Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
Multi-dialect phonetisation
for Irish text-to-speech synthesis:
a modular approach

Brian Ó Raghallaigh

A thesis submitted for the degree
of Doctor of Philosophy

Supervisor: Prof. Ailbhe Ní Chasaide
Co-supervisor: Dr. Pauline Welby

School of Linguistic, Speech and Communication Sciences
Trinity College Dublin
September 2010
Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I agree that the Library of Trinity College Dublin may lend or copy the thesis upon request.

[Signature]

September 30, 2010

Brian Ó Raghallaigh
Summary

This thesis describes the development of a modular multi-dialect phonetiser for the Irish language. Phonetisation (or grapheme-to-phoneme conversion), the process in which ordinary text is converted into a transcription of how that text might be pronounced, is a key step in the process of text-to-speech (TTS) synthesis, the conversion of written text to artificial speech. Phonetisation for Irish poses a number of challenges, including a complex (but transparent) orthographic system, the absence of a standard dialect, and the lack of prerequisite technological resources for the language. For example, phonetisation typically uses a combination of letter-to-sound (LTS) rules and a pronunciation lexicon or dictionary, but no dialect-specific pronunciation lexicon is available for Irish.

The phonetiser described here was developed as part of a project to build a TTS system for Irish dialects and has been successfully implemented as part of that system (publicly available at aabair.ie). An important objective of the larger synthesis project was to produce a synthesiser for at least one subdialect of each of the three main dialect groups of Irish: Ulster (U), Connaught (C) and Munster (M). In developing the phonetiser, a two-stage modular approach was adopted to produce the multi-dialect system required. The modular architecture of the phonetiser consists of a single global module for all dialects, and one local module for each dialect implemented. The global module converts orthographic text to an abstract inter-dialectal transcription, and the local rules convert this abstract representation to a dialect-specific transcription. As a result, the system can easily be extended to additional dialects. I present here phnetisation for three dialects: Gaoth Dobhair (Co. Donegal, Ulster), Conamara (Co. Galway, Connaught), and Corca Dhuibhne (Co. Kerry, Munster).

The approach adopted is rule-rich and lexicon-light. It takes advantage of the fact that Irish has a largely transparent orthographic system, with letter-to-sound correspondences that are mostly regular, although somewhat complex. For exam-
pie, the word \textit{acadamh} `academy' is pronounced \([\text{ak}\text{\textordbar {\textasciicircum d}l}\text{\textordbar {\textasciicircum v}\text{\textordbar {\textasciicircum v}}}]\) in Connaught Irish, ending with a \([\text{v}\text{\textordbar {\textasciicircum v}}]\), and \([\text{ak}\text{\textordbar {\textasciicircum d}l}\text{\textordbar {\textasciicircum u}\text{\textordbar {\textasciicircum u}}}]\) in Ulster Irish, ending with the vowel \([\text{u}\text{\textordbar {\textasciicircum u}}]\). The LTS correspondences are clearly not one-to-one and they often differ markedly across dialects, as well as from the correspondences in the spelling systems of other languages (i.e., in the world's languages, the trigraph `amh’ does not often stand for \([\text{\textordbar {\textasciicircum v}\text{\textordbar {\textasciicircum v}}}]\) or \([\text{u}\text{\textordbar {\textasciicircum u}}]\)). Yet Irish orthography is largely transparent. These correspondences are regular: within a dialect, words with similar spellings will have similar pronunciations (for example, \textit{annamh} `rare’ C: \([\text{\textordbar {\textasciicircum n}\text{\textordbar {\textasciicircum l}}\text{\textordbar {\textasciicircum a}\text{\textordbar {\textasciicircum v}}}]}\), U: \([\text{\textordbar {\textasciicircum n}\text{\textordbar {\textasciicircum l}}\text{\textordbar {\textasciicircum u}\text{\textordbar {\textasciicircum u}}}]}\), \textit{talamh} `land’ C: \([\text{\textordbar {\textasciicircum t}\text{\textordbar {\textasciicircum l}}\text{\textordbar {\textasciicircum a}\text{\textordbar {\textasciicircum m}}}]}\), U: \([\text{\textordbar {\textasciicircum t}\text{\textordbar {\textasciicircum l}}\text{\textordbar {\textasciicircum u}\text{\textordbar {\textasciicircum u}}}]}\)). The phonetiser therefore relies heavily on LTS rules in conjunction with a small exceptions lexicon, in contrast to the more widely used technique of using LTS rules only for words not in the lexicon.

Following an introductory chapter, Chapter 2 provides background information on text-to-speech synthesis and locates the phonetisation component within the larger enterprise. Chapter 3 addresses aspects of the sound system of Irish that are relevant for our research on phonetisation for the development of TTS synthesis and for applications. In particular we focus on language-general and dialect-specific aspects relevant to the modular two-stage multi-dialect approach developed here. Chapter 4 discusses the current orthographic conventions of Irish, reviews past developments in the standardisation of the written language, and introduces the reading rules of Irish. Chapter 5 presents the architecture of a modular multi-dialect phonetiser for Irish and its components: the letter-to-sound rules and the exceptions lexicons. A detailed description of the global and local rules is given in Chapter 6. Chapter 7 describes procedures for testing and evaluating the phonetiser and the results of this evaluation. Chapter 8 discusses some further potential applications, beyond TTS synthesis, of the Irish phonetiser. Chapter 9 concludes and offers some directions for future research. Additional information is given in a number of appendices, including actual rule files used in the implementation of the phonetiser, as well as other relevant information.
Acknowledgements

I would like to sincerely thank my principal supervisor, Prof. Ailbhe Ní Chasaide, for the considerable support, encouragement, expertise and inspiration she gave me throughout this process. During this thesis I received financial support from the European Union and Foras na Gaeilge for which I am very grateful.

I may not have completed this thesis were it not for the support I received from my co-supervisor Dr. Pauline Welby. Pauline’s passion for Irish, linguistics, and academic scholarship inspired me to persevere. I am sincerely grateful to her for the long hours she dedicated to ensuring that I completed this thesis.

I would also like to thank Harald Berthelsen, who acted as consultant advisor to my research. His expert contributions to both my research and the Irish synthesis projects are invaluable.

Many thanks to Dr. Caoilfhionn Nic Pháidín and Dr. Peadar Ó Flatharta at Fiontar, Dublin City University, for supporting my research goals. I would also like to thank everyone else at Fiontar, in particular Michal Boleslav Méchura for helping me to settle into my new role and also for his fantastic work on user-friendly Internet resources for the Irish language.

Thanks also to Dr. Christer Gobl, Amelia Kelly, Christoph Wendler, and my brother Dr. Eoghan Ó Raghallaigh, for their valuable feedback on earlier drafts.

I wish to thank my fellow doctoral students and researchers in the Phonetics and Speech Lab, Trinity College Dublin, for their friendship and wish them luck in their future careers: Martha Dalton, Amelie Dorn, John Duggan, Nadia Genserovskaya, Raya Kalaldeh, John Kane, Áine Ní Bhriain, Maria O’Reilly, Michelle Tooher, John Wogan, Irena Yanushevskaia, and Eric Zoerner. My thanks also to various staff members of the School of Linguistic, Speech and Communication Sciences, of the School of Computer Science and Statistics, and of Scoil na Gaeilge, Trinity College Dublin. I would also like to thank the members of the Laboratoire Parole et Langage (CNRS and Université de Provence), and the
researchers at Canolfan Bedwyr, University of Bangor, in particular Dr. Briony Williams. Thanks also to everyone who assisted me with examples and references.

I wish to thank the Irish Times and the Royal Irish Academy Library, for kind permission to use images from their collections. I also wish to thank Institiúid Teangeolaíochta Éireann and An Gúm in relation to the use of *Foclóir Póca*.

I would like to pay tribute to Prof. Dónall Ó Baoill and Prof. Máirtín Ó Murchú for their contributions to the field.

Gabhaim buíochas le Daid agus le hEoghainín as ucht mé a spreagadh i dtreo na Gaeilge.

It goes without saying that I owe a great debt of gratitude to my parents, Eoghan and Ríona Ó Raghallaigh, and to my family and friends. Daracha Nic Philibín and Sadhbh Nic Philibín-Ó Raghallaigh have endured the most during the completion of this thesis, and therefore deserve my sincere thanks and enduring appreciation.
Do m'inión, Sadhbh Nic Philibín-Ó Raghallaigh

Grianghraf: Eric Luke/The Irish Times
## Contents

1 General Introduction 1
   1.1 Research goals ................................................................. 1
   1.2 The Irish language today .................................................... 3
   1.3 Phonetisation ................................................................. 5
   1.4 A multi-dialect strategy for Irish TTS .................................. 6
   1.5 A modular approach to phonetisation .................................. 8
   1.6 Thesis overview ............................................................. 10

2 Irish text-to-speech (TTS) synthesis 11
   2.1 What is text-to-speech synthesis? ........................................ 11
   2.2 Text-to-speech synthesis software ...................................... 13
   2.3 Concatenative speech synthesis ......................................... 15
      2.3.1 Non-uniform unit selection synthesis ............................. 15
      2.3.2 Diphone synthesis ......................................................... 18
   2.4 The basic steps in TTS ...................................................... 19
      2.4.1 Text processing .......................................................... 19
      2.4.2 Synthesis ................................................................. 20
   2.5 Phonetisation ................................................................. 22
   2.6 Text-to-speech synthesis for Irish ....................................... 26
      2.6.1 Progress in Irish text-to-speech synthesis ....................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6.2 The need for multi-dialect synthesis for Irish</td>
<td>30</td>
</tr>
<tr>
<td>2.6.3 Applications of TTS for Irish</td>
<td>34</td>
</tr>
<tr>
<td><strong>3 Phonetisation and the sound system of Irish</strong></td>
<td>37</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>37</td>
</tr>
<tr>
<td>3.2 A two-stage approach to Irish phonetisation</td>
<td>38</td>
</tr>
<tr>
<td>3.3 Consonants</td>
<td>44</td>
</tr>
<tr>
<td>3.4 Vowels</td>
<td>50</td>
</tr>
<tr>
<td>3.5 Initial mutation</td>
<td>55</td>
</tr>
<tr>
<td>3.6 Lexical stress and vowel reduction</td>
<td>59</td>
</tr>
<tr>
<td>3.7 Devoiced sonorants</td>
<td>61</td>
</tr>
<tr>
<td><strong>4 Phonetisation and Irish spelling</strong></td>
<td>63</td>
</tr>
<tr>
<td>4.1 The text in Irish text-to-speech</td>
<td>63</td>
</tr>
<tr>
<td>4.2 Orthographic depth and orthographic distance</td>
<td>64</td>
</tr>
<tr>
<td>4.3 The Irish alphabet</td>
<td>67</td>
</tr>
<tr>
<td>4.4 The modernisation of Irish spelling</td>
<td>68</td>
</tr>
<tr>
<td>4.5 Phonetisation and the new standard</td>
<td>76</td>
</tr>
<tr>
<td>4.6 Reading rules for Irish</td>
<td>78</td>
</tr>
<tr>
<td>4.6.1 Consonantal graphemes</td>
<td>78</td>
</tr>
<tr>
<td>4.6.2 Vocalic graphemes</td>
<td>84</td>
</tr>
<tr>
<td><strong>5 A modular multi-dialect phonetiser for Irish</strong></td>
<td>91</td>
</tr>
<tr>
<td>5.1 A standard approach to phonetisation</td>
<td>91</td>
</tr>
<tr>
<td>5.2 Methods for developing LTS rules</td>
<td>95</td>
</tr>
<tr>
<td>5.3 Multi-dialect phonetisation</td>
<td>97</td>
</tr>
<tr>
<td>5.4 A phonetiser suited to Irish: a global-local modular approach</td>
<td>99</td>
</tr>
<tr>
<td>5.5 Abstract transcription vs. Láráchnúint</td>
<td>100</td>
</tr>
</tbody>
</table>
5.6 Global transcriptions ................................................................. 103
5.7 Multi-dialect phonetisation step-by-step .............................................. 104
5.8 The structure of the lexicon .............................................................. 106
  5.8.1 Global exceptions lexicons .......................................................... 108
  5.8.2 Dialect-specific local exceptions lexicons ...................................... 109
5.9 The letter-to-sound (LTS) rules ......................................................... 110
5.10 Interaction between the phonetisation and linguistic analysis com-
  ponents of TTS synthesis ................................................................. 111
  5.10.1 Morphological analysis and part-of-speech tagging ...................... 111
  5.10.2 Stress assignment in compound words and borrowings ............... 112
  5.10.3 Syllabification ............................................................................. 114

6 A detailed description of the global and local rules .............................. 117
  6.1 Introduction ..................................................................................... 117
  6.2 The structure of the phonetiser......................................................... 118
  6.3 Rule syntax and interpretation .......................................................... 118
    6.3.1 Regular expressions .................................................................... 119
  6.4 Linguist-friendly letter-to-sound rules .............................................. 121
  6.5 The global rule set ............................................................................ 122
    6.5.1 Phonemically velarised and phonemically palatalised con-
        sonants .......................................................................................... 124
    6.5.2 Initial mutations .......................................................................... 128
    6.5.3 Epenthesis ................................................................................... 129
    6.5.4 Future, conditional, and verbal adjective endings ......................... 129
    6.5.5 Rule ordering and grouping ......................................................... 130
  6.6 The local rule sets ............................................................................. 132
  6.7 Adding a new dialect ....................................................................... 133
6.8 Tracing the application of the phonetiser: a detailed example . . . 135

7 Testing and evaluating the phonetiser 140
7.1 Testing that the rules work together: internal rule compatibility . 140
7.2 Syntax of the tests ................................................................. 142
7.3 Evaluating the output ............................................................. 144
   7.3.1 A point of comparison: dictionaries vs. phonetised corpora 144
   7.3.2 Test dialect: the Lárchanúint ............................................ 146
   7.3.3 Hand-written rules versus automatically trained rules . . . 148
7.4 Improving the results ............................................................. 149

8 Irish phonetisation: applications beyond synthesis 151
8.1 Introduction ............................................................................. 151
8.2 Pronunciation guides ............................................................. 152
8.3 Theoretical and applied psycholinguistics .............................. 157
8.4 Conclusion .............................................................................. 159

9 Conclusion 160
9.1 Goals and results ................................................................. 160
9.2 Future work .......................................................................... 162

B Bibliography 165

A Irish phoneset comparison tables 181

B Letter-to-sound (LTS) rule files 185
   B.1 Global/lts.lff ........................................................................... 186
   B.2 Global/global-units.txt ......................................................... 212
   B.3 Global/lex-exceptions.txt ...................................................... 216
   B.4 Global/lex-function-words.txt .............................................. 217
### Contents

| B.5   | Global/lex-non-native.txt | 224 |
| B.6   | Global/test.txt           | 231 |
| B.7   | Larchanuint/lts.lff       | 232 |
| B.8   | Larchanuint/lex-exceptions.txt | 237 |
| B.9   | Larchanuint/test.txt      | 238 |
| B.10  | GaothDobhair/lts.lff      | 242 |
| B.11  | GaothDobhair/lex-exceptions.txt | 248 |
| B.12  | GaothDobhair/test.txt     | 260 |
| B.13  | Conamara/lts.lff          | 266 |
| B.14  | Conamara/lex-exceptions.txt | 272 |
| B.15  | Conamara/test.txt         | 274 |
| B.16  | CorcaDhuibhne/lts.lff     | 275 |
| B.17  | CorcaDhuibhne/lex-exceptions.txt | 282 |
| B.18  | CorcaDhuibhne/test.txt    | 286 |
| B.19  | Template/lts.lff          | 287 |
| B.20  | Template/lex-exceptions.txt | 292 |
| B.21  | Template/test.txt         | 293 |

