A Corpus of Irish English - Past, Present, Future

Gosia Barker and Anne M. O'Keeffe

University of Limerick

Acknowledgements: The authors would like to express their gratitude to Prof. Michael McCarthy for his encouragement and inspiration.

Correspondence: Gosia Barker, Department of Languages and Cultural Studies, University of Limerick, Limerick, Ireland. E-mail: gosia.barker@ul.ie

Anne M. O'Keeffe, Arts Departments, Mary Immaculate College, University of Limerick, South Circular Road, Limerick, Ireland. E-mail: anne.okeeffe@mic.ul.ie

A Corpus of Irish English - Past, Present, Future

Abstract

To date, no corpus of Irish English exists. Most previous research has focused on the syntactic and phonological peculiarities of Irish English showing how it differs from standard British English, and in the same way, many studies show how certain patterns have parallel forms in the Irish language. The dearth of synchronic study of Irish English has been noted in the past. This paper will examine the pattern of research into the English language in Ireland and it will suggest that in order to develop adequate paradigms for description and research, corpus-based methods are essential. Corpus linguistics offers the study of Irish English the opportunity to become part of a global body of research and in so doing will enhance the description of the English language.

Introduction

A corpus is a collection of language data brought together for linguistic analysis (Crystal 1995: 450). In the 1700s, when Dr Johnson was compiling the first comprehensive dictionary of the English language, he manually collated a corpus of language data based on samples of usage from the period 1560 to 1660. Almost three centuries later, corpora are vast, methodical collections of both spoken and written texts, stored on computer and designed so as to provide systematic representation of different types of language in use, such as: sermons, everyday conversations, lectures, advertisements, works of fiction and academic prose. This collection of computerised language can be used to form and to test hypotheses about language. It can facilitate the investigation of how people use language in particular situations, examine how usage differs between spoken and written language, compare language varieties, look at variation within language and provide empirical evidence of language usage and language change.

According to many, corpus linguistics heralds a new era of the scientific study of language. Leech (1991: 9) states that the value of the corpus as a source of systematically retrievable data and as a testbed for linguistic hypotheses has become widely recognised and exploited. Large quantities of 'raw' text can be processed directly allowing a researcher to examine objective evidence of the language. Machine-readable text collections have grown from one million to almost a thousand million words in thirty years. Sinclair (1991: 1) states that thirty years ago, when this type of research started, it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular.

Background to Corpus Linguistics

Post-Bloomfieldian structural linguists in the USA may have sown the seeds of modern corpus linguistics.[1] Some linguists, such as Harris and Hill in the 1950s, under the influence of positivism and behaviourism, regarded 'corpus' as the primary explicandum of linguistics (Leech 1991: 8). By 1959, a new school of corpus linguistics had come into existence when Randolph Quirk announced his plan for a corpus of both spoken and written British English. This one-million-word corpus, now known as the Survey of English Usage (SEU), was compiled under the direction of Quirk and Greenbaum. Subsequently, in 1961, Brown University set up a one-million-word corpus of written American English. The London-Lund Corpus (LLC) initiated in 1975 by Jan Svartvik is still regarded as an unmatched resource for the study of spoken English (Leech 1991: 9).

In recent years, the scope of corpora has broadened through the addition of new corpora for British, American, Indian, Australian, Singaporean and Malaysian English and so on (Taylor, Leech and Fligelstone 1991: 321). Another development has been the creation of domain-specific corpora such as: CHILDES, a corpus of children's spoken and written language, COLT, a corpus of teenage language, ICLE, the International Corpus of Learner English and the Melbourne-Surrey Corpus of Australian newspapers etc. The value of language corpora is now well established and has already had an enormous impact on the empirical study of the English language.

For almost forty years, a number of linguists have been collecting and computerising spoken and written samples of the English language. The advances in computer technology have led to the stage where computerised language corpora of almost a thousand million words can be stored. It is estimated that a word count of one million million could be achieved by 2021 (Leech 1991: 10). Currently, there are signs of growing recognition that the comprehensive study of language must be corpus-based (Sinclair 1991: 6). Svartvik (1991: 8) states that

linguistic competence and performance are too complex to be adequately described by introspection and elicitation alone or as Sinclair (1991: 6) quips 'one does not study all of botany by making artificial flowers'. By using a corpus of natural language in action, more objective statements can be made than introspective observation permits (Svartvik 1991: 9). Language corpora not only facilitate the empirical analysis of language synchronically but also provide a record of language for diachronic research. The benefits of being able to conduct empirical research on language are obvious to the cause of furthering linguistics as the scientific study of language.

