Issues in Designing a Corpus of Spoken Irish

Elaine Uí Dhonnchadha, Alessio Frenda, Brian Vaughan
Centre for Language and Communication Studies,
Trinity College Dublin, Ireland.
E-mail: {uidhonne; frendaa; bvaughan}@tcd.ie

Abstract

This paper describes the stages involved in implementing a corpus of spoken Irish. This pilot project (consisting of approximately 140K words of transcribed data) implements part of the design of a larger corpus of spoken Irish which it is hoped will contain approximately 2 million words when complete. It hoped that such a corpus will provide material for linguistic research, lexicography, the teaching of Irish and for development of language technology for the Irish language.

Keywords: spoken language, corpus design, Irish

1. Introduction

This paper describes the design of a corpus of spoken Irish. The proposed spoken corpus, consisting of approximately 2 million words will provide material for linguistic research, lexicography, the teaching of Irish and for development of language technology for the Irish language. Also described are various stages involved in the pilot implementation of part of the proposed corpus.

In order to create a comprehensive corpus of spoken Irish, the design includes dialectal and chronological variation, as well as different registers and contexts of language use. In addition to new recordings, material will also be drawn from existing collections and archives (radio and TV broadcast and folklore archives). The corpus is in line with current standards in terms of time-alignment of transcripts, XML formatting and part-of-speech tagging for electronic searchability and querying. It will also be available online.

2. Linguistic Background

Irish is the first official language of Ireland with English being the second official language. In practice Irish is spoken as a first language in only a small number of areas known as *Gaeltachtaí* which are mainly on the western seaboard. For the remainder of the population Irish is learned at school (compulsorily) as a second language. While 1.6 million[1] of the 3.9 million population report proficiency in the spoken language, the number of native speakers is much lower, at 64 thousand[2] and dwindling in the *Gaeltachtaí* (although numbers are increasing in urban areas). These sociolinguistic conditions mean that a comprehensive spoken corpus has a vital role to play in promoting and preserving the spoken language.

[1] Census 2006 http://www.cso.ie/en/newsandevents/pressreleases/2007pressreleases/2006censusofpopulation-volume9-irishlanguage/

[2] Census 2006 http://census.cso.ie/Census/TableViewer/tableView.aspx?ReportId=96447

3. Corpus Design

In order to design a corpus that is representative and authoritative, it is useful to take into account the design adopted by recent, state-of-the art corpora for other languages. We examined the design of a number of corpora (London-Lund Corpus of Spoken English[3], Lancaster/IBM Spoken English Corpus (SEC)[4], Corpus of Spoken New Zealand English[5] British National Corpus[6], COREC (Corpus oral de referencia del Español Contemporáneo)[7], CLIPS (Corpora e Lessici dell'Italiano Parlato e Scritto)[8], ICE (The International Corpus of English)[9] and CGN (Corpus Gesproken Nederlands)[10]
One common feature shared by the more recent corpora surveyed here is the extent of naturalistic conversational material they include.
There is no existing corpus of spoken Irish which meets our criteria of including dialectal and chronological variation. The most substantial collection of spoken language transcripts, *Caint Chonamara* (Wigger, 2000) (1.2 million words approx.) relates to one dialect only (Conamara) and one year, 1964, and is not linguistically annotated.

Our design considers the following variables:

- time frame: we aim to create a diachronic corpus by including spoken Irish from the earliest available recordings to the present day. We have decided upon

[3] London-Lund Corpus of Spoken English http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

[4] Lancaster/IBM Spoken English Corpus (SEC) http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM.

[5] Corpus of Spoken New Zealand English http://icame.uib.no/wsc/index.htm.

[6] British National Corpus http://www.natcorp.ox.ac.uk/corpus/index.xml.

[7] A reference corpus for contemporary spoken Spanish http://www.lllf.uam.es/~fmarcos/informes/corpus/corpulee.html.

[8] Corpora and Lexica for Spoken and Written Italian http://www.clips.unina.it/it/.

[9] International Corpus of English http://ice-corpora.net/ice/.

[10] Spoken Dutch Corpus http://lands.let.kun.nl/cgn/ehome.htm.

the three time periods, P1: 1930-1971, P2: 1972-1995 and P3 1996-present. In our pilot corpus we concentrated on contemporary speech (P3 1996-present).

