PaddyWaC: A Minimally-Supervised Web-Corpus of Hiberno-English

Brian Murphy
Centre for Mind/Brain Sciences,
University of Trento
38068 Rovereto (TN), Italy
brian.murphy@unitn.it

Egon Stemle
Centre for Mind/Brain Sciences,
University of Trento
38068 Rovereto (TN), Italy
egon.stemle@unitn.it

Abstract

Small, manually assembled corpora may be available for less dominant languages and dialects, but producing web-scale resources remains a challenge. Even when considerable quantities of text are present on the web, finding this text, and distinguishing it from related languages in the same region can be difficult. For example less dominant variants of English (e.g. New Zealander, Singaporean, Canadian, Irish, South African) may be found under their respective national domains, but will be partially mixed with Englishes of the British and US varieties, perhaps through syndication of journalism, or the local reuse of text by multinational companies. Less formal dialectal usage may be scattered more widely over the internet through mechanisms such as wiki or blog authoring. Here we automatically construct a corpus of Hiberno-English (English as spoken in Ireland) using a variety of methods: filtering by national domain, filtering by orthographic conventions, and bootstrapping from a set of Ireland-specific terms (slang, place names, organisations). We evaluate the national specificity of the resulting corpora by measuring the incidence of topical terms, and several grammatical constructions that are particular to Hiberno-English. The results show that domain filtering is very effective for isolating text that is topic-specific, and orthographic classification can exclude some non-Irish texts, but that selected seeds are necessary to extract considerable quantities of more informal, dialectal text.

1 Introduction

For less dominant language variants, corpora are usually painstakingly constructed by hand. This results in high quality collections of text, classified and balanced by genre, register and modality. But the process is time-consuming and expensive, and results in relatively small resources. For example the International Corpus of English (ICE) project (Greenbaum, 1996) has already resulted in the publication of corpora covering ten dialects of English, following a common schema, but the individual corpora are limited to approximately one million words.

An alternative is to use automatic methods to harvest corpora from the Web. Identification of major languages is a robust technology, and where the regional boundaries of a language or dialect correspond closely to a national top-level internet domain, very large collections (of several billion words) can now can be produced easily, with close to no manual intervention (Baroni et al., 2009). These methods can also deal with some issues of text quality found on the web, successfully extracting coherent pieces of running text from web pages (i.e. discarding menu text, generic headings, copyright and other legal notices), reducing textual duplication, and identifying spam, portal pages and other files that do not contain linguistically interesting text.

Corpora of minor languages that lack their own domain, but that have clear orthographic differences from more dominant neighbouring languages can be collected automatically by using a small set of seed documents, from which language-specific search terms can be extracted (Scannell, 2007). These methods, combined with automated language identification methods, can quickly produce large, clean collections with close to no manual intervention.

However for language variants that do not have their own domain (e.g. Scots, Bavarian), it is less clear that such web corpora can be automatically constructed. Smaller or politically less dominant countries that do have their own domain (e.g. Belgium, New Zealand), may also find the language of their "national" web strongly influenced by other language varieties, for example through syndication of journalistic articles, or materials published by foreign companies.

In this paper we use minimally supervised methods (Baroni and Bernardini, 2004; Baroni et al., 2009) to quickly and cheaply build corpora of Hiberno-English (English as spoken in Ireland), which are many times larger than ICE-Ireland, the largest published collection

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 22–29,
Edinburgh, Scotland, UK, July 27–31, 2011.

currently available (Kallen and Kirk, 2007). We investigate several combinations of strategies (based on domain names, and on regional variations in vocabulary and orthography) to distinguish text written in this minor language variant from related dominant variants (US and UK English). We validate the specificity of the resulting corpora by measuring the incidence of Ireland-specific language, both topically (the frequency with which Irish regions and organisations are mentioned), and structurally, by the presence of grammatical constructions that are particular to Hiberno-English. We also compare our corpus to another web-corpus of Hiberno-English that is in development (*Crúbadán*, Scannell, personal communication) that relies on domain filtering of crawled web-pages.

The results show that filtering by national domain is very effective in identifying text that deals with Irish topics, but that the grammar of the resulting text is largely standard. Using a set of seed terms tailored to the language variant (Irish slang, names of Ireland-based organisations, loanwords from Irish Gaelic), yields text which is much more particular to Hiberno-English usage. At the same time, such tailored seed terms increase the danger of finding "non-authentic" uses of Irishisms (sometimes termed *paddywhackery* or *oirish*), either in fictional dialogues, or in documents discussing distinctive patterns in Irish English. The application of a British/American spelling filter has less clear effects, increasing topical incidence slightly, while reducing structural incidences somewhat.

