PBML

logo

The Prague Bulletin of Mathematical Linguistics

NUMBER ??? JUNE 2017 1–12

A Crowd-sourcing Approach for Translations of Minority Language User-Generated Content (UGC)

Meghan Dowling, Teresa Lynn, Andy Way

ADAPT Centre, Dublin City University

Abstract

Data sparsity is a common problem for machine translation of minority and less-resourced languages. While data collection for standard, grammatical text can be challenging enough, efforts for collection of parallel user-generated content can be even more challenging. In this paper we describe an approach to collecting English↔Irish translations of user-generated content (tweets) that overcomes some of these hurdles. We show how a crowd-sourced data collection campaign, which was tailored to our target audience (the Irish language community), proved successful in gathering data for a niche domain. We also discuss the reliablity of crowd-sourcing English↔Irish tweet translations in terms of quality by reporting on a self-rating approach along with qualified reviewer ratings.

1. Introduction

Irish is the first official language of Ireland, an official language of the European Union, and a recognised minority language in both Northern Ireland and the European Union. However, despite its status, the 2012 META-NET White Paper Series report classifies the Irish language as having "weak/no support" with regards to machine translation resources (Judge et al., 2012). Recently, in response to this, there has been notable progress in terms of gathering parallel data for English↔Irish (EN↔GA) machine translation (Arcan et al. (2016), Dowling et al. (2015)).

Of course, a robust statistical machine translation (SMT) system, which is data-driven, relies on the availability of a significant amount of parallel data suitable for

Corresponding author: meghan.dowling@adaptcentre.ie

Cite as: Meghan Dowling, Teresa Lynn, Andy Way. A Crowd-sourcing Approach for Translations of Minority Language User-Generated Content (UGC). The Prague Bulletin of Mathematical Linguistics No. ???, 2017, pp. 1–12.
doi: HERE-WILL-BE-A-DOI-NUMBER

the translation domain. However, there is still only a relatively small amount of parallel data available for the English↔Irish language pair,[1] the vast majority of which contains curated, grammatical and carefully translated content. As expected, this type of text differs greatly from the characteristics of user-generated content (UGC). Lynn et al. (2015) report on some interesting nuances of Irish language UGC such as code-switching, verb drop and phonetic spelling, all of which can cause challenges for text processing. Their work is part of a recent growth of interest in automated processing of Irish UGC, which also saw the creation of the only known user-generated corpus of English↔Irish (EN↔GA) text, gathered as part of the Brazilator sentiment analysis and machine translation project (CNGL-DCU Team, 2014). This project provided Twitter with a number of SMT systems (including Irish) that allowed for real-time translation of tweets during the 2014 World Cup.[2]

There are a number of difficulties associated with collecting parallel data for machine translation of EN↔GA UGC content. Firstly, UGC content is relatively new and has really only become prevalent in the past 10 years following the growth in popularity of social media platforms such as Facebook, Twitter, Instagram, and so on. This means that translated data is not as readily available as it may be with other content types (e.g. public documents, educational materials, etc.). Secondly, the domains of UGC content vary so much that a system would need to be tuned to the specific terminology use of that topic or domain. This indeed was the case with the MT system used in the Brazilator project, due to soccer-related tweets carrying a particular register and terminology usage. Finally, compounding the challenges for Irish UGC translation is the lack of high-quality human translators in general who are available to translate English content into Irish, and vice-versa. Moorkens (2016) reports that there is a relatively small number of accredited Irish language translators available, and that the demand for translation exceeds the availability of quality translators to such an extent that there is a derogation on Irish translation in the European Commission until 2022. With all these obstacles facing the development of domain-specific EN↔GA parallel data, we are faced with formulating an alternative approach to data collection.

A well-attested solution to gathering translation content is through crowd-sourcing platforms (e.g. Ambati et al. (2010), Zaidan and Callison-Burch (2011)). However, as is the case for many minority languages, English↔Irish translation requires a niche set of skills, which contributors to well-known global crowd-sourcing platforms such as Crowd Flower[3] or Mechanical Turk[4] are unlikely to hold. In this paper, we describe a

[1] Approx. 348,964 sentences of parallel text publicly available, according to Arcan et al. (2016). It should also be noted that a large portion of these 'sentences' in fact contain word to word translations, similar to a terminology database.

