

Software & Tools

Hyphenation patterns for minority languages

Kevin P. Scannell

Abstract

We present some techniques used in developing hyphenation patterns for the Irish language that we hope will be applicable to other languages with limited computational resources.

1 Introduction

Irish is one of six languages in the Celtic branch of the Indo-European family (the others are Scottish Gaelic, Manx Gaelic, Welsh, Cornish, and Breton). Typesetting enthusiasts might be familiar with the so-called “Gaelic” fonts (in Irish, *seanchló*, literally “old type”) used to print the language until the early part of the 20th century, and which trace their roots back to the exquisite illuminated manuscripts

produced by Irish monks in the centuries preceding the Norman conquest [4, 5, 6]:

Δοιῖνν βελεῖα ἀν ῥολῶιηε
 ὅϊορ ἄς οἔληλῶη ἄ λῆίξινν;
 ιρ ῥολῶρ οἴβ, ἄ ὄλοιηε,
 ζυηῶδ οῶ ιρ λοἴβηε ιη Ἐηηηη. ¹

Once spoken by several million people, there are now perhaps only 50,000 native speakers, mostly in remote regions of the west of Ireland.² English remains a constant presence throughout, especially in the contexts of technology and computing. This has had the unfortunate tendency to reinforce the view, especially prevalent among the young, that Irish is “irrelevant” for modern life.

Since 1999 the author has been engaged in the development of Irish language software as a way of helping to stem the tide of language shift. Already completed are a general purpose web crawler, spell checker, and grammar checker, with a monolingual thesaurus and various localization projects currently in progress.³ Recently, a set of T_EX hyphenation patterns for Irish was completed as part of this work.⁴ This topic forms the focus of this paper.

In most languages, the choice of font has no effect on hyphenation, but typesetting Irish in Romanized script (as is standard nowadays) implies some important *orthographic* changes, most notably the use of the letter ‘h’ to indicate “lenition”, or softening, of the preceding consonant. This is indicated by a *ponc* (dot) over the lenited consonant in Gaelic type, as can be seen in the excerpt above. The current version of the patterns is designed to work with the modern orthography, but if desired can be modified to work for *seanchló* as well.

In the following sections some of the techniques used in developing the patterns will be described in the hope that they will be applicable to other minority languages. Indeed the main goal in writing this article is to encourage further work on hyphenation and other natural language processing (NLP) tools for marginalized and under-resourced languages. Some relevant statistics are presented in the final section for further reflection.

The so-called “information bottleneck” of NLP is especially acute for minority languages. It is accompanied in most cases by a lack of skilled

¹ “Lovely the life of the scholar, diligently working; you know well, good people, his is the sweetest lot in Ireland.” This Gaelic font, produced using METAFONT by Ivan Derzhanski, is called *eiad* and is available from CTAN.

² The language has the nominal support of the Irish government and is taught in the schools, so a somewhat larger number of people claim fluency.

³ <http://borel.slu.edu/gaeilge.html>

⁴ <http://borel.slu.edu/fleiscin/>

software engineers, linguists, or both. This induces one to exert effort in those areas where the materials produced can be deployed as widely as possible. In this context, it is worth noting (even to an audience of \TeX devotees) that Liang’s hyphenation algorithm has come into much wider use over the last two or three years. The \TeX hyphenation files themselves can be used directly by GNU Troff,⁵ and slightly modified versions are now used by the free software packages OpenOffice⁶ and Scribus.⁷ There are, in addition, at least two implementations of XSL-FO processors (these convert XML data plus style sheet information into PDF and other formats) that employ \TeX hyphenation (including Apache’s FOP⁸).

2 General techniques

Much of what is said here is well known; in particular, bootstrapping a database of hyphenated words using PATGEN is a well-established technique; see, e.g. [8]. Readers interested in creating new patterns are encouraged to read Petr Sojka’s papers (especially [8] and [9] with Pavel Ševeček), Yannis Haralambous’ PATGEN tutorial [2], and the Master’s thesis of David Antoš [1].