The paper proceeds as follows: in the next section we introduce Hiberno-English, situating it relative to other variants of English, and concentrating on the characteristic features that will be used as metrics of "Irishness" of text retrieved from the Web. Next we describe the process by which several candidate corpora of Hiberno-English were constructed (section 3), and the methods we used to quantify incidence of distinctive usage (section 4). In the final two sections we compare the incidence of these markers with those found in corpora of other variants of English (UK, US), Scannell's IE-domain filtered corpus, and a hand-crafted corpus of Hiberno-English (ICE-Ireland), and reflect on the wider applicability of these methods to variants of other languages and orthographies.

2 Structures and Lexicon of Hiberno-English

Hiberno-English differs in a range of ways from other varieties of English. In broad terms it can be grouped with British English, in that its lexicon, grammar and orthographic conventions are more similar to that of Great Britain, than to that of North America. For example with lexical variants such as *bumper/fender, rubbish bin/trash can, lift/elevator* and *zed/zee* it shares the former British usage rather than the latter American usage, though there are exceptions (in Irish usage the North Americans term *truck* is replacing the British *lorry*). Similarly in syntax it tends to follow British conventions, for instance *He's familiar with X* rather than *X is familiar to him*, *write to me* rather than *write me* and the acceptability of singular verbal marking with group subjects, as in *the team are pleased* – though there are counterexamples again, in that Irish English tends to follow American dialects in dispensing with the *shall/will* distinction. Most obviously, Irish writing uses British spellings rather than American spellings.

However, there are still dialectal differences between Irish and British English. Beyond the usual regional differences that one might find between the words used in different parts of England, the English spoken in Ireland is particularly influenced by the Irish language (Gaelic, *Gaeilge*) (Kirk and Kallen, 2007). While English is the first language of the overwhelming majority of residents of Ireland (estimates of Irish mother-tongue speakers are of the order of 50,000, or about 1% of the population), Irish retains status as the first official language of the Republic of Ireland, maintained as a core subject at all levels of school education, and through state-maintained radio and television channels. As recently as the early 19th century, Irish was the majority language, and so many traces of it remain in modern Hiberno-English, in the form of Irish loan-words (e.g. *slán* 'goodbye', *gaelscoil* 'Irish (speaking) school'), Anglicizations (e.g. 'gansey', jumper, from Irish *geansaí*), and composites (e.g. 'jackeen', a pejorative term for Dubliners, combining the Irish diminutive *-ín* with the English 'Jack').

In this paper we take a series of characteristic terms and structures from Hiberno-English, mostly inspired by (Kirk and Kallen, 2007), and use them as markers of the Irishness of the text we assemble from the web. While there are many more interesting grammatical differences between Hiberno-English and other variants (e.g. perfective use of the simple present: *I know that family for years*), we restrict ourselves to those that can be automatically identified in a corpus through searching of plain text, or of shallow syntactic patterns (parts of speech).

The first marker we use is to measure the incidence of a set of terms that are topically related to Ireland: proper names of Ireland-based organisations, and geographical terms. The method for assembling this list is described in section 4.

The most simple structure that we use as a marker of Hiberno-English is the contraction *I amn't* (*I'm not* or *I ain't* in other varieties). The next is the "after" perfective, which often expresses immediacy, and a negative outcome:

(1) I'm after losing my wallet
'I just lost my wallet'

A further structure that is novel from the point of view of other variants of English is a particular use of verbs that take a complement that expresses a question (most commonly *ask, wonder, see* and *know*), without the use of a complementizer such as *if* or *whether* and with an inversion of subject-verb order (typical of interrogatives):

(2) I wonder is he coming"
'I wonder if/whether he is coming'

Finally we consider the expanded usage of reflexive pronouns in Hiberno-English, where they may be used for emphasis, in any argument position, and without being anaphorically bound, as is usually required. Here we limit ourselves to subject position reflexives, which can be identified from word order patterns, without any deeper semantic analysis:

(3) himself is in big trouble
'he is in big trouble'

With the exception of the *amn't* contraction, all of these phenomena are demonstrated by (Kirk and Kallen, 2007) to be common in the ICE-Ireland corpus, though somewhat less common in Northern Irish portion of that collection, and to be very rare or completely absent in the ICE-GB corpus of the English of Britain (Nelson et al., 2002). Significantly, these constructions are found predominantly in the spoken language portion of the ICE-Ireland corpus, suggesting that speakers are perhaps aware that they are not "standard" English, and so not considered appropriate in the written register.