[2] Sentiment analysis was carried out on tweets so that the change in polarity of tweets could be viewed in real time over the course of each game depending on fans' views of the match.

[3] https://www.crowdflower.com

[4] http://www.mturk.com/mturk/

crowd-sourcing campaign which allowed us to develop a user-generated Irish↔English tweet dataset (for the purposes of a study in Sentiment Analysis within MT (Afli et al., 2017)) by directly attracting altruistic contributions from the Irish-speaking community.

The remainder of this paper is divided as follows: in Section 2, we describe the motivation behind this project. In Section 3, we provide details of the design of the crowd-sourcing interface. In Section 4, the public and media responses relating to this project are discussed. Finally, in Section 6, we provide some conclusions on this project.

2. Motivation

2.1. Resource collection motivation

In recent years, there has been an increased awareness of the usefulness of NLP analysis of social media content when reporting on significant societal events or topical discussion (e.g. the analysis on Twitter of rioting (Lukasik et al., 2015), fake news (Gupta and Kumaraguru (2012), Mitra et al. (2017)), rumours (Jin et al., 2013) and elections (O'Connor et al. (2010), Bakliwal et al. (2013)). Particularly relevant to this work, sentiment analysis helps to provide both governments and the general public with an overview of the online community's opinions or feelings towards events or people (for example election candidates (e.g. Ceron et al. (2014)). In Ireland, the national broadcaster (RTÉ - Raidió Teilifís Éireann), through sentiment analysis of tweets with the hashtag #GE16, reported on opinion trends in the lead-up to the 2016 General Election.[5] One shortcoming of this report, however, is that it only reported on the English language tweets. In other words, the sentiment of the Irish-speaking online community was not represented.

Subsequent to this work, a study was carried out to investigate the sentiment of Irish language tweets from this period of time, and containing the same #GE16 hashtag (Afli et al., 2017).[6] The study focused on analysing sentiment analysis of Irish language tweets and assessing whether sentiment holds across languages through translation. In order to carry out the study, a parallel corpus of EN↔GA tweets was required, on which sentiment polarity are annotated. Here, we describe the crowd-sourcing method used in the collection of data for the creation of this parallel corpus of EN↔GA tweets.

[5] https://analysis.rte.ie/business/2016/02/29/ge16-the-first-social-media-election/

[6] Irish tweets around the General Election (*olltoghchán*) tended to also incude the English language hashtag #GE16, along with #togh16, #olltoghchán or #OT16

2.2. Motivation for crowd-sourcing

As discussed, Irish language research is low in resources, both in terms of funding and in terms of skilled translators. For these reasons, professional translation of the dataset would be beyond the budget of this project. Considering the positive disposition towards Irish language promotion (Darmody et al., 2015), an approach that benefits from the altruistic nature of Irish speakers seemed more realistic and more feasible. Members of the Irish-speaking community, both on and off-line, are passionate and proactive about Irish language promotion. In recent times, it has been noted that the community have a strong presence on social media (Lackaff and Moner, 2016). Social network platforms such as Facebook and Twitter have proven beneficial for online campaigns related to the Irish language (e.g. The Twitter campaign known as #AchtAnois - '(Legislative) Act Now').[7] These platforms are also positively exploited in 'spreading the word' about Irish language related activites (e.g. #PopUpGaeltacht tweets have helped the promotion of informal Irish-speaker social meetups both in Ireland and in cities around the world).[8]

Given the positive disposition online towards the cultivation and growth of the Irish language, it is unsuprising to note that previous crowd-sourcing campaigns have proved successful. For example, through crowd-sourcing, 1000 English tweets related to the 2014 World Cup were translated into Irish for the Brazilator project (CNGL-DCU Team, 2014).[9] In addition, Meitheal Dúchas is a larger, more recent and ongoing campaign that has shown how this approach engages the community at large to contribute to language conservation. The Meitheal Dúchas transcription project allows the general public to transcribe The School's Collection section of the digitization of the National Folklore Collection. [10] The project's website site provides up-to-date statistics on the contributions to the collection so far. Given these previous successes, we created an online translation interface open to the public and re-enforced its promotion with a social media campaign to elicit participant involvement from the online Irish-speaking community.

