

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 sociolinguistic 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.

c 2019 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

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

¹Note 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 Proceedings of the Celtic Language Technology Workshop 2019 Dublin, 19–23 Aug