Describing Irish English with the ICE Ireland Corpus

Schneider, Gerold

Abstract: The investigation of specific features of Irish English has a long tradition. Yet, with the arrival of large corpora and corpus tools, new avenues of research have opened up for the discipline. The present paper investigates features commonly ascribed to Irish English on the basis of the ICE Ireland corpus in comparison with ICE corpora representing other varieties of English. We use several corpus tools to access the ICE corpora. First, an offline concordance program, AntConc V 3.3 (Anthony 2004). Second, Corpus Navigator, an online corpus query tool allowing researchers to query regular expressions on the surface texts. Third, we are in the process of writing a version of Dependency Bank (Lehmann and Schneider 2012) which contains a selection of ICE corpora, and which will be called ICE online. This research methodology allows us to reassess how specific features found in Irish English are in comparison with other international varieties of English and illustrates that even simple corpus-based search patterns can produce powerful results.

Originally published at:
Schneider, Gerold (2013). Describing Irish English with the ICE Ireland Corpus. Cahiers de l'institut de linguistique et des sciences du langage, 38:137-162.

INVESTIGATING IRISH ENGLISH WITH ICE-IRELAND

Gerold SCHNEIDER
Universität Zürich
gschneid@es.uzh.ch

Abstract

The investigation of specific features of Irish English has a long tradition. Yet, with the arrival of large corpora and corpus tools, new avenues of research have opened up for the discipline. The present paper investigates features commonly ascribed to Irish English on the basis of the ICE Ireland corpus in comparison with ICE corpora representing other varieties of English. We use several corpus tools to access the ICE corpora. First, an offline concordance program, *AntConc* V 3.3 (Anthony 2004). Second, *Corpus Navigator*, an online corpus query tool allowing researchers to query regular expressions on the surface texts. Third, we are in the process of writing a version of *Dependency Bank* (Lehmann and Schneider 2012) which contains a selection of ICE corpora, and which will be called *ICE online.* This research methodology allows us to reassess how specific features found in Irish English are in comparison with other international varieties of English and illustrates that even simple corpus-based search patterns can produce powerful results.

1. INTRODUCTION

1.1 THE INTERNATIONAL CORPUS OF ENGLISH (ICE)

The International Corpus of English[1] (Greenbaum 1996) is a collection of corpora of national or regional varieties of English with a common corpus design and a common scheme for grammatical annotation.[2] The primary aim of the corpus series has been to provide material for comparative studies of English worldwide. Each ICE corpus consists of one million words of spoken and written English produced after 1989, and comprises 500 texts (300 spoken and 200 written) of approximately 2,000 words each.

[1] **http://ice-corpora.net/ice/index.htm**

[2] I am very grateful to Hans Martin Lehmann for writing the *ICE online* tool, and to John Kirk, Patricia Ronan and Shane Walshe for many inspiring linguistic discussions.

ICE-Ireland (Kallen & Kirk, 2008) contains both data from Northern Ireland and the Republic of Ireland, allowing us on the one hand to compare Irish English to other varieties of English, and on the other hand to investigate differences between Northern and Southern Irish English. For our comparison to Irish English, we have used most of the ICE corpora that are currently complete and available. We have included the following ICE Corpora in our comparison:

- ICE Canada
- ICE Great Britain
- ICE Hong Kong
- ICE India
- ICE Ireland
- ICE Jamaica
- ICE New Zealand
- ICE Philippines
- ICE Singapore

This selection includes four varieties that are known as Inner Circle varieties, i.e. first language varieties of English (Canada, Great Britain, New Zealand and Ireland itself) and five Outer Circle varieties, varieties from countries where English is used as a second language (Kachru, 1992). On the one hand, Irish English is certainly a classical Inner Circle variety, and has contributed to the variety formation in a large number of countries that were at some stage English colonies and in other territories. On the other hand, Ireland itself has been under British rule for several centuries, and had extensive contact with Celtic languages (Ronan, this volume). For long periods, the majority of the speakers of Irish English had Irish as their first language, which means that substrate influences which are typical of L2 languages and Outer Circle English varieties, can be expected and have been described in linguistic research. We give an overview of those features of Irish English that are of interest for this paper in section 1.2 and present detailed results of our own research into these features in ICE Ireland in section 3.

Our aim in this study is threefold. First, we would like to contribute to research on Irish English by showing which of the many of the described features can indeed be found in relatively small corpora such as ICE Ireland, and which others may be too rare, and which are potentially receding or do not hold up to empirical scrutiny. Second, we would like to illustrate, using ICE Ireland as a show-case, how corpus software can be used easily, also by the less computer-savvy, to investigate regional variation and operationalize linguistic features. Third, we will give a preview of

ICE online, an advanced online tool supporting syntactic queries and statistical tests. We discuss our retrieval approach in section 2.

1.2 IRISH ENGLISH FEATURES: OVERVIEW

Many features have been claimed or described for Irish English. The list of phenomena we investigate using the ICE Ireland corpus in this study is not comprehensive. We use Trudgill and Hannah (2002), Hickey (2007) and Filppula (1999) as starting points. Trudgill and Hannah (2002: 106-108) describe pronunciation differences, lexical features, and amongst others the following morphosyntactic characteristics of Irish English:

a. Low frequency of *shall* (section 3.2; see also McCafferty, 2011)
b. Habitual aspect with *do* (section 3.7; see also Filppula, 1999: 130ff.)
c. *After* perfect (section 3.5; see also Filppula, 1999: 99ff.)
d. Clefting with copular verbs (section 3.4; see also Filppula, 1999: 243 ff.)
e. Indirect questions with inversion (section 3.8; see also Filppula, 1999: 167ff.)

We discuss one lexical feature, and the listed morphoysntactic features in section 3. Except for the low frequency of *shall,* all of the morphosyntactic characteristics are also listed in Hickey (2007: 146-147), who also mentions additional features. We will discuss a selection of them, in particular

f. *For to* infinitive (section 3.3; see also Filppula, 1999: 185)
g. *Be* as auxiliary with past participle (section 3.11; see also Filppula, 1999: 114 ff.)
h. Singular existential with plural NP (section 3.6; see Hickey, 2005: 121 and Walshe 2009)

We further include two additional features from the extensive description of Filppula (1999):

i. reflexive pronouns in place of non-reflexive pronouns (section 3.9; Filppula, 1999: 77-8 calls them unbound reflexives)
j. medial object perfect (section 3.10; Filppula, 1999: 107 ff.)

2. METHODS

In order to retrieve instances of the features under discussion, we use *AntConc*, *Corpus Navigator*, and *ICE online*. We explain the retrieval queries that we use in detail, aiming to show that formulating queries is not difficult after an initial learning step. We discuss simple word-based queries, slightly more tricky regular expressions and powerful syntactic queries. Except in syntactic queries, the aim is typically to achieve an operationalization which is often a crude approximation. Typically, the queries retrieve many hits (the instances that are found and displayed to the user), but often the majority of them do not contain the feature under investigation, and we need to filter the results, separating the wheat (true positives) from the chaff (false positives, also referred to as garbage). Generally, this is a two-step procedure:

1. We formulate a permissive corpus search query, which should contain most of the instances of the phenomenon under investigation. In other words, it should have high recall, but it may have low precision.
2. We do a manual filtering and inspection to select those matches which really are instances. This step repairs the low precision from step 1.