3. Design considerations

Our aim was to provide a practical crowd-sourcing translation platform with a suitable user interface tailored to the contributors. Two important factors we needed to bear in mind were that the translators would be (1) unpaid and (2) un-acknowledged

[7] #AchtAnois is used by the Irish language community to show their annoyance at the standard of Irish language legislation in Northern Ireland

[8] http://www.gaelchultur.com/en/newsletters/NewsletterArticle.aspx?id=543

[9] These 1000 tweets were translated in sets of 100 tweets by 10 volunteers, whose help was enlisted through an online campaign.

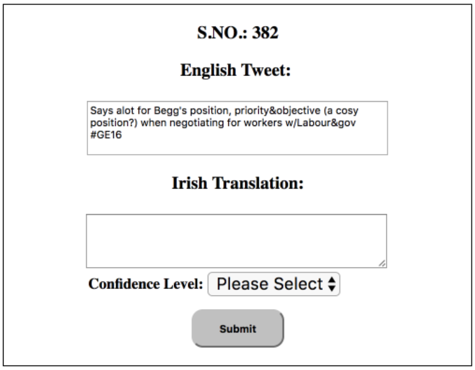
[10] http://www.duchas.ie/en/meitheal/

(anonymous). In order to optimise contributions, the design, therefore, needed to ensure that the translation request did not feel like a project or tedious task. This consideration in particular arose from lessons learned from the Brazilator data collection, where the provision of a shared spreadsheet with a large list of tweets for translation resulted in procrastination by some of the volunteers.

3.1. Design criteria

The criteria identified for this website is as follows:

- The website needed to be user-friendly and casual, with clear instructions.
- Users should feel as though they had a trivial task, and could complete as many translations as they felt comfortable with.
- Given that the translators may not be qualified or accredited, it was vital that they could provide some feedback on their measure of quality of the translations they provided.
- While low-quality translations would not be included in the dataset, all contributions were to be deemed valuable.
- Both native and non-native speakers should be able to contribute to the translation effort.
- An effort should be made to maintain consistency across the translations (i.e. approaches to dealing with Twitter-style language).
- Administrators should be able to easily view and access the translations and metadata.



S.NO.: 382

English Tweet:

Says alot for Begg's position, priority&objective (a cosy position?) when negotiating for workers w/Labour&gov #GE16

Irish Translation:

Confidence Level: Please Select

Figure 1: A portion of the translation interface. Image has been slightly altered for printing purposes

3.2. Implementation

Given the specific criteria identified as necessary for this crowd-sourcing platforms, the following features were implemented:

- The landing page of the website has a 'no-fuss' appearance, with just four options to choose from (two translation direction options, guidelines and an Admin login option).
- Users were presented with just one tweet to translate at a time, creating a casual opt-in/ opt-out environment.
- Users were required to assign a confidence level from 0–10. The purpose of the scoring was to allow for retranslation on lower-scored tweets in an effort to achieve a high quality translation corpus.
- It was possible for users to skip any translations that they did not feel confident translating, allowing for another user to undertake instead.
- The language direction (English→Irish or Irish→English) could be chosen and switched between at any time.
- Non-native Irish speakers could still contribute by choosing to translate into English, their (presumably) native language with more ease.
- A set of translation guidelines, outlined in Section 3.2.1, were provided to aid users and ensure consistency.
- The Admin user-interface provided a spreadsheet view of all tweets, their translations, and their confidence scoring.

Figure 1 shows a screenshot of the translation inferface for the translation of English→Irish tweets.

3.2.1. Translation guidelines

The following are the translation guidelines provided to users to aid them in their translation.

- **Placeholders:** #hashtags and @twitterhandles are to be left untranslated. Emoticons have been replaced by the placeholder [emoticon]. Please retain these placeholders in your Irish translation (or English translations) also.
  e.g. My Dad [emoticon] soaked but smiling #ge16 → *M'athair [emoticon] fliuch báite ach fós gealgháireach #ge16*
- **Case:** Please keep translations case sensitive where possible.
  e.g.: FULL HOUSE Great night tonight @SorchaNicC #GE16 launch. → *TEACH LÁN Oíche iontach anocht ag seoladh #GE16 @SorchaNicC.*
- **Text speak:** Where possible, please translate English text speak to Irish text speak (and vice versa), where there are equivalents.
  e.g. tnx (thanks) → *grma (go raibh maith agat)*. If there is no shortened Irish/English equivalent that you are aware of, translate the word into its full form.