3 Constructing a Web-Corpus of Hiberno-English

Within the WaCky initiative (Web-as-Corpus kool ynitiative) (Baroni and Bernardini, 2006) a community of linguists and information technology specialists developed a set of tools to selectively crawl sections of the Web, and then process, index and search the resulting data. Contributions like BootCaT (Baroni and Bernardini, 2004), an iterative procedure to bootstrap specialised corpora and terms from the Web, have been successfully used in a range of projects: first in the construction of the *WaCky corpora*, a collection of very large (>1 billion words) corpora of English (ukWaC), German (deWaC) and Italian (itWaC); and subsequently by other groups, e.g. noWaC and jpWaC (Baroni et al., 2009; Guevara, 2010; Erjavec et al., 2008).

Here we use BootCaT to build seven prototype corpora of Hiberno-English, and evaluate the dialect-specificity of each by measuring the incidence of proper terms and constructions that are associated with this language variant. Additionally, we use ukWaC as the de-facto standard British English Web corpus, and construct a medium size web-corpus of the US domain to represent American usage. Each corpus is preprocessed and formatted for the IMS Open Corpus Workbench (CWB, (Christ, 1994; Web, 2008)), a generic query engine for large text corpora that was developed for applications in computational lexicography.

BootCaT first takes a set of manually assembled seed terms, these (possibly multi-word) terms are randomly combined, and then are used as search queries with a Web search engine; the HTML documents of the top results are downloaded and cleaned to extract running text and discard all web-markup. Preprocessing and formatting for the CWB consists of tokenising, lemmatising, and part-of-speech tagging the corpus, and then converting the result into CWB's internal format; we replicated the processing stages employed for ukWaC.

The construction of the nine corpora differs on three dimensions:

Seeds: two seed sets were used namely, an Hiberno-English one (IEs), and the original ukWaC list of mid-frequency terms (UKs) from the British National Corpus (Burnard, 1995); the Irish seeds were used in pairs and triples to attempt to vary the degree of regional specificity.

TLDs: two types of top-level internet domain (TLD) restrictions were imposed during (or after) the construction of the corpora; either no restriction was imposed (.ALL), or a corpus was filtered by a specific national TLD (e.g. .ie).

Spelling: two types of spelling filter were imposed; either none, or an 'orthographic convention factor' (OCF) was calculated to detect American and British spellings, and a corpus was filtered accordingly (BrEn).

The IE seeds contained 81 seed terms, gathered using one author's native intuition, and words indicated as being specific to Irish English by the Oxford English Dictionary, and from various Web pages about Hiberno-English. 76 single-word and 5 two-word terms were used falling into three main categories: Irish place names, regional variant terms (mostly slang), and load words from Irish Gaelic (many being state institutions). The full listing of terms is given here:

Place names: Dublin, Galway, Waterford, Drogheda, Antrim, Derry, Kildare, Meath, Donegal, Armagh, Wexford, Wicklow, Louth, Kilkenny, Westmeath, Offaly, Laois, Belfast, Cavan, Sligo, Roscommon, Monaghan, Fermanagh, Carlow, Longford, Leitrim, Navan, Ennis, Tralee, Leinster, Connaught, Munster, Ulster

Regional variants: banjaxed (wrecked), craic (fun), fecking (variant of fucking), yoke (thing), yer man/one/wan (that man/woman), culchie (country dweller), da (father), footpath (pavement),

gaff (home), gobshite (curse), gurrier (young child), jackeen (Dubliner), jacks (toilet), janey mac (exclamation), jaysus (variant of exclamation "jesus"), kip (sleep; hovel), knacker (Traveller, gypsy), knackered (wrecked), langer (penis; idiot), langers/langered (drunk), scallion (spring onion), skanger (disgusting person), strand (beach, seaside), scuttered (drunk), boreen (small road), gob (mouth; spit), eejit (variant of idiot), lough (lake), fooster (dawdle), barmbrack (traditional Hallow'een cake), shebeen (unlicensed bar), bogman (contry dweller), old one (old lady), quare (variant queer), gansey (pullover)

Loan words: garda, gardaí (police), taoiseach (prime minister), dáil (parliament), Sláinte ("cheers"), Gaeltacht (Irish speaking areas), Seanad (senate), Tánaiste (deputy prime minister), ceol ((traditional Irish) music), slán ("goodbye"), grá (affection, love for), gaelscoil (Irish speaking school)

These seed terms were combined into a set of 3000 3-tuple (3T) and a set of 3000 2-tuple (2T) search queries, i.e. two-word terms were enclosed in inverted commas to form one single term for the search engine. For 3T this resulted in over 80% 3-tuples with 3 single-word terms, and slightly over 17% with 2 single-word terms, and the remaining percentages for 3-tuples with 1 single-word and no single-word terms; for 2T this resulted in almost 88% 2-tuples with 2 single-word terms, almost 12% with only 1 single-word terms, and less than 1% with no single-word terms. The UK seeds were the original ones used during the construction of the ukWaC corpus and they were combined into 3000 3-tuple search queries.