The two evaluation measures *precision* and *recall* are defined as follows. Precision expresses how many of the returned matches are positive samples. Recall expresses how many of all the positive samples in the corpus are returned by the retrieval query. While we can easily increase precision by manual filtering of the hits, we can never be sure that our query has maximally high recall. Often, one is even ready to use queries which explicitly only find a subset of the instances of the phenomenon. We then need to make the assumption that the subset is representative of the complete set. Not every measurable difference between occurrence groups (for example Irish English versus British English or written versus spoken) is meaningful if counts are small, therefore we need to do significance tests to separate random fluctuation from significant differences.

An approach related to ours is Kirk and Kallen (2006), who searched ICE Ireland for the *after* perfect (section 3.5), the medial object perfect (section 3.10), unbound reflexives (section 3.9), and indirect questions with inversion (section 3.8). In the present study we include more features, and also use syntactically analysed corpora. The syntactic analysis is done with Pro3Gres (Schneider, 2008), a Dependency Grammar parser. The full annotation pipeline, and an introduction to syntactic queries, are given in Lehmann and Schneider (2012). A case study similar

to this has been carried out on Indian English (Schneider, 2013). There, the approach is different, however, because a corpus-driven approach has been used, while here we test well-known features.

3. RESULTS

3.1 LEXIS : THE EXAMPLE OF *WEE*

In place of the description of a host of lexical items, we restrict the discussion to one example: the word *wee* (with the meaning of « small »), which can be used by speakers of Irish English (see e.g Walshe, this volume). Finding lexical items in corpora using concordance software is trivial: typing the word from into the search field retrieves all its occurrences. Figure 1 shows the result of the word form query for <wee> using the concordance program *AntConc* V 3.3 (Anthony 2004), after loading the ICE Ireland corpus raw text, as distributed. Indeed, *AntConc* and *Corpus Navigator* report 220 hits, so as a first approximation we may assume that *wee* is an Irish English feature.

It could of course be, however, that *wee* is also frequent in other varieties of English. Loading some other ICE corpora reveals that *wee* is known in most Englishes, but is much less frequent. For the Inner Circle varieties, the counts are: 8 in ICE-Canada, 61 in ICE-New Zealand, 4 in ICE-GB. Even if we consider that the sizes of the corpora vary a little, the Irishness of the feature seems to be confirmed, and after applying a significance test which reveals that the difference is highly significant, we might stop here. However, *wee* is considered to be a typically Scottish dialectal word so it is no wonder then that it appears in ICE-GB and Irish and Scottish emigrant countries. If we split ICE-Ireland into its Northern (NI) and Southern (RoI) part, a stark difference appears: Northern Irish English, which has been much influenced by Scottish immigrants, dominates the counts, as figure 2 shows. The concordance plot also shows that the word is more frequent in the beginning of the corpora, which is due to the fact that the first 3/5ths of the corpora are spoken language. In conclusion, *wee* is mainly a Scottish feature which has come to Northern Ireland (and to other places, such as New Zealand) via Scottish settlers. However, it also enjoys relative popularity in Southern Irish English, and the feature distribution also shows that Scottish English is very narrowly represented in ICE-GB.