C The Láircheannúint  

294
List of Figures


1.2 A schema of the global-local modular multi-dialect phonetiser. ............................................. 9

2.1 A visualisation of a sound file used in TTS synthesis with phone, syllable, and word tiers, for the recorded sentence *Ar maidin an lá arna mbárách bhí sé ar shiúl go hAlbain* ‘The following morning he was away to Scotland’ (Gaoth Dobhair corpus). ............................................. 17

2.2 Concatenative TTS system overview. Components of phonetiser are shaded in grey. ............................................. 21

2.3 17 of the 34 different forms of the Irish verb *bris* ‘to break’ generated automatically using a morphological analyser. Note that *Len* stands for *lenition* and *Ecl* stands for *nasalisation/eclipsis* (see Chapter 3). ............................................. 23


2.5 A screenshot of the phonetisation tool available on *abair.ie*, which uses the phonetiser described in this thesis. ............................................. 28
2.6 A map of Ireland, showing the Gaeltachtaí (Irish speaking districts). Gaeltacht regions to the north are in Ulster, those to the west are in Connaught, and those to the south are in Munster. [Image author: Angr. Wikimedia Commons.] ..................................... 31

3.1 The vowel phonemes of Gaocht Dobhair Irish (based on Ní Chasaide, 1999). .......................................................... 51

4.1 Modern Irish alphabet. Note that there are only 18 letters (5 of which can be accented). The letters j, k, q, v, w, x, y, z are not part of traditional alphabet, although they are used in loan words (e.g. jab ‘job’). ................................................................. 67

4.2 An excerpt from the Myles na gCopaleen novel An Béal Bocht published in 1941 in Cló Gaelach. .......................................................... 68

4.3 The Irish alphabet in a modern Cló Gaelach computer font. .... 68

4.4 An excerpt from Theobald Stapleton’s 1639 catechism in Latin and Irish. The Irish text was written using a simplified system of spelling and typeset in Roman script. (Reproduced with kind permission of the Royal Irish Academy Library.) ......................... 70

4.5 The poem Dómhnall Bán in classical spelling and in Letiriu Shimplí ‘Simplified Spelling’ (O’Leary & Bergin, 1911). ......................... 72

5.1 An schematic illustration of text being passed through a normaliser and a phonetiser in turn. ...................................................... 92

5.2 An excerpt from the Carnegie Mellon University (CMU) Pronouncing Dictionary for North American English (presented earlier as Figure 2.4). ................................................................. 94
5.3 A schematic representation of the difference between the (theoretic) coverage of the Lárcanúint artificial dialect (A), the phonetiser’s global module (B), and the phonetiser’s local module (C). In each diagram, the circles represent regional dialects of Irish. The shaded area in A represents how the Lárcanúint in theory relates to regional dialects. The shaded area in diagram B represents the way in which the global module covers all regional dialects, and the shaded area in C shows how local modules cover only features of the regional dialects.

5.4 Phonetisation step-by-step. Separate components that interact with the phonetiser are shaded in grey.

8.1 A screenshot of the online terminology database focal.ie.
Chapter 1

General Introduction

1.1 Research goals

Text-to-speech (TTS) synthesis, the transformation of ordinary text into artificial speech, plays a vital and increasingly important role in modern society. For example, people with visual impairments use screen readers integrating TTS synthesis, and so are able to read web pages, e-mail, text messages, electronic textbooks and other documents independently. People with vocal impairments use devices integrating TTS synthesis to communicate. Speech synthesis is also used to automate services such as directory enquiries, telephone banking, public announcements, and in car satellite navigation (GPS) systems.
Yet for speakers of endangered or less commonly used languages, these types of resources are often not available. For example, without Irish language communication devices or screen readers, speakers of Irish with visual or vocal disabilities may not be able to fully participate in their communities. The research described here is part of a larger project at Trinity College Dublin, one of the principal goals of which is to develop TTS synthesis upon which these types of adaptive technology for Irish can be built.¹

The Irish language today continues to be an integral (although complicated) part of Irish life, and TTS synthesis is and will continue to be part of the landscape of modern society. Our research helps bring together these two aspects of modern Irish society. This thesis focuses on developing an essential component of TTS synthesis, grapheme-to-phoneme conversion or phonetisation, conversion from ordinary spelling to pronunciation, that is, from letters to sounds. The thesis has several inter-related goals: to design a working phonetiser for multi-dialect Irish TTS synthesis, to defend the two-stage modular approach adopted, to argue for the soundness of the hand-written rule approach used, and to demonstrate the potential for phonetisation in applications beyond synthesis. Before turning to more technical aspects of the endeavour, I outline the current sociolinguistic context of the Irish language in Ireland.

¹The project to develop TTS synthesis and associated speech technology resources for Irish is ongoing at the Phonetics and Speech Lab, Trinity College Dublin (http://www.tcd.ie/slacs/clcs/phonetics/).
1.2 The Irish language today

Today, Irish is spoken as a community language in several remote regions of the western and southern seaboards of Ireland. These areas are referred to as the *Gaeltacht* and enjoy official status. According to the Irish Government Department of Community, Equality and Gaeltacht Affairs, “The key underlying objective of Government policy...is to maintain the primacy of the Irish language and its related culture in the Gaeltacht”.

Historically, Irish was spoken throughout the island of Ireland, but the language has been in decline since the Flight of the Earls 400 years ago, which saw the Gaelic aristocracy in Ireland replaced by an English-speaking elite. The language was dealt a second catastrophic blow by the Great Famine of the 1840s, which caused the loss of over a quarter of the population (or approximately 2.5 million people), mostly Irish speakers, to hunger, disease, and emigration (Ó Cuív, 1969). From this point on, Irish declined rapidly and was replaced by English almost throughout the country. There is no standard spoken variety of Irish. The areas in which the language is still spoken as a community language were largely cut off from one another, due to distance and difficulty of travelling, which allowed the emergence of three major dialects, Ulster, Connaught, and Munster (Ó Siadhail, 1981), corresponding to the provinces in which they are spoken (see Figure 1.1) and communities of speakers who are attached to their particular dialects (see § 2.6.2).

---

2Outside of Ireland, the language is often referred to as *Gaelic* (from the Irish word *Gaeilge*), but in Ireland, it is always called *Irish*. *Irish* is also the standard term used by linguists, lexicographers, and the European Union.

3See: [http://www.pobail.ie/ie/AnGhaeltacht/](http://www.pobail.ie/ie/AnGhaeltacht/)
Of a population of over 6 million people on the island of Ireland, fewer than 100,000 use Irish on a daily basis outside the school system. Even in the Gaeltacht, only a little more than a third of all people use Irish on a daily basis outside the school system. This number falls to just over a quarter for young people between the ages of 5 and 24 (Central Statistics Office, 2007).

In addition to the demand for TTS synthesis common to other languages (for example, adaptive technologies), the central place of the Irish language in the Irish education system and the fact that most speakers are not native speakers create certain challenges, some of which may be partially addressed by TTS synthesis. Irish is a compulsory subject for almost all pupils in the Republic of Ireland, and the language is also taught in many schools in the North of Ireland. Most Irish people study Irish for 13 or 14 years, though with generally poor results.
(see, for example, Harris & Murtagh, 1999). Many learners have little contact with native speakers and learners often report difficulty understanding unfamiliar dialects and native speakers in general. Another challenge for learners is the Irish spelling system, which is quite complex and very different from that of English, the native language of most learners (see also § 4.2). For example, the grapheme (th) in English is pronounced [θ] or [ð] (as in *thought* and *the*), while the Irish grapheme (th) is pronounced [h], as in *Áth Cliath* [aːh ˈcíːʃ] ‘Dublin’. TTS synthesis and its components (including phonetisation) can help to bridge the gap between written and spoken Irish and allow the development of applications in accessibility and in computer-assisted language learning (CALL) technologies.

1.3 Phonetisation

Phonetisation (or grapheme-to-phoneme conversion) is the process in which regular text is converted to phonetic transcription, an abstract representation of how the text might sound if spoken. These phonetic transcriptions resemble pronunciations often listed next to headwords in dictionaries. Different dictionaries use different systems, but some, including the Oxford English Dictionary use systems closely based on the International Phonetic Alphabet (IPA) (International Phonetic Association, 1999).

To convert regular Irish spelling to a sequence of sounds (a phonetic transcription), a system of conversion rules has to be developed. These rules have to map

---

4 The term *grapheme* is used here to mean all of the letter combinations (including a single letter) that represent a phone. Here and throughout, graphemes are enclosed in angle brackets. For typesetting reasons, angle brackets are often substituted with the binary operators ‘<’ (less than) and ‘>’ (greater than). Actual angle brackets, ‘(’ and ‘)’, will be used here. Orthographic letters are enclosed in single quotes.
letters to sounds (i.e. \((X) \rightarrow /Y/\), where \((X)\) is a grapheme consisting of one or more letters and \(/Y/\) is a consonant or a vowel sound). The spelling system of Irish (as we will see in Chapter 4) is quite complicated, with many-to-one letter-to-sound mappings, but is also largely regular, which we will argue makes the language well-suited to the development of hand-written letter-to-sound (LTS) rules. For example, considering the dedication at the beginning of this thesis, the grapheme \((adh)\) is pronounced \([\text{a}i]\) and the grapheme \((bh)\) is pronounced \([\text{v}^\text{y}]\) (as in the first name Sadhbh), the grapheme \((ph)\) is pronounced \([p]\) (as in the surname Nic Philibín), and the grapheme \((aigh)\) is pronounced \([i:I]\) (as in Ó Raghallaigh). For phonetisation, this graphemic complexity is unproblematic, since the grapheme-to-phoneme mapping is largely regular. This complexity is comparable to that of a language like French, which also has a quite regular orthography that contains many multi-letter graphemes.\(^5\) As we will see however, the orthographic distance between Irish and English, i.e. the differences between the two orthographic systems, very likely contributes to the difficulties many second language speakers and learners of Irish encounter in reading the language (see § 4.2).

1.4 A multi-dialect strategy for Irish TTS

Given the lack of a standard spoken variety of Irish, the necessity for a multi-dialect approach to TTS synthesis for Irish was recognised and planned for from

\(^5\)For example, the phoneme /o/ can be represented by many different graphemes, including: \((au)\) au ‘at the.MASC.SG’, \((aux)\) aux ‘at the.PL’, \((eau)\) eau ‘water’, \((eaux)\) eaux ‘waters’, \((ot)\) pot ‘pot’, \((ots)\) pots ‘pots’, \((ôt)\) impot ‘tax’, \((ôts)\) impôts ‘taxes’, \((op)\) sirop ‘syrup’, \((ops)\) sirops ‘syrups’, \((aut)\) saut ‘jump’, and \((auts)\) sauts ‘jumps’. The grapheme-to-phoneme mapping here is in fact more complex, since the final consonants are pronounced in certain liaison contexts (e.g., aux États-Unis ‘in the U.S.’). For more information on phonetisation for French, see Catach, 1984. The basic point remains the same.
the outset of the larger Trinity College Dublin synthesis project.

Dialect-specific pronunciation dictionaries are often used to develop phonetisers. For many of the languages for which synthesis systems have been so far developed, a dialect-specific pronunciation dictionary was already available for the standard variety. This can be used as a basis for the development of letter-to-sound (LTS) rules, using either statistical machine learning methodologies or hand-written rules. As a phonetiser involves both LTS rules and a pronunciation dictionary, an existing dialect-specific pronunciation dictionary may in itself also form part of the phonetiser.

But, as noted above, Irish has no standard spoken variety and such prerequisite resources were lacking for Irish. The Ó Dónaill (1977) Irish-English Dictionary, which constitutes a *de facto* extension of the official written standard, does not include pronunciations. No pronunciation dictionary exists for any of the dialects of Irish. The only current Irish dictionary with pronunciations of any kind is *Focloir Póca* ‘pocket dictionary’ (Ó Baoill, 1986a). *Focloir Póca* gives pronunciations in an artificial dialect known as the *Lárchanúint* (lár ‘centre/core’ + canúint ‘dialect’) (Ó Baoill, 1986b). While the Lárchanúint, which was designed to be equally relatable to the various regional varieties, is a valuable resource, its pronunciations are not dialect-specific and are therefore not suitable for our purposes.

Our approach provides a system that accounts for regional variation in the absence of an accepted spoken standard, but takes advantage of the underlying phonological similarities among the Irish dialects. This modular multi-dialect approach to phonetisation is introduced in the next sections.
On the face of it, it could be argued that the simplest possible approach to developing a phonetiser for Irish is to treat the dialects as separate languages and have different sets of letter-to-sound rules and a pronunciation dictionary (or lexicon) for each of the dialects for which a TTS voice is envisaged. However, given the resources (or lack of resources) available to us, and the desire not only to provide for the short-term development of TTS for one dialect of Irish, but also to lay a foundation for eventual multi-dialect synthesis, a multi-dialect strategy was adopted for the development of the phonetiser as well. Similar types of approaches have been successfully used for synthesis in other languages (see § 5.1) and it is in line with the philosophy of TTS synthesis that emphasises the development of components or modules that can be used and re-used for different dialects or languages. The project to develop TTS for Irish started with the development of a synthetic female voice for the dialect of Gaoth Dobhair, Co. Donegal (Ulster). Our approach has recently been applied to build a Conamara voice (Co. Galway, Connaught) and phonetisation developed here will permit the development of a Corca Dhuibhne voice (Co. Kerry, Munster). For a discussion on the choice of these dialects, see § 2.6.2.

1.5 A modular approach to phonetisation

To achieve the goal of multi-dialect phonetisation for Irish, a set of global rules, which apply to all dialects, form the core of the phonetisation system. These global rules take text as input and convert it to an intermediate inter-dialectal linguistic representation which is passed to a set of local rules that convert the

---

6Note that we use the Irish spelling Conamara, a region that includes both Gaeltacht and non-Gaeltacht (Gàilltacht) areas. This use parallels the Irish-only placenames of other Gaeltacht areas.
input they receive to dialectal transcriptions. A set of local rules is required for each dialect. However, since much of the work is done by the global module, adding a new dialect should be relatively straightforward. In this way, the various dialectal rule sets should only require minor dialect-specific modifications. A schematic representation of the modular multi-dialect phonetiser is shown in Figure 1.2.

```
Input text
  ↓
GLOBAL MODULE
  ↓
Intermediate inter-dialectal linguistic representation
  ↓
LOCAL MODULE
  ↓
Dialect-specific phonetic transcription
```

Figure 1.2: A schema of the global-local modular multi-dialect phonetiser.

