a common phenomenon in user-generated content for many languages. The verb is usually understood and can be interpreted through the context of the tweet.

**Spacing** Spacing after punctuation is often overlooked (i) in an attempt to shorten messages or (ii) through carelessness. In certain instances, this can cause problems when tokenizing tweets; Li,Ciarrai => Li, Ciarrai.

**Phonetic spelling** Linguistic innovations often result from tweeters trying to fit their message into the 140 character limit. Our dataset contains some interesting examples of this phenomenon occurring in Irish. For example *t7ain* is a shortened version of *tseachtain* 'week'. Here the word *seacht* 'seven' is shortened to its numeral form and the initial mutation *t* remains attached. Other examples are *gowil (go bhfuil), beidir (b'fhéidir), v (bhí)*.

**Abbreviations** Irish user-generated text has its own set of frequently used phrase abbreviations – referred to sometimes as text-speak. Forms such as *mgl:maith go leor*, 'fair enough' and *grma:go raibh maith agat* 'thank you' have been widely adopted by the Irish language community.

The linguistic variation of Irish that is used in social media is relatively unexplored, at least not in any scientific manner. We expect therefore that the part-of-speech tagged corpus and taggers that we have developed for Irish language tweets will contribute to further research in this area.

## 3 Building a corpus of annotated Irish tweets

Unlike rule-based systems, statistical data-driven POS-taggers require annotated data on which they can be trained. Therefore, we build a gold-standard corpus of 1537 Irish tweets annotated

[1] http://indigenoustweets.com/ga/

[2] The tag G is used for foreign words, abbreviations, items and unknowns, as shown in Table 1.

with a newly defined Twitter POS tagset. The following describes this development process.

### 3.1 New Irish Twitter POS tagset

The rule-based Irish POS-tagger (Uí Dhonnchadha and van Genabith, 2006) for standard Irish text is based on the PAROLE Morphosyntactic Tagset (ITÉ, 2002). We used this as the basis for our Irish Twitter POS tagset. We were also inspired by the English-tweet POS tagset defined by Gimpel et al. (2011), and have aimed to stay closely aligned to it in order to facilitate any future work on cross-lingual studies.

We started by selecting a random sample of 500 Irish tweets to carry out an initial analysis. From our analysis of these tweets we concluded that our new Twitter-specific POS tagset would not require the granularity of the original standard Irish POS set. For example we do not need to differentiate between a locative adverb and a temporal adverb, or between a vocative particle and an infinitive particle. While our tagset is also closely aligned with the English-tweet POS tagset, we introduce the following tags that the English set does not use:

- **`VN`: Verbal Noun** Progressive aspectual phrases in Irish are denoted by the preposition *ag* followed by a verbal noun (e.g. *ag rith* 'running'). We choose to differentiate between N and VN to avoid losing this verbal information in what would otherwise be a regular prepositional phrase.

- **`#MWE`: Multiword hashtag** These are hashtags containing strings of words used to categorise a text (e.g. #godhelpus). We retain information on the multi-word nature of these hashtags in order to facilitate future syntactic analysis efforts.

We also adapt the `T` particle to suit Irish linguistic features.

- **`T`: Particle** We extend the `T` tag to not only cover verb particles, but all other Irish particles: relative particles, surname particles, infinitive particles, numeric particles, comparative particles, the vocative particle, and adverbial particles.

We do not use the following tags from the English set: `S, Z, L, M, X, Y,` as the linguistic cases they apply to do not occur in either standard or non-standard Irish. The final set of 21 POS-tags is presented in Table 1.

Most of the tags in the tagset are intuitive to an Irish language speaker. However, some tags require specific explanation in the guidelines. Hashtags and at-mentions can be a syntactic part of a sentence or phrase within a tweet. When this is the case, we apply the relevant syntactic POS tag. For example, *Beidh mé ar chlár @SplancNewstalk$_\wedge$ anocht ag labhairt leis @AnRonanEile$_\wedge$ faoi #neknomination$_N$* 'I will be on @SplancNewstalk tonight speaking to @AnRonanEile about #neknomination'. Otherwise if they are not part of the syntactic structure of the tweet (typically appended or prepended to the main tweet text), they are tagged as `@` and `#` (or `#MWE`). In our gold standard corpus, 554 out of 693 hashtags (79.9%), and 1604 out of 1946 at-mentions (82.4%) are of this non-syntactic type.