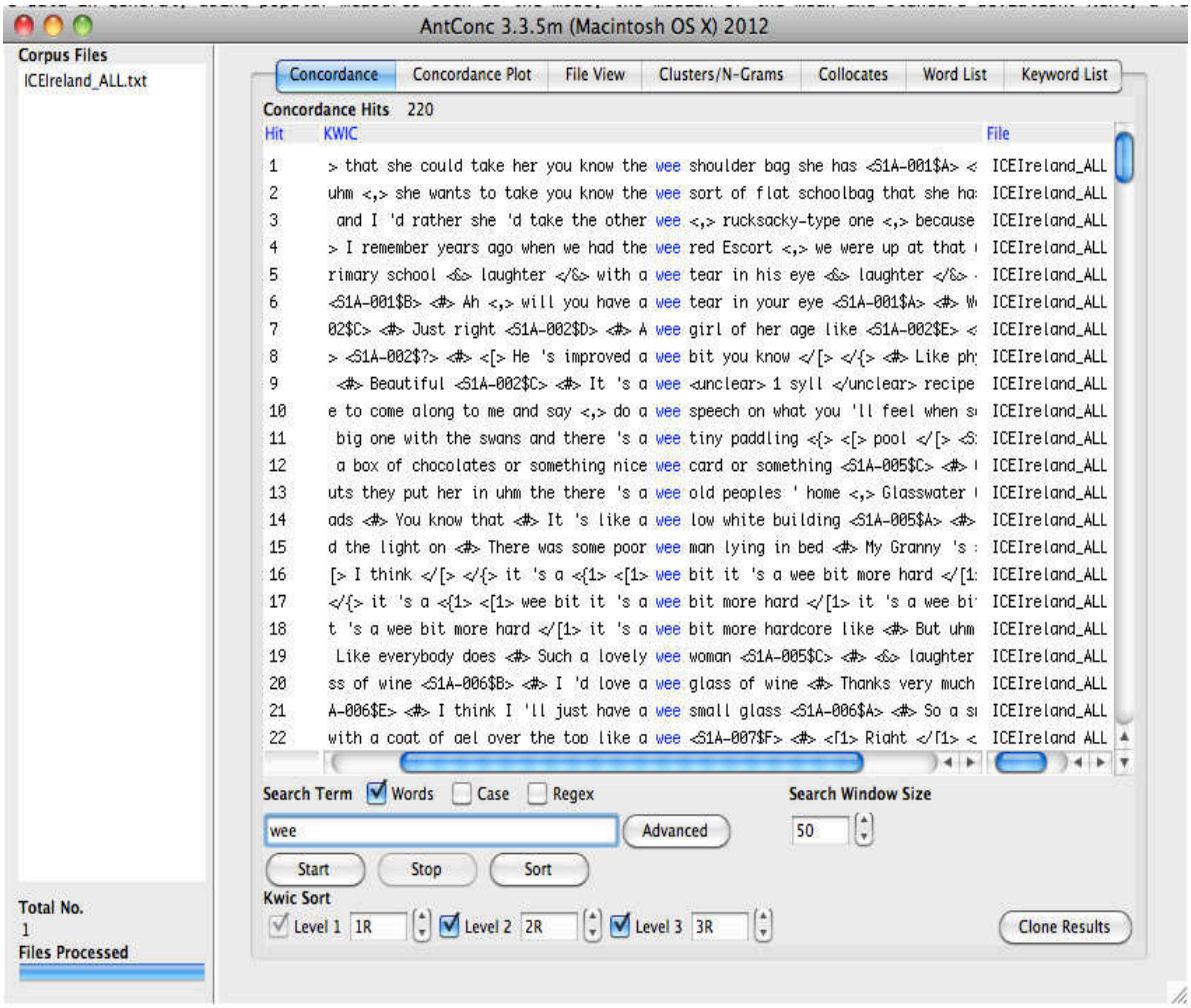


Figure 1: query for *wee* in ICE Ireland using the concordance program *AntConc*

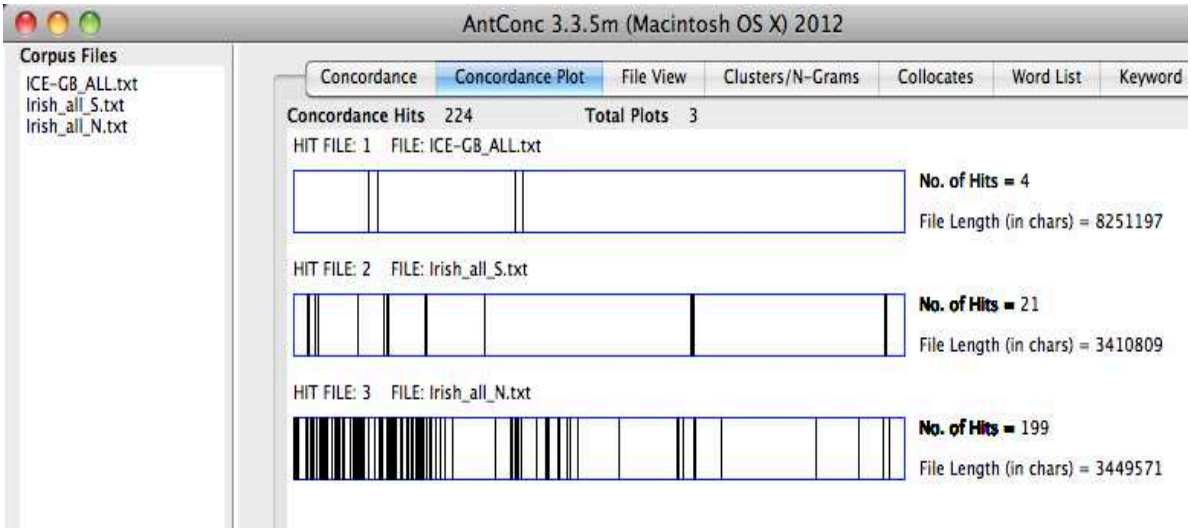
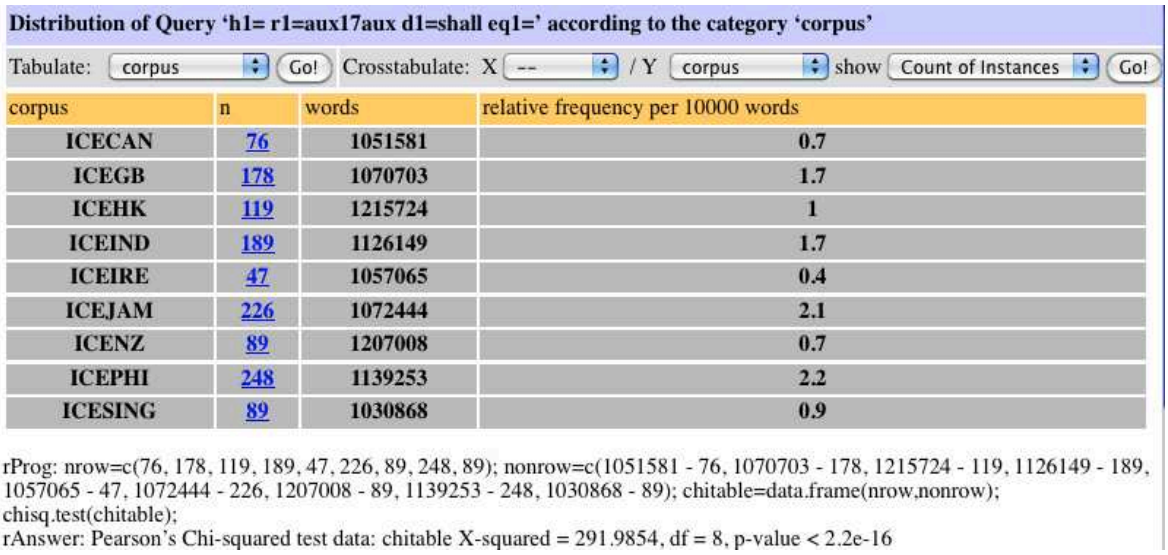


Figure 2. Concordance plot of *wee* in ICE-GB, ICE Ireland South and North

As described by Kirk and Kallen (2010), there are similarities between Scottish and Irish English. While they conclude that the influence was relatively small, there seem to be areas in lexis where there was considerable import, as is indeed visible in the use of *wee* in Irish English.

3.2 LOW FREQUENCY OF *SHALL*

In Irish English, *shall* is used very infrequently, most forms have been replaced by *will.* McCafferty (2011) traces the history of the decline of the auxiliary *shall* in Irish English. When comparing ICE-Ireland to other corpora from the ICE family, this feature is immediately apparent. The frequency of the auxiliary *shall* is lower than in any other 8 ICE corpora used in our study, and the difference is highly significant (chi-square contingency test, p<0.01%). In *ICE online*, the necessary chi-square test of significance is done automatically, which is useful when doing exploratory research.



Distribution of Query 'h1= r1=aux17aux d1=shall eq1=' according to the category 'corpus'

Tabulate: corpus Go! Crosstabulate: X -- / Y corpus show Count of Instances Go!

corpus	n	words	relative frequency per 10000 words
ICECAN	76	1051581	0.7
ICEGB	178	1070703	1.7
ICEHK	119	1215724	1
ICEIND	189	1126149	1.7
ICEIRE	47	1057065	0.4
ICEJAM	226	1072444	2.1
ICENZ	89	1207008	0.7
ICEPHI	248	1139253	2.2
ICESING	89	1030868	0.9

rProg: nrow=c(76, 178, 119, 189, 47, 226, 89, 248, 89); nonrow=c(1051581 - 76, 1070703 - 178, 1215724 - 119, 1126149 - 189, 1057065 - 47, 1072444 - 226, 1207008 - 89, 1139253 - 248, 1030868 - 89); chitable=data.frame(nrow,nonrow); chisq.test(chitable);
rAnswer: Pearson's Chi-squared test data: chitable X-squared = 291.9854, df = 8, p-value < 2.2e-16

Figure 3. Screenshot of the results of the query for shall as auxiliary in ICE online

3.3 *FOR TO* INFINITIVE

The features that we have described so far are lexical or morphological, and thus easy to find with word form searches. But there are many morphosyntactic features which are described as being frequent in Irish English. On of these is the *for to* infinitive (Filppula, 1999: 185). This often, but not necessarily, expresses a purpose infinitive. As it involves the two-word complementizer *for to*, surface word form

searches will probably still find most occurrences, in other words achieve high recall. A search for <for to> in AntConc reveals 4 hits in ICE Ireland, one of which is a false positive, though, arising from a false start or correction in spoken data.