This approach also allows us to exploit past linguistic research on dialect variation, and also to exploit as far as possible the ideas and content of the Lárchamúint and *Foclóir Póca*. At a phonological level, and with respect to the correspondences between orthography and pronunciation, all Irish dialects are structurally very similar. As unit selection synthesis, the technique we are currently using in the development of Irish TTS voices (see Chapter 2), requires only broad transcription of a target sound, information about surface realisation does not need to be specified.
1.6 Thesis overview

Following this introductory chapter, Chapter 2 provides background information on text-to-speech (TTS) synthesis and locates the phonetisation component within the larger enterprise. Chapter 3 discusses the sound system of Irish in the context of phonetisation, addressing some similarities across the dialects chosen to represent the three main dialect groups. Chapter 4 discusses modern standard Irish orthography and reviews past developments in the standardisation of the written language. Chapter 5 presents the architecture of a modular multi-dialect phonetiser for Irish and its components: the letter-to-sound (LTS) rules and the exceptions lexicons. The phonetiser is modular in that it contains a global module containing LTS rules and an exceptions lexicon for all dialects of Irish, as well as local modules containing dialect-specific rules and lexicons. A detailed description of the global and local rules is given in Chapter 6. Chapter 7 describes procedures for testing and evaluating the phonetiser and the results of this evaluation. Chapter 8 discusses some further potential applications, beyond TTS synthesis, of the Irish phonetiser. Chapter 9 concludes and offers some directions for future research. Additional information is given in a number of appendices, including actual rule files used in the implementation of the phonetiser, as well as other relevant information.
Chapter 2

Irish text-to-speech (TTS) synthesis

2.1 What is text-to-speech synthesis?

Speech synthesis is where synthetic speech is generated by computer software and hardware systems. These systems use a variety of techniques that have developed since research and development in the area of speech synthesis began in the 1950s. A text-to-speech system for speech synthesis takes ordinary text as input and produces synthetic speech as output.

Some people may think of synthetic speech as being computerised speech and therefore likely to be robotic sounding. However, current speech synthesis output can sometimes be mistaken for real human speech. The main reason for this is that the most popular method for producing synthetic speech is based on recordings of real human speech. In this method, concatenative speech synthesis, pieces
Irish text-to-speech (TTS) synthesis

of recorded speech are concatenated, or pieced together, to produce synthetic speech output. With advancements to date in recording and concatenation techniques, synthetic speech produced in this way can be almost indistinguishable from real speech.

The development of concatenative speech synthesis has been driven in part by the need to automate certain services. Automation of services such as telephone banking and directory enquiries, as well as public announcements such as those in train stations that inform passengers of delays to scheduled services can result in a reduction in costs. This can be achieved by replacing human workers with synthesisers during the unsocial and therefore costly hours during which such round-the-clock services have to operate. The need for synthesis in computer applications that aid the visually and vocally disabled (screen readers, document readers, voice output communication aids, etc.), as well as people with learning disabilities such as dyslexia, has also in part driven the development of speech synthesis.

While concatenative speech synthesis is at present the most effective way of producing intelligible natural sounding speech output, it is not an ideal solution. Speech databases are expensive to produce and are very restricted in the type of output they can produce. Expressive speech synthesis is an example of the type of synthesis that is difficult to produce using corpus-based synthesis, although work is being carried out in this area (Eide et al., 2004). The ultimate goal of research in speech synthesis technology is to develop speech synthesis where every aspect of the voice being produced can be controlled to produce both expressive and emotional synthetic speech (Schröder, 2001). However, it is generally accepted that concatenative synthesis techniques are restricted in ways that will never
allow this type of synthesis (Sproat, 1998).

The alternative to concatenative synthesis is parametric synthesis. Parametric synthesis encompasses a variety of existing techniques such as formant synthesis and articulatory synthesis (Taylor, 2009). The basic idea is that linguistic knowledge about how a speech signal is produced at the voice source (vocal cords) and modified by the filter (vocal tract and lips, and nasal cavity) should allow us to generate truly synthetic speech, indistinguishable from that produced by a human, using signal processing techniques.

In our work on Irish we use concatenative speech synthesis, but regardless of how the speech output is generated, the applications of speech synthesis remain the same.

2.2 Text-to-speech synthesis software

Most modern computer operating systems ship with text-to-speech (TTS) synthesis software pre-installed. The pre-installed system voices provided can be used directly or accessed via other software such as a screen reader which will pass text sentences from the screen to the system's TTS engine.

For example, as of this writing, the Microsoft Windows XP operating system has one system voice installed by default called Microsoft Sam. To try out this voice directly, select "Speech" from the "Sounds, Speech, and Audio Devices" menu in the Control Panel. If screen reading software such as "Speak It", the optional add-on for the Firefox web browser has been installed, text can be

\footnote{http://www.mozilla.com/en-US/firefox/} Add-ons currently available, may not be available in the future, or may not be compatible with
selected within the browser’s window and read aloud using the Microsoft Sam voice by typing Ctrl+Alt+V (i.e., these three keys at the same time). An entire web page can be read aloud by typing Ctrl+Alt+B and so on.

On other platforms such as Linux, developmental concatenative TTS synthesis systems exist where developers can build new voices, by recording new speech databases. As synthetic voices are language-specific, voices for new languages can also be added if additional language-specific text processing and linguistic information is provided.

The eSpeak\(^8\) speech synthesiser is an example of a TTS system that allows adding new languages without recording anything in that language, and it can be run on both Linux and Windows. At the time of writing, there was no Irish voice available with eSpeak.

The system most commonly used to develop new voices for regional and minority languages is the Festival Speech Synthesis System (Clark et al., 2004), which was developed at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh.\(^9\) The Festival system is free and provides a number of different speech engines for developing new voices using concatenative techniques. One such technique is known as unit selection synthesis, a type of concatenative speech synthesis. This technique is the most popular technique in use today as it generates the most natural sounding speech output.

---

future versions of Firefox.

\(^8\)http://espeak.sourceforge.net/

\(^9\)http://www.cstr.ed.ac.uk/
2.3 Concatenative speech synthesis

Concatenative synthesis creates speech output by concatenating pieces of recorded speech from a speech database. The two primary variations on concatenative synthesis are non-uniform unit selection synthesis and diphone synthesis. As the work described here was undertaken as part of a project to develop concatenative speech synthesisers for Irish, I will elaborate on these two types of synthesis in the following sections.

2.3.1 Non-uniform unit selection synthesis

A non-uniform unit selection\(^{10}\) synthesiser uses a very large database of recorded speech, typically of the order of many hours of speech. Creating a speech database for use in a unit selection synthesiser involves a number of steps. Firstly, a speech corpus of sufficiently high quality must be recorded. This corpus must then be annotated. Annotation involves marking the boundaries of sentences, words, syllables and segments (phones). This process can be automated using a technique known as "forced alignment". Forced alignment is similar to automatic speech recognition (ASR) where Hidden Markov Models (HMMs) are used to model the characteristics of individual phones in a language. In forced alignment, however, both the text, in the form of a text file containing the phone string\(^{11}\)

---

\(^{10}\) Non-uniform unit selection will be abbreviated to unit selection throughout the thesis.

\(^{11}\) Researchers use different terms for describing this input string, with some using phonemic transcription. I use the term phoneme in the sense of "the smallest unit of phonological structure that can signal a difference in meaning". The phonemic status of a sound is therefore determined by its participation in minimal pairs within a given variety of a language. I choose the more neutral term phone for two reasons. First, as Thatham & Morton (2005, p. 94) note, "...labeling an acoustic database with phoneme symbols makes nonsense of attempts at alignment: to align a representation this abstract with the actual acoustic signal is not a sound practice scientifically". Second, the phonological status of some of the labelled phones in our Irish corpora is not clear (for example, the devoiced sonorants described in § 3.7) (Ni Chasaide,
of the sentences recorded, and the sound files themselves are fed to the aligner. The process is therefore that of alignment rather than that of recognition. Forced alignment can be done using various software toolkits such as HTK (Odell et al., 1995) and eHMM (Black & Lenzo, 2007). The database is then indexed according to the phone labelling. Attributes of the segments such as fundamental frequency, duration, and context (segments before and after) are also taken into account during indexing. It is important to take segmental context into account because of the way in which coarticulation affects the way segments are produced. Target utterances are then generated by selecting the best path through a matrix of candidate units of varying lengths (hence non-uniform) and concatenating the units selected.

A unit selection system can produce excellent quality, natural sounding synthesis, and unit selection is currently the state of the art in synthesis. Two of the most significant disadvantages to unit selection synthesis, synthesis, are the large size of the database (typically many hours of speech must be read and recorded) and the slow speed of the synthesis due to the large size of the database that has to be searched during synthesis. There is also an inconsistency in terms of the quality of the output with such a large database as automatic techniques used to annotate speech corpora tend not to be 100% accurate. Units are concatenated and the joins (seams) between them will be more or less audible or noticeable depending on factors including the accuracy of the annotation and acoustic differences between adjacent units. Another challenge in developing a unit selection speech database is ensuring that there is full coverage of diphones in various contexts (such as word-initial, word-final, phrase-initial, phrase-final) in the database (Clark et al., 2007). Figure 2.1 shows an example of the annotations that accompany speech
waveforms in a typical speech corpus.

Figure 2.1: A visualisation of a sound file used in TTS synthesis with phone, syllable, and word tiers, for the recorded sentence *Ar maidin an lá arna mháirach bhí sé ar shiúl go hAlbain* 'The following morning he was away to Scotland' (Gaoth Dobhair corpus).
2.3.2 Diphone synthesis

Diphone synthesis relies upon a corpus of recorded speech, but a much smaller one. The diphone corpus contains all possible diphones that might occur in the language. A diphone is a unit of sound that begins halfway through one sound (a consonant or a vowel) and ends halfway through another sound segment, capturing the transition between the two sounds. A diphone corpus should have at least one occurrence of each diphone. The size of the corpus depends on the number of phonemes in the target language variety as well as the rules governing how phonemes can be combined. For example, an ⟨ng⟩ [ŋ] sound can occur word initially in Irish as in *i nGaillimh* ‘in Galway’, but not in English. A language with a large phoneme inventory such as Irish (see Chapter 3) would have diphones numbering in their thousands, but nonetheless, the corpus will still be considerably more compact than a unit selection corpus.

A diphone system can produce good quality synthesis but will not sound as natural as unit selection synthesis. Diphone synthesis is however still preferable to unit selection synthesis in certain applications as its output tends to be more predictable. Another significant advantage diphone synthesis has over unit selection synthesis is speed. Fast generation of speech output is essential when synthesis is used in screen reading applications for the visually impaired or in speech devices for the vocally impaired. Such applications are used to perform tasks such as quickly scanning e-mail and therefore screen readers often “talk” at speeds very much above that of normal speaking rate. Changing the speech rate in this way is possible with diphone synthesis. Speed is also important for the vocally impaired users, so that they can effectively participate in conversation.

The output of diphone synthesis can also be manipulated at runtime using
digital signal processing techniques. This facilitates the superimposition of an appropriate prosody (including intonation and relative segment durations). So for example, a user of a voice output communication aid that uses diphone synthesis might be able to instruct the device to output a question-type intonation if required. Yet, unless carefully implemented, the use of digital signal processing may degrade the signal leading to speech output that sounds less natural.

### 2.4 The basic steps in TTS

To illustrate the basic steps involved in unit selection text-to-speech (TTS) synthesis, I will describe the process in terms of the Festival speech synthesis system discussed earlier in § 2.2. As described in § 2.6, Festival was used in the development of the Irish TTS system to which the work described in this thesis has contributed. A more detailed account of Festival is available in the Festival manual (Black et al., 2002).

#### 2.4.1 Text processing

Unit selection TTS synthesis with Festival involves a number of basic steps. The first step involves taking ordinary text as input and converting it to a linguistic representation similar to a pronunciation one might find in a dictionary. This process is known as *text processing*. The text processing module first separates the input into words and tokens. Tokens are items such as ordinal numbers, cardinal numbers, time and date expressions, acronyms and abbreviations. The next stage involves the conversion of such tokens into words. For example, in the phrase *3 dogs*, the digit *3* will be converted to the word *three*. This process is known
as normalisation (or sometimes tokenisation). The normalisation module uses a set of language-specific token-to-word rules. Festival incorporates a scripting language called Scheme that allows developers to write such rules.

After the text is normalised, each word is then looked up in a lexicon (pronunciation dictionary). The lexicon contains grammatical information (part-of-speech) and a pronunciation for each entry. Transcriptions also contain information about syllable structure and word stress. The lexicon is described in more detail in §2.5. If a word is found in the lexicon, its transcription is sent for synthesis. If a word is not found in the lexicon, as will often be the case with words such as proper nouns, novel compounds, inflected forms or newly coined words, Festival runs a set of letter-to-sound (LTS) rules over the word to produce a transcription. This transcription is then sent for synthesis. The process of converting orthographic words into an abstract linguistic representation using the lexicon and LTS rules will be referred to here as phonetisation (grapheme-to-phoneme conversion). The output from the phonetiser constitutes what is searched for in the corpus, marked-up so that appropriate audio segments can be located and concatenated to produce synthetic speech output. Phonetisation is discussed in more detail in §2.5.

2.4.2 Synthesis

A speech engine, or synthesiser, takes the phone sequence that the text processing module outputs and produces synthetic speech that matches it. It does so by concatenating sound segments from a linguistically annotated speech corpus. The unit selection speech engine in Festival is called Multisyn. As a detailed description of the synthesis engine is beyond the scope of this thesis, I refer readers to Clark et al. (2007) for a more detailed description of this part of the process.
of TTS.

A detailed schematic illustration of the two basic steps in TTS, text processing, and synthesis, are illustrated in detail in Figure 2.2, with the steps pertaining to phonetisation highlighted in grey. For the rest of this thesis, the focus will shift to phonetisation.

```
INPUT
   Text input

TEXT PROCESSING
   Normalisation
       Turns numbers and symbols into regular words
   Syntactic and morphological analysis
       Analyses part-of-speech information and word structure
   Lexicon look-up
       Gets pronunciations for words that are listed in the lexicon
   Letter-to-sound rules
       Creates pronunciations for words not in the lexicon

ABSTRACT LINGUISTIC REPRESENTATION
   Pronunciation of the input text

SYNTHESIS
   Unit selection
       Gets units from the recorded corpus that match phone strings in the text
   Concatenation
       Puts the selected units together to create a sound file

OUTPUT
   Synthetic speech to match the text input
```

Figure 2.2: Concatenative TTS system overview. Components of phonetiser are shaded in grey.
2.5 Phonetisation

Phonetisation (grapheme-to-phoneme conversion) of normalised text, can be done in two ways. The first and most straightforward way is to use a pronunciation dictionary, referred to as the lexicon.

The lexicon is a vital part of most TTS systems. If the number of words in a language were limited, all the words could be stored in the lexicon with their corresponding pronunciations. In practice this is not possible as content words do not form a closed class, and languages allow new content words to be created on the fly from scratch or by concatenating existing words and morphemes in novel ways. For example, Béarlíóireacht ‘(act of) speaking English (Béarláir ‘English speaker’ + -eacht)’ or Gaeilgeoireacht ‘(act of) speaking Irish (Gaeilgeoir ‘Irish speaker’ + -eacht)’. Other examples of this are commonplace in well-known company and brand names, for example, Symantec. Such words would cause a TTS system that relied wholly on a lexicon for phnetisation to fail, since these novel words would not be found.