No TLD restriction means that the search engine was not instructed to return search results within a specific domain, and hence, documents originate from typical English-language domains (.com, .ie, .uk, etc.) but also from .de and potentially any other. A restriction meant that the documents could only originate from one TLD.

No spelling filter means that nothing was done. The OCF indicates the degree to which terms within a document are predominantly spelled according to one predefined word list relative to another. The number of term intersections with each list is counted and OCF is calculated as the difference between counts over their sum. To simplify matters, we utilised a spell-checker to return the list of known words from a document, this corresponds to checking a document for spelling errors and only keeping the non-erroneous words. In our case we used an en_GB dictionary, an en_US one, and the two together. The three lists yield the needed numbers of words only known by one of the two dictionaries, and, hence unknown by the other dictionary, and the ratio in the range of $[-1, +1]$ can be calculated.

The search engine we used for all queries was Yahoo (Yahoo! Inc., 1995); for all search queries English results were requested, that is we relied on the search engine's built-in language identification algorithm[1], and from all search queries the top 10 results were used. Cleaning of the Web pages (termed *boilerplate removal*) was accomplished by BootCaT's implementation of the BTE method (Finn et al., 2001); it strives to extract the main body of a Web page, that is the largest contiguous text area with the least amount of intervening non-text elements (HTML tags), and discards the rest.

Several corpora were constructed from the Irish seeds using 2- or 3-tuple search terms: either without restricting the TLDs; subsequent restriction to the .ie TLD; or subsequent filtering according to spelling. Corpora were also constructed with the search engine instructed to directly return documents from the .us or the .ie TLD, respectively, where the latter one was later also filtered according to spelling. The ukWaC corpus is restricted to the .uk TLD.

4 Evaluating Variety Specificity of the Corpus

To evaluate the dialectal specificity of the text in each putative corpus of Hiberno-English, we measured the incidence of several characteristic terms and structures. The same phenomena were counted in corpora of US and UK English (identified as that found under the .us and .uk TLDs respectively) to establish baseline frequencies. All corpora were HTML-cleaned, lemmatised and part-of-speech tagged using the same methods described above, and searches were made with identical, case-insensitive, queries in the CQP language.

First we quantified topical specificity by searching for a set of Irish geographical terms (towns, counties, regions), and Ireland-based organisations (companies, NGOs, public-private bodies), to identify text which is "about Ireland". There were 80 terms, evenly split between the two categories. In this list we avoided proper names which are orthographically identical to content words (e.g. Down, Cork, Clones, Trim, Limerick, Mallow, Mayo), given names (Clare, Kerry, Tyrone), place names found in other territories (Baltimore, Skibbereen, Newbridge, Westport, Passage West), or names that might be found as common noun-phrases (e.g. Horse Racing Ireland, Prize Bond Company, Electricity Supply Board). While political terms might have been appropriate markers (e.g. the political party Fianna Fáil; the parliamentary speaker the Ceann Comhairle), the seed terms we used contained many governmental institutions, and so this could be considered an unfairly biased diagnostic marker. The full list of terms is given below.

Topical terms: ActionAid, Aer, Aer, Allied, An, Arklow, Athlone, Athy, Balbriggan, Ballina, Ballinasloe, Bantry, Bord, Bord, Bord, Buncrana, Bundoran, Bus, Carrick-on-Suir, Carrickmacross, Cashel, Castlebar, Christian, Clonakilty, Clonmel, Cobh, Coillte, Comhl(ála)mh, Connacht, C(ólo)ras, Donegal, Dublin, Dublin, Dungarvan, Eircom, EirGrid, Enniscorthy, Fermoy, Fyffes, Glan-

[1] This restriction is very effective at distinguishing non-English from English content, but returns content from any English variant.

bia, Gorta, Grafton, Greencore, Iarnr(ólo)d, IONA, Irish, Irish, Irish, Kerry, Kilkee, Kilrush, Kinsale, Laois, Leixlip, Letterkenny, Listowel, Listowel, Loughrea, Macroom, Mullingar, Naas, Nenagh, Oxfam, Paddy, Portlaoise, Radi(oló), Ryanair, Telif(íli)s, Templemore, Thurles, Tipperary, Tramore, Trinity, Tr(ólo)caire, Tuam, Tullamore, Tullow, Vhi, Waterford, Youghal

For the structural markers we used more conservative query patterns where appropriate, to minimise false positives. For this reason the incidence figures given here should be considered lower estimates of the frequency of these structures, but they allow us to establish an independent metric with a minimum of manual intervention.