1. S1A-043:1:33:A We'll do the talking **for to** for you re too drunk.

It is essential to inspect the matches – or a random subset of them if numbers are large – to obtain an assessment of the precision of one's queries, which in this case is 3/4 = 75%. Importantly, also seemingly unambiguous queries, such as <for to> can retrieve false positives. The other 3 occurrences are true positives, for example:

2. S1A-014:152:B_ No it's no it was two hundred no it's two hundred and twenty from Gatwick and then I haven't paid **for to** get over to London yet but then the rest of it was all insurance.

Checking in other Inner Circle varieties (ICE-GB, ICE-NZ, and ICE-Can), we observe that this construction can also be found there. The 3 occurrences returned from ICE-GB are all true positives:

3. icegb:S1A-074:5:326:C Uh when should I pop back **for to** sign those and read them through and send them off.
4. icegb:W1B-030:1:8 I do not think I have changed the meaning at all, but perhaps it would be as well **for to** cast a final eye over it anyway.
5. icegb:w1b-030:6:120 We would be grateful if you would consider providing funding **for to** attend the course as outlined above and look forward to hearing from you with any queries you might have.

We conclude that *for to* is generally rare but permissible, it intuitively seems old-fashioned, but the claim that it is an Irish feature merits reassessment.

3.4 *IT* CLEFTING WITH THE COPULA

Another syntactic feature described by Filppula (1999) is *it* clefting with a copula (Filppula, 1999: 243 ff.). The placing of the preposition towards the end of the sentence is quite popular in sentences with *it* clefting in Irish English. For example «It's Glasgow he's going to tomorrow» is more likely to occur than «It's to Glasgow he's going tomorrow».

This feature is much more difficult to operationalise than any discussed before in this paper. It is probably impossible to formulate a word-based search string. As clefting is a syntactic long-distance dependency, the distance between the copula

and the preposition can be very long – unlike in the non-clefted version. We first tried to examine this theory by looking at the position of prepositions in such sentences in the corpus, but we were overwhelmed with hits and extremely low precision. During our inspection of the examples discussed in the literature we noticed though, first, that there are occurrences where the sentence-initial *it is* gets contracted to *'tis* or *'twas*, and second, that the distance between *it* and the main verb remains quite short.

Based on the first observation we queried for *'Tis* (returning 14 hits) and *'Twas* (returning 4 hits). Note the capitalization, which ensures sentence-initial position. Such a pattern is explicitly a serious compromise, it will have very low recall, all non-contracted forms and all non-sentence-initial forms are missed in principle. On top of these disadvantages, it turns out that, first, all 14 matches are false positives, none contains a moved preposition; second we observed that the spelling variants *'tis* and *'twa*s are unique to ICE-Ireland, and mostly from spoken categories, which means that we are probably dealing with transcribing conventions rather than linguistic features. Although the ICE corpora are intended for comparison and share standardizing guidelines, when it comes to the nitty-gritty details, standardization is often insufficient, and care needs to be exerted constantly.

Based on the second observation, there is a chance to bring down the 1543 hits of sentence-initial *It is* in ICE Ireland to a size that is manageable for manual inspection. If we additionally take into consideration that many examples are in the progressive, a simple operationalisation is to search for *It is* and then *is* a few words later. In the AntConc query language, the hash symbol (#) can be used to stand for any word, a so-called *wildcard* word. We can formulate queries such as «It 's ## is» for two intervening words, and «It 's ### going» for three intervening words. These queries returned many false positives, but still no true positives.

In addition to dealing with very high levels of false positives, it is cumbersome to have to deal with many queries. There is a query language that is much more powerful, and is available both in *AntConc* and *Corpus Navigator*: so-called *regular expressions*. Regular expressions allow for fast searches, because they can be translated into finite-state automata whose search time is linearly correlated with corpus length.

```
Regular Expression Primer:
http://marvin.cs.uidaho.edu/~heckendo/Handouts/regex.html

Most important regular expressions:

a?          optional a
a*          0 to infinite a's
a+          at least one a
(aa|bb)     aa or bb
[abc] a or b or c, e.g. s[iauo]ng
\w          any word character = [A-Za-z]
\b          word boundary
\n          newline
\t          tab
\s          whitespace = [ \t\n\r\f]
[^a]        anything but a, e.g. [^_]+_N is a noun-tagged word
a{1,5}      between 1 and 5 times a
```

Figure 4: Regular Expression Primer

Using the powerful regular expressions in *Corpus Navigator* or *AntConc* we can e.g. formulate the following query (for *AntConc* and ICE Ireland):

```
It 's ([\w']+ ){1,5}\w+ing
```

A word consists of one or several word characters (\w) or the apostrophe ('), we are looking for between 1 and 5 five words intervening between the sentence-initial *It's* and a word ending in *–ing* as it can be expected of verbs in the progressive. The query brings 129 hits, but only one true positive.

6. S1B-015:46:A_ **It's the introduction you're looking for** okay introduction

We can also use variants of the above query, restricting to typical proper names, typical NP beginnings, and extended the window of intervening words.

```
It 's [A-Z]([\w']+ ){1,8}\w+ing

It 's a ([\w']+ ){1,8}\w+ing

It 's the ([\w']+ ){1,8}\w+ing
```

But we found no further instances with these queries. On qualitative grounds, it can be argued that the non-clefted version of this sentence sounds equally or less acceptable in all English variants, *It's for the introduction you're looking …*

On quantitative grounds, a single example is insufficient evidence. Similar queries on other ICE corpora return very few true positives, we have found none in the other inner circle varieties. ICE Hong Kong furnishes the following example:

7. icehk:S2A-028:1:89:A **It s the countries that they are selling to** that is so sensitive with Washington.

Thus, we cannot draw any conclusion regarding the use of *it*-clefts in ICE Ireland. This feature is too rare to be found in the one-million word corpus, particularly as we had to use an operationalisation with low recall. We have learnt though, that alternative surface operationalisations can at least furnish examples, and that it is worth testing as many formulations as possible. To complicate matters further, we should point out that the above patterns only work for ICE Ireland and in *AntConc*. In ICE-GB and ICE Canada *It's* is not pre-tokenized in the distributed version, the space between the two words therefore needs to be omitted. Again, comparisons across the ICE corpora need to consider that not everything is standardised. In *Corpus Navigator*, the pre-tokenization of ICE Ireland *'s* has the effect that the apostrophe is deleted by the corpus tokenizer. The apostrophe thus needs to be omitted in the query.