- **Tweet length:** Although the original tweets have been limited to 140 characters, your translations do not have to adhere to this.
- **Pre-translate options:** It is acceptable to use Google Translate to pre-translate the tweets and correct the output – if you find it helpful. If it is too much of a hindrance, translation from scratch might work better. Note that the translations do not have to be 100% sound. Remember that the quality of Twitter language is questionable at the best of times, so your best shot is enough. Where there is ambiguity, go with your intuitive translation.
- **Confidence level:** After having translated the tweet, you are asked to indicate how confident you are that your translation is accurate. Please rate your translation on a scale of 1–10 from the drop-down menu provided.
- **Skip translation:** If you want to skip a tweet leave the translation field blank and submit a confidence level of 0

4. Dissemination and Public Response

4.1. Dissemination

Given the previous positive reactions to Irish language social media campaigns, a call for participation on social media sites was a natural starting point for gathering prospective translators. This approach also takes into account that this is a non-conventional[11] crowdfunding platform, and therefore participants must be actively sought out. A web-based approach is most suitable in order to spread the word rapidly and reach a wide audience. In addition, it is worth noting that due to the fact that the translation platform was new and entirely web-based, it was more effective to direct users to the website through digital means (i.e. through sharing a hyperlink).

As mentioned earlier, the Irish language community is highly active on social media, particularly Twitter[12] and Facebook.[13] Participants with knowledge and regular use of the Irish language on social media were especially valuable to this project, as it related to translation of a specific genre of language. The language used on Twitter by Irish language users often takes a different shape to language from other domains (Lynn et al., 2015). For instance, in Example (1), taken from our collected Twitter corpus, the term *fér plé* is used, which is an Irish phoneticisation of the phrase ‘fair play’ (‘well done to...’) as well as *fé* which uses non-standard orthography based on the dialectal pronunciation of the word *faoi* ‘about/on’.

[11] As opposed to Mechanical Turk or Crowdflower where frequent users visit the site to seek work, we needed to invite people to visit our site.

[12] 1,681,291 Irish language tweets to date according to Indigenous Tweets, a website which provides statistics on minority language tweeting: http://indigenoustweets.com/

[13] For example, the public group ‘Gaeilge Amhain’ Available at https://www.facebook.com/groups/166677873392308

(1) *Fér plé do @RTERnaG as leanúint leis an gcraoltóireacht fé #GE16!*
'Fair play to @RTERnaG for following reports on #GE16!'

4.2. Public Response

The press and broadcasting media play a central role in the Irish language community, both in Ireland and among the diaspora overseas. It was fortunate, therefore, that this crowd-sourcing campaign was picked up, endorsed and distributed by a variety of Irish-language digital media outlets, e.g. Raidió na Gaeltachta, Raidió na Life, and Tuairisc.[14] This happened mainly through promotion on Twitter, through the tagging of such media bodies in tweets or retweets. Endorsements from such public-facing outlets undoubtedly helped to shape the positive public response we received towards the campaign and thus broadened the reach for soliciting contribution.

Feedback from users, however, suggested that it would have been helpful for this platform to be available as a mobile application. One possible assumption that could be made from this is that users did indeed feel as though single tweet translation was a trivial task that could be carried out on the move and during a moment of downtime.

5. Results and Evaluation

Through our crowd-sourcing platform, over 1000 tweet translations were collected from 4th July, 2016 until 18th August, 2016 (see Table 2). A larger number of GA→EN tweets were collected (720) than EN→GA (324).

Language direction	**Translations collected**	**Average confidence value**
English→Irish	324	8.04
Irish→English	720	8.70

Table 1: Crowd-sourced translations, including average self-score rating

A natural question that arises in a study like this is the question of reliability of crowd-sourcing as a method for translation, and ultimately for data set creation. As the translation contributions are anonymous and the link through which the tweets are translated is available to the general public, how can we assess that the translations we solicit are reliable? We took two approaches to answering this question:

(1) We asked the translators to score themselves, and as such rate their own translation quality. The purpose of this was two-fold. Firstly, it allowed for lower-scored translations to be re-presented to another user for translation as part of a quality control measure. Secondly, to assess the reliability of self-scoring as a method for evalu-

[14] Tuairisc is an online Irish language periodical of a news/journal/magazine nature. http://tuairisc.ie

ating (or roughly evaluating) the quality of the crowd-sourced translations. It can be seen from the results in Table 1 that the average confidence value is above 8 (out of a range of 1-10) for crowd-sourced translations in both language directions.