With some Twitter clients, if a tweet exceeds the 140 character limit, the tweet is truncated and an ellipsis is used to indicate that some text is missing. We leave this appended to the final (usually partial) token, which was often a URL. We marked these cases as `G`. For example *http://t.co/2nvQsxaIa7…*.

Some strings of proper nouns contain other POS elements, such as determiners and common nouns. Despite being a proper noun phrase syntactically, we tag each token as per its POS. For example, *Cú$_\wedge$ na$_D$ mBaskerville$_\wedge$* 'The Hound of the Baskervilles'.

### 3.2 Tweet pre-processing pipeline

About 950,000 Irish language tweets were posted between Twitter's launch in 2006 and September 2014 by approximately 8000 users identified and tracked by the Indigenous Tweets web site. Non-Irish tweets from these users were filtered out using a simple character-trigram language identifier. We selected a random sample of 1550 tweets from these 950,000 tweets and processed them as follows:

(1) We tokenised the set with Owoputi et al. (2013)'s version of `twokenise`[3], which works well on web content features such as emoticons and URLs.

(2) Using a list of multiword units from Uí Dhonnchadha (2009)'s rule-based Xerox FST tokeniser[4], we rejoined multiword tokens that had

[3] Available to download from `http://www.ark.cs.cmu.edu/TweetNLP/#pos`

[4] Available to download from `https://github.com/stesh/apertium-gle/tree/master/dev/`

| Tag | Description (PAROLE TAGS) |
|---|---|
| N | common noun (Noun, Pron Ref, Subst) |
| ∧ | proper noun (Prop Noun) |
| O | pronoun (Pron Pers, Pron Idf, Pron Q, Pron Dem) |
| VN | verbal noun (Verbal Noun) |
| V | verb (Cop, Verb*) |
| A | adjective (Adj, Verbal Adj, Prop Adj) |
| R | adverb (Adv*) |
| D | determiner (Art, Det) |
| P | preposition, prep. pronoun (Prep*, Pron Prep) |
| T | particle (Part*) |
| , | punctuation (Punct) |
| & | conjunction (Conj Coord, Conj Subord) |
| $ | numeral, quantifier (Num) |
| ! | interjection (Itj) |
| G | foreign words, abbreviations, item (Foreign, Abr, Item, Unknown) |
| ~ | discourse marker |
| # | hashtag |
| #MWE | multi-word hashtag |
| @ | at-mention |
| E | emoticon |
| U | URL/email address/XML (Web) |

Table 1: Mapping of Irish Twitter tagset to PAROLE tagset. (* indicates all forms of the fine-grained set for that tag.)

been split by the language-independent tokenizer (e.g. the compound preposition *go_dtí*).

(3) Using regular expressions, we then split tokens with the contractions *b' (ba), d' (do), m' (mo)* prefixes. For example *b'fhéidir* 'maybe'; *d'ith* 'ate'; *m'aigne* 'my mind'.

(4) We took a bootstrapping approach by pre-tagging and lemmatising the data with the rule-based Irish POS-tagger first, and then mapped the tags to our new Twitter-specific tagset.

(5) In cases where the rule-based tagger failed to produce a unique tag, we used a simple bigram tag model (trained on the gold-standard POS-tagged corpus from Uí Dhonnchadha (2009) – see Section 5.1) to choose the most likely tag from among those output by the rule-based tagger.

(6) Finally, we manually corrected both the tags and lemmas to create a gold-standard corpus.

### 3.3 Annotation

The annotation task was shared between two annotators. Correction of the first 500 tweets formed a basis for assessing both the intuitiveness of our tagset and the usability of our annotation guide. Several discussions and revisions were involved at this stage before finalising the tagset. The next 1000 tweets were annotated in accordance with the guidelines, while using the first 500 as a reference. At this stage, we removed a small number of tweets that contained 100% English text (errors in the language identifier). All other tweets containing non-Irish text represented valid instances of code-switching.