Letter-to-sound (LTS) rules, are thus always required, even if the phnetiser includes a comprehensive lexicon.\footnote{This term letter-to-sound rules, the standard term in the TTS literature, should not be interpreted to mean that there is a one-to-one mapping between letters and sounds.} Rule-based systems should in fact be able to handle any word they are given, including newly coined words, even without a lexicon. Rule sets can, however, become very complex if required to handle spelling exceptions. For this reason, phnetisers usually consist of a lexicon and a system of LTS rules that work in conjunction with each other.

In addition to this, some languages such as Irish are more heavily inflected than others, and storing all inflected forms of Irish words in a lexicon would
entail significant redundancy. Consider for example, the list in Figure 2.3, which shows some of the many forms of the regular verb *bris* ‘to break’.

\[
\begin{align*}
\text{bris+Verb+VTI+PresInd} & \quad \text{brisann} \\
\text{bris+Verb+VTI+PresInd+1P+Pl} & \quad \text{brisimid} \\
\text{bris+Verb+VTI+PastInd+Len} & \quad \text{bhrís} \\
\text{bris+Verb+VTI+PastInd+Auto} & \quad \text{briseadh} \\
\text{bris+Verb+VTI+FutInd+1P+Pl} & \quad \text{brisfimid} \\
\text{bris+Verb+VTI+Cond+Len} & \quad \text{bhrísfeadh} \\
\text{bris+Verb+VTI+Cond} & \quad \text{brísfeadh} \\
\text{bris+Verb+VTI+Cond+2P+Sg+Len} & \quad \text{bhrísfí} \\
\text{bris+Verb+VTI+Cond+3P+Pl+Len} & \quad \text{bhrísidís} \\
\text{bris+Verb+VTI+PastImp+Len} & \quad \text{bhríséadh} \\
\text{bris+Verb+VTI+PastImp+2P+Sg+Len} & \quad \text{bhristí} \\
\text{bris+Verb+VTI+PastImp+3P+Pl+Len} & \quad \text{bhrísidís} \\
\text{bris+Verb+VTI+Imper+1P+Sg} & \quad \text{brísim} \\
\text{bris+Verb+VTI+Imper+3P+Sg} & \quad \text{briseadh} \\
\text{bris+Verb+VTI+Imper+2P+Pl} & \quad \text{brisigí} \\
\text{bris+Verb+VTI+Imper+Auto} & \quad \text{bristear} \\
\text{bris+Verb+VTI+PresSubj+1P+Pl+Ecl} & \quad \text{mbrisimid}
\end{align*}
\]

Figure 2.3: 17 of the 34 different forms of the Irish verb *bris* ‘to break’ generated automatically using a morphological analyser. Note that *Len* stands for *lenition* and *Ecl* stands for *nasalisation/eclipsis* (see Chapter 3).

To avoid this type of redundancy in the lexicon, phonetisation could be combined with morphological analysis. In this case, only root words and their pronunciations would be stored in the lexicon. However, since such redundancy is not a problem in terms of performance (indeed searching a lexicon is faster than performing morphological analysis) it is perfectly acceptable to include all inflected forms in the lexicon. The argument against such an approach is that the lexicon can be difficult and costly to maintain. It is on this basis an approach to phonetisation which minimises use of the lexicon is preferred in the work presented here.

---

\(^{13}\)This list was generated automatically using an Irish morphological analyser available [https://www.cs.tcd.ie/Elaine.UiDhonnchadh/irish.htm](https://www.cs.tcd.ie/Elaine.UiDhonnchadh/irish.htm) and described in Ui Dhonnchadh (2003)
on Irish phonetisation, and the future addition of a morphological analysis component to the Irish TTS system is assumed (Úi Dhonnchadha, 2003). In addition, the lexicon is used only for exceptions, and exceptions common to all dialects are stored only once. This approach is described in Chapter 5.

Information about lexical stress and syllable structure may be needed by a phonetiser to perform accurate conversions and thus should be marked, if possible, prior to the LTS rule application. In the case of Irish, vowel reduction (weakening) occurs in unstressed syllables containing short vowels. For example, in a simple word such as *seanchas* ‘folklore’ [ˈcaɲəxəs], the final vowel is a reduction of the short vowel /a/, as only the first syllable in the word is stressed. In the compound word *seanfhear* (sean ‘old’ + fear ‘man’) [ˈcaɲərfəɾ], however, the /a/ in the second syllable is not reduced, since in Irish, words prefixed with non-intensifying prefixes (such as sean-) take secondary stress in the resulting compound. For LTS rules to predict the correct vowel quality in these cases, stress information would have to be available and taken into account by the LTS rules. In the case of the Irish phonetisation module described here, additional linguistic rules are used to syllabify words and mark stress before the application of the LTS rules, however these rules are not central to this thesis. Transcriptions in the lexicon should also be syllabified with stress marked. Stress rules in Irish are described in more detail in Section 3.6.
Figure 2.4: An excerpt from the Carnegie Mellon University (CMU) Pronouncing Dictionary for North American English.

Phonetic transcriptions of words in a TTS lexicon usually use a machine-readable version of the International Phonetic Alphabet (IPA) such as X-SAMPA\(^{14}\) since non-ASCII (ASCII = 7-bit plain text) characters and diacritics (e.g. \(\ddot{a}, \ddot{e}, \ddot{f}, \ddot{y}\), etc.), may not be compatible with synthesis software or may not display correctly. Eventually, if Unicode is fully adopted and incorporated into all computer systems (Méchura, 2007), actual IPA transcriptions encoded in Unicode may be used in lexicons. Figure 2.4 shows an excerpt from the CMU American English lexicon. This lexicon is in a tab-separated format where each headword and its pronunciation are separated by a tab space, with one entry per line. The transcriptions use a machine-readable American English phoneset.\(^{15}\) Lexical stress is

\(^{14}\text{http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm}\)
\(^{15}\text{http://www.speech.cs.cmu.edu/cgi-bin/cmudict}\)
indicated by either a 0, 1 or 2, where 0 stands for no stress, 1 for primary stress, and 2 for secondary stress.

The phonetiser for Irish described in this thesis is comprised of lexicons and letter-to-sound rules and is described in detail in Chapters 5 and 6.

2.6 Text-to-speech synthesis for Irish

2.6.1 Progress in Irish text-to-speech synthesis

The work being described here has been undertaken within projects carried out at the Phonetics and Speech Laboratory, Trinity College Dublin (TCD), to develop speech technology resources and text-to-speech (TTS) synthesis for Irish dialects. The WISPR (Welsh and Irish Speech Processing Resources) project involved a Welsh/Irish collaboration with European Union funding, between Bangor University, Wales, and TCD, with collaboration from University College Dublin, Dublin City University, and Instituitid Teangeolafochta Éireann (‘The Linguistics Institute of Ireland’). The aim was to develop the prerequisite resources required for the development of TTS synthesis for Irish and Welsh (Ní Chasaide et al., 2006).

The work on Welsh built upon previous research (Williams, 1994a). The most significant product of the Irish side of the WISPR project was a large annotated speech corpus of the Gaoth Dobhair dialect of Irish. Gaoth Dobhair is a Gaeltacht region in northwest Co. Donegal, in Ulster.

16http://www.tcd.ie/slscs/clcs/phonetics/
Further research and development is ongoing in the Phonetics and Speech Laboratory, Trinity College Dublin, and is being funded by Foras na Gaeilge, a North/South body that funds initiatives related to the Irish language. In an initial project, the goal was to take the resources relating to Irish produced under the WISPR project and develop a full unit selection (see § 2.3.1) speech synthesiser for Donegal Irish that included a dialect-specific normaliser and phonetiser (an early version of the phonetiser described here). This has been completed and has resulted in the first Irish unit selection voices. The first artificial voice speaks the Gaoth Dobhair variety of Irish (Co. Donegal, Ulster variety) and is based on a speech corpus recorded and annotated during the WISPR project. The goal of the current project, Cabógaí II (‘chatterboxes’), is to extend TTS synthesis for further dialects of Irish using various synthesis techniques. To date, the system has been extended to include a unit selection voice for Conamara Irish (Co. Galway, Connaught variety), and a HTS voice (Masuko et al., 1996; Yoshimura et al., 1999) for Conamara Irish. These voices use the phonetiser described in this thesis.

The Gaoth Dobhair Irish unit selection synthesiser has been made publicly available through a web application called abair.ie (abair ‘say.imp.sg’).\(^{17}\) It was launched to facilitate online use of the synthetic voices produced as part of the project, and to collect feedback on the project. The text normalisation and phonetisation tools, which were developed as text processing modules for the synthesisers can also be used through the web site. The current architecture involves an entirely new text processing engine written in the Python programming language (Ní Chasaide et al., 2006). The rationale behind this development was to simplify and streamline the development of new voices for further dialects of Irish, and to provide a framework within which research, such as this, could

\(^{17}\)http://www.abair.ie/
Irish text-to-speech (TTS) synthesis

be undertaken, tested, and evaluated. The phonetiser (lexicon and LTS rules), which I developed and describe here, plugs into this new text processing engine and thus constitutes an integral part of the abair.ie TTS system. The speech synthesis engines currently being used include Festival Multisyn (Clark et al., 2007), and the HTS HMM-based speech synthesis system (Masuko et al., 1996; Yoshimura et al., 1999).

Figure 2.5: A screenshot of the phonetisation tool available on abair.ie, which uses the phonetiser described in this thesis.

Prior to the series of projects carried out at Trinity College Dublin, some research into speech synthesis for Irish had been carried out elsewhere. Two separate studies on Irish diphone synthesis were carried out by Charonnat et al.
29

Irish text-to-speech (TTS) synthesis

(1998) and Hogan et al. (2003); Robertson & Timoney (2003). Charonnat et al. (1998) developed an Irish diphone synthesiser, and Hogan et al. (2003) identified steps needed to develop an Irish diphone synthesiser using the Festival speech synthesis system (see also Clark et al., 2004). These preliminary investigations highlighted the absence of any existing TTS synthesiser for Irish, as well as the lack of prerequisite resources for Irish TTS, such as letter-to-sound rules and pronunciation dictionaries. A diphone synthesiser for Scottish Gaelic, a sister language of Irish, was developed by Wolters (1997).

As noted in this early research, there was a need for TTS for Irish, but a lack of prerequisite resources necessary for developing a TTS system for the language. Another challenging aspect of the development of TTS speech synthesis for Irish is the lack of a standard spoken variety and the existence instead of three major dialects of roughly equal status that differ considerably. Our approach has been to address both of these issues. In addition, instead of the diphone synthesis envisaged by earlier work on TTS for Irish, we chose to develop unit selection TTS systems, which is the state of the art in terms of producing natural-sounding speech (although the technique, of course, has other limitations). The current stage of development, namely the development of voices for further dialects of Irish is benefiting from the multi-dialect strategy adopted from the outset during the WISPR project. Most recently, work done for the Gaoth Dobhair voice (including work on the phonetiser) has been successfully reused for the development of a Conamara voice.
2.6.2 The need for multi-dialect synthesis for Irish

Unlike many languages for which TTS has been developed, Irish has no standard spoken variety. It is generally agreed that Irish has “three main dialect groups” (Wagner, 1958, p. IX), corresponding to the provinces in which they are spoken: Ulster (U), Connaught (C), and Munster (M). These dialect groups are of roughly equal status with no dialect being dominant on economic, political or cultural grounds. They are geographically (see Figure 2.6) fragmented and separated by English-speaking districts, thus making them easier to define along geographical lines (Wigger, 1979). Ulster currently includes dialects in Co. Donegal, Connaught includes dialects in the counties of Mayo and Galway, and Munster includes dialects in the counties of Kerry, Cork and Waterford.

\[^{18}\]O’Rahilly (1932/1972) posits only two dialects, a northern grouping and a southern grouping, but this analysis is not generally accepted. Most importantly, it would not be acceptable to contemporary native speakers. Moreover, Williams (1994b, p. 471) explicitly questions O’Rahilly’s analysis, stating “Is ar éigean is fior don Rathileach é nuair a deir sé nach raibh ach dhá mhórlimistéar canína Gaeilge in Éirinn. Is fearr a réiteodh sé leis an bhfianaisé dá n-áiteofaí gur trí mhórchamhúint a bhí sa tír.” ‘O’Rahilly is unlikely to be correct when he posits only two major Irish language dialect regions in Ireland. It would better reflect the evidence to posit that there were three major dialects in the country.’

\[^{19}\]There is a Gaeltacht in the province of Leinster, Ráth Cairn, that was artificially created in 1935 by a government programme to transplant Irish speakers from impoverished rural areas to more fertile lands in Co. Meath. Although the people came from a variety of Gaeltachtaí, the Irish spoken there today is most similar to Conamara (Galway) Irish (Ó Conghaile, 1986).
The Ulster, Connaught, and Munster varieties of Irish differ markedly along many dimensions, including syntax, lexicon and phonology. The richness and complexity of these groupings can be appreciated through a review of the dialect maps compiled by Wagner (1958, 1964, 1966, 1969) in his fieldwork data gathered through dialect questionnaires administered to Irish-speaking informants throughout the island of Ireland.²⁰ Taken together, the pattern of differences in

²⁰Some scholars have criticised the Linguistic Atlas, since many of Wagner’s informants were people who did not speak Irish fluently (see Wagner’s own detailed description of his informants Wagner, 1958, pp. XI-XXI): “some of them were not even very good Irish speakers” (Wagner, 1969, p. VIII), or who did not use Irish regularly: “most of our subjects have not spoken the language for a long time” (Wagner, 1958, p. X). Wagner himself acknowledged these shortcomings, but stressed the need to act to collect linguistic data while the language still survived, even though he was “dealing with...the ruins of a language” (Wagner, 1958, p. X). He writes: “[p]eople who consider our work as a rather hasty affair do not know or do not want to know that collecting native Irish speech is a matter of greatest urgency" (Wagner, 1969, p. VIII). That the differences among the three main dialect groups emerge despite the imperfect nature of the data speaks to the robustness of these dialect groupings.
the hundreds of questionnaire items supports the traditional grouping into three main dialect groups. For example, along the morphosyntactic dimension, the dialects differ in terms of the prevalence of synthetic verb forms, where person is marked on the verb in the synthetic forms of Munster dialects (*do dhiolás 'I sold*'), but indicated with a pronoun (*mé 'I*) in the Ulster and Connaught dialects (*dhiol mé*) (Wagner, 1958, p. 1). The future is marked differently in Ulster than in Connaught and Munster (*ceanglóidh 'tie.FUT' [caŋɡl̂ɔʰiː] vs. [caŋɡl̂ɔːɾ]) (Wagner, 1958, p. 15). Lexical differences include common words such as *scairteadh*21 (U) and *glaoch* (C, M) 'calling' (Wagner, 1958, p. 17), and *tinn* (U, C) and *breoite* (M). One phonological difference between the dialects is the effect of stress on vowel reduction. For example, in Munster Irish, a word such as *cosán* 'footpath' has the prosodic pattern V-V:, with stress on the second syllable and a long vowel in the second syllable. In Connaught Irish, the prosodic pattern is 'V-V:', with stress on the first syllable and a long vowel in the second syllable. In Ulster Irish, the prosodic pattern is 'V-V, with stress on the first syllable and a short vowel in the second syllable.