As mentioned above, for the emphatic use of reflexives, we searched only in the subject verb configuration, even though these are possible in other argument positions also (e.g. *I saw himself in the pub yesterday*). The query was restricted to reflexive pronouns (other than *itself*) found at the start of a sentence, or immediately after a conjunction, and directly before a finite verb (other than *have* or *be*). The CQP query (4) yields examples such as (5)-(7).

(4) [pos="CC" | pos="SENT"] [lemma=".+self" & lemma!="itself"] [pos="VV[ZD]?"];

(5) ... more commonplace or didactic, less imaginative? **Himself added**, "You are a romantic idiot, and I love you more than...

(6) ... Instruments in Lansing, Michigan, where Val and Don **and myself taught** bouzouki, mandolin, guitar and fiddle workshops. It is a...

(7) ... game of crazy golf, except this time it was outdoor. Conor **and myself got** bored straight away so we formed our own game while Mike ...

For the "after" perfective construction, we searched for a pattern of a personal pronoun (i.e. not including *it, this, that*), the lexeme *after*, and a gerund form of a common verb (other than *have, be*). The query (8) allowed for a modal auxiliary, and for intervening adverbs, as illustrated in (9)-(11).

(8) [pos="PP" & word!="it" %c & word!="that" %c & word!="this" %c] [pos="RB.*"]* [lemma="be"] [pos="RB.*"]* [word="after"] [pos="RB.*"]* [pos="V[VH]G"]

(9) ... the holy angels on your head, young fellow. I hear tell **you're after winning** all in the sports below; and wasn't it a shame I didn't ...

(10) ... MICHAEL – Is the old lad killed surely? PHILLY. **I'm after feeling** the last gasps quitting his heart. MICHAEL – Look at ...

(11) ... placards with the words "Blind as a Batt" and "Batman **you are after robbing** us". They came from as far away as Wexford and called ...

The use of embedded inversions in complements was queried for the same four verbs identified by (Kirk and Kallen, 2007): *ask*, *see*, *wonder* and *know*. Other verbs were considered, by expansion from these four via Levin verb classes (Levin, 1993), but preliminary results gave many false positives. The query used search for one of these four verbs, followed by a form of the verb *be*, and then a personal pronoun specific to the subject position (12). Examples of the instances extracted are given below (13)-(15).

(12) [pos="VV.*" & lemma="(ask|know|see|wonder)" %c] [lemma="be"] [word="(I|he|she|we|they)" %c];

(13) ... but that is the reality. I remember as a young child being **asked was I** a Protestant or a Catholic: that's the worst thing ...

(14) ... unless I get 170+, there isn't a chance. And then I **wonder am I** mad even applying for medicine. Anyway anyone else who's...

(15) There was the all important question and she was dying to **know was he** a married man or a widower who had lost his wife or some ...

Finally, examples of the *amn't* contraction (17)-(19) were extracted with the simple case-insensitive query (16).

(16) "am" "n't";

(17) Hi I'm relatively new to CCTV but work in IT and so **amn't** 100 % lost ! Anyway, I have already set up a personal ...

(18) ... and plaster, with some pride.) It was he did that, and **amn't** I a great wonder to think I 've traced him ten days with ...

(19) "I will indeed Mrs. R, thanks very much, sure **amn't** I only parchin?" Ye needn't have gone to the trouble of ...

It should be noted that these structural usages differ in the degree to which they are perceived as distinctive. While speakers of Irish English may not be aware that *amn't* and the embedded inversion construction are dialectally restricted, many do know that the *after* and reflexive constructions are particular to Ireland. Hence by searching for these constructions our evaluation is biased towards colloquial language and consciously dialectal usage.

5 Results

As can be seen in the first two rows of table 1, considerably large Irish corpora were gathered with ease, and even after applying several subsequent filtering strategies, the smallest corpus was several times the size of the manually assembled ICE-Ireland corpus.

Figure 1 (left panel) further shows that the strategy of searching by random seed combinations yielded pages in many domains, with a considerable proportion being in the .ie domain, but by no means the majority. This suggests that Ireland specific usage of English is not restricted to the national internet domain, i.e. the .ie TLD. The relative proportion of .ie domain pages (see right panel of same figure) was increased by selecting only pages which had predominantly British orthography, suggesting that this has some efficacy in eliminating texts written in American English.