3.5 THE *AFTER* PERFECT

The possibly best investigated feature of Irish English is the *after* perfect (e.g. Filppula, 1999: 99 ff.), which is considered to have developed due to contact with Irish Gaelic. Surface search string operationalisations to find this feature in corpora fortunately are simple:

```
after \w+ing
```

```
(am|'m|m|be|was|is|'s|s|are|'re|re) after \w+ing
```

The first query has very low precision. The second query finds frequent forms of the auxiliary *be* followed by *after* and an *–ing* form. It returns 9 hits from ICE Ireland, 8 of which are true positives, given in the following.

8. S1A-046:100:A_ A new fella is **after taking** over uhm one of the pubs at home
9. S1A-046:100:A_ And he's **after coming** back from England you know
10. S1A-055:145:E_ They thought he was **after going** into a coma with diabetes
11. S1A-067:111:D_ The wife and children are **after going** off there the other day

12. S1B-017:99:D_ I'm **after booking** one
13. S2A-012:2:A_ In the opening round I thought for a while that Walsh was going to win inside the distance but he's **after running** into a couple of hard ones here from Barrett
14. S2A-047:2:A_ Okay and here it's **after listing** the command that it's executed
15. S2B-014:1:4:C And we ve had no word or phonecall or anything you know we we ve we **re after** we re after being trying in Waterford city all with a pile of guest houses down there.

The first query (*after \w+ing*) has the disadvantage that it returns 71 hits, mostly false positives, but it also finds one occurrence which the second query does not find:

16. S1B-077:90:A_ There's nothing new **after coming** in anyway so

7 of these 9 occurrences were also found by Kirk and Kallen (2006: 95-98). As the next step, we need to test if the *after* perfect might occur in other English variants. Manually filtering hundreds of false positives after using the first query would be cumbersome. In order to get a different perspective on the data, in order to start with a high precision base, and thus in order not to need to do very much filtering, we also used the syntactic query function which *ICE online* and *Dependency Bank* (Lehmann and Schneider, 2012) provides. The full results are given in figure 5. All 4 hits from ICE Ireland are true positives, one of them (line 16) has escaped our surface searches. All hits from the other ICE corpora are false negatives, except for line 2 from ICE Canada.

With counts of 10 in ICE Ireland against 0 in the other corpora (or 1 in ICE Canada), the differences are highly significant, according to the binomial test.

We looked for *after* perfect constructions by using key words like <after + ing>, and by a syntactic query. But *after* perfects can also occur in noun phrases, e.g. *We're just after our dinner*. Using the following query, we found one such instance.

```
after (\w+){1,5} (dinner|breakfast|tea|lunch|sleep|nap|beer|walk)
```

17. S1A-008:112:A_ I'm not not that long **after my dinner**

Your Query:'h1=be r1=pobj d1= eq1=h2= r2=prep d2=after eq2=depID=headID' returned 22 results in ICE9_t6571.

|< << >> >| Show Page: 1 Show chunks Show Tags Frequency Distribution Go!

No	Reference	Solutions 1 to 22 Page 1/1 Processed for gerold at 130.60.155.214
1	ICECAN:S2B-001:1:13:A	He 's back among the best in the world **after winning** a bronze medal.
2	ICECAN:W2B-008:1:94	" They 're **after chang** ing the music " says 79 -_: year-old Joe Kennedy as he takes a break from playing me some tunes in his cluttered house deep in the woods near Inverside.
3	ICEGB:S2A-001:1:198:A	Again the England defence can do the mopping up and again it 's back with Chris Woods back in the England goal **after missing** the last three internationals.
4	ICEGB:S2A-016:1:22:A	He 's now back in his element **after coming** through the mountains and just about getting through the mountains.
5	ICEGB:S2B-010:1:31:A	Throughout the day members of the Security Council have been in one huddle **after** another **trying** to agree on whether at midnight tonight diplomacy is officially declared dead.
6	ICEHK:S1B-075:1:307:A	Oh I see this one uhm latest statement yih lihng lihng yat ji well okay this is **after gau** uh nine eleven.
7	ICEHK:S2B-006:2:131:A	Defending champion Steffi Graf will be the will be there **after whipping** Natalie Zvereva of the Commonwealth of the Independent States.
8	ICEHK:W2C-010:5:81	It had been reported that Mr Zhou had been in poor health soon **after taking** up the post.
9	ICEINDIA:S1B-003:2:208:A	Nerve tonic that is **after reading** lesson.
10	ICEINDIA:S1B-038:1:130:B	And has the mosque been built on that site **after detroying** the temple ?
11	ICEINDIA:S2A-019:1:17:A	So it's a great day for me to be here **after winning** the last national being played in Chandigarh to come here and commentate today on this mens ` and the womens ` singles.
12	ICEINDIA:W2C-017:1:41	It was only **after rolling** for about 185 feet that both the front and rear wheels were touching the ground.
13	ICEIRE:S1A-046:1:100:A	A new fella is **after taking** over uhm one of the pubs at home.
14	ICEIRE:S1A-055:1:145:E	They thought he was **after going** into a coma with diabetes.
15	ICEIRE:S1A-067:1:111:D	The wife and children are **after going** off there the other day.
16	ICEIRE:W1B-007:2:30	This was **after** him **sending** me a mushy letter to work on Tuesday.
17	ICEJAM:S2B-014:0:16:A	This is **after admitting** to importing and possessing marijuana with a street value of more than twenty -_: seven thousand U S dollars.
18	ICENZ:S2B-019:1:17:R	and the former kiwi half back clayton friend is back in the country **after playing** for carlisle in england.
19	ICENZ:W1B-006:4:131	I do sympathise with the long distance relationship and you 're right **after seeing** them again it's always harder !.
20	ICENZ:W2C-013:6:120	Clark went in the 11th over for 40 while Bracewell was still there at stumps **after compiling** 59.
21	ICESING:S1B-075:1:101:B	That must be **after paying** back.
22	ICESING:W1A-001:1:26	And " I was equally confounded at the Sight of so many Pignies ;_: for such I took them to be , **after having** so long accustomed my eyes to the monstrous objects I had left ..

Figure 5 : *after* in the ICE corpora

We conclude that the *after* perfect, one of the best known features of Irish English, can be observed very well in relatively small corpora such as the ICE series of one million carefully sampled running words. We also conclude that, despite lower recall, syntactic queries are a useful approach to data exploration.

3.6 SINGULAR EXISTENTIAL WITH PLURAL NP

The *singular existential* (see Hickey, 2005: 121 and Walshe, 2009) is another well-known Irish feature. A simple surface operationalisation can be made by the following query:

```
there ?'?s \w+s
```

This query returns 53 hits, about two thirds are false positives. The true positives contain examples like the following.

18. S1A-027:177:C_ I'm sure **there's loads** of cafes saying that they're the they're
19. S1A-028:52:C_ But **there's lots** of uhm like I mean say if you were going to analyse a a rock face I mean there's probably only one way you can actually analyse it
20. S1A-064:1:16:E You know when like when you ve got all these tractors and all I suppose on the road at home and I was there going oh yeah **there's tractors** on all the roads.
21. W1B-003:3:56 Talking of which, **there's soldiers** all over the show here -_: everywhere ya fuckin' look!