Of course, as the examples above show, there is not always a three-way pattern corresponding to the three provinces. Individually, these patterns may apply to more than one dialect (e.g. the lexical item *glaó* is used in both Connaught and Munster dialects) or may not follow province boundaries (e.g. for 'helping' *cúiditú* is found in Ulster and northern Connaught, *cúnamh* in southern Connaught and much of Munster, with *cabhrú* also found in Munster). As this example shows, and as with any language, there is often variability within any given dialect region for any given item. For example, most Munster speakers, but not all, use the synthetic form of the verb 'sold'. The term *eallach* is used for 'cattle'.

---

21Here and throughout, standard spelling is used in place of dialectal spellings found in the Wagner’s Linguistic Atlas.
in Ulster, *beithigh* is used in Munster, but both terms are found in Connaught (Wagner, 1958, p. 2). Some areas, particularly border areas, show a mix of patterns associated. A particular case is that of Achill Island in Co. Mayo, which contained certain Ulster features due to migration of Ulster speakers to the island. As Wagner writes “it is well known that the dialect of Achill and Ballycroy, the latter being represented by our Inishhiggle material, shows a most interesting mixture of Ulster and Connaught Irish” (Wagner, 1958, p. XVIII). When the totality of the dialect pattern evidence is considered as a whole, however, three main dialect groups (Ulster, Munster, Connaught) emerge.

A dialect or variety of a language is defined not only by geographical factors but also by social factors such as socio-economic class, age, and gender (Chambers, 1995). Our goal in phonetisation for TTS synthesis is not to reflect the full range of linguistic complexity present in a given speech community, but rather to produce a synthetic voice that includes salient aspects of the varieties spoken by members of a given speech community and that is accepted by members of that community. The synthetic voice in unit selection synthesis often sounds very much like the source voice and thus many aspects of dialect produced by the synthesiser are determined by the choice of speaker recorded. However other aspects of the dialect are determined by the phonetiser. For example, a synthetic Conamara voice would produce a word like *staighre* ‘stairs’ as /sʰaɪtʰaɪlə/, and a synthetic Gaith Dobhair voice will produce *staighre* as /sʰaɪtʰeɪlə/, with a different stressed vowel, as determined by the phonetiser rules. However the exact realisation of the /r/ would differ across dialects ([ɾ] and [ɾʰ]), and depending on what the speaker recorded produces, /ɾ/ could be pronounced as [ɾʰ] or [ɾ] in Gaith Dobhair (these variants are in free variation in this dialect).
Catering for all major varieties of Irish within TTS using unit selection (see § 2.3.1) involves selecting a speaker suitably representative of the salient aspects of each dialect, and whose speech is acceptable to other members of the dialect community in question as being of that dialect. One possibility for achieving this in the case of Irish is to choose a speaker of one of the stronger subdialects, and this was the approach we adopted.

For Ulster Irish we used the dialect of Gaoth Dobhair, one of the strongest Gaeltachtai in Donegal (and indeed, in Ireland) (Ó Giollagáin et al., 2007). For Connaught and Munster Irish we used the dialects of Conamara (Co. Galway) and Corca Dhuibhne (Co. Kerry) respectively. The dialect of the Conamara region is one of the strongest and best known dialects of Connaught (de Bhaldraithe, 1945), and the Conamara region and its Irish is almost synonymous with Connaught Irish. Likewise, Kerry is the Munster county with the largest Gaeltacht. And the Corca Dhuibhne Gaeltacht is the strongest regional dialect in Kerry, with some of its constituent electoral divisions being assigned “Category A” Gaeltacht status in Ó Giollagáin et al. (2007), as more than 67% of the total population (3 years+) are daily speakers of Irish. Choosing a strong dialect such as this has the added benefit of there being more speakers to choose from when searching for a speaker to base a synthesiser on.

2.6.3 Applications of TTS for Irish

Many of the applications of TTS synthesis introduced in § 2.1 are relevant to Irish. However, it is unlikely that there will ever be a significant commercial demand for Irish language TTS synthesis. That said, limited domain applications may be developed from time to time such as the recently updated and now fully bilingual
Irish text-to-speech (TTS) synthesis

onboard announcements on certain national rail services (Iarnród Éireann), and partially bilingual onboard announcements on Dublin light rail services (DART and LUAS). These applications would be developed using commercial synthesis software and would not require an Irish unit selection synthesiser capable of generating novel sentences.

TTS applications such as screen and document readers for the visually impaired and for people with learning disabilities such as dyslexia are, however, urgently required for Irish. As mentioned in Chapter 1, the vast majority of Irish children attending school, including those with visual and learning disabilities, study Irish, as it is a core curricular subject at both primary and post-primary levels in the state (Ó Murchú, 2001), and studied by thousands of students in Northern Ireland. When studying Irish, these students currently have no access to the same type of assistive technologies that would routinely be available for English. The situation is even worse for students who attend Irish-medium schools (Gaelscoileanna and schools in Gaeltacht areas) where all subjects except English are studied through the medium of Irish. The research being carried out at TCD is in part motivated by the needs of these students, including reading textbooks, doing research on the Internet, and asking questions in class. A beta version of a screen reader has been developed and is being pilot-tested. This tool uses the HTS Conamara voice described in § 2.6.1. HTS synthesis is appropriate for screen readers in large part because it is fast and consistent. This is important because screen readers often “talk” at speeds much higher than normal speaking rate, to allow users to quickly scan documents.

There are also a number of other potential applications of Irish TTS synthesis and its associated speech technology resources, which include pronunciation dic-
tionaries (lexicons), normalisers (see § 2.4.1) and phonetisers (see § 2.5). It has also been suggested that synthesis can be used as a method for long-term preservation of dialects that may be on the verge of dying out (Ni Chasaide et al., 2006). The recording of speech corpora in the case of such dialects in itself is a useful exercise towards archiving these dialects, but if these speech corpora are suitable for the development of synthesisers, the potential arises for these dialects to “live on” through synthesis. There are also a number of potential pedagogical applications, such as multimedia guides to the pronunciation of Irish in different dialects (Ó Néill & Mac Lochlainn, 2007). Another possible application is that the phonetiser being described here be used in the development of dialectal pronunciation dictionaries for Irish. Further applications of the Irish phonetiser, beyond text-to-speech synthesis, are discussed in Chapter 8.
Chapter 3

Phonetisation and the sound system of Irish

3.1 Introduction

This chapter addresses aspects of the sound system of Irish that are relevant for our research on phonetisation for the development of text-to-speech (TTS) synthesis and for applications. In particular we focus on language-general and dialect-specific aspects relevant to the two-stage multi-dialect approach developed here. For an in-depth account of Irish phonology, see Ó Siadhail & Wigger (1975), Ó Siadhail (1989), and Ní Chiosáin (1991).

The Chapter is structured as follows. In § 3.2 I describe the motivation for my approach, a two-stage approach involving a conversion from orthographic forms to abstract non-dialectal forms in the first stage, followed by a conversion to dialect-specific forms in the second stage.
3.2 A two-stage approach to Irish phonetisation

The approach adopted here is to derive dialect-specific representations from orthographic text using a two-stage model (see Figure 1.2). The particular two-stage approach adopted involves the conversion of orthographic forms to an abstract dialect-neutral representation based on phone-level and morpheme-level information, which is in turn converted to a dialect-specific representation.

The abstract intermediate representation (the global module) we use will remind some readers of approaches proposing unified phonological representations for dialects of a given language suggested by earlier researchers working on Irish and other languages. Our approach has some similarities with these approaches, but it differs from them in several important ways. We will therefore discuss this earlier work in some detail.

In the late 1960s, an effort was made to create a pronunciation norm for use in the preparation of Irish language teaching resources (Ó Murchú, 1969), a need motivated by the lack of a standard variety of the language. Ó Murchú suggested that “one solution to the problem—which may or may not be the best one—is to construct an artificial norm which as far as possible would be equally relatable to the various regional varieties” (p. 43) and set out to accomplish this goal by employing the concepts of common core and underlying forms, basic to the living varieties of the language and common to all dialects.

This approach is similar in many ways to work in generative dialectology which seeks to provide a single phonological system unifying all dialects. In this panlectal phonology approach “variations between speakers, or more generally between language varieties, can largely be accounted for by positing differences in relatively
superficial rules or their ordering, the underlying representations being common to all varieties of the language” (Wells, 1982, p. 69). Classic examples of this type of approach include Newton (1972) for Greek and Brown (1972) for the Bantu language Lumasaaba.

In such approaches, Irish past tense verbs like *cheannaigh* ‘bought’, pronounced [cənˠaː] in Munster, [cənˠa] in Connaught, and [cənˠiː] in Ulster, which end in a consonant in Munster, but in a vowel in other dialects, might have an underlying phonological representation in all dialects containing a final consonant. This final consonant could then be deleted by a rule in Connaught and Ulster. As Wells writes “[s]uch a panlectal grammar is justified, it is claimed by its adherents, not merely as a descriptive device enabling us to summarize a variety or differences in a convenient form but more significantly as furnishing a model of the adult speaker’s ability to understand, interpret and even predict the speech of those who speak a different accent from his own” (p. 69). It is established that historically the form *cheannaigh*, and similar past tense forms, ended with a consonant (Ó Sé, 2005), and this is reflected in the spelling. But whether it is possible or desirable to posit a unified phonological system is controversial. Wells gives a number of examples from English where it is difficult or impossible to account for forms in different dialects with a single underlying phonological representation (e.g. the intrusive [r] of Received Pronunciation).

Some researchers working within this framework argue that these proposed unified representations are psychologically real and form part of speakers’ phonological competence. Ó Murchú (1969) explicitly does not make this claim. He writes “[t]his notion of underlying form differs somewhat from that of Sol Saporta etc... [Saporta’s] approach is often more realistic, both from the point of view
of using ordered rules as a descriptive device in dealing with divergent forms and
as a reflection of speakers’ intuitive feeling for the relationships involved, but it
would encounter complications in handling some of the problems discussed in
this paper...” (pp. 47,48). Kallen (1997) argues for at least a “qualified” role
for generative dialectology writing “[t]he rule systems which I envisage in a new
generative dialectology will not necessarily claim ‘psychological validity’ for all
speakers... it is quite clear that such systems will not be based on any single
dialect of a language” (p. 97).

These approaches address important and thus far unresolved questions about
the relationship between dialects and phonological representation. Similar ques­tions have been addressed in the recent phonetics and psycholinguistics literature.
For example, in an eye-tracking study (Dahan et al., 2008), show that listeners
dynamically alter their representation of speech sound categories, that they learn
from a speaker’s dialect-specific phonological differences and use this new knowl­edge when evaluating subsequent utterances from the same speaker. Specifically
these experiments showed that listeners who heard a speaker producing a raised
\( /æ/ \) before \( /g/ \) in words like \emph{bag}, but not before \( /k/ \) (in words like \emph{back}), a pat­tern found in some varieties of American English, were less likely to misinterpret
words with a \emph{back}-like vowel than listeners who heard the speaker pronounce a
non-raised \( /æ/ \) in this context.

Some of the insights from panlectal phonology and similar approaches, and in
particular \( Ó \) Murchú’s common core and its extension to the Lárchamhúint com­promise dialect, were useful in the development of our approach to phonetisation
for Irish, as described in the next paragraph. Our goal, however, was to produce
a phonetiser that satisfied the requirements of corpus-based unit selection syn-
thesis. It was not to propose a unified phonological system for all the dialects of Irish. In fact, our phonetisation includes at least one element that is arguably not present in the phonological system of any dialect, that is, the devoiced sonorants discussed in § 3.7. Including these segments was necessary for accurate unit selection from a technical point of view.

Our approach was inspired by Ó Murchú’s common core, and the pronunciation system known as Lárácháinúint (lár ‘centre/core’ + canúint ‘dialect’), which builds on the foundations laid down in Ó Murchú (1969) (see Ó Baoill, 1986b), was also valuable in its development. Lárácháinúint was intended as a useful resource for learners and teachers of Irish, and was not meant to replace living dialects but to supplement them, and is used in Foclóir Póca (Ó Baoill, 1986a), a pocket English-Irish/Irish-English pronunciation dictionary. Ó Baoill (1986b) also includes a fully developed set of rules for deriving Lárácháinúint pronunciations from standard spelling. The Lárácháinúint is based on the essential contrasts found in the three main dialects and correspondence between graphemes and phonemes is systematic. However, although it is claimed that given the broad nature of the transcription system, actual pronunciations can be derived via the application of dialect specific pronunciation rules, this is not actually the case, and thus the Lárácháinúint pronunciations were not sufficient as an intermediate representation for the multi-dialect task of phonetisation.

For example, the inventory of the Lárácháinúint transcription system is problematic for our purposes. Since it omits phonemic contrasts that are not found in all dialects, the Lárácháinúint inventory cannot be used by the current system if it is to produce output in any regional dialect of Irish. The global inventory required in our case, rather than being a subset of phonemes, must be a superset of
phonemes. The Lárchanúint phonemic inventory does not account, for example, for the potential four-way and robust three-way opposition amongst laterals and nasals in Donegal Irish (Ní Chasaide, 1977), nor do they account for the extra long back vowel phoneme (/ɔː/ʃ/) in these varieties.

The approach adopted was preferred over other alternative two-stage approaches. One such approach might take the historical development of the language into account. It has been suggested that it might be possible to derive modern dialectal realisations by rule from the spelling system of early Modern Irish, which reflects historical pronunciations. It could be argued that this intermediate representation would be less abstract than the intermediate representation used in the system described here since it would in many cases reflect a real historical pronunciation. However, on closer examination, it is clear that such an approach would be unworkable. There is no clear link between the modern spelling system, upon which the phonetiser depends, and historical forms, and such an intermediate representation would therefore be difficult to derive. For example, it is impossible to derive by rule the historical form cruaidh of the modern form crua ‘hard’ (Ó Siadhail, 1981). There are in fact many examples that show that the mapping between early Modern Irish spelling and modern standard spelling is not commutative (i.e. the fact that you can get from Classical orthography to modern orthography does not mean that you can get from modern orthography to Classical orthography). For example, the modern grapheme (İo) in diúlta ‘revenge’ would map to (ogha) in early Modern Irish spelling, i.e. dioghaltas, but the same modern grapheme in the word inúon ‘daughter’ would map to (gea), i.e. inghean. Another example is the grapheme (û) which maps to either (umha) (e.g. cúng ~ cumhang ‘narrow’), (abh) (e.g. ceilíúradh ~ ceileabhradh ‘celebration’), (amhadh) (e.g. ceathrú ~ ceathramhadh ‘a quarter’), (adh) (e.g. bunús
Phonetisation and the sound system of Irish

~ bunadhas ‘basis’), (ughadh) (e.g. foilsiú ~ foillsiughadh ‘publication’), or (ugh) (e.g. údar ~ ughdar ‘author’). These and other examples can be found in Ahlqvist (1994), and further examples can be found by consulting Dineen (1927). Furthermore, as Ahlqvist (1994, p. 40) states “[N]ior choras seasta e coras litrithe na Nua-Ghaeilge Moiche ar a Ian slite” ‘The spelling system of Early modern Irish was not a stable system in numerous ways’. Additionally, such an intermediate representation would in fact be very abstract except to a very limited number of people with detailed knowledge of Irish historical linguistics. For example, the modern Irish word saol ‘life’ would have the intermediate representation /sɪɾʊːjəl/, reflected in the Classical spelling of the word. A TTS synthesis developer trying to extend the phonetisation to a new dialect can understand a system in which the monosyllabic word saol is represented as a monosyllabic word with one vowel, which maps to either [i:] or [e:] (the approach the phonetiser described here used). Using an intermediate representation with two syllables and an intervocalic fricative would needless complicate the task. As discussed in Ahlqvist (1994) and Ó Murchú (1977) the loss of intervocalic fricatives was a pervasive sound change, as the difference between the Classical orthography and the modern orthography indicates. Even if it were possible to map from the modern standard orthography to early Modern Irish pronunciations, this type of intermediate representation would work against our goal of building a user-friendly system to facilitate the addition of new dialects.