Table 1 also shows the absolute incidence of each of the five characteristic phenomena considered. All matches returned by the CQP search queries were manually evaluated, to ensure that they were authentic examples of the constructions in question (for the larger ukWaC corpus only a random sample were examined). Numbers of false positives that were excluded are shown in brackets, such as the examples from ukWaC below:

(20) ... just as they were **after** receiving secret briefings from Health Commission Wales officers.

(21) All I **know is they**'re getting cold.

The bars in sets one and two show figures for the manually compiled ICE-Ireland corpus, and the Crúbadán web-corpus. The ICE-Ireland numbers differ somewhat from those reported in that paper (Kirk and Kallen, 2007), since we used more selective search strategies (note that the cut-off reported relative incidences reach about 21 per mil. tokens), which would miss some examples such as those below which have the after construction without a personal pronoun, and have the non-reflexive use in object position, respectively:

(22) There's nothing new **after** coming in anyway so

(23) Again it's up to **yourself** which type of pricing policy you use

It should also be noted that ICE-Ireland, following the standard scheme for the International Corpus of English project (Greenbaum, 1996), is biased towards spoken language, with written text only making up only 40% of the total text.

The relative incidence (per million tokens) of Ireland-specific topics and constructions is summarised in figure 2. The bars in sets three and four demonstrate that these same characteristics, very common in Hiberno-English as evidenced by the ICE-Ireland, appear to be exceedingly rare in UK and US English. Unsurprisingly, web authors in the US and UK domains do not write often about Irish places and organisations. But constructions that are putatively exclusive to Hiberno-English are seldom found. Those that are found might be explained by the effect of language contact with Irish immigrants to those countries, and the fact that text by Irish authors may be found in these domains, whether those people are resident in those countries or not. For instance in the example below, the given name *Ronan* suggests that the author might be of Irish extraction:

(24) At about that point Cardinal Cormac of Westminster walked right past us and Ronan and **myself** went to say hello to him and tell him we were up here from his diocese.

The sets headed ".ie" show the figures for the corpora we constructed by querying seed terms within the Irish national domain. The incidence of characteristic features of Hiberno-English grammar are higher than those seen in the US and UK domains, similar to that seen in the Crúbadán corpus, and lower than in the ICE-Ireland corpus, perhaps reflecting the fact that these constructions are less common in written Hiberno-English. Subsequent filtering out of pages with dominance of American English spelling (".ie, BrEn") does not have much effect on the numbers.

The "Irish Seeds (IEs)" bars show that the use of tailored seed terms returns text which has a similar topical specificity to that in the .ie domain generally, but which shows more structural characteristics of Hiberno-English. These results can also be improved upon, first by concentrating on the .ie domain portion of the tailored-seeds extracted pages ("Irish Seeds (IEs), IE Dom (.ie)") which boosts topical specificity. Filtering instead by orthography ("IEs, BrEn") seems to strike a happy medium, increasing incidence in all categories.

However returning to table 1, it is apparent that there are many false positives among the constructions found using Irish seed terms. This was caused by the search strategy retrieving a small number of pages on the topic of Hiberno-English, that contained many constructed examples of the structures of interest. The same corpora contained smaller numbers of examples from theatre scripts and other fiction.

6 Discussion

The results show us that our methods can be effective in extracting text that is both specific to Irish topics, and includes instances of constructions that are particular to the variety of English spoken in Ireland. The incidences relative to corpus size are not as high as those seen in the

Table 1: Corpora sizes, incidences of Ireland terms and constructions; absolute numbers (false positives in brackets)

	ICE-Ireland	Crubadan	ukWaC	UKs, 3T, .us	UKs, 3T, .ie	UKs, 3T, .ie, BrEn	IEs, 3T, .ALL	IEs, 3T, .ALL, .ie	IEs, 3T, .ALL, BrEn	IEs, 2T, .ALL	IEs, 2T, .ie
Size (in 10^6 Tokens)	1.1	46.3	2119.9	74.7	17.8	15.0	25.2	2.6	17.3	18.4	6.4
Size (in 10^3 Docs)	0.5	43.0	2692.6	4.6	2.0	1.6	3.4	0.7	2.5	7.3	2.3
Ireland Terms	194	17330	12743	82	14199	13802	23527	7264	22071	12454	9935
"after" Construction	7 (-4)	12 (2)	48 (72)	1 (2)	11 (1)	7 (1)	26 (50)	2 (1)	11 (47)	14 (38)	9 (1)
"amn't" Construction	0 (0)	0 (0)	32 (0)	0 (0)	0 (0)	0 (0)	5 (45)	1 (1)	2 (43)	6 (36)	0 (0)
embedded Inversions	24 (-18)	18 (5)	42 (309)	0 (15)	5 (2)	5 (0)	20 (4)	2 (1)	17 (2)	4 (1)	5 (0)
Subject Reflexives	22 (-19)	33 (0)	1797 (115)	35 (8)	15 (1)	10 (0)	39 (0)	2 (0)	30 (0)	17 (3)	8 (1)