2.1 Parallel development

Just five years ago, there were virtually no Irish language lexical resources in machine-readable form. Based on the author’s experience, developing a full suite of resources in parallel is easier than attempting each individually. In short, what is advocated here is a *synergistic approach* to the development of multiple NLP tools that exploits “feedback loops” and synergies between them.

As a simple illustration, consider the simultaneous development of a web crawler, text corpus, and spell checking database. The web crawler reported here uses the Google API⁹ and words from the spell checking database to search for potential Irish language documents on the web; the spell checker and some statistical techniques are used to determine which documents (or sections thereof) are actually in the Irish language. These are added to the text corpus, and more statistical analyses (frequencies, character n -grams) are used to find reliable candidate words for the spell checker [7]. Such a system can be bootstrapped from a small word list, and for much of its life cycle requires

no human intervention. Each of the three subsystems improves over time. Like many individuals working with small languages, I was led to employ unsupervised, statistically-based NLP not from any *a priori* fundamental belief in its effectiveness, but simply because it appeared the most viable method to achieve reasonable results in a limited timeframe.

2.2 Hyphenation and spell checking

In the context of hyphenation, feedback between the hyphenation patterns and the morphological and phonological data encoded in advanced spell checkers like `aspell` can be exploited.¹⁰

At the most basic level, a spell checker is really just a word list stored in some kind of hash table for efficient lookup. The standard UNIX spell checkers also offer *affix compression*. This is a way of encoding portions of the morphology of a language in an “affix file”; then, instead of storing all variants of a given word in the word list, the approach is simply to store something like a dictionary headword and a flag indicating the rules that govern inflection of the word. For heavily inflected languages like Irish, this compresses the hash table by around 70%. Below is shown a tiny chunk of the Irish affix file, showing three future endings of one kind of verb. The left hand side gives the ending to which the rule applies (as a regular expression in general) and the right hand side indicates which letters to strip off and which to add:

```
A Í M > -AÍM,ÓIDH
A Í M > -AÍM,ÓIMID
A Í M > -AÍM,ÓFAR
```

Since Irish is generally hyphenated according to morphological rules, the spell checker offers a powerful means to insert all of these hyphen points into the database at one go. For this, a “fake” affix file was created containing all the same rules, but which permitted an additional non-alphabetic character to appear on either side of a rule (here the use of ‘!’ is adopted since Irish has quite a few explicitly hyphenated words). A command line option to the spell checker allows one to expand all affix flags from the (unmodified!) word list according to these rules, resulting in a rich initial set of hyphen points. One bootstrapping iteration with PATGEN handled all irregular verbs which weren’t encoded in the affix file.

To convince the reader of the importance of parallel development, this subsection is closed with an example of feedback from the hyphenation patterns

⁵ <http://www.gnu.org/software/groff/groff.html>

⁶ <http://www.openoffice.org/>

⁷ <http://web2.altmuehlnet.de/fschmid/>

⁸ <http://xml.apache.org/fop/index.html>

⁹ <http://www.google.com/apis/>

¹⁰ <http://aspell.net/>.

back to the spell checker. The easiest example of this is the input to the “metaphone algorithm” implemented as part of `aspell`. This algorithm depends on the existence of a “coarse” phonetic encoding of your language that can be used to improve suggestions when a misspelling is encountered.¹¹ Having an accurate hyphenation database in place before attempting such an encoding offers a significant efficiency. For instance, the words *garbhuille* (*gar* + *bhuille* lit. “near + stroke”—an approach shot in golf) and *garbhghlórach* (*garbh* + *ghlór* + *ach* lit. “rough + voice + ish”—raucous) share their first five letters, but these are pronounced quite differently in each case. Having the hyphenations in place while constructing the phonetic rules allows one to avoid numerous special cases dealing with situations like this.