- dialectal variation: we aim to cover the three main dialects of Irish in equal measure: i.e. not proportionally to the number of speakers of each dialect, given that the corpus is diachronic and the relative proportions might have varied over the years, but to provide equal documentation of each dialect insofar as possible.
- sociolinguistic variation: we aim to include Irish speakers from all linguistic backgrounds (a) 'traditional' native speakers, (b) non-native speakers and (c) 'non-traditional' native speakers, i.e. those who describe themselves as native speakers having being raised through Irish by L1 or L2 parents, typically in a non-*Gaeltacht* setting, and who have subsequently attended Irish-medium schools (Ó Giollagáin & Mac Donnacha, 2008, p. 111f.).
- gender and age: we aim to represent both males and females proportionally, and to include a spread of ages (e.g. young adults, middle aged and elderly).
- context and subject matter: we aim to include conversations recorded in a variety of contexts (home, work, leisure, education etc.) and cover a variety of topics.

The corpus design for Period 3 (1996-present), which is inspired by the ICE and CGN corpus designs, proposes 70% dialogic speech (420 X 8-10 min recordings) and 30% monologic speech (180 X 8-10 min recordings). Each of the 600 recording transcriptions will contain 1500-2000 words giving a corpus of between 900,000 and 1,200,000 words. The dialogic speech is further categorised into 'private' (e.g. face-to-face conversations, phonecalls and interviews) and public (e.g. broadcast discussions and interviews, parliamentary debates, classroom lessons, business meetings etc.) speech. The monologic speech is categorised as either scripted (news broadcasts, speeches etc.) or unscripted (e.g. sports commentaries, unscripted speeches, demonstrations, legal presentations etc.) speech.
For Period 1 (1930-1971) and Period 2 (1972-1995), not all of the required types of material will be available. We will aim to keep the same proportions as for Period 3 but the quantities will necessarily be less. In order to ensure that as many of the design categories are repesented as fully as is possible, a thorough investigation of available archival material will have to be undertaken.

4. Data Collection and Recording

In the case of dialogic speech (70%) there is ample public broadcast material available in the form of radio podcasts and archives. Other categories such as classroom lessons and business meetings will have to be recorded. All private dialogue speech will have to be recorded in the various dialectal regions. In the case of monologic speech (30%), the majority of this can be sourced from broadcast media and archives, with some categories such as legal presentations being recorded.

Funding was obtained from Foras na Gaeilge (the cross-border body responsible for promoting the Irish language on the island of Ireland) to carry out a pilot study. We decided to concentrate our initial efforts in the contemporary period (P3) and on dialogic speech. As time and resources were limited we used readily available public broadcast dialogues (radio interviews and discussions).

We also carried out a small amount of video recording of private dialogue conversations. Four pairs of volunteers agreed to be video recorded in informal conversation in the Speech Communications Laboratory[11], TCD. The interactions were video recorded using a Sony HDR-XR500v High Definition Handycam. The audio was recorded in two ways: 1) using the onboard camera microphone and 2) using two Sennheiser MKH-60 shotgun microphones and an Edirol 4-channel HD Audio recorder. Audio was recorded at a sampling rate of 96KhZ with a bit rate of 24 bits. For practical purposes, the audio was bounced down to a sampling rate of 44.1KhZ with a bit rate of 16bits (the Redbook audio standard), with the higher 96KhZ files being used for archiving.

In total, 70 x 8 min. recording extracts were transcribed giving 102,000 words of transcribed speech. By also aligning and formatting some existing transcripts[12], the overall total is currently 140,000 words (approximately).

5. Transcription

Spoken and written language differ in a number of important respects. The syntactic structure of spontaneous spoken utterances is usually simpler, but any faithful transcript of a spoken conversation will not look as orderly as a written dialogue. It is natural in spontaneous speech to produce repetitions, make false starts, to hesitate or simply to leave part of a message unfinished, relying instead on non-verbal communication such as a gesture or the tone of voice. This together with dialectal pronunciations which deviate substantially from standard orthographical representations means that transcribing spoken language presents immediate challenges.