The true positives are dominated by *loads* and *lots*, which can be seen as lexicalised predeterminers, but also abstract nouns and even some animate nouns, as in the last example, can be found. The last example is from the written part, but intends to represent spoken language.

The same query reports 44 hits in ICE-Can, 58 in ICE-NZ, 31 in ICE-GB, for example:

22. icecan:S1A-006:114:1:B You know **there's things** to do
23. icegb:S2A-025:91_1:A_1 And **there's examples** of the damage to those which required the building to be closed with the possibility of demolition involved in that case

Also the levels of false positives do not seem higher at a first glance. We use a syntactic query in ICE online to obtain higher precision finds. The query can be seen in figure 6. The verb *be* needs to have a subject *there* and an object which is in the plural, which we constrain by requiring the object dependent to have the part-of-speech tag *NNS*. We restrict the verb to its singular third person form by requiring it to have the tag *VBZ*.

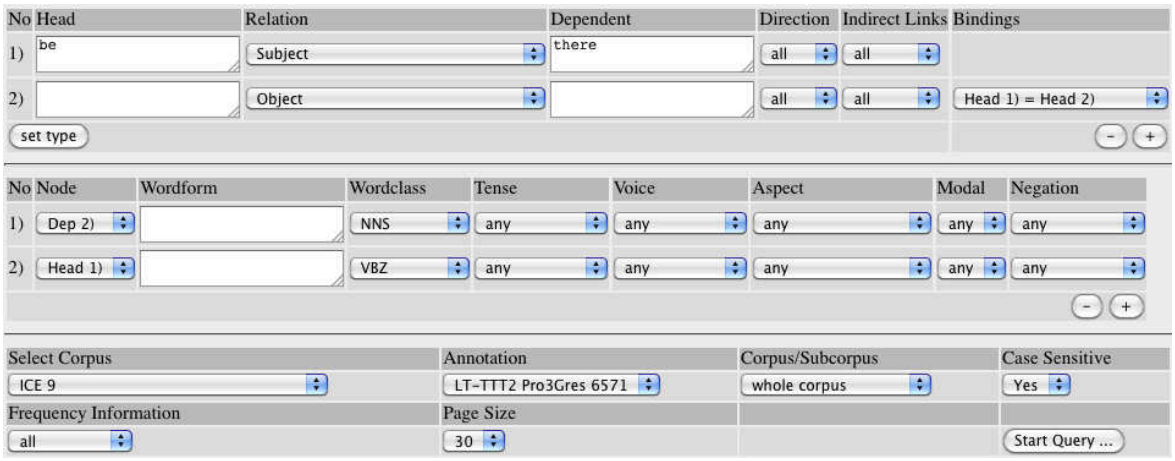


Figure 6. Query for copula-complement concord violation in ICE online.

The distribution across the 9 ICE corpora in ICE online can be seen in figure 7. The occurrences are cross-tabulated between the corpora and spoken/written. We can see that singular existential is mainly a feature of spoken language, and that there is some regional variation. ICE Ireland even displays this feature sparingly in comparison with ICE Canada, ICE New Zealand and even ICE-GB.

Crosstabulation of Query 'h1=be r1=subj d1=there eq1=h2= r2=obj d2= eq2=headID=headID' according to categ
Absolute Frequency.

Tabulate: corpus Go! Crosstabulate: X iceCat1 / Y corpus show Count of Instances

corpus/iceCat1	spoken	written	total
icecan	121	7	128
icegb	65	5	70
icehk	21	10	31
iceind	18	4	22
iceire	13	1	14
icejam	24	4	28
icenz	122	6	128
icephi	10	3	13
icesing	32	5	37
total	426	45	471

Figure 7. Distribution of singular existentials across 9 ICE corpora

We conclude that singular existentials are probably globally on the rise in spoken language. Inner Circle varieties appear to be in the lead in this process, and their use does not seem to be a particularly Irish feature.

3.7 HABITUAL ASPECT FORMS

Habitual aspect forms with *do* or with *be(e)s* have been described as an Irish feature (Filppula, 1999: 130 ff.; Ronan, 2011), which may have arisen in the contact situation with Irish Gaelic. Expected examples are «They do be there every Friday» or «They be(e)s there every Friday». These forms have been described as getting rarer, but still used in rural Ireland. We used surface queries for *bees* (which returned interesting insights into the species *apis mellifera*, but no habituals) and *bes,* which returned a single hit (which, given the context, may be a performative use):

24. S1A-032:37:A_ He just stands there and **bes** Frankenstein

The query *does be* also returns exactly one hit:

25. S1A-087:139:B_ That that buck that **does be** on the television on the video

Both these matches are also found by Ronan (2011). Queries on the other ICE corpora returned no hits, as expected. For queries including all verbs we used the tagged version available in Corpus Navigator and ICE online for the surface queries

```
doe?s?_VB[ZP] \w+_VB.?
```

The hits are overwhelmingly cases of emphatic rather than habitual *do*, or at best ambiguous. In the other ICE corpora, we find similar levels of hits, typically higher in Inner Circle varieties than in Outer Circle ones (counts are highest in ICE New Zealand, followed by ICE-GB, and then ICE Ireland). In conclusion, the feature is still used, but it is too infrequent for quantitative claims in the case of *bes*, and disambiguation between emphatic and ambiguous is difficult in many cases, and the emphatic form seems to be less used in Outer Circle varieties.

3.8 INVERSION IN INDIRECT QUESTIONS

In Irish English, indirect questions can retain the inversion of direct questions, resulting in sentences such as «I asked him was he going home». According to Bliss (1984: 148), indirect questions can take two forms. The first form is that of indirect simple questions, which in Standard English require an introductory *if* or *whether* and which can often be answered by a simple *yes* or *no*. In Irish English, however, the *if* or *whether* is omitted and the inversion, also known as «embedded inversion» (Filppula, 1999: 167), is retained. Typical verbs which introduce this type of indirect question include *ask*, *wonder*, *know* and *see*. A second form that these

indirect questions can take is as indirect complex questions. These preserve the interrogative word (*who*, *what*, *when*, *where*, *which* and *how*) and unlike in Standard English they again retain the word order of a direct question. Filppula observed that embedded inversion was more likely to occur in simple *Yes*/*No* questions than in complex WH-questions.

We have used the following surface query, focusing on the question word *ask*. As we only allow one intervening word, the query has a preference for finding the first type.

```
\bask\w* \w+ (was|is|are|will|would)
```

Among the 21 hits for this query we get 11 true positives, for example:

26. S1A-035:62:B_ And then I **asked her would** she let him and she said no
27. S1A-088:173:D_ And Medbh **asked me would** I come over and would I bring Jane with me right
28. S1A-088:173:D_ So I Lara rang today and I **asked Lara would** she do it
29. S2B-021:6:D_ So she **asked her would** she have any food that she could give her some and feed the baby
30. W2C-012:5:p_D Quoting the handwritten note - "If pressed on this question keep repeating the above" - Ms Harney said she had repeatedly **asked who was** the Minister who had written it.