2.3 Print sources and human informants

Another serious problem worth mentioning, as it surely faces most other minority languages, is the lack of explicit standards for, or printed dictionaries of, Irish hyphenations. The only general observation beyond debate is that Irish is best hyphenated according to etymological and morphological rules. Knowing this, it becomes quickly and painfully apparent when a given text has been hyphenated according to an English (syllabic) computer algorithm, or, as apparently happened quite often during the early Irish revival, by a monolingual English-speaking compositor.

The most abominable examples of this result from the the convention, noted above, of using an ‘h’ in Roman type to indicate lenition; this rule has the corollary that one should never split the ‘h’ from the preceding consonant. Unfortunately, examples like *com-halta* or *bót-har* can be found in abundance in printed books. Matters are somewhat complicated by the fact that ‘h’ appears occasionally in loanwords and in such contexts often *is* a good hyphenation point: *Bóí-héam-ach* (“Bohemian”).

Nevertheless, the author was able to assemble enough suitable printed material to populate the initial hyphenation database manually.¹² Originally it was hoped to extract, automatically, hyphenated words from the many online PDF documents produced by the Irish government, but (presumably be-

cause of the lack of proper hyphenation technology) many of these are set with a ragged right!

This work also benefited greatly from the input of many Irish speakers who checked over the hyphenations produced by early versions of the patterns on the top 1000 most frequent Irish words; see <http://borel.slu.edu/fleiscin/mile.html>.

2.4 PATGEN esoterica, final results

One of the most difficult aspects of using PATGEN effectively is the choice of correct parameters. Some good heuristics are offered in [1], with actual examples (size-optimized, precision-optimized, etc.) in [9]. I found that with (the usual) five levels of hyphenation, I consistently ended up with a couple hundred bad hyphenations; adding a sixth (inhibiting) level with parameters (1,1000,1) disposed of these and only added 1K or so to the final pattern file. Here is the full set of parameters which worked best:

| Level | Lengths | Parameters |
|-------|---------|------------|
| 1 | 2 ... 4 | (1,2,30) |
| 2 | 2 ... 5 | (1,2,30) |
| 3 | 3 ... 6 | (1,2,6) |
| 4 | 3 ... 7 | (1,2,6) |
| 5 | 3 ... 8 | (1,1000,1) |
| 6 | 3 ... 9 | (1,1000,1) |

The result is a large set of patterns (about 6000) but an extremely accurate one: no bad hyphens and just 10 missed hyphens from a database of 314,639 possible hyphen points in 234,789 words.

3 Other languages

According to the Ethnologue database (<http://www.ethnologue.com/>), there are more than 6800 living languages; at least 2000 have some form of writing system (based on a count of the number of languages with at least partial Bible translations).¹³ By one rough count, however, there are only 36 languages having a reasonably complete desktop computing environment available.¹⁴ Only half of the world’s population are native speakers of one of these 36 languages, meaning some three billion people have no way of using a computer in a native language context (ignoring the more fundamental problems of poverty and illiteracy for many of these same three billion).

¹¹ This is especially important for Irish, which has many silent consonants and which underwent a major spelling reform in the 1950’s. For example, the top suggestion made by `aspell` for the pre-standard form *imfhiosach* is, correctly, *iomasach*, which has the same phonetic encoding.

¹² For the Irish speakers among the readership, the best choices were books published by Sáirséal agus Dill during the 1950’s and 1960’s.

¹³ For a nice discussion of this question, see <http://www.omgios.org/117.htm>.

¹⁴ For instance, version 3.1 of the KDE desktop for GNU/Linux is at least half translated for exactly 36 languages; Windows and Mac localizations make up a (small) proper subset of these.