The annotators were also asked to verify and correct the lemma form if an incorrect form was suggested by the morphological analyser. All other tokeniser issues, often involving Irish contractions, were also addressed at this stage. For example *Tá'n* $->$ *Tá an*.

## 4 Inter-Annotator Agreement

Inter-Annotator agreement (IAA) studies are carried out during annotation tasks to assess consistency, levels of bias, and reliability of the annotated data. For our study, we chose 50 random Irish tweets, which both annotators tagged from scratch. This differed from the rest of the annotation process, which was semi-automated. However, elimination of possible bias towards the pre-annotation output allowed for a more disciplined assessment of agreement level between the annotators. We achieved an agreement rate of 90% and a $\kappa$ score (Cohen, 1960) of 0.89.

Smaller tagsets make an annotation task easier due to the constraint on choices available to the annotator, and is certainly one reason for our high IAA score. This result also suggests that the tagging guidelines were clear and easy to understand. A closer comparison analysis of the IAA data explains some disagreements. The inconsistency of conflicts suggests that the disagreements arose from human error. Some examples are given below.

**Noun vs Proper Noun** The word *Gaeilge* 'Irish' was tagged on occasion as `N` (common noun) instead of ∧ (proper noun). This also applied

`irishfst`

to some proper noun strings such as *Áras an Uachtaráin* (the official name of the President of Ireland's residence).

**Syntactic at-mentions** A small number of at-mentions that were syntactically part of a tweet (e.g. *mar chuid de @SnaGaeilge* 'as a part of @SnaGaeilge') were incorrectly tagged as regular at-mentions (`@`).

**Retweet colons** One annotator marked ':' as punctuation at random stages rather than using the discourse tag `~`.

## 5 Experiments

### 5.1 Data

We took the finalised set of Irish POS-tagged tweets and divided them into a test set (148 tweets), development set (147 tweets) and training set (1242 tweets). Variations of this data are used in our experiments where we normalise certain tokens (described further in Section 5.2.)

We also automatically converted Uí Dhonnchadha (2009)'s 3198 sentence (74,705 token) gold-standard POS-tagged corpus using our mapping scheme. This text is from the New Corpus for Ireland – Irish[5], which is a collection of text from books, newswire, government documents and websites. The text is well-structured, well-edited, and grammatical, and of course lacks Twitter-specific features like hashtags, at-mentions, and emoticons, thus differing greatly from our Twitter data. The average sentence length in this corpus is 27 tokens, diverging significantly from the average tweet length of 17.2 tokens. Despite this, and despite the fact the converted tags were not reviewed for accuracy, we were still interested in exploring the extent to which this additional training data could improve the accuracy of our best-performing model. We refer to this set as `NCII_3198`.

### 5.2 Taggers

We trained and evaluated three state-of-the-art POS-taggers with our data. All three taggers are open-source tools.

**Morfette** As Irish is an inflected language, inclusion of the lemma as a training feature is desirable in an effort to overcome data sparsity. Therefore we trained Morfette (Chrupala et al., 2008), a lemmatization tool that also predicts POS tags and uses the lemma as a training feature. We report on experiments both with and without an optional dictionary (`Dict`) information. We used the dictionary from Scannell (2003), which contains $350,418$ surface forms. Our baseline Morfette data (`BaseMorf`) contains the token, lemma and POS-tag. The lemmas of URLs and non-syntactic hashtags have been normalised as $< URL >$ and $< \# >$, respectively.

We then evaluated the tagger with (non-syntactic) $< \# >$, $< @ >$ and $< URL >$ normalisation of both token form and lemma (`NormMorf`). Both experiments are re-run with the inclusion of our dictionary (`BaseMorf+Dict`,`NormMorf+Dict`).

**ARK** We also trained the CMU Twitter POS-tagger (Owoputi et al., 2013), which in addition to providing pre-trained models, allows for re-training with new languages. The current release does not allow for the inclusion of the lemma as a feature in training, however. Instead, for comparison purposes, we report on two separate experiments, one using the surface tokens as features, and the other using only the lemmas as features (`ArkForm, ArkLemma`). We also tested versions of our data with normalised at-mentions, hashtags and URLs, as above.

