Applications of parallel corpora to the development of monolingual language technologies

Kevin P. Scannell
Department of Mathematics and Computer Science
Saint Louis University
Saint Louis, Missouri, USA

Abstract

We describe the development of an aligned parallel corpus of English and Irish texts, along with a simple application enabling the standardization of documents written in prestandard or dialect forms of Irish.

1 Introduction

Parallel corpora have many applications in natural language processing for problems involving multiple languages: machine translation, multilingual information retrieval, etc. We believe they may play an equally important role in the development of *monolingual* language technologies, particularly for minority languages, which often lack basic tools for NLP. In such cases, our strategy has been to construct a parallel corpus with a global language (often English or French) and to exploit the resources that already exist for the global language in constructing analogous resources for the minority language.

This paper describes a simple application of this strategy to the Irish language. With a dwindling base of roughly 50,000 native speakers embedded in a thoroughly English-speaking island of five million, Irish fits any reasonable definition of "minority language". On the other hand, a combination of factors (positive economic conditions in Ireland, the constitutional status of Irish as an official language, etc.) have placed the language in a relatively strong position for the development of advanced NLP tools, particularly when compared with some of the languages of Africa and the Pacific to which we hope to apply our techniques, including several with tens of millions of speakers (e.g. Hausa, Yoruba).

In §2 we describe the development and contents of an aligned parallel corpus of English and Irish texts, and in §3 we discuss an application of (a subset of) the corpus; namely, a program to standardize Irish documents written in either prestandard or dialect forms of the language.

A number of other projects have involved minority/global language parallel corpora in one way or another, including (but certainly not limited to) the following:

- The EMILLE project (McEnery et al., 2000) developed large corpora for several "non-indigenous" minority languages of Britain (Indic languages) with smaller parallel corpora on the order of 200,000 words.

- In (Scannell, 2003), a monolingual Irish thesaurus is constructed by statistical means, by exploiting the various English thesauri already available in electronic form to transfer semantic relationships over to Irish.

- The OPUS corpus (Tiedemann and Nygaard, 2004) contains over 30 million words of parallel texts in 60 languages (including several minority languages) harvested from the translation compendia of various open source software projects.

- The STRAND project (Resnik and Smith, 2003) created databases of parallel texts harvested automatically from the web (including 59 Basque-English document pairs).

- (Trosterud, 2002) discusses how parallel corpora can be used for minority language plan-

ning of terminology and other lexicographical purposes.

2 The Irish-English Parallel Corpus

This section gives a brief overview of the *Corpas Comhthreomhar Gaeilge-Béarla* (Irish-English Parallel Corpus), including the techniques used to assemble and align the texts, and the available interfaces.

2.1 Construction

All governmental bodies in Ireland are required to release documents such as annual reports in both English and Irish, and this provides a wealth of material for alignment. Indeed, the largest single source of such material comes from the full-text database of all legislation enacted by the Oireachtas (Parliament) since 1922[1].

We gathered the corpus as follows. First, a large (about 25 million word) monolingual Irish corpus was harvested using a web crawler that has the ability to target particular languages. Essentially it works by bootstrapping a list of high-frequency words and a (character) 3-gram model from previously crawled text. These data are used to create search engine queries and to recognize documents in the target language. The corpus constructed in this way is accurate (we have found no non-Irish documents among those determined to be Irish by the crawler) and comprehensive (combinatorially generated search engine queries rarely turn up new documents not already downloaded). Such broad coverage is clearly only possible with languages having a limited presence on the web.

The documents in the monolingual corpus were then searched for a number of heuristic indicators that a translated English version might be available ("siblings" following the terminology of (Resnik and Smith, 2003)). These indicators were mostly naive: hypertext links labeled "English Version", "English", "Béarla", etc. and a number of heuristics based on the URL, for instance those matching patterns like `/_ga/` or containing a directory named "`irish`" or "`ga`" in the path. When corresponding candidate English translations were found, these were downloaded, checked manually, and prepared for alignment when appropriate.

The corpora were aligned at the sentence level using a slightly-modified version of the original Gale-Church algorithm (Gale and Church, 1993). The best results were obtained by using special code for handling sentence segmentation of the Irish texts. In addition, we found it worthwhile to add (manually) a number of "hard delimiters" before alignment, particularly for noisy texts and for some text pairs that were not perfect translations of one another. As the accuracy of the Gale-Church algorithm has been well-established in earlier work, and because our focus is on end-user applications we have made no attempt at systematically evaluating the quality of the alignments.