5.1 Guidelines

We examined a number of transcription conventions already in use including CHAT[13], LINDSEI[14], and LDC[15]. The CHAT (Codes for the Human Analysis of Transcripts) System is a comprehensive standard for transcribing and encoding the characteristics of spoken language (MacWhinney, 2000). These guidelines were developed for the transcription of spoken interactions between children and their carers in order to study child language

[11] http://www.tcd.ie/slscs/clcs/scl/
[12] Frenda (2011) material transcribed for PhD research TCD (20K); Wigger (2000) Caint Chonamara (10K); Dillon, G., material transcribed for PhD research TCD (5K).
[13] CHAT http://childes.psy.cmu.edu/manuals/chat.pdf
[14] Louvain International Database of Spoken English Interlanguage Transcription guidelines http://www.uclouvain.be/en-307849.html
[15] Linguistic Data Consortium http://www.ldc.upenn.edu/Creating/creating_annotated.shtml#Transcription

acquisition. They give detailed guidelines for marking up such phenomena as inaudible segments, phonetic fragments, repetitions, overlaps, interruptions, trailing off, foreign words, proper nouns and numbers etc. While the guidelines are very comprehensive there are a few drawback to implementing the guidelines in full; it can slow down the transcription process considerably; some are quite subjective (short, medium and long pauses) while others are difficult to implement (retracings and reformulations).

At the other end of the scale, the LDC guidelines advocate simplicity. The philosophy here is to keep the rules to a minimum in order to make transcription as easy as possible for the transcriber, which increases transcription speed, accuracy and consistency. In addition automatic procedures are used when possible.

We also consulted researchers in other universities and research institutes[16] who have worked on the transcription of spoken Irish and obtained advice and good-practice guidelines with regard to the orthographic rendition of dialect-specific features of spoken Irish.

From our experience, it takes on average 30 minutes to orthographically transcribe 1 minute of audio material. Considering that transcription is a slow and painstaking process we believe that in order to achieve a sufficient quantity of accurately transcribed material, the transcription process must be as straightforward and intuitive as possible. This means that codes should be kept to a minimum and those codes which are necessary should use a minimum number of keystrokes.

Some aspects of speech do not need to be recorded in the transcription as they can be automatically generated at a later stage, e.g. the length of pauses. The standard orthography is morphologically transparent, i.e. it shows the internal structure of a word, which is a distinct advantage for the automatic treatment of the text, e.g. for part-of-speech tagging and the generation of a broad phonemic transcription (Ó Raghallaigh, 2010, p. 76).

We have chosen to use standard orthographic representation, for which *Caighdeán Oifigiúil* (1979) and *Foclóir Gaeilge-Béarla* (Ó Dónaill, 1977) are taken as references, and to avoid invented and ad hoc spellings at all times. There are a number of advantages to using standard orthography:

- It makes the job of transcription easier and quicker for transcribers
- It helps mimimise spelling inconsistencies among transcribers as only standard spelling is used, apart from a predefined lists permitted exceptions
- Attempting to represent actual pronunciation in orthography is difficult and prone to inconsistency. It requires specialist knowledge and can be more accurately captured in a separate phonetic transcription layer (which may be partially generated from the orthography).
- Standard orthography facilitates corpus querying and lexical searches
- Standard orthography facilitates automatic text processing, such as part-of-speech tagging and parsing
- Transcription codes for some linguistic features (e.g. co-articulation effects, ellision etc.) would require specialist training for transcribers, in order to ensure accuracy and consistency, and are better undertaken as a separate task.

Based on the above principles a set of guidelines for the transcription of Irish was developed which is available online on the project website.[17] In addition to these general guidelines are lists of prescribed spellings for filled pauses, contracted wordforms, multi-word fixed phrases (including several English fixed phrases, e.g. you know, so, just etc.) and some common dialectal forms not included in the reference dictionary (Ó Dónaill, 1977).

5.2 Software

Creating a corpus of spoken language requires transcribing audio or video recordings (spoken conversations, interviews, speeches etc). These transcriptions should ideally be time-aligned with the speech signal. There are a variety of freely available software packages to carry out this task and to aid the transcription process in general.

We tested several pieces of freely-available transcription and annotation software (e.g. Praat, ELAN, Anvil, CLAN, Xtrans, Transcriber) and chose *Transcriber*[18] as the most suitable software for the orthographic transcription of audio speech at this stage of the project, for the following reasons:

- It has a straightforward user interface which means transcribers can become proficient users is a short amount of time;
- It facilitates alignment of the audio and text transcription in XML format;
- It provides audio duration and word count information at a glance;
- Transcripts can be conveniently exported as text;
- It handles a variety of audio file types, including wav, mp3 (podcasts) and ogg which were used in this project.
- The later version of the software (TranscriberAG) can handle video as well as audio;
- It facilitates the annotation of various features of spontaneous speech (overlap, interruptions, coughs, laughs, etc.) as well as linguistics categories (e.g. proper nouns, human/animate etc. etc.) if desired.
- It can be used with foot pedals for increased speed if necessary;

This decision will be kept under review in future phases as new and inproved software regularly becomes available,

[16] Pauline Welby, Laboratoire Parole et Langage CNRS - Aix-Marseille Université (personal communication); Brian Ó Raghallaigh, Fiontar, Dublin City University (p.c.) ; Eoghan Ó Raghallaigh (Doegen Project, http://dho.ie/doegen/); McKenna, M. (2005).