Another alternative approach considered would convert directly from orthographic to dialectal forms. Rules to map from one dialect to another could then be applied. This is not the best approach because it is not always possible to satisfactorily map from the pronunciations of one dialect to the pronunciations of another. For example, the word buile ‘rage’ is pronounced [bˈɪlə] in all dialects,
but the word *buille* ‘hit’ is pronounced [bʰɪːlə] in Connaught and Ulster dialects but pronounced [bʰɪlə] in Munster dialects. Were a phonetisation system to map to the Munster transcription first, it would not then be possible to reproduce the contrast between the two types of lateral consonants found in Connaught and Ulster dialects via an inter-dialectal second-stage mapping from a Munster transcription. This could be overcome by mapping first to the dialect with the highest number of contrasts, however, this would entail treating that dialect differently to the other dialects in the first stage, and free mapping from dialect to dialect in any direction would not be possible. To give another example, the verbal ending *-adh* is pronounced [a] in Munster and Connaught dialects, but as [uː] in Ulster in words like *bualadh* ‘hitting’ for example, but not all instances of Munster and Connaught schwa ([ə]) are pronounced [uː] (for example *buile* ‘rage’ is pronounced with a final schwa in all dialects: [bʰɪlə]). These are not exceptional examples and many more could be given (e.g. *tapa* ‘fast’ [tʰapə] in Connaught and Munster, [tʰapɪː] in Ulster).

### 3.3 Consonants

All Irish dialects include a rich inventory of consonants. The consonants phonemes\(^\text{22}\) of Gaóth Dobhair (the Ulster dialect we have chosen, see § 2.6.2) are shown in Table 3.1. All dialects of Irish include a phonological opposition between velarised and palatalised consonants. In Table 3.1, this opposition is accommodated by subdividing the cells and placing phonemically velarised consonants in the upper

\(^{22}\)Recall that the term *phoneme* is used here to refer to segments that act as the smallest unit of phonological structure that can signal a difference in meaning within the phonological system of a given variety of a language.
half of the cell and phonemically palatalised consonants in the lower half. When primary place of articulation is the same, both members of the pair will be in the same cell. That the members of each pair are in contrastive distribution is made clear by the presence of minimal pairs such as *bád* /bˠaːdˠ/ ‘boat.NOM.SG’ vs. *báid* /bˠaːdʲ/ ‘boat.GEN.SG’ and *mún* /mˠuːnˠ/ ‘urinate’ vs. *múin* /mˠuːnʲ/ ‘teach’.

<table>
<thead>
<tr>
<th>Labial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Alveolo-palatal</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>pʰ</td>
<td>bʰ</td>
<td>tʰ</td>
<td>dʰ</td>
<td>t̠ʲ</td>
<td>d̠ʲ</td>
</tr>
<tr>
<td>Fricative/ Approximant</td>
<td>fˠ</td>
<td>w</td>
<td>sˠ</td>
<td>c</td>
<td>ç</td>
<td>j</td>
</tr>
<tr>
<td>Nasal</td>
<td>mˠ</td>
<td>n</td>
<td>n̠</td>
<td>ń̠</td>
<td>ń̠</td>
<td>η</td>
</tr>
<tr>
<td>Tap</td>
<td>rˠ</td>
<td>l</td>
<td>j</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral Approximant</td>
<td>l̠</td>
<td>ĺ̠</td>
<td>j̠́</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Consonant phonemes of Gaoth Dobhair Irish (Ní Chasaide, 1999).

As the phoneme chart is based on a description of Gaoth Dobhair Irish, the phonetic value of the symbols used will relate more closely to this dialect.

In Gaoth Dobhair Irish, the phonemically palatalised coronal plosives /t̠ʲ d̠ʲ/ are realised as alveolo-palatal affricates. For example, *(ar) tí* /t̠ʲiː/ ‘about to’ is realised as [t̠ʲiː]. The consonants /w v j/ which are fricatives in many other dialects, may be realised as approximants in this dialect. For example, *(mo) ghiall* /jiaːlˠ/ ‘(my) hostage’ is realised as [jiaːlˠ] (Ní Chasaide, 1999).

The /x/ phoneme in Gaoth Dobhair Irish tends to be realised as a voiceless

---

23 When discussing this phonemic opposition we will use the terms *phonemically velarised* and *phonemically palatalised* since the actual phonetic realisation of these segments differs across dialects. Many researchers use the traditional terms *broad* and *slender* to refer to this phonemic opposition and use a prime (e.g. b’) to mark the slender/phonemically palatalised segments. Historically, consonant quality was “conditioned by the following vowel, being (a) palatal before ŋ and ē, (b) neutral before ā and ō, (c) u-quality before ų.” (Thurneysen, 1946, p. 97). The claim of the last conditioning environment (c), is controversial.
velar fricative [x] only when it occurs word-initially (e.g. (mo) chóta /xoːtə/ ‘(my) coat’ [xoːtə]), and is usually realised as a glottal fricative [h] when it occurs non-initially (e.g. teach /təx/ ‘house’ realised as [təh]). The Connaught and Munster dialects also have a /x/ phoneme, but the [x] allophone is found in all contexts in these dialects. In Gaoth Dobhair, when /x/ occurs word-finally before /tʃ/, it may be realised as [x], as in seacht [ɔxʃə] ‘seven’. This allophone is in free variation with [x] in this context ([ɔxʃx] is also heard). The /r/ phoneme is realised as [ɾ] or [r] when it occurs non-initially, as in amárach /əməɾax/ ‘tomorrow’. The /ɾ/ phoneme may be realised as either [ɾ] or [ɾ] in postvocalic and intervocalic contexts. The [ɾ] allophone (e.g. Máire /məɾə/ (girl’s name) realised as [məɾə]) of this phoneme is one of the most distinctive features of this dialect. Word-initially, the contrast between /r/ and /ɾ/ has collapsed in this and in most Irish dialects and is realised as [ɾ] in this context, as in Róisín [ɾɔsɨn] (girl’s name) and Ríona [ɾiʃə] (girl’s name), and when it occurs before the following phonemically palatalised consonants: /dɾ, nɾ, pɾ/ (Ó Sé, 1995; Ó Baoill, 1996; Ó Murchú, 1998; Ó Buachalla, 2003).

In terms of phonemic inventory, the Gaoth Dobhair dialect retains a three-way phonemic opposition amongst laterals and nasals, which historically was a four-way opposition (Quiggin, 1906; Sommerfelt, 1922; Ó Searcaigh, 1925). The historically velarised alveolars have merged with the velarised dentals leaving /nɾ/ realised as [ɾ], and /lɾ/ realised as [ɾ]. A phonemic distinction between the phonemically palatalised alveolars /n l/ and the phonemically palatalised alveolopalatalas /nɾ lɾ/ remains, and the alveolar phonemes are realised with a fairly neutral quality (Ní Chasaide, 1977, 1999).

In Conamara Irish (the Connaught dialect we have chosen, see § 2.6.2, and refer
to de Bhaldraithe, 1945), the phonemically palatalised plosives /tʲ dʲ/ are realised as palatalised alveolar affricates [tʃ dʒ] (e.g. (ar) tí /tʃiː/ ‘about to’ is realised as [tʃiː]). The voiced consonant /w/ is realised as a velarised voiced labiodental fricative [vʲ] when it occurs word-finally or before a consonant (e.g. slíabh /cʲiav/ ‘mountain’ [cʲiavʲ], bhraith /vʲr⁷əh/ ‘feel.PST’ [vʲr⁷əh]). In other contexts, it is realised as a labial-velar approximant [w] (e.g. mhaígh /wiː/ ‘boast.PST’ [wiː]). The phonemically palatalised dorsal /j/ is realised in this dialect as a palatal fricative [j], as in (mo) ghiall /jia F / ‘(my) hostage’, which is realised as [jia F].

The phonemically palatalised voiceless coronal consonant /p/ is realised as [J] (e.g. seacht /caxtʲ/ ‘seven’, which is realised as [caxtʲ]) (de Bhaldraithe, 1945; Murchú, 1998; Curnain, 2007).

In Conamara Irish, the phonemically velarised dorsal consonant /x/ may be realised as either [x] or [χ] in all contexts, as in (mo) chóta /xoːtʲə/ ‘(my) coat’ [xoːtʲə], [χoːtʲə]. Word-initially, the phonemically palatalised dorsal consonant /ç/ is nearly always realised as [ç], while medially, the allophones [ç], [P] and [h] are found, as in cheap /caPʲ/ ‘think.PST’ [caPʲ], fiche /Pʲə/ ‘twenty’ [Pʲə], cluiche /kʲɾʲə/ ‘game’ [kʲɾʲə], and droichead /dʲɾʲɪcaɾʲdʲ/ ‘bridge’ [dʲɾʲɪcaɾʲdʲ].

The phonemically palatalised rhotic /ɾʲ/ may be realised as a fricative similar to [ʒ] (e.g. aire ‘caution’ [aʒə]).

While a three-way phonemic opposition amongst laterals and nasals is also present in Conamara Irish, it is not present in Corca Dhuibhne Irish (the Munster dialect we have chosen, see § 2.6.2, and refer to Ó Sé, 2000). Is this dialect, the phonemically palatalised alveolar and alveolo-palatal consonants have merged, leaving a robust two-way opposition amongst laterals and nasals. For example, while leas ‘benefit’ and cáil ‘fame’ would be realised as [lʲasʲ] and [kaːl]
respectively in Conamara Irish, they would be realised as \( [\text{Pas}^\text{\textasciitilde}] \) and \( [\text{ka}\text{r}^\text{\textasciitilde}] \) respectively in Corca Dhuibhne Irish, with the lateral consonants having a colouring that is somewhere between a plain and a palatalised alveolar. The /\eta/ phoneme is sometimes realised as \( [\text{n}^\text{(\textasciitilde)}] \) word-finally, as in \textit{fairsing} /\text{f}^\text{\textasciitilde}\text{ar}\text{c}\text{\textasciitilde}\text{\textasciitilde}/ ‘extensive’ \( [\text{f}^\text{\textasciitilde}\text{ar}\text{\textasciitilde}\text{\textasciitilde}\text{\textasciitilde}] \).

Most importantly, the larger phonemic inventories of the Conamara and Gaorth Dobhair dialects need to be accounted for by the phonetiser. Although, as described above, there are major differences across dialects in the allophonic realisation of the consonant phonemes of Irish, these differences are not relevant to the primary application of phonetisation, i.e. TTS synthesis. A broader level of transcription without a lot of phonetic detail is appropriate and desirable for unit selection speech synthesis. As phonetic detail is often determined by context, the unit selection algorithm, which takes context into account, will select units from appropriate contexts and therefore produce synthetic speech output that is phonetically accurate. For example, the transcription used by the current speech synthesis system uses a single symbol for phonemically velarised labial consonant in all dialects, however, in Donegal, before a vowel, this segment is realised as as a voiced rounded labial-velar approximant \([\text{w}]\) (e.g. \textit{mo mhac} /\text{m}^\text{\textasciitilde}\text{\textasciitilde}\text{wak}/ ‘my son’ \( [\text{m}^\text{\textasciitilde}\text{\textasciitilde}\text{wak}] \)), as a labiodental fricative (\([\text{v}]\)) in Corca Dhuibhne (e.g. \textit{mo mhac} \( [\text{m}^\text{\textasciitilde}\text{\textasciitilde}\text{v}^\text{\textasciitilde}\text{ak}] \)), and as \([\text{w}]\) or \([\text{v}]\) in Conamara depending on the context. Another example is the raising of the vowel /\text{o}:/ to [\text{u}:] following nasals in Conamara Irish, as in \textit{Nóra} /\text{n}^\text{\textasciitilde}\text{o}\text{\textasciitilde}\text{\textasciitilde}\text{\textasciitilde}/ (girl’s name) \( [\text{n}^\text{\textasciitilde}\text{u}:\text{\textasciitilde}\text{\textasciitilde}] \), where the TTS synthesis selection algorithm will find a unit containing a phone labelled /\text{o}:/ but pronounced [\text{u}:] by the native Conamara speaker recorded. It would therefore be a waste of resources (time, effort, and money) to narrowly annotate a speech corpus intended for use in this type of synthesis. That said, this level of detail would be important
for other potential applications of phonetisation, such as pronunciation training, as discussed in Chapter 8. For example, a learner trying to emulate a native speaker of this dialect might benefit from a transcription of the name Nóra that illustrated this pronunciation rule.

Just as the realisation of individual phonemes varies from dialect to dialect, so does the realisation of the phonological opposition between the series of phonemically velarised and phonemically palatalised consonants. The convention of classifying consonants as velarised or palatalised does not reflect the full complexity of the realisation of the opposition in different dialects. If we look first at the velarised series of consonants, grouping them according to place of articulation, the picture becomes clearer.

Phonemically velarised labials in all Irish dialects are characterised by labialisation and a following [u]-glide vowel when followed by a front vowel. Despite this being systematic, it is not reflected in a standard phonemic transcription. For example, as the phonemic transcription conventions in Ní Chasaide (1999) are followed, the word but 'yellow' is phonemically transcribed as /b̠iː/. However, a narrow phonetic transcription would be [b̠̂ri:]. Similarly, phonemically velarised coronals in Irish are characterised by a following [u]-glide vowel when followed by a front vowel. For example, a standard phonemic transcription of the word tut 'hay' would be /t̠̂iː/, but the actual realisation would be [t̠̂uiː]. Phonemically velarised dorsal consonants preceding front vowels are characterised by the [u]-glide, but crucially are not labialised. A classic example of this can be found in the word Gaeilge /g̠e̠j̠ə/ 'the Irish language', realised as [g̠ue̠i̠ja], where learners often substitute labialisation of the consonant for the [u]-glide.

24The only deviation from the phonemic transcription conventions in Ní Chasaide (1999) is the use of the /a/ symbol for all low vowels and the consistent marking of length (e.g. /aː/) on phonemically long vowels (see § 3.4).
These glide vowels are known as off-glides. When a velar consonant is followed by a back vowel, off-glides do not apply. For example, *gallda* ‘foreign’, would be transcribed phonetically as [ɡæl^d' ^ a]. When phonemically velarised consonants are preceded by a front vowel, they are characterised by a preceding [ə]-glide, known as an on-glide. On-glides are less noticeable than off-glides and are thus transcribed in superscript in narrow phonetic transcription, as in *crios* /cr^is^v/ ‘belt’ [cr^i^s^v].