Because one of the applications we intended for this database was to assist translators working on translating software into Irish, we augmented the corpus by including translations of several software packages (KDE, OpenOffice, Mozilla) as well as various online terminology databases.

In all, the corpus contains nearly 6.45 million words of English and 6.56 million words of Irish.

2.2 User interface

Since the majority of the documents in the database were harvested from the web and are covered by copyright restrictions, we are unable to redistribute the corpora *in toto*. Nevertheless, a large subset, including all of the open source software translations, is freely available and can be accessed either through a web search interface or an XML-RPC interface.

The search engine accepts queries in either English or Irish and returns all sentence pairs containing the search terms as a TMX-compliant [2] XML document that can be used by various translation memory applications. In addition, there is an option for converting the TMX output into readable HTML when the results are intended for direct human consumption.

The XML-RPC web service accepts queries and generates TMX output in much the same way as the web interface. In this case though, queries can be made programmatically from within client applications such as translators' tools. Because XML-RPC

[1] `http://www.achtanna.ie/`

[2] `http://www.lisa.org/tmx/`

clients are available in Java, Python, Perl, C, etc., the parallel corpus can be easily integrated into existing applications no matter what language they are implemented in, or what platform they run on.

In fact, as a proof of technology, we have written a simple text-based translation memory tool (in Perl) for translating open source software into Irish. An English string to be translated is sent as an XML-RPC call to the central server containing the parallel corpus, and the 25 closest sentences found in the English half of the corpus are returned along with their Irish counterparts. When the string has been translated, this new pair can be added to the parallel corpus if desired.

We envision a whole suite of similar client programs used for translation, language learning, and lexicography all sharing (and contributing to) a common database.

3 Standardization of Irish texts

3.1 Background

Relatively few books were published in Irish before the late 19th century, and those that were exhibit very little regularity in spelling. For instance, the Royal Irish Academy's *Irish Language Corpus* (CNG, 2004) of texts published between 1600 and 1882 shows more than a dozen spellings for the word we now know as *aoibhinn* (pleasant), with acute accents placed more or less at random on any of the available vowels. Dinneen's Irish-English dictionary, published in 1927 (Dinneen, 1927), brought some order, if only by selecting in most instances a single form as headword, but still retained many archaic forms and silent consonants. In the 1940's, there was a major spelling reform known as the *Caighdeán Oifigiúil* (Official Standard) initiated and implemented by *An Rannóg Aistriúcháin* (the governmental body in charge of official translations) that was solidified in the decades that followed through the publication of the two primary bilingual dictionaries (de Bhaldraithe, 1959), (Ó Dónaill, 1977).

It is worth noting at this point that there are three major dialects of Irish (Connacht, Munster, and Ulster, corresponding roughly to the west, south, and north of the island), each with its own idiosyncrasies of vocabulary and orthography. Advocates for these dialects often feel that their own dialect is not sufficiently well-represented in the standard form of the language, and as a consequence one still finds quite a bit of text in newspapers, online discussion groups, etc. not adhering to the standard. On the other hand, nearly all of the documents published by the government (and, as noted earlier, these form the great majority of the parallel corpus) are standard.

3.2 Standardizer

Our goal in developing software for standardizing Irish texts is not to impose a sterile or artificial standard on speakers of the language, but primarily as a tool for indexing and information retrieval of prestandard texts. Databases of prestandard texts can be passed through the standardizer and indexed according to the standard forms, allowing a search for *aoibhinn* to return documents containing any of the dozen or so forms appearing in the Irish Language Corpus. Conversely, terms entered into a search engine can be standardized, allowing searches for *aóibhinn*, *aoidhbhinn*, and so on, to return modern documents containing the standard form. In fact, the standardizer is used in precisely this way as part of the search engine interface to the very parallel corpus used in its construction.

Many of the spelling reforms are given by simple rules that can be reliably implemented with some regular expressions. For example, all instances of /sg/ are replaced by /sc/ in the standard: *sgéal* becomes *scéal*, etc. Unfortunately, other reforms were applied inconsistently, so such an approach is error prone. For example, the silent /dh/ is generally removed from words such as *biadh*, *cruadh*, and *comhrádh* in the standard, but not for *ádh*, *cogadh*, *margadh*, etc.

Because the parallel corpus contains large amounts of text from both before and after the implementation of the standard, we were able to create a database of Irish orthographical changes, essentially by collating pre- and post-standard (bilingual) terminology lists induced with standard mutual information techniques. This analysis was performed using the subcorpus of translations produced by *An Rannóg Aistriúcháin* consisting of the texts of all legislation enacted by the Irish government since 1922.