**Stanford tagger** We re-trained the Stanford tagger (Toutanova et al., 2003) with our Irish data. We experimented by training models using both the surface form only (`BestStanForm`) and the lemma only (`BestStanLemma`). The best performing model was based on the feature set `left3words, suffix(4), prefix(3), wordshapes(-3,3), biwords(-1,1)`, using the `owlqn2` search option.[6]

**Baseline** Finally, to establish a baseline (`Baseline`), and more specifically to evaluate the importance of domain-adaptation in this context, we evaluated a slightly-enhanced version of the rule-based Irish tagger on the Twitter dataset. When the rule-based tagger produced more than one possible tag for a given token, we applied a bigram tag model to choose the most likely tag, as we did in creating the first

[5] New Corpus for Ireland - Irish. See `http://corpas.focloir.ie`

[6] All other default settings were used.

draft of the gold-standard corpus. In addition, we automatically assigned the tag U to all URLs, # to all hashtags, and @ to all at-mentions.

### 5.3 Results

| Training Data | Dev | Test |
|---|---|---|
| **Baseline** | | |
| Rule-Based Tagger | 85.07 | 83.51 |
| **Morfette** | | |
| BaseMorf | 86.77 | 88.67 |
| NormMorf | 87.94 | 88.74 |
| BaseMorf+Dict | 87.50 | 89.27 |
| NormMorf+Dict | 88.47 | 90.22 |
| **ARK** | | |
| BaseArkForm | 88.39 | 89.92 |
| ArkForm#@ | 89.36 | 90.94 |
| ArkForm#URL@ | 89.32 | 91.02 |
| BaseArkLemma#URL | 90.74 | 91.62 |
| ArkLemma#URL@ | **91.46** | **91.89** |
| **Stanford** | | |
| BestStanForm | 82.36 | 84.08 |
| BestStanLemma | 87.34 | 88.36 |
| **Bootstrapping Best Model** | | |
| ArkLemma#URL@+NCII | **92.60** | **93.02** |

Table 2: Results of evaluation of POS-taggers on new Irish Twitter corpus

The results for all taggers and variations of data-setup are presented in Table 2.

Firstly, our best performing single model (ArkLemma#URL@) on the test set achieves a score of 91.89%, which is 8 points above our rule-based baseline score of 83.51%. This confirms that tailoring training data for statistically-driven tools is a key element in processing noisy user-generated content, even in the case of minority languages. It is worth noting that the best-performing model learns from the lemma information instead of the surface form. This clearly demonstrates the effect that the inflectional nature of Irish has on data sparsity. The Twitter-specific tokens such as URLs, hashtags and at-mentions have been normalised which demonstrates the impact the relative uniqueness of these tokens has on the learner.

All of our results are comparable with state-of-the-art results produced by Gimpel et al. (2011) and Owoputi et al. (2013). This is interesting, given that in contrast to their work, we have not optimised our system with unsupervised word clusters due to the lack of sufficient Irish tweet data. Nor have we included a tag dictionary, distribution similarity or phonetic normalisation – also due to a lack of resources.

We carried out a closer textual comparison of Owoputi et al. (2013)'s English tweet dataset (daily547) and our new Irish tweet dataset. After running each dataset through a language-specific spell-checker, we could see that the list of highly ranked OOV (out of vocabulary) tokens in English are forms of text-speak, such as *lol* 'laugh out loud', *lmao* 'laugh my ass off' and *ur* 'your', for example. Whereas the most common OOVs in Irish are English words such as 'to', 'on', 'for', 'me', and words misspelled without diacritics. This observation shows the differences between textual challenges of processing these two languages. It may also suggest that Irish Twitter text may follow a more standard orthography than English Twitter text, and will make for an interesting future cross-lingual study of Twitter data.

Finally, we explored the possibility of leveraging from existing POS-tagged data by adding NCII_3198 to our best performing model ArkLemma#URL@. We also duplicated the tweet training set to bring the weighting for both domains into balance. This brings our training set size to 5682 (117,273 tokens). However, we find that a significant increase in the training set size only results in just over a 1 point increase in POS-tagging accuracy. At a glance, we can see some obvious errors the combined model makes. For example, there is confusion when tagging the word *an*. This word functions as both a determiner and an interrogative verb particle. The lack of direct questions in the NCII corpus results in a bias towards the D (determiner) tag. In addition, many internal capitalised words (e.g. the beginning of a second part of a tweet) are mislabelled as proper nouns. This is a result of the differing structure of the two data sets – each tweet may contain one or more phrases or sentences, while the NCII is split into single sentences.