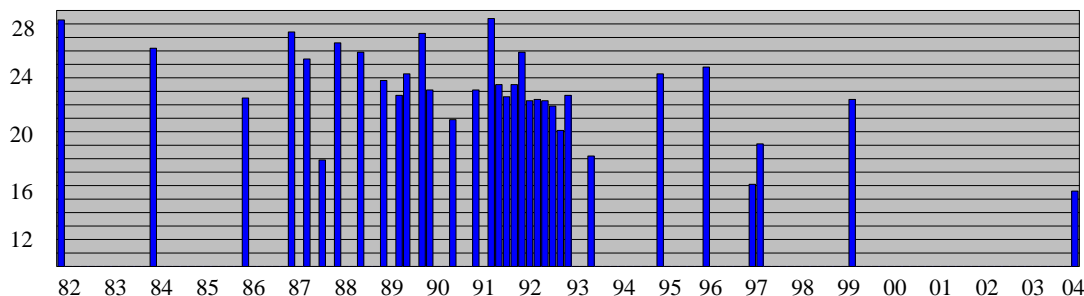


Figure 1: TeX hyphenation patterns by initial release date

The numbers are similar for TeX hyphenation patterns; using the table in [9], the CTAN archives, and some web searching, I found at least some mention of the existence of patterns for 34 natural languages (not surprisingly, 28 of these appear on the list of 36 above). It seems, though, that production is slowing down (although other explanations are possible, e.g. lack of publicly released materials). In figure 1, the horizontal axis represents time (labeled with two-digit years). Each bar indicates the initial release of TeX hyphenation patterns for a new language, with the height of the bar given by the base 2 logarithm of the number of native speakers. For example, Estonian (patterns released in 1992) has a hair over 2^{20} or one megaspeaker. The far left represents Liang’s thesis [3] (assuming $2^{28.3} = 340,000,000$ native English speakers) while the far right represents the patterns for Irish (assuming $2^{15.6} = 50,000$ native speakers). Note that in the past decade, patterns for just six new languages have been released (Romanian, Indonesian, Sorbian, Basque, Mongolian, and Irish) with just one in the past five years.

There is a small research community concerned specifically with NLP for minority languages (see, for instance, <http://193.2.100.60/SALTMIL/>), but their work is confined largely to the European sphere, encompassing perhaps thirty languages beyond the three dozen or so noted above. The author has undergraduate students currently applying some of the techniques described in this paper to the Maori and Inuktitut languages, but this, of course, is just a drop in the bucket.

It is worth emphasizing, in closing, the importance of an open source approach to these problems, leveraging the collective effort of small communities, and transcending the purely market-driven approach that has led to the current dismal state of affairs. It is hoped that the work reported in this

paper makes a small contribution to addressing this situation.

References

- [1] David Antoř. Generation of Patterns with the OPatGen Program. <http://www.fi.muni.cz/~xantos/patlib/thesis.html>, 2001.
- [2] Yannis Haralambous. A small tutorial on the multilingual features of PatGen2. <http://www.ctan.org/tex-archive/info/patgen2.tutorial>.
- [3] Franklin M. Liang. *Word Hy-phen-a-tion by com-puter*. Ph.D. thesis, Stanford University, 1983.
- [4] E. W. Lynam. *The Irish Character in Print*. Barnes and Noble Inc., New York, 1969.
- [5] Dermot McGuinne. *Irish Type Design*. Irish Academic Press, Baile Átha Cliath, 1992.
- [6] Timothy O’Neill. *The Irish Hand*. The Dolmen Press, Port Laoise, 1984.
- [7] K. P. Scannell. Automatic thesaurus generation for minority languages: An Irish example. In *Actes de la 10e conference TALN  Batz-sur-Mer*, volume 2, pages 203–212. ATALA, 2003.
- [8] Petr Sojka. Hyphenation on Demand. *TUGboat*, 20(3):241–247, 1999.
- [9] Petr Sojka and Pavel řevecek. Hyphenation in TeX — Quo Vadis? *TUGboat*, 16(3):280–289, 1995.

◊ Kevin P. Scannell
 Department of Mathematics and
 Computer Science
 Saint Louis University
 St. Louis, MO 63017
 USA
scannell@slu.edu
<http://bore1.slu.edu/>