Figure 1: Domain composition of Irish-Seed based Corpora

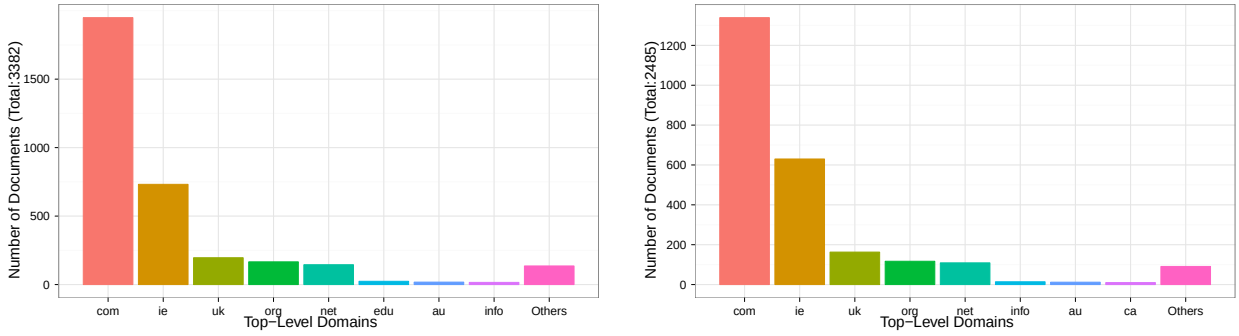
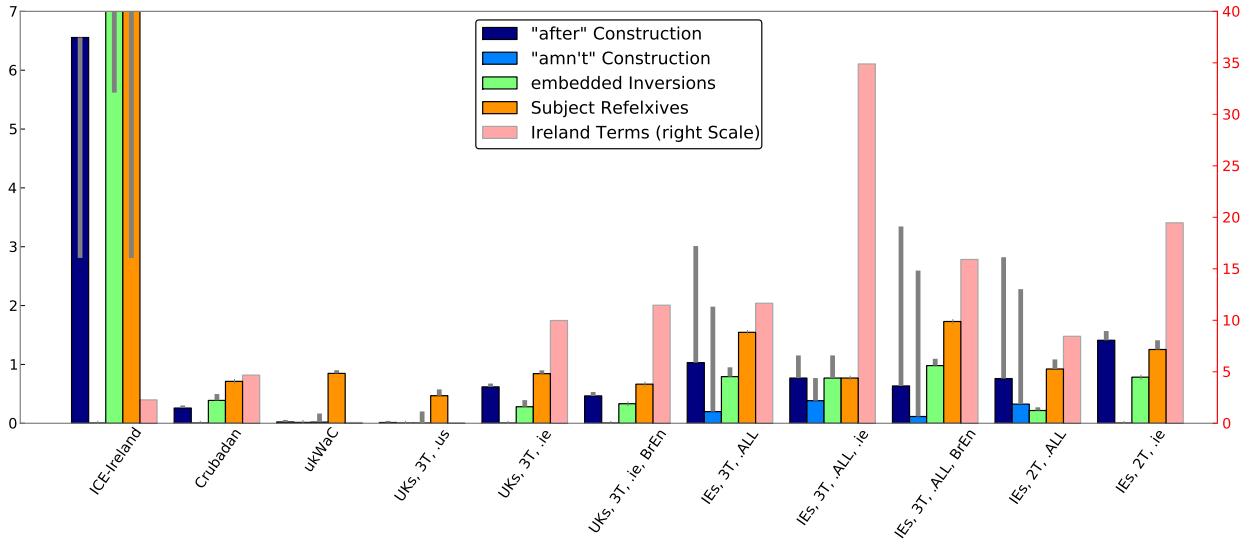