Phonemically palatalised labials in Irish are not palatalised (i.e. they are “plain”) when followed by a front vowel, as in the word *pictiúr* /p^i^k^t^i^ur^v/ ‘picture’, which is realised as [pict^i^ur]. Phonemically palatalised labials followed by a back vowel will for the most part be phonetically palatalised, however, in certain dialects, such as the dialect of Torr, Co. Donegal, phonemically palatalised labials are phonetically “plain” (Sommerfelt, 1922), as in *peann* ‘pen’ /p^a^n^v/ [p^a^n^v]. What is interesting about this is that the phonemic contrast that exists between phonemically velarised and phonemically palatalised consonants is carried by the vowel. This will be discussed in more detail in Section 3.4. When phonemically palatalised consonants are preceded by a back vowel, they are characterised by an [i] on-glide, as in *baic* /b^i^ac/ ‘nape’ [b^i^a^c].

### 3.4 Vowels

In all dialects of Irish, there is a phonological opposition between long and short vowels. In all dialects there are at least five long vowel phonemes /i: e: a: o: u:/ and five corresponding short phonemes /i  e  a  a  u/.\(^{25}\) Unstressed short vowels are

\(^{25}\)IPA symbols are used here to represent vowel phonemes in line with the treatment of consonants. Readers may be more familiar with the traditional Celticist notation: /i: e: a: o:
realised as schwa in all dialects. There are at least three diphthong phonemes /âu ëa uə/. The exact vowel inventory differs according to dialect.

The vowel inventory of Gaoth Dobhair Irish is: /i: e: a: o: u: i e a a y âu ëa uə/. Approximate pronunciations for the Gaoth Dobhair dialect’s inventory of monophthongs are given in the vowel chart in Figure 3.1, adapted from that in Ní Chasaide (1999). As seen in Figure 3.1 there is a sixth long monophthong phoneme /ɔ:/ as illustrated by the minimal pair cró /krɔː/ ‘hut’ vs. cnó /krɔː/ ‘nut’. Example words illustrating Gaoth Dobhair Irish vowel contrasts are given in Table 3.2.

Figure 3.1: The vowel phonemes of Gaoth Dobhair Irish (based on Ní Chasaide, 1999).
Phonetisation and the sound system of Irish

Table 3.2: Example words illustrating Gaoth Dobhair Irish vowel contrasts.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>/iː/</td>
<td>crích ('finish.DAT') /críː(ʃ)/</td>
</tr>
<tr>
<td>/ɪ/</td>
<td>críth ('shake') /críː(h)/</td>
</tr>
<tr>
<td>/eː/</td>
<td>cré ('soil') /créː/</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>creid ('believe') /crɛːd/</td>
</tr>
<tr>
<td>/æː/</td>
<td>leann ('ale') /lɛːŋ/</td>
</tr>
<tr>
<td>/ɑː/</td>
<td>creathe ('shake') /krɛː(h)/</td>
</tr>
<tr>
<td>/ɔː/</td>
<td>cnó ('nut') /krɔː/</td>
</tr>
<tr>
<td>/oː/</td>
<td>cró ('hut') /krɔː/</td>
</tr>
<tr>
<td>/ʌ/</td>
<td>croch ('hang') /krɔːx/</td>
</tr>
<tr>
<td>/uː/</td>
<td>crú ('milking') /krʊː/</td>
</tr>
<tr>
<td>/ɛː/</td>
<td>cruth ('shape') /krɛː(h)/</td>
</tr>
<tr>
<td>/aː/</td>
<td>leamhan ('moth') /lɛːmʰ/</td>
</tr>
<tr>
<td>/iː/</td>
<td>Cian ('boy’s name') /kiːⁿ/</td>
</tr>
<tr>
<td>/uː/</td>
<td>cuan ('bay') /kuaⁿ/</td>
</tr>
</tbody>
</table>

In the Conamara (Connaught) and Corca Dhuibhne (Munster) dialects of Irish there is a fourth diphthong phoneme /ʌi/ as illustrated by the minimal pair cadhan /kʰʌiːn/ ‘brent goose’ vs. cuan /kʰuaⁿ/ ‘bay’. Conamara and Corca Dhuibhne Irish both have vowel inventories with five long monophthongs, five short monophthongs, and four diphthongs: /iː æː oː iː aː yː aːi aː ʌi aː/. The /ʌi/ is not present in Ulster Irish, where a word like cadhan would map to the following phonemic transcription: /keːn/. The articulation of the other long vowels varies very little from one dialect to another, with /iː/ being realised close to cardinal vowel [i], /eː/ being realised close to cardinal vowel [e], /oː/ being realised close to cardinal vowel [o], and

For the most part, there are only minor differences in the realisations of vowels across dialects, however there are some striking differences. The most significant of these is the articulation of the long low vowel phoneme /aː/, which is realised as [æː] (a front vowel) in Gaoth Dobhair Irish, but as [ɑː] (a back vowel) in many other Irish dialects, including the dialects of Conamara and Corca Dhuibhne. The articulation of the other long vowels varies very little from one dialect to another, with /iː/ being realised close to cardinal vowel [i], /eː/ being realised close to cardinal vowel [e], /oː/ being realised close to cardinal vowel [o], and
/uː/ being realised close to cardinal vowel [u] (although this may be advanced in Gaóth Dobhair Irish) across the dialects (Ó Cuív, 1944; de Bhaldraithe, 1945; Breatnach, 1947; de Búrca, 1958; Mhac an Fhailigh, 1968). Additionally, in Gaóth Dobhair Irish, long vowels tend to be reduced in final unstressed syllables, as in cosán ‘footpath’ /ˈkaːs̪aːn̥ˠ/ (although this may be advanced in Gaóth Dobhair Irish) across the dialects (Cuív, 1944; de Bhaldraithe, 1945; Breatnach, 1947; de Búrca, 1958; Mhac an Fhailigh, 1968). Additionally, in Gaóth Dobhair Irish, long vowels tend to be reduced in final unstressed syllables, as in cosán ‘footpath’ /ˈkaːs̪aːn̥ˠ/, realised as [ˈkaːs̪aːn̥ˠ]. In Conamara and Corca Dhuibhne Irish, the second element of the diphthongs /iːa/ and /uːa/ are reduced, being realised as [iːa] and [uːa] respectively.

In contrast to the long vowels, which are relatively stable, the short vowels in all dialects are characterised by instability. Ní Chiosáin (1995, p. 158) discusses the interaction between short vowels and consonants, citing de Bhaldraithe (1945); Ó Siadhail & Wigger (1975); Ó Siadhail (1989), and stating “I. Non-low vowels get their backness specification from a following consonant. II. If there is no consonant following the vowel, the vowel gets its backness specification from the preceding consonant”. This accounts for the vowel quality in words like clog /kl̪̊oːɡ/ ‘clock. NOM. SG’ [kl̪̊oːɡ] versus cloig /kl̪̊oːɡ/ ‘clock. GEN. SG’ [kl̪̊oːɡ], and bith /b̪̊iː/ ‘existence’ [b̪̊iː] versus puth /p̪̊uː/ ‘a breeze’ [p̪̊uː]. Similarly, in all dialects, unstressed short vowels are reduced, being realised as [a] or [i] depending on the quality of the following consonant (i.e. velarised vs. palatalised, e.g. tacar /ˈt̪̊aːkəɾ̥ˠ/ ‘set. NOM. SG’ realised as [ˈt̪̊aːkəɾ̥ˠ], tacair /ˈt̪̊aːkəɾ̥ˠ/ ‘set. NOM. PL’ realised as [ˈt̪̊aːkəɾ̥ˠ]), or the quality of the preceding consonant when there is no following consonant (e.g. úlla /ˈuːlaː/ ‘apples’ [ˈuːlaː], and file /ˈfiːlaː/ ‘poet’ [ˈfiːlaː]). Reduced short vowels are always transcribed as /a/ in broad transcription here, as well as when labelling speech data for synthesis. The realisation of stressed short vowels depends on the dialect. Even within any given dialect, however, the

---

26 According to O’Rahilly (1972, p. 105) there was historically some variation within the Ulster dialects in terms of this vowel reduction. He writes “it was probably not until the [18th] century that the Ulster vowel-shortening, long scorned as a vulgarism, gained complete ascendancy.”
actual realisation of the vowels will fluctuate according to consonantal context, thus making them difficult to define. As Wagner (1958, p. XXII) states “the [short] vowels have a wider phonetic radius. Free interchange of central vowels ..., all very close to the ‘irrational’ vowel a, is a common feature of Irish dialects, and is often the despair of the phonetician who tries to define them exactly.” For example, while /i/ may be realised close to cardinal vowel [i], it may be retracted when it occurs following a phonemically velarised consonant, as in cuid /kʰud/ ‘part’ [kʰud]. In all dialects, the stressed short vowel /a/ is realised as a front vowel close to [a] when it occurs following a phonemically palatalised consonant (e.g. gear /gaer/ ‘cut’ realised as [gaer]), and is realised as a back vowel close to [a] when it occurs following a phonemically velarised consonant (e.g. gar /gar/ ‘favour’ realised as [gar]).

If a velarised consonant, with secondary articulation involving the back of the tongue, occurs next to a front vowel, a glide vowel (or semivowel) will occupy the transition between consonant and vowel. For example, gui /gʰui:/ ‘prayer’ is realised as [gʰui:]. Similarly, if a palatalised consonant, with secondary articulation involving the front of the tongue, occurs next to a back vowel, the consonant and vowel are separated by a glide. For example, giúl /ju:l/ ‘joule’ is realised [ju:l]. Glide vowels and semivowels are not marked explicitly in phonemic transcription. Glide vowels are however marked in the orthography and as thus are discussed in Chapter 4 and handled by the phonetiser (see Chapter 5).

Most importantly for TTS synthesis, the correspondence between graphemes and phonemes differs from dialect to dialect, as in the example cadhan ‘brent goose’ mapping to /kʰain/ in Conamara and Corca Dhuibhne Irish, but to /ke:n/ in Gaath Dobhair Irish. As with the consonants discussed in § 3.3, the exact
nature of the allophonic realisation of these phonemes is relevant only to a limited extent to the task of phonetisation. It will not be relevant if phonetisation is for unit selection synthesis, but will be relevant if phonetisation is for pedagogical applications such as pronunciation training. For example, if learners wish to approximate Gaoth Dobhair Irish, they will need to know that the long low vowel phoneme in the word *bód* ‘boat’ will be realised as a front vowel [æː], and not as the back vowel [ɑː] and that the initial consonant is produced with lip rounding, differences not reflected in the broad transcription /bədə/. Similarly, transcribing the effect of velarisation and palatalisation on the realisation of the vowels is not necessary for phonetisation for unit selection synthesis, but may be relevant for other applications of phonetisation.

3.5 Initial mutation

An important feature of Irish phonology is its system of initial consonant mutation, which affects about one in five of all words in the language. Initial mutation, which is found in all Celtic languages, is a morphophonemic process in which an initial consonant changes in quality, or an initial vowel is prefixed with a consonant, given certain contexts. The process has been analysed in a number of theoretical frameworks (Ó Siadhail & Wigger, 1975; Ní Chiosáin, 1991), and affects nouns, verbs and adjectives. The two most common types of initial mutation are *lenition* and *nasalisation/eclipsis*. In lenition, the initial consonan-

---

27 According to calculations by Kevin Scannell of St. Louis University (U.S.) based on a 1.7 million word corpus.

28 The Irish language terms for these processes, *séimhíú* and *urá*, respectively, are well-defined, but finding satisfactory English terms is problematic. For example, *lenition* is often used to describe the process of *séimhíú*, but the change from a voiceless consonant to a voiced consonant, which is a type of *urá*, is also a lenition in the standard linguistic definition of the term. Similarly, the term *nasalisation* also has a specific meaning in phonetics. As Jaskula
nant is weakened (e.g., stops like /b\^/ become approximants or fricatives, e.g., *b\^d* [b\^\'ax\^]\] ‘boat’ → *mo bh\^d* [wa\^]\] ‘my boat’, and labiodental fricatives are deleted).

(2006) points out “[t]he very term ‘nasalization’ is also somewhat imprecise when we look at what happens to the radical consonantal segments in nasalizing environments. In particular, only the radical voiced stops are nasalized, while the voiceless obstruents cease to be voiceless”. In addition, using nasalisation to describe the prefixation of a nasal consonant to a vowel may be particularly confusing since a nasalised vowel is one produced with a lowered velum to allow air to pass through the nasal passages. This is not the case in the initial mutation in question. To avoid this confusion, “…the more neutral term ‘eclipsis’ is frequently used…” (Jaskula, 2006, p. 36). Taking these problems into account, the two types of initial mutation in Irish will be referred to here as lenition and nasalisation/eclipsis, following McCloskey (2010).
In nasalisation/eclipsis, voiceless stops become voiced, voiced stops become nasals, and a coronal nasal is prefixed to vowel-initial words (e.g., *bád* [b̠请输入您的音标] ‘boat’ → *i mbád* [m̠请输入您的音标] ‘in a boat’).

Historically these were phonological processes, with lenition triggered by vowel-
Phonetisation and the sound system of Irish final proclitics and nasalisation/eclipsis triggered by proclitics ending in a nasal (Thurneysen, 1946). But as Jaskula (2006, p. 23) notes initial mutations “stopped being phonological a long time before Old Irish [700–900 A.D.]”. Thurneysen (1946, p. 140) also notes that initial mutations “have sometimes spread by analogy”.

<table>
<thead>
<tr>
<th>Unmutated</th>
<th>Mutated</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>/pˠ/</td>
<td>/bˠ/</td>
<td>póca ‘pocket’ /pˠəoka/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>i bpóca ‘in a pocket’ /ə bˠəoka/</td>
</tr>
<tr>
<td>/p³/</td>
<td>/b³/</td>
<td>pictiúr ‘picture’ /pʰɪkˈtɹərˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>i bpictiúr ‘in a picture’ /ə bʰɪkˈtɹərˠ/</td>
</tr>
<tr>
<td>/bˠ/</td>
<td>/mˠ/</td>
<td>bás ‘boat’ /bˠəsˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a mbás ‘their boat’ /ə mˠəsˠ/</td>
</tr>
<tr>
<td>/b³/</td>
<td>/m³/</td>
<td>Béarla ‘English’ /bʲəɾˠəlˠa/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a mBéarla ‘their English’ /ə mʲəɾˠəlˠa/</td>
</tr>
<tr>
<td>/tˠ/</td>
<td>/dˠ/</td>
<td>tóg ‘take’ /tˠəgˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>an dtógha ‘would you take’ /ə dˠəghə:/</td>
</tr>
<tr>
<td>/t³/</td>
<td>/d³/</td>
<td>tiocfaidh ‘will come’ /tʲəkʰaθ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>an dtiocfaidh ‘would you come’ /ə dʲəkʰə:/</td>
</tr>
<tr>
<td>/dˠ/</td>
<td>/nˠ/</td>
<td>dóigh ‘burn’ /dˠəiː/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nach ndóigh ‘wouldn’t you burn’ /nˠəx nˠəiː/</td>
</tr>
<tr>
<td>/d³/</td>
<td>/n³/</td>
<td>déan ‘do’ /dˠənˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nach ndéanfhá ‘wouldn’t you do’ /nˠəx nʲənˠəθə:/</td>
</tr>
<tr>
<td>/k/</td>
<td>/ɡ/</td>
<td>ag cur ‘putting’ /aɡ kɜrˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>á gcur ‘putting them’ /ə gɜrˠ/</td>
</tr>
<tr>
<td>/c/</td>
<td>/ɟ/</td>
<td>ag ciontú ‘convicting’ /aɡ kʲɪɔntuː/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>á gciontú ‘convicting them’ /aɡ kʲɪɔntuː/</td>
</tr>
<tr>
<td>/ɡ/</td>
<td>/ŋ/</td>
<td>glúin ‘generation’ /ŋlˠɥiː/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ár glúin ‘our generation’ /aŋ rˠ ŋlˠɥiː/</td>
</tr>
<tr>
<td>/ʃ/</td>
<td>/ɹ/</td>
<td>geata ‘gate’ /ʃətə/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ár geata ‘our gate’ /aɾˠə ʃətə/</td>
</tr>
<tr>
<td>/ʃˠ/</td>
<td>/ɹˠ/</td>
<td>fás ‘growth’ /ʃˠəsˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bhur bhfás ‘your.PL growth’ /wɜɾˠ bʰʃəsˠ/</td>
</tr>
<tr>
<td>/fi/</td>
<td>/vi/</td>
<td>fion ‘wine’ /fʲiːnˠ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bhur bhfion ‘your.PL wine’ /wɜɾˠ fʲiːnˠ/</td>
</tr>
</tbody>
</table>

Table 3.4: Consonant nasalisation/eclipsis in Irish.