A quick glance through this material shows

spelling changes beginning to occur, as expected, sometime in the mid-1940's; for instance the prestandard form *cialluíonn* occurs only prior to 1945 while the modern version *ciallaíonn* occurs only after 1943 (and similarly for all other verbs of this declension). We therefore defined the prestandard corpus to consist of the text pairs dated 1922–1943 and the standard corpus to consist of those dated 1945–1998 (1944 is something of a mixed bag and was left out).

The texts for both languages were tokenized and indexed with all letters converted to lowercase. In addition, because Irish has a number of "initial mutations" that are applied to words based on local context[3] these were stripped off before indexing the Irish terms. No stemming was performed. We then extracted all pairs of words with mutual information at least $\frac{1}{\sqrt{2}}$ and frequency at least four, restricting first to the 1922–1943 texts (yielding 2433 pairs), and then repeating this computation restricting to the 1945–1998 texts (3569 pairs).

Because English spelling has remained constant during this brief time span, the two terminology lists were easily collated, giving a list of 1482 putative standardizations. The table shows a small subset of these data, with the English "bridge word" provided in the third column, though this is not used in the final application.

1922–1943	1945–1998	English
árthach	soitheach	vessel
ath-achtú	athachtú	re-enactment
athair	athair	father
atharuithe	modhnuithe	modifi cations
athchomhairc	achomharc	appeal
ath-dhíol	athdhíola	resale
athghairmtear	aisghairtear	repealed
athnuachaint	athnuachan	renewal
ath-shocrú	idirmheasctha	re-arrangement
ath-shuidheamh	athfhostú	re-instatement
ath-thógáil	opera	rebuilding
ath-thógainte	athghabhála	resumption
átomaitigiúil	uathoibríoch	automatic
atúrnae	aturnae	solicitor
atúrnaethe	aturnaetha	solicitors

This selection exhibits many of the phenomena we see in the larger list: some words are correctly unchanged (*athair*), others have undergone more or less standard spelling reform (*aturnae*), and others have changed terminology entirely. There are very few pairs in the full list that are completely incorrect; the example *ath-thógáil/opera* above is one of them (several places where the word "rebuilding" is used in the 1945–1998 English texts refer to the rebuilding of the Opera House in Cork).

The resulting standardization software combines a set of 340 regular expression rules with a large database of replacements generated in this way. We are currently performing some tests on the text of the Irish Constitution, which was recently published in parallel prestandard/standardized form, and can be used as a gold standard for our software (Ó Cearúil, 2003).

3.3 Future Work

This application of the parallel corpus is straightforward but quite important for correctly processing prestandard Irish texts and we hope will be useful for information retrieval purposes and lexicographical projects, including the *Foclóir Stairiúil na Nua-Ghaeilge*[4] and the new English-Irish dictionary project[5]. In addition we expect this approach to be useful for some of the many other languages that have undergone a spelling reform or standardization.

References

Corpas na Gaeilge: 1600–1882. Acadamh Rí oga na hÉireann, Baile Átha Cliath.

Tomás de Bhaldraithe, editor. 1959. *English-Irish Dictionary*. An Gúm, Baile Átha Cliath.

Patrick S. Dinneen, editor. 1927. *Foclóir Gaedhilge agus Béarla*. Irish Texts Society, Baile Átha Cliath.

W. Gale and K. W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102.

A. McEnery, P. Baker, and L. Burnard. 2000. Corpus Resources and Minority Language Engineering. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, pp. 801–806.

M. Ó Cearúil. 2003. *Bunreacht na hÉireann: An téacs Gaeilge arna chaighdeánú*. Coiscéim, Baile Átha Cliath.

[3]The word *bean* also appears as *bhean* (lenition) and *mbean* (eclipsis), but in all three cases would appear in parallel with the same English form "woman", or "wife".

[4]http://www.ria.ie/projects/fng/

[5]http://www.focloir.ie/

Niall Ó Dónaill, editor. 1977. *Foclóir Gaeilge-Béarla*. An Gúm, Baile Átha Cliath.

P. Resnik and N. A. Smith. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

K. P. Scannell. 2003. Automatic thesaurus generation for minority languages: an Irish example. In *Actes de la 10e conférence TALN à Batz-sur-Mer*, volume 2, pages 203–212.

J. Tiedemann and L. Nygaard. 2004. The OPUS corpus – parallel and free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.

T. Trosterud. 2002. Parallel corpora as tools for investigating and developing minority languages. In L. Borin, editor, *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*, pages 111–122. Rodopi, Amsterdam-New York.