Language direction	**Translations reviewed**	**Average reviewer score**
English→Irish	180	8.68
Irish→English	180	9.22

Table 2: Reviewer quality rating for subset of crowd-sourced data: average score for both language directions

(2) A native Irish speaker reviewed a portion of the tweet translations (n=180) and assigned them a quality rating (1–10).[15] This scoring gave us a true indication of translation quality.

In order to assess the reliablity of the crowd-sourced self-scoring method, we compare the reviewer's rating to the translators' self ratings. The reviewer's average quality rating is higher (by more than 0.5) than the average rating of the translators in both language directions (see Table 2). Furthermore, in 71% of English→Irish translations and 82% of Irish→English translations, the reviewer deemed the translations either the same or of a higher quality than the original self-rated score (see Figure 2).

6. Conclusions and Future Work

6.1. Conclusions

We have shown that for a minority language such as Irish, while traditional crowd-sourcing platforms may not be an option for the collection of data, it is possible instead to benefit from the altruistic nature of the community towards language cultivation – in a way that would not be possible for a majority language. We have presented a web interface that is tailored to the needs of this project – user-friendly, casual, and accessible.[16] The success of this platform is evident in the 1000+ tweet parallel corpus of user–generated content that has been collected and quality-assessed, as well as the positive public and media response that the project received. It is clear that when presented with a project that has clear benefits for the Irish language, speakers will donate their time and efforts to participate.

[15] The same scoring system as the original translator: 1 being incomprehensible, and 10 being fully acceptable in terms of fluency and adequacy.

[16] The code for this platform is open-source and it available from `https://github.com/saurabhgpta20/Deep-Senti-Analytics/tree/master/Translator`

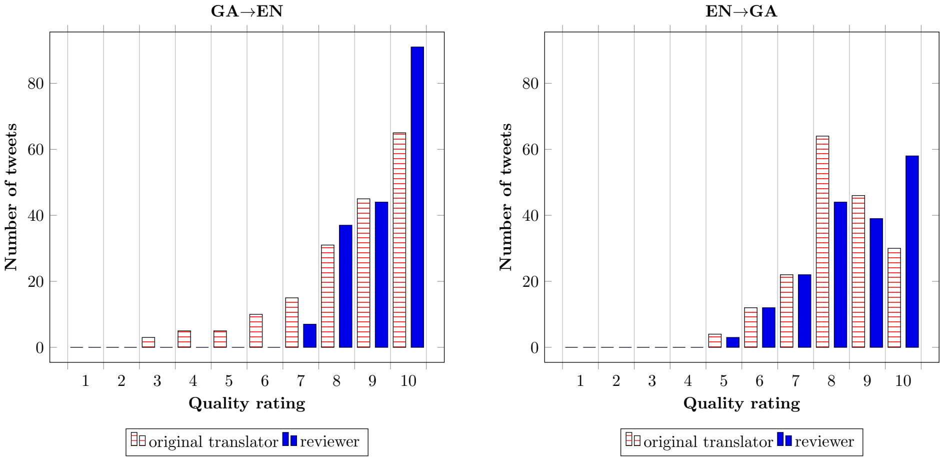


Figure 2: Quality ratings for Irish→English and English→Irish translations provided by the original translator and the reviewer

We have also shown that a high quality of translation can be acquired through a crowd-sourcing campaign amongst the Irish-speaking online community. Our preliminary study has shown that a self-rating approach to evaluation can be a reliable indicator of the general quality of a crowd-sourced data set. Further extensive studies of course are required before more definite conclusions can be drawn on this.