## 6 Future Work

Limited resources and time prevented exploration of some options for improving our POS-tagging results. One of these options is to modify the CMU (English) Twitter POS-tagger to allow for inclusion of lemma information as a feature. Another

option, when there is more unlabelled data available (i.e. more Irish tweets online), would be to include Irish word cluster features in the training model. This approach has also been taken by Rehbein (2013) for POS tagging German tweets.

The resources we provide through this study are a valuable contribution to the Irish NLP community. Firstly, we expect that this new data resource (the POS-tagged Twitter corpus) will provide a solid basis for linguistic and sociolinguistic study of Irish on a social media platform. This new domain of Irish language use can be analysed in an empirical and scientific manner through corpus analysis by means of our data. The authors are currently working towards this follow-up study.

From a tool-development perspective, we expect this corpus and the derived POS-tagging models could be used in a domain-adaptation approach to parsing Irish tweets, similar to the work of Kong et al. (2014). This would involve adapting Lynn et al. (2012)'s Irish statistical dependency parser for use with social media text. Our corpus could provide the basis of a treebank for this work.

Following our discovery of the extent that code-switching is present our Irish Twitter data, we feel future studies on this phenomenon would be of interest to various research disciplines (e.g Solorio et al. (2014)). In order to do that, we suggest updating the corpus with a separate tag for English tokens (that is, a tag other than `G`, which is also used for abbreviations, items and unknowns) before carrying out further experiments in this area.

## 7 Conclusion

We present the first dataset of gold-standard POS-tagged Irish language tweets and we have produced training models for a selection of POS-taggers.[7] We have also shown how we have leveraged from existing work to build these resources for a low-resourced language, to achieve state-of-the-art results. We also confirm that the NLP challenges arising from noisy user-generated text can also apply to a minority language.

## 8 Acknowledgments

The authors would like to thank the three anonymous reviewers for their helpful feedback. We also would like to thank Kevin Gimpel for his support with using the CMU English Twitter POS tagger, Djamé Seddah for his support with Morfette, and Elaine Uí Dhonnchadha and Francis Tyers for their support with the Irish rule-based POS tagger. This work was funded by the Fulbright Commision of Ireland (Fulbright Enterprise-Ireland Award 2014-2015), and supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University. The second author was partially supported by US NSF grant 1159174.

[7] Our data is available to download from `https://github.com/tlynn747/IrishTwitterPOS`

## References

Tetske Avontuur, Iris Balemans, Laura Elshof, Nanne van Noord, and Menno van Zaanen. 2012. Developing a part-of-speech tagger for dutch tweets. *Computational Linguistics in the Netherlands Journal*, 2:34–51, 12/2012.

Grzegorz Chrupala, 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*.

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

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *RANLP*, pages 198–206. RANLP 2011 Organising Committee / ACL.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369. Association for Computational Linguistics.

J. Foster, Ö. Çetinoglu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, D. Hogan, J. Van Genabith, et al. 2011. # hardtoparse: Pos tagging and parsing the twitterverse. In *Proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pages 20–25.

Kevin Gimpel, 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.

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

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October. Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Elaine Uí Dhonnchadha. 2012. Active learning and the Irish treebank. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 23–32.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In *First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Ines Rehbein. 2013. Fine-grained pos tagging of german tweets. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *GSCL*, volume 8105 of *Lecture Notes in Computer Science*, pages 162–175. Springer.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin P. Scannell. 2003. Automatic thesaurus generation for minority languages: an Irish example. *Actes de la 10e conférence TALNa Batz-sur-Mer*, 2:203–212.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung, 2014. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, chapter Overview for the First Shared Task on Language Identification in Code-Switched Data, pages 62–72. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elaine Uí Dhonnchadha and Josef van Genabith. 2006. A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Elaine Uí Dhonnchadha. 2009. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. Ph.D. thesis, Dublin City University.