[17] GaLa Project http://www.tcd.ie/slscs/assets/documents/research/gala/Treoirlinte_agus_Transcriber.pdf

[18] http://trans.sourceforge.net/

and project requirements may change.

5.3 Transcribers

As there were no available experienced transcribers of Irish, it was necessary to recruit and train transcribers in the use of the transcription software and transcription guidelines.

Notices were posted in both Irish and Linguistics departments of universities around Ireland and a good response was received. We required a panel of transcribers covering the various dialects, therefore applicants were asked to nominate their preferred dialect and their second choice (if any). A dialect-specific test workpackage (consisting of a one minute audio file and transcription guidelines) was sent to all suitable applicants. Based on the results of the test piece, a panel of twenty-two transcribers was established.

We organised a transcription workshop which was attended by a number of the transcribers, together with interested parties from Foras na Gaeilge, the Royal Irish Academy as well as post graduate researchers. This proved to be very beneficial to all and discussions about transcriptions issues lead to modifications in the transcription guidelines.

Audio segments of 8 min. in duration containing broadcast discussions and interviews were selected mainly from *Raidio na Gaeltachta* podcasts. Workpackages were sent via e-mail to members of the panel of transcribers who worked from home. They returned a time-aligned transcription and timesheet for each workpackage completed.

5.4 Checking and Anonymising

Each transcript was checked for accuracy against the audio file by a member of the project team. In the case of new video-recordings, the transcripts were also anonymised, i.e. names and places which could identify the participants were replaced by fictitious names to ensure anonynity. Anonymising is not carried out for existing recordings which are available on the internet as podcasts or which have been broadcast on radio or TV.

6. Corpus Processing

6.1 Corpus Metadata

All relevant details related to speakers, transcripts and transcribers are recorded in a database. Each speaker is given a speaker code which is used in the transcript in place of the speaker's name, in order to make speakers less recognisable. Speaker attributes such as dialect, language acquisition type, i.e. whether native Gaeltacht speaker (L1 Gaeltacht), native non-Gaeltacht speaker (L1 non-Gaeltacht) or a non-native speaker (L2), gender and age, etc are recorded where known.

This data is used to generate XML corpus headers, and to facilitate onging monitoring of word counts of the various corpus design categories.

6.2 Corpus Encoding Standards

For each transcript, the output of the Transcriber software was transformed into TEI compliant XCES (XML Corpus Encoding Standard) format using a Perl script and data from the corpus database. The script also computed word counts per speaker which were fed back into the database.

All of the transcripts to date are conversations or interviews involving at least two participants. It is quite common, particularly in radio interviews, for spoken interactions to take place between speakers with different dialects or between native and non-native speakers. As we would like to be able to create sub-corpora on the basis of dialect, native/non-native status, speaker, age, gender etc. then these features must be recorded at the level of speaker-turn rather than for the transcript as a whole.

Therefore, as well as having a detailed transcript header which includes time of recording and source of audio/video file etc. we also include speaker attributes on the <speaker_turn> tag, as shown in Figure 1.

```
<doc id = "irbs0012" title =
"Barrscéalta 08 October 2010" period
= "1996-pres" medium =
"broadcast-radio"spokentype =
"interview" text_source = "GALA-TCD"
av_source = "RnaG podcast">

<speaker_turn id = "200" code =
"RNG_ANC" dialect = "Ulaidh" gender =
"Bain" actype = "L1 Gaeltacht" year =
"2010">
caidé méid airgid a chosnódh sé na bádaí
seo a thabhairt suas chun dáta agus
cloígh lena rialacha úra atá tagtha
isteach?
</speaker_turn>

<speaker_turn id = "559" code =
"RNG_LCI" dialect = "Mumhan" gender =
"Fir" actype = "L1 Gaeltacht?" year =
"2010" >
Bhuel ehm braitheann sé sin ar
chaighdeán an bháid, abair, agus níl
aon dabht faoi ach go bhfuil sé
costasach, abair, [tá tá] tá tuairiscí
faighte agamsa ar daoine go raibh orthu
eh [céad míle ar] céad míle euro a
chaitheamh eh ag tabhairt a mbád suas
chun caighdeáin. …
</speaker_turn>
```