As this was a exploratory work, we have also been able to identify some learnings that should be considered in future crowd-sourcing efforts. While generating awareness online is invaluable for the initial promotion of such a project, it became clear that the "hype" can die down relatively quickly if there is not a concerted effort to continue with the promotion drive. This is understandable, as the public will assume that (without reminders) all required translations have been collected. One option to mitigate against this is to provide a progress bar on the site to indicate the percentage that has already been translated, and how much is outstanding.

6.2. Future Work

In the future, we aim to extend this study in a number of ways. Firstly, we would like to further investigate Irish speakers' self-perceptions of their translation abilities in comparison to the actual professionally-rated quality. To this end, we would ask two professional translators to provide a quality rating of all tweet translations, and compare their scores to the original self-rated scores.

It would also be interesting to analyse more closely the tweets where the self-rating score differed significantly to that of the reviewer. By analysing the disagreements, we would have an insight into whether the reasons were due to major grammatical errors, problems with adequacy, fluency or merely typos or misuse of elements such as hashtags. This would give us a better insight into how reliable self-rating is as a metric for evaluation.

It is also our aim to perform preliminary MT experiments using the crowd-sourced data with the view to creating a UGC-specific MT system for the translation of English↔Irish text.

Acknowledgements

This work is 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 like to acknowledge the work Saurabh Gupta put into the development of this crowd-sourcing platform, and to thank him for his support in our data analysis. We would also like to thank Yvette Graham, Áine Monk, Aoife Mitchell, and Abigail Walsh for their valuable comments and advice. We would also like to thank the two anonymous reviewers for their useful comments.

Bibliography

Afli, Haithem, Sorcha McGuire, and Andy Way. 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, 2017.

Ambati, Vamshi, Stephan Vogel, and Jaime G. Carbonell. Active Learning and Crowd-Sourcing for Machine Translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2169–2174. European Language Resources Association, 2010.

Arcan, Mihael, Caoilfhionn Lane, Eoin Ó Droighneáin, and Paul Buitelaar. IRIS: English-Irish Machine Translation System. In *Language Resources and Evaluation Conference*. Special Interest Group on the Design of Communication (SIGDOC), 2016.

Bakliwal, Akshat, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Ceron, Andrea, Luigi Curini, and Stefano M. Iacus. Using Sentiment Analysis to Monitor Electoral Campaigns. *Social Science Computer Review*, 33(1):3–20, 2017/03/31 2014.

CNGL-DCU Team. Brazilator. In *The 11th Conference of the Association for Machine Translation in the Americas*, 2014. URL `http://www.mt-archive.info/10/AMTA-2014-showcase.pdf`.

Darmody, Merike, Tania Daly, et al. Attitudes towards the Irish Language on the Island of Ireland. *The Economic and Social Research Institute*, 2015.

Dowling, Meghan, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge. Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish. 2015.

Gupta, Aditi and Ponnurangam Kumaraguru. Credibility Ranking of Tweets During High Impact Events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, New York, NY, USA, 2012. ACM. doi: 10.1145/2185354.2185356.

Jin, Fang, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological Modeling of News and Rumors on Twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, SNAKDD '13, pages 8:1–8:9, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2330-7. doi: 10.1145/2501025.2501027.

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha. *The Irish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, 2012.

Lackaff, Derek and William J. Moner. Local languages, global networks: Mobile design for minority language users. In *Proceedings of the 34th ACM International Conference on the Design of Communication*, page 14. ACM, 2016.

Lukasik, Michal, Trevor Cohn, and Kalina Bontcheva. Classifying Tweet Level Judgements of Rumours in Social Media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2590–2595, 2015.

Lynn, Teresa, Kevin Scannell, and Eimear Maguire. Minority language Twitter: Part-of-speech tagging and analysis of Irish tweets. *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2015.

Mitra, Tanushree, Graham P. Wright, and Eric Gilbert. A Parsimonious Language Model of Social Media Credibility Across Disparate Events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 126–145, New York, NY, USA, 2017. ACM. doi: 10.1145/2998181.2998351.

Moorkens, Joss. Irish Translator Survey Report. Dublin City University, 2016.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Cohen, William W. and Samuel Gosling, editors, *ICWSM*. The AAAI Press, 2010.

Zaidan, Omar F. and Chris Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Address for correspondence:
Meghan Dowling
meghan.dowling@adaptcentre.ie
ADAPT Centre, Dublin City University, Glasnevin, Dublin 9, Ireland