The Nature of Research into Irish English[2]

Most previous research into the English language in Ireland has focused on the lexical, syntactic and phonological features of Irish English showing systematically how it differs from standard British English. In the same way, many studies seek to show how sometimes infrequently used patterns have parallel forms in the Irish language. While investigating the origin and development of this speech variety is both interesting and worthwhile, it can also be seen as highly introspective or even anachronistic that this remains the dominant paradigm for research.

Kallen (1985: 1) identifies three major 'points of view' distinct in Irish English research: historical-descriptive, bilingual and theoretical. This offers an insight into how we have sought to explain why we speak differently to other English speakers either: as a consequence of our history, as a result of having spoken Irish or as a phenomenon within the theoretical framework of language change (cf. Weinreich 1953). The dearth of synchronic study of Southern Irish English was noted in Kallen (1985: 11) and he suggested that further study 'may be of value not only to a segment of Irish life, but for the development of general linguistics' (Kallen 1985: 12). Kallen warned, however, that this value could only be realised through the development of adequate paradigms for description and research.

Ostensibly, hitherto analysis of English in Ireland as a whole has been based on intuition rather than empirical observation. Although important localised synchronic research, based on observed rather than intuitive data, has yielded very interesting results, such as Henry (1957), Barry (1981), Lunny (1981), Finlay and McTear (1986), Pitts (1986) and Moylan (1996), conclusions cannot be generalised. This leaves the empirical results at a local level while major research questions go unanswered.

Research to date has also neglected to discriminate between spoken and written language usage. This lack of distinction is common to the study of the English language in general where the models of grammar used to represent and to examine spoken English are rooted in descriptions of the grammar of written English (Carter and McCarthy 1995: 141, Summers 1997: 100, Glisan and Drescher 1993: 30, Willis 1997: 105).

Kallen (1985: 11) concluded that there is virtually no area of research on Irish English which can be said to be completed. Over a decade later, areas of research have by no means been exhausted. Corpus linguistics offers the study of Irish English the opportunity to become part of a global body of research and in so doing will enhance the description of the English language.

Current Movements in Corpus-based research into Varieties of English

A Corpus of Spoken Northern Ireland English is commercially available and comprises 400,000 words of spoken material from 42 grid-referenced localities in Northern Ireland over three age groups. Another important development is that of the International Corpus of English (ICE), which will include an Irish English component. Corpora of one million words were aimed at in each variety and it was limited to adults of eighteen years or older who were educated speakers, that is, those who have completed secondary school at least.

ICE seeks to compare spoken and written English of the same period within and between national varieties of English including countries where English is not the first language for the majority of the population, for example, India. Greenbaum (1991: 84) makes the point

that in both first and second language countries, there is a continuum of competence in English. Internationally, speakers along this cline of competence are developing their own linguistic norms and Greenbaum points out that sociolinguistically this is an interesting time for the investigation of new institutionalised national varieties.[3]

In the context of research into Irish English, the Irish English component within ICE will be of enormous interest to researchers. It must be remembered, however, that the project's aim is not to provide every participant country with a national corpus, but rather, to set up a corpus as a basis for comparing spoken and written varieties from educated native and non-native speakers.

Corpus-Based Research – Practical benefits

Corpora have a wide application in the field of linguistics, such as: discourse structure analysis, machine translation, lexicography, morphological analysis, phrase structure analysis, word sense disambiguation and language modelling etc. An Irish English corpus would be extremely beneficial as it would further empirical analysis as well as establishing Irish English as part of the description of English as a world language.