Figure 2: Relative Incidences of Ireland terms and constructions, per million words (grey bars indicating the original counts before manual inspection), in each copus



manually constructed ICE-Ireland corpus. We can speculate on the reasons for this. It may be in part due to "pollution" of our corpus with non-Irish English, via syndicated journalism (e.g. some Irish newspapers are repackaging of British newspapers with added Irish content), or via multinational organisations with bases in Ireland. In our view the main explanatory factor is that of modality and register. The ICE-Ireland corpus is predominantly spoken (~60%), with many texts coming from informal settings (unscripted speeches, face to face and telephone conversations). One reading of the figures which is consistent with this viewpoint is that the .ie domain corpora contain proportionally more high register, edited text (e.g. from governmental and commercial organisations, for which the use of the .ie domain may be an important part of corporate identity), and that the tailored-seed corpora contain more text contributed by individuals (forums, blogs, etc), for whom domain endings are of little consequence. Nevertheless, the use of Hiberno-English specific seed terms did reveal higher incidences of distinctive Irish usages than simple domain filtering.

But despite these lower incidences, in absolute terms our corpora provide many more examples of Hiberno-English than that were hitherto available. For example the ICE-Ireland corpus contains a total of seven examples of the "after" construction, while with our Irish-seeds derived corpus, and using a fairly restrictive query pattern, we isolated 26 examples of this structure. Further the size of these pilot corpora were kept intentionally limited, a small fraction of the approximately 150 million .ie domain pages indexed by Google. Much larger corpora could be constructed with relative ease, by using a larger seed set, or with an interactive seed-discovery method, where the text from the first round of web-harvesting could be analysed to identify further terms that are comparatively specific to Hiberno-English (relative to corpora of other varieties of English), in a similar fashion to the methods discussed in (Scannell, 2007).

In terms of wider implications, the fact that seeds tailored to a particular region and language variant is as effective as filtering by domain, is encouraging for dialects and minority languages that lack a dedicated internet domain. This suggest that for less-dominant language variants without distinctive established orthographies (e.g. Scots, Andalusian, Bavarian), large corpora displaying characteristic features of that variant can be constructed in a simple automatic manner with minimal supervision (a small set of seeds provided by native speakers). Our methods might also prove useful for dialects in which a standard variant is dominant in the written language (e.g. Arabic, Chinese). One might expect that the written Arabic in the .ma (Morocco) domain would differ little from that in the .qa domain (Qatar) despite the large differences in vernacular speech. Similarly the grammar and vocabulary of Chinese written in Mainland Chinese, Taiwanese, Hong Kong and Singaporese domains (ignoring orthography) might be less representative of the variation in everyday language. The use of regional slang and proper names may help one to collect more examples of this more natural language usage, and less of the dominant standard variant.

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In (ELRA), E. L. R. A., editor, *Proceedings of LREC 2004, Lisbon: ELDA.*, pages 1313–1316.

Baroni, M. and Bernardini, S., editors (2006). *Wacky! Working papers on the Web as Corpus*.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Burnard, L. (1995). *Users Reference Guide, British National Corpus, Version 1.0*. Oxford University Computing Services/British National Corpus Consortium, Oxford.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Papers in Computational Lexicography (COMPLEX '94)*, pages 22–32.

Erjavec, I. S., Erjavec, T., and Kilgarriff, A. (2008). A web corpus and word sketches for Japanese. *Information and Media Technologies*, 3:529–551.

Finn, A., Kushmerick, N., and Smyth, B. (2001). Fact or fiction: Content classification for digital libraries.

Greenbaum, S. (1996). *Comparing English Worldwide*. Clarendon Press.

Guevara, E. (2010). NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the Sixth Web as Corpus Workshop (WAC6)*, pages 1–7. The Association for Computational Linguistics.

Kallen, J. and Kirk, J. (2007). ICE-Ireland: Local variations on global standards. In Beal, J. C., Corrigan, K. P., and Moisl, H. L., editors, *Creating and Digitizing Language Corpora: Synchronic Databases*, volume 1, pages 121–162. Palgrave Macmillan, London.

Kirk, J. and Kallen, J. (2007). Assessing Celticity in a Corpus of Irish Standard English. In *The Celtic languages in contact: papers from the workshop within the framework of the XIII International Congress of Celtic Studies, Bonn, 26-27 July 2007*, page 270.

Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press, Chicago.

Nelson, G., Wallis, S., and Aarts, B. (2002). *Exploring natural language: working with the British component of the International Corpus of English*. John Benjamins.

Scannell, K. (2007). The Crúbadán project: Corpus building for under-resourced languages. In Fairon, C., Naets, H., Kilgarriff, A., and de Schryver, G.-M., editors, *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Web (2008). The IMS Open Corpus Workbench (CWB).

Yahoo! Inc. (1995). The Yahoo! Internet search engine.