Figure 1 Fragment XCES formatted spoken transcript

6.3 Part-of-speech Tagging

The XML transcripts have been part-of-speech tagged. Additional codes and lexical items were added to the finite-state tokenizer and morphological analyser (Uí Dhonnchadha, 2006) to handle some features specific to spoken language such filled pauses (em, eheh etc,) fixed

phrases (*an dtuigeann tú* 'do you understand', *mar a déarfá* 'as you say' etc.), as well as codes for non-verbal events (coughs, laughs, sneezes etc.), phonetic fragments (*b- b- bosca* 'b- b- box') and indecipherable material (xxx). Dialectal varriants (Ó Dónaill, 1977) e.g. *gleamaigh* 'lobster', *aoinne* 'anyone' etc. proved useful as these forms are perhaps more common in spoken language than written language.

Spoken transcripts contain more English words than would be found in written Irish, therefore a list of English vocabulary items would be useful addition to the morphological analyser, but this was not carried out in the current phase project. Detailed analysis of the accuracy of the POS tagging on spoken language as compared to accuracy on written language also has not yet been carried out.

6.4 SketchEngine Corpus Query Engine

All POS tagged transcripts have been converted to vertical format and loaded into the SketchEngine [19] Corpus Query System. For each transcript the following information is available: document id, title, time period, text_source (source of transcription) and av_source (source of audio/video file). For each speaker turn the following information is available: speaker code, dialect, actype (language acquisition type), gender, year of recording. Sub-corpora can be created by selecting particular values for any selection of the above variables, i.e. dialect = Ulster, actype=L1, etc.

7. Conclusion

In this paper we have outlined the issues involved in designing a spoken corpus, including data collection and transcription and initial stages of corpus processing. Through implementing a pilot corpus, we believe that we have overcome most difficulties likely to be encountered in a fullscale project, and are in a position to make infomed decisions about costings and timings of a larger scale project.

8. Future Work

The main tasks for the future, are to collect additional data particularly through the recording of spontaneous conversations from volunteers in various *Gaeltacht* locations around the country, and also to improve the part-of-speech tools to better handle the particular characteristics of spoken langauage. Quality control measures would also need to be put in place to ensure the quality and consistency of future transcriptions.

9. Acknowledgements

We would like to thank Foras na Gaeilge for funding this phase of the project.

[19] http://the.sketchengine.co.uk/

10. References

Caighdeán Oifigiúil, An. (1979[1958]). *Gramadach na Gaeilge agus litriú na Gaeilge: An caighdeán oifigiúil.* Baile Átha Cliath: Oifig an tSoláthair.

Frenda, A. (2011). *Gender in Insular Celtic: A functionalist account of variation and change in Irish and Welsh.* PhD thesis, Trinity College Dublin.

Ó Curnáin, B. (2007). *The Irish of Iorras Aithneach County Galway*. Dublin: Dublin Institute for Advanced Studies.

Ó Dónaill, N. (1977). *Foclóir Gaeilge-Béarla*. Baile Átha Cliath: Roinn Oideachais agus Eoilaíochta.Ó Giollagáin, C. & S. Mac Donnacha (2008). The Gaeltacht today. In C. Nic Pháidín and S. Ó Cearnaigh (Eds.), A new view of the Irish language, pp. 108–120. Dublin: Cois Life.

Ó Raghallaigh, B. (2010). *Multi-dialect phonetisation for Irish text-to-speech synthesis: A modular approach.* PhD thesis. Trinity College Dublin.

MacWhinney, B. (2011). *The CHILDES Project: Tools for Analyzing Talk*. Electronic Edition. Available online at http://childes.psy.cmu.edu/manuals/chat.pdf.

McKenna, M. (2005). *Seanchas Rann na Feirste: is fann guth an éin a labhras leis féin*, pp.169-180. Dublin: Coiscéim.

Uí Dhonnchadha, E. and van Genabith, J. (2006). Scaling an Irish FST morphology engine for use on unrestricted text, In: Yli-Jyrä, A., Karttunen, L., Karhumäki, J. (Eds.). *Finite-State Methods in Natural Language Processing* (Book Series: Lecture Notes in Artificial Intelligence), Springer-Verlag, pp. 247 – 258.

Wigger, A. (Ed.) (2000). *Caint Chonamara: Bailiúchán Hans Hartmann. Imleabhar IX. Ros Muc*. Universität Bonn.