The same query on ICE-GB obtains 9 hits, and 3 true positives are among them, for example:

31. icegb:S1B-015:1:89:A In a sense you 're **asking what is** the next stage.

The number of hits (21) is highest in ICE Ireland of all the 9 ICE corpora in our investigation. The comparison across the ICE corpora based on the number of hits does not reach statistical significance yet. We added further question words like *wonder*, which only returns few hits in all ICE variants, and *know*, which returns very many hits. As most true positives of *know* are negated, we restricted the search to *not know* and *n't know*:

```
n['?o]t know\w* \w+ (was|is|s|are|will|ll|would)\b
```

The results have high precision, 12 of the 18 hits from the spoken part of ICE Ireland (and all 4 in ICE-GB) are true positives. There are strong, and statistically significant differences across the ICE corpora. Outer Circle varieties, particularly ICE Hong Kong and ICE Philippines show high frequency, while low frequency

can be observed in all Inner Circle varieties except for ICE Ireland, as we can see in figure 8.

Crosstabulation of Query 'n['?o]t know\w* \w+ (was|is|s|are|will|ll|would)\b' accordi Frequency.

Tabulate: corpus Go! Crosstabulate: X iceCat1 / Y corpus

corpus/iceCat1	spoken	written	total
icecan	4	0	4
icegb	4	0	4
icehk	35	6	41
iceind	15	2	17
iceire	18	6	24
icejam	12	3	15
icenz	9	1	10
icephi	27	1	28
icesing	18	2	20
total	142	21	163

Figure 8. Frequency of matches using negated *know* and inverted indirect questions.

We can also see that the feature, as expected, is largely a spoken language feature. Interestingly, in some Outer Circle varieties, and in ICE Ireland, it is also sometimes used in the written part. 5 of the 6 hits from the written part of ICE Ireland are true positives, as can be seen in figure 9. We conclude that inversion in indirect questions is indeed an Irish feature, and the patterning of ICE Ireland between an Inner and Outer Circle variety (Kachru, 1992) is particularly interesting, and merits further investigation.

We have also used patterns including more intervening words between the matrix and subordinate verbs. A query for up to 5 intervening words is:

```
\bask\w*( \w+){1,5} (was|is|are|will|would)
```

Although precision for this query is very low, it returns some new true positives, for example:

32. S1B-057:35:C_ I just want to **ask the Taoiseach is** he going to start

Your Query:'n['?o]t know\w* \w+ (was|is|s|are|will|ll|would)\b' returned 163 results in Reduced to 6 with restriction ICE9_meta.iceCat1='written' and ICE9_meta.corpus='iceire'

|< << >> >| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 30 Page 1/1 Processed for gerold at 178.192.45.239
1	ICE9:iceire:W1B-004:2:34	I just don't know what s the best thing to do.
2	ICE9:iceire:W2B-004:2:20	I didn't know what was going on, but I crept back upstairs, because if my mother had caught me down there she d have hit me!.
3	ICE9:iceire:W2C-002:5:12	The house was full of petrol fumes and I really did not know what was happening.
4	ICE9:iceire:W2C-003:1:24	My girlfriend and son had to fly home not knowing what was happening to me..
5	ICE9:iceire:W2F-012:1:58	I don't know what s been on my mind really since the accident, Aunt Cissie.
6	ICE9:iceire:W2F-020:1:76	The way things are today, I don't know what s happening.

Figure 9. The 6 hits from the written part of ICE Ireland

Kirk and Kallen (2006: 106 ff.) search parts of ICE Ireland for inversions in indirect questions. Except for the question word *don't know* their frequencies remain low and do not reach statistical significance.

3.9 REFLEXIVE PRONOUNS IN PLACE OF NON-REFLEXIVE PRONOUNS

Irish English can use reflexive pronouns instead of non-reflexive pronouns. Filppula (1999: 77-8) refers to them unbound reflexives. A simple query for surface forms of reflexives in sentence-final position reveals 5 occurrences among 30 hits.

```
(your|her|him)self\b ?\.
```

33. W1B-006:39:p_A So how's it going **yourself .**
34. W1B-006:34:p_A What's the crack with **yourself .**
35. W1B-022:7:p_N I'm sure this offer will be attractive to local motorists like **yourself** .
36. W2A-007:96:p_A The other females in the book are all doubles for **herself .**
37. W2B-020:131:p_A Particularly so for someone of maturing age such as **himself .**

We found no unbound reflexives among the hits in ICE-GB. The probability (5 versus 0) according to the two-tailed binomial test is 6%, which would only be sufficient to reach lowest significance levels. Kirk and Kallen (2006: 103 ff.) have a more detailed investigation of unbound reflexives, and considerably more hits, probably based on manually filtering all occurrences of reflexive pronouns. In their investigation, only subject conjunction (e.g. *mum and myself*) emerges as

statistically significant, while the category that we have searched for here (*object* in Kirk and Kallen's terminology) is also found in ICE-GB and therefore not restricted to the Irish English data.

3.10 THE MEDIAL OBJECT PERFECT

The medial object perfect is described in detail in Filppula (1999: 107 ff.). It involves a non-canonical word order, placing the object between the auxiliary and the participle. A simple lexis-based surface operationalization to find these in a corpus is

```
have \w+ done
```

It has very low precision but allows us to find a first true positive:

38. S1A-002:48:A_ They **have obituaries done** for William and Harry

Using the tagged version of ICE-Ireland in Corpus Navigator or ICE online we can also formulate more restricted queries, for example requiring the intervening word to be a pronoun, which increases precision.

```
ha(s|ve)_VB. \w+_PRP \w+_VB[ND]
```

This query allowed us to find e.g.

39. S1A-067:73:D_ I **have it wrote** down.

40. S1A-087:295:A_ They probably **have him chained** so he won't get out.

A similar query restricting the intervening word to be a singular noun is:

```
ha(s|ve)_VB. \w+_NNS? \w+_VB[ND]
```

This query returned the following hit, among many false positives:

41. S1A-006:1:144:C But he cos I cos when he said last night then I was saying I was thinking och no maybe he **has something organised** cos he was saying aw you know

Kirk and Kallen (2006: 98 ff.) have a detailed investigation of the medial object perfect in ICE Ireland, including the discussion of many occurrences and semantic interpretations, therefore this section may be considered a technical footnote to their contribution.

3.11 *BE* AUXILIARY WITH PAST PARTICIPLE

Filppula (1999: 114ff.) describes the use of the *be* perfect. In Early Modern English, *be* was still preferred to *have* for perfect-formation and in some dialects it is still used in rare cases. The prototypical participle is *gone*, although other dynamic verbs can be used. Semantically, this perfect stresses the resultative aspect, and it is difficult to distinguish it from adjectival uses. We concentrate on *gone*, using the simple surface search *is gone* which returns the following instances.