Accurate information about how Irish people use English towards the end of the twentieth century would be available for diverse research purposes. In the future, this could form part of a diachronic study of English used in Ireland over a period of time and it would help to account for language change. Empirical work such as Filppula (1986) could be used in parallel with a corpus to investigate language change. Retrospectively, previous research in the area of Irish English could be verified or falsified. For instance, the extensive description of the grammar of Irish English found in the work of Harris (1993) could be tested against a corpus. Such an investigation would give quantitative and qualitative results, which would further validate such a description.

Biber, Conrad and Reppen (1994: 169) identify the two main strengths of corpus-based research as: providing large databases of naturally-occurring discourse, enabling empirical analyses of the actual patterns of use in a language and, when coupled with semi-automatic computational tools, the corpus-based approach, enables a scope not otherwise feasible. They point out that corpus-based analyses frequently show that earlier conclusions, based on intuition, are inadequate or incorrect. Even the notion of core grammar is brought into question by corpus evidence. Biber et al (1994: 169) point out that the notion of core grammar needs qualification since investigations of the patterns of structure and use in large corpora reveal important, systematic differences across registers at all linguistic levels.[4]

A corpus forms an immediate teaching resource. Language teachers could use the corpus of natural language data to generate concordances or to download segments of conversation for use in the language classroom. In this manner, it would prove an invaluable resource for increasing language awareness in native speakers and as a teaching resource for the estimated 107,000 foreign students studying English as a Foreign Language in Ireland each year.[5]

A corpus of Irish English would have a dramatic effect on the level and scope of research in the area of Irish English. It would also facilitate diverse research to be conducted both nationally and internationally. Fields, such as, forensic linguistics, natural language processing, computational linguistics etc. would all benefit from a commercially available corpus of Irish English. In the short term, the greater benefit would be to fill the lacuna in synchronic study of the English language spoken in Ireland as a whole.

It would offer a new paradigm, distinct from historical description or the theoretical framework of language change. Furthermore, it would provide the answers to global research questions, such as, what grammatical categories constitute a variety as distinct from other varieties, not just standard British English. Research into hitherto unexplored or under-

researched areas such as modality, tense/aspect systems, voice and deixis could be undertaken.

Leech (1991: 20) announced that we are now in a position where corpus-based research has truly taken off, not only as the paradigm for linguistic investigation but as a key contribution to the development of natural language processing software. He identifies the following areas as priorities for immediate and for longer-term development: (1) basic corpus development; (2) corpus tools development and (3) development of corpus annotations. Leech predicts that these areas are likely to attract not only academic attention but also the necessary governmental and industrial funding.

Conclusion

Clearly, corpus linguistics offers a worthwhile paradigm for quantifying and qualifying Irish English as a discrete variety amongst world Englishes. Diachronic and synchronic approaches to linguistic analysis merged with corpus linguistics will enhance the description of Irish English. If resources could be pooled in an effort to build a state-of-the-art corpus of Irish English, those who have fostered research over the years, internationally as well as within Ireland, would benefit greatly in many diverse ways.

In corpus linguistics, language and computer science are linked in an unprecedented and irreversible way, but while hardware technology is advancing by leaps and bounds, software tools are struggling to keep up with the pace, and scientific methodology is noticeably lagging behind. As a small country, with a rapidly growing technology-oriented economy, in a climate of increasing investment in third-level research by industrial partners, we must find our place amidst the global research surge currently underway in corpus linguistics.

Notes

[1] See for example, Harris (1951), and Hockett (1948).

[2]The term Irish English will be used to refer to the English language used by people from the island of Ireland. It is favoured over the term Hiberno-English as it is in line with contemporary international terminology in corpus-based studies of language varieties, eg. South African English, Kenyan English, American English, Australian English, British English and so on. This term takes into account that more accurate subdivisions can be made into Northern Irish English and Southern Irish English. These terms, in turn, as mentioned in Trudgill and Hannah (1982: 102), are not coterminous with the political division of Ireland.

[3]Another Irish English corpus-based project is in progress as part of the Cambridge-Nottingham Corpus of Discourse in English CANCODE (personal communication with Prof. Michael McCarthy, University of Nottingham).

[4] The notion of core grammar discussed in Biber et al (1994) refers to the pedagogical consensus about the areas of English grammar presented and prioritised in EFL/ESL grammars. Using corpus evidence, Biber et al illustrate how such an assumption does not reflect reality in the case of postnominal modifiers. They point out that the notion of a core grammar for pedagogical purposes cannot be reconciled with the existence of many different situationally-defined text varieties, which they refer to as registers.

[5] This figure is calculated by Bord Fáilte in *Survey of Overseas Travellers* (forthcoming); this estimate is based on language students, over the age of sixteen, visiting Ireland in 1997 (personal communication with Bord Fáilte).

Bibliography

Barry M.V. (1981). Aspects of English dialects in Ireland, vol.1, Queen's University Belfast, Belfast.

Biber, D., S. Conrad and R. Reppen (1994). Corpus-based approaches to issues in Applied Linguistics Applied Linguistics 15(2), 169-189.

Carter, R. and M. McCarthy (1995). Grammar and the spoken language. Applied Linguistics 16(2), 141-158.

Crystal, D. (1995). The Cambridge encyclopedia of the English language. Cambridge University Press, Cambridge.

Filppula, M. (1986) Some aspects of Hiberno-English in a functional sentence perspective. University of Joensuu Publications in the Humanities, 7., Joensuu.

Finlay, C. and M. F. McTear (1986). Syntactic variation in the speech of Belfast schoolchildren. In J Harris, D. Little, D. Singleton (eds.) Perspectives on the English language in Ireland, Centre for Language and Communication Studies, Trinity College Dublin, Dublin, pp.175-186.

Glisan, E. and V. Drescher (1993). Textbook grammar: does it reflect native speaker speech? The Modern Language Journal 77(1), 23-33.

Greenbaum, S. (1991). The development of the international corpus of English. In K. Aijmer, B. Altenberg (eds.) English corpus linguistics. Longman, London, pp. 83-91.

Harris, J. (1993) The grammar of Irish English. In J. Milroy, L. Milroy (eds.) Real English: the grammar of English dialects in the British Isles, Longman, New York, pp. 139-186.

Henry, P.L. (1957). An Anglo-Irish dialect of North Roscommon, University College Dublin, Dublin.

Kallen, J. (1985). The study of Hiberno-English. In D. Ó Baoill (ed.) Papers on Irish English, Irish Association for Applied Linguistics, Dublin, pp.1-15.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer, B. Altenberg (eds.) English corpus linguistics, Longman, London, pp. 8-29.

Lunny, A. (1981). A linguistic interaction: English and Irish in Ballyvourney, West Cork. In M.V. Barry (ed.) Aspects of English dialects in Ireland, vol.1, Queen's University Belfast, Belfast, pp. 118-141.

Moylan, S. (1996). The language of Kilkenny: lexicon, semantics, structures. Geography Publications, Dublin.

Pitts, A. H. (1986). Differing prestige value for the (ky) variable in Lurgan. In J Harris, D. Little, D. Singleton (eds.) Perspectives on the English language in Ireland, Centre for Language and Communication Studies, Trinity College Dublin, Dublin, pp.209-224.

Summers, D. (1997). Credibility gap? The language we teach and the language we use. In P. Grundy (ed.) IATEFL 1997 Brighton conference selections, International Association of Teachers of English as a Foreign Language, Kent, pp. 99-105.

Sinclair, J. (1991). Corpus, concordance, collocation, Oxford University Press, Oxford.

Svartvik, J. (1992). Corpus linguistics comes of age. In J. Svartvik (ed.) Directions in corpus linguistics: proceedings of Nobel symposium 82, Mouton de Gruyter, Berlin, pp.7-13.

Taylor, L., G. Leech and S. Fligelstone (1991). Survey of English machine-readable corpora. In S. Johansson, A. B. Stenström, (eds.) English computer corpora: selected papers and research guide, Mouton de Gruyter, Berlin, pp. 319-354.

Trudgill, P. and J. Hannah (1982). International English: a guide to varieties of standard English, Edward Arnold Publishers, London.

Weinreich, U. (1953). Languages in contact, Mouton, The Hague.

Willis, J. (1997). Exploiting task recordings for the grammar of spoken English. In P. Grundy (ed.) IATEFL 1997 Brighton conference selections, International Association of Teachers of English as a Foreign Language, Kent, pp. 105-106.