42. S2B-033:4:B_ In his famous dialogue in Hybernian Stile Swift noted the use of many Gaelic phrases carried over into English I wonder what **is gone** with them meaning I wonder what has happened to them

43. S2B-049:2:A_ That time **is gone**

We have also used a syntactic query to obtain more forms. We restrict our search to verbs which have a realised surface subject. The results in figure 10 reveal, however, that using the auxiliary *be* with the participle *gone* is an option in most ICE corpora. The 4 hits in ICE Ireland are

44. S1A-073:1:23:A But they but they if they bring the divorce in before they amend that it ll fail again and then **it ll be gone** for a whole generation not just ten or five years the next time.

45. S1A-078:1:177: **That ripple icecream ll be gone** soft.

46. S1B-068:1:25:B All **that** would have **to be gone** into.

47. W2F-018:1:50 **Anything** that has to get gone into **can be gone** into in the morning Rose said.

Only one of them (45) is neither adjectival nor passive. On quantitative grounds, the use of the *be* perfect with *gone* is thus too sparse to reach significance levels in ICE Ireland.

Distribution of Query 'h1=go r1=subj d1= eq1=h2= r2=aux1 d2=be eq2=headID=headI

Tabulate: corpus Go! Crosstabulate: X -- / Y corpus

corpus	n	words	relative frequency per 10000 words
ICECAN	1	1051581	0
ICEGB	4	1070703	0
ICEHK	2	1215724	0
ICEIND	2	1126149	0
ICEIRE	4	1057065	0
ICENZ	6	1207008	0
ICEPHI	3	1139253	0
ICESING	3	1030868	0

Figure 10. Frequencies of *be* + *gone* in a syntactic query on 9 ICE corpora.

4. CONCLUSIONS

We have investigated 11 salient features of Irish English using relatively simple retrieval strategies on the ICE Ireland corpus. We have compared the results to other ICE corpora. The features commonly ascribed to Irish English can be divided into three categories. First, the category where ICE Ireland offers enough evidence, and where a quantitative comparison to other ICE corpora shows statistically significant differences:

a. Low frequency of *shall* (section 3.2; see also McCafferty, 2011)
c. *After* perfect (section 3.5; see also Filppula, 1999: 99 ff.)
e. Indirect questions with inversion (section 3.8; see also Filppula, 1999: 167ff.)

Second, a category in which sparse data does not allow us to draw any conclusion, if we base our investigation on the ICE corpora only, and using only our simple, coarse retrieval patterns:

d. Clefting with copular verbs (section 3.4; see also Filppula, 1999: 243 ff.)
b. Habitual aspect with *do* (section 3.7; see also Filppula, 1999:130 ff.)
g. *Be* as auxiliary with past participle (section 3.11; see also Filppula, 1999: 114 ff.)

Third, the category of Irish features where we think that their status as specific features of Irish English needs re-assessment.

f. *For to* infinitive (section 3.3; see also Filppula, 1999: 185)

h. Singular existential with plural NP (section 3.6; see Hickey, 2005:121 and Walshe 2009)

In our discussions of the non-reflexive pronouns (section 3.9) and the medial object prefect (section 3.10) we pointed out that Kirk and Kallen (2006) offer more detailed investigations based on more manual filtering than in our study, which is intended to be a show case.

We have shown how *AntConc*, *Corpus Navigator* and *ICE online* can be used with relatively simple search queries by any corpus linguist. We have also shown that syntactic queries, which are offered in Dependency Bank and ICE online, can be used with ease for exploratory research.

BIBLIOGRAPHICAL REFERENCES

ANTHONY, Laurence (2004), «AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit». Proceedings of *IWLeL 2004: An Interactive Workshop on Language e-Learning,* pp. 7-13.

FILPPULA, Markku (1999), *The grammar of Irish English. Language in Hibernian style*, London, Routledge.

GREENBAUM, Sidney (ed.) (1996), Comparing English Worldwide: The International Corpus of English. Oxford: Oxford University Press.

HICKEY, Raymond (2005), *Dublin English: Evolution and Change. Varieties of English around the world.* Amsterdam/Philadelphia: Benjamins.

HICKEY, Raymond (2007), *Irish English. History and Present-Day Forms.* Cambridge, Cambridge University Press.

KACHRU, Braj. B. (1992), «World Englishes: approaches, issues and resources». *Language Teaching* 25.1, pp 1-14.

KALLEN, Jeffrey L. & John M. KIRK (2008), *ICE-Ireland. A user's guide*, Belfast, Cló Ollscoil na Banríona.

KIRK, John & Jeffrey L. KALLEN (2006), «Assessing Celticity in a Corpus of Irish Standard English». In Hildegard L. C. Tristram (ed.), *The Celtic Languages in Contact.* Potsdam: Potsdam University Press, pp. 270–288.

KIRK, John & Jeffrey L. KALLEN (2010), «How Scottish is Irish Standard English?», in R. McColl Millar : *Northern Lights, Northern Words: Selected Papers from the FRLSU Conference, Kirkwall 2009*, Aberdeen, Forum for Research on the Languages of Scotland and Ireland, pp. 178-213.

LEHMANN, Hans Martin & Gerold SCHNEIDER (2012), «BNC Dependency Bank 1.0». In Signe Oksefjell Ebeling, Jarle Ebeling, & Hilde Hasselgård.(eds.), *Studies in Variation, Contacts and Change in English, Volume 12: Aspects of corpus linguistics: compilation, annotation, analysis*. Helsinki: Varieng.
[http://www.helsinki.fi/varieng/journal/volumes/12/]

MCCAFFERTY, Kevin (2011), «English grammar, Celtic revenge? First-person future shall/will in Irish English». In Raymond Hickey (ed.), *Studying the languages of Ireland. A Festschrift for Hildegard L.C. Tristram*. Uppsala, Uppsala University Press, pp. 223-242.

RONAN, Patricia (2011), «Irish English Habitual do be: More on Origins and Use», *Groninger Arbeiten zur Germanistischen Linguistik* 53.2 (December 2011), 105-118.

SCHNEIDER, Gerold (2008), *Hybrid long-distance functional dependency parsing.* Doctoral Thesis. University of Zurich, Faculty of Arts.

SCHNEIDER, Gerold (2013), «Using automatically parsed corpora to discover lexico-grammatical features of English varieties». In Fryni Kakoyianni Doa (Ed.). *Penser le Lexique-Grammaire, perspectives actuelles*. (Papers from the 3rd International Conference on Lexis and Grammar, Nikosia, Cyprus, 5-8 October, 2011). Paris, Éditions Honoré Champion.

TRUDGILL, Peter & Jean HANNAH (2002), *International English: A Guide to the Varieties of Standard English* (4th ed). London, Arnold.

WALSHE, Shane (2009), *Irish English as Represented in Film*, Frankfurt, Peter Lang.