The conditioning environments for initial mutations vary across dialects. For
example, a noun in the dative case following the definite article will be lenited in Gaoth Dobhair (Ulster) Irish, but nasalised/eclipsed in Conamara and Corca Dhuibhne Irish, for example, *ar an bhád* (Ulster) ‘on the boat’ but *ar an mbád* (Connaught, Munster). In addition, within a dialect, initial mutation is not always found in contexts where it might be expected. As we see in the charts and as we will see in the next chapter, initial mutation is clearly marked in the orthography.

Currently the phonetiser uses the same symbols to represent consonants occurring as a result of mutation, and their corresponding base forms (e.g. The [b] of *ar an bpus*, and the [b] of *an bus*). We do not know however if there are acoustic differences between these base consonants and their mutated counterparts. We are currently investigating this question, inspired by work on French showing that phonemically identical consonants in similar sandhi contexts have systematic acoustic differences (e.g. the liaison [t] in *petit ami* [patitiːmɪ] ‘boyfriend’ vs. the [t] in *petit tamis* [patitiːmɪ] ‘little sieve’). If we find systematic differences, there would be an argument for using different symbols in the phonetisation.

3.6 Lexical stress and vowel reduction

In all Irish dialects, stress typically falls on the first syllable of words (e.g. ‘*siolla* ‘syllable’). In Corca Dhuibhne (and in Munster more generally), the situation is

---

29 People cite this phenomenon as *drochGhaeilge* ‘bad Irish’, evidence of the decline of the language. It is interesting to note therefore that Thurneysen (1946, p. 140) writes of variability in Old Irish [700–900 A.D.]: “[initial mutations] occur most consistently within a word-group the members of which, closely connected in speech, form a notional unit. The looser the connection, the less frequently and regularly do the mutations appear.”

30 This research is part of the project the Production and Perception of Irish Initial Mutations, funded by the Ulysses Ireland/France research exchange programme (Maire Ni Chiosáin and Pauline Welby, co-principal investigators, Brian Ó Raghallaigh, participant).
Phonetisation and the sound system of Irish

more complicated, since there has been a historical stress shift from initial to non-initial syllables in certain circumstances (*tuir'seach* ‘tired’). In Munster, stress falls on a non-initial syllable in the following circumstances: when the second syllable contains a long vowel, the second syllable is stressed (e.g. *amh'rán* ‘song’); when the first syllable contains a short vowel and does not end in /h/, and the second syllable ends in /ax/, the second syllable is stressed (e.g. *ba'cach* ‘lame’); when the first two syllables contain short vowels and the third syllable contains a long vowel, the third syllable is stressed (e.g. *achai'ní* ‘appeal’) (Ó Sé, 1995).

The addition of prefixes has the same effect on stress in all the dialects and this effect varies with the type of prefix. The addition of an intensifying prefix add a primary stress (e.g. *lán'cinnte* ‘absolutely sure’ *lán* ‘full’ + *cinnte* ‘sure’, *príomh'oide* ‘principal’ *príomh* ‘principal’ + *oide* ‘teacher’). Other prefixes add a secondary stress (e.g. *do'thuigthe* ‘incomprehensible’ *do* ‘not’ + *tuigthe* ‘comprehensible’). Other compounds give rise to a primary-secondary stress pattern (*aer'fhórsa* ‘air force’, *béal,oideas* ‘folklore’ *béal* ‘mouth’ + *oideas* ‘recipe/education’).

There is a closed class of words which are historical compounds “[which] would all seem to have originated in the fusion of two or more words” (Thurneysen, 1946, p. 30), where a non-initial syllable has primary stress (e.g. *a'raile* ‘another’, *a'nonn* ‘thither’, *las'tuaidh* ‘on the north side’). Finally, in loan words, the stress pattern from the source language is adopted (*sio'raf* ‘giraffe’).

Information about word-level stress is important for optimising unit selection synthesis. Information about stress would ideally be available to the phonetiser for it to deal correctly with the reduction of unstressed short vowel which occurs in all dialects of Irish. Obtaining stress information is dependent on morpho-
logical analysis, an important component of any TTS system, but one that is separate from phonetisation (Úi Dhonnchadha, 2003, 2009; Úi Dhonnchadha & Van Genabith, 2006). The term reduction refers to the weakening of a vowel. In Irish, this amounts to the phonetic centralisation of vowels and reduced vowels differ in quality from the vowel before reduction. This will be discussed in further detail in Chapter 5.10.2.

3.7 Devoiced sonorants

There is a debate among linguists as to the phonemic status of a series of devoiced sonorants (liquids, nasals, and rhotics) that occur in certain contexts in Irish (Ni Chasaide, 1999). Sonorants are devoiced following a word initial /t̪ˠ t̪ˠ s̪ˠ c/ under lenition. For example, trnúthán /t̪ˠn̪ˠu:ha:ŋˠ/ ‘expectancy’ becomes trnúthán /n̪ˠu:ha:ŋˠ/ when lenited (e.g. a trnúthán ‘its expectancy’), being pronounced with an initial /ŋˠ/. These consonants are also devoiced preceding certain future or conditional verb endings, for example seolfaidh /xoːlˠa/ ‘will send’, chluinfeadh /xlˠin9/ ‘would hear’. For the purposes of corpus-based speech synthesis, it is necessary to incorporate these segments in the corpus. The reason for this is that the traditional treatment of these segments, which occur only in certain grammatical morphemes, as a sequence of /h/ plus sonorant or sonorant plus /h/ (e.g. /hn̪ˠu:haːŋˠ/, /coːlˠhə/, /xlˠmhə/) rather than as a single consonant (see, for example, Sommerfelt, 1922), poses the potential problem of incorrectly labelled units in the speech corpus: these are phonetically devoiced segments that do not have the same spectral characteristics as, for example, a nasal followed by a glottal fricative. As this is the case, provision for their annotation is required in the phoneset used for annotation.
We decided to develop a complete phoneset that included devoiced sonorants corresponding to the sonorant inventory of Gaoth Dobhair Irish, a dialect with a maximal set of sonorants. This complete phoneset is shown in Table 3.5. Readers more familiar with traditional Celticist notation as exemplified in the Dublin Institute for Advanced Studies monographs on Irish dialects (Ó Cuív, 1944; de Bhaldraithe, 1945; Breatnach, 1947; de Búrca, 1958; Mhac an Fhailigh, 1968) are referred to Appendix A where a table comparing the phonesets is provided. Transcriptions based on this phoneset are produced by the phonetiser and these transcriptions were used in the annotation of Gaoth Dobhair and Conamara speech corpora which were successfully used to develop unit selection speech synthesisers.

<table>
<thead>
<tr>
<th></th>
<th>Labial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Alveolopalatal</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p^v</td>
<td>b^v</td>
<td>t^v</td>
<td>d^v</td>
<td>t^j</td>
<td>d^j</td>
<td>k</td>
</tr>
<tr>
<td></td>
<td>p^j</td>
<td>b^j</td>
<td></td>
<td></td>
<td>c</td>
<td>j</td>
<td></td>
</tr>
<tr>
<td></td>
<td>f^v</td>
<td>w</td>
<td></td>
<td></td>
<td>x</td>
<td>y</td>
<td>h</td>
</tr>
<tr>
<td></td>
<td>f^j</td>
<td>v^j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m^v</td>
<td>m^v</td>
<td>n</td>
<td>n</td>
<td>n^j</td>
<td>n^j</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>m^j</td>
<td>m^j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>y^v</td>
<td>y^v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>y^j</td>
<td>y^j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap</td>
<td>t^v</td>
<td>t^v</td>
<td>t^j</td>
<td>t^j</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>l^v</td>
<td>l^v</td>
<td>l</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>l^j</td>
<td>l^j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral</td>
<td>l^v</td>
<td>l^v</td>
<td>l</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5: Consonant phoneset of Irish dialects using IPA transcription.
Chapter 4

Phonetisation and Irish spelling

Aigh nó a mean thu ios só léasaigh dat thi slíps in this clós, bhears a bíord, and dos not smóc bícás obh de trobal obh straigeing a meaits.

It is só long sins thi did an anasth dea’s bhorc dat thi thincs ‘manuil leabear’ is de neim obh a Portuguis arditeitear.

–Myles na gCopaleen, The Best of Myles.

‘I know a man who is so lazy that he sleeps in his clothes, wears a beard, and does not smoke because of the trouble of striking a match. It is so long since he did an honest day’s work that he thinks “manual labour” is the name of a Portuguese agitator.’

4.1 The text in Irish text-to-speech

I will now turn to the text part of Irish text-to-speech synthesis, that is, the spelling system of Irish. This chapter has several goals. I hope to help readers
understand some of the historical forces that have shaped the spelling system and to place Irish orthography in the context of other orthographic systems. I focus on the modern standard spelling system established in 1958, discussing some of the motivations for the development of this standard. I also highlight the morphological transparency of the Irish spelling system and discuss the advantages this gives to human readers and synthesis systems. I then introduce readers to the reading rules of Irish. These rules should be of some value to readers on their own, and they lie at the heart of the phonetiser presented in the next chapter.

4.2 Orthographic depth and orthographic distance

Researchers on reading and the relationship between the written form of a word and its phonological representation discuss writing systems with respect to their orthographic depth, “the degree of regularity in symbol-sound correspondences. In ‘shallow’ orthographies, the symbol-sound relationships are regular, and thus transparent.... [A] phonologically ‘deep’ system... tends to preserve morphological information at the expense of phonological transparency” (Hamada & Koda, 2008, p. 5); see also van den Bosch et al. (1994) on metrics for quantifying orthographic depth. A good example of a phonologically transparent writing system is that of Spanish, which is highly regular and has mostly one-to-one letter-to-sound correspondences (e.g., *plátano* ['platano] ‘banana’; *juntos* ['xuntos] ‘together’; *dañino* [da'nino] ‘harmful’). English, on the other hand, has a famously opaque writing system; often a single grapheme can correspond to more than one phoneme in unpredictable ways (e.g. ⟨gh⟩ in words like *enough, ugh, and*
As we will see in this chapter, in terms of phonological transparency, Irish lies somewhere between the extremes of English and Spanish, probably closer to Spanish. Most importantly for the task at hand, phonetisation for text-to-speech (TTS) and related resources, the letter-to-sound correspondences of the Irish writing system are fairly regular and can thus be formulated as a set of rules.

Another relevant concept is that of orthographic distance or orthographic congruity, the degree of similarity between two writing systems. Irish and English are orthographically distant; many comparable phonemes in Irish and in English are not represented by the same graphemes. The Irish novelist and satirist Myles na gCopaleen often poked fun at this orthographic distance. In the passage opening this chapter, he uses Irish spelling conventions to write English words, for example, the four-letter grapheme ⟨aigh⟩ (aígh I) for the sound [aɪ] (a grapheme that is uncommon in English, but found in the word straight for the sound [ɔɪ]), ⟨bh⟩ for [v] or [w] (bhears wears, obh of, bhorc work), or ⟨th⟩ for [h] (thú who, thi he). Even the pen name he used Myles na gCopaleen (his real name was Brian Ó Nualláin) pokes fun at the oddities of Irish spelling—with lowercase letters preceding uppercase ones. This would have resonated with his readers, many of whom would have perceived Irish spelling as absurd. Considering just the consonantal graphemes of Irish (see § 4.6.1), only a handful of the dozens of multi-letter graphemes have roughly equivalent grapheme-to-phoneme correspondences in English (the double letter graphemes ⟨ll⟩, ⟨nn⟩, ⟨rr⟩, ⟨ng⟩, and ⟨ph⟩, and the grapheme ⟨ch⟩ in some words like loch).

While this orthographic distance is not directly relevant to phonetisation for TTS synthesis, it is an important consideration motivating envisaged pedagogical applications. There is a large body of research on the relationship between
the orthographic form of a word and its phonological representation and on the influence of orthography on spoken word recognition and word learning (see for example, Perre & Ziegler, 2008, Perre et al., 2009, and references therein). There is also work on the effect of orthographic distance in second language learning. Most learners of Irish are native speakers of English and so are most familiar with the spelling system of English. Studies have shown that a greater orthographic distance between a first language (L1) and a second language (L2) poses an increased challenge to L2 learners in phonological decoding and word retention (Koda, 1996; Hamada & Koda, 2008). It seems likely that the orthographic distance between English and Irish could interfere with the learning of Irish as an L2.

The results of another study suggest, however, that the opaque orthographic system of their native language may confer at least some benefit on English speakers learning Irish. Erdener & Burnham (2005) compared the performance of native speakers of Australian English (a language with an opaque orthography) and native speakers of Turkish (a language with a transparent orthography) in a nonword (pseudo-word) repetition task with stimuli from Spanish and Irish. They report “Turkish participants were better than Australian participants for Spanish stimuli but worse for Irish stimuli... Thus, for speakers of the transparent Turkish, orthography was beneficial for the transparent Spanish, but detrimental for the opaque Irish. For speakers of the opaque Australian English, there was little difference in performance on Spanish and Irish” (p. 209). There was clear evidence of orthographic interference, however. For example, Australian English listeners were much more likely (5 to 10 times more likely, depending on the condition) to produce the stop [d] for the affricate [d̪] when the orthographic form of the word was given (in the Irish nonword deaip, for example).
The results of these studies will come as no surprise to teachers of Irish. In my own experience teaching Irish, I routinely hear pronunciations of Irish that respect the spelling system of *English*. As described in Chapter 8, one of the applications for phonetisation of Irish is in the development of pedagogical applications that will partly address this problem.

### 4.3 The Irish alphabet

Today Irish is written almost exclusively in Roman letters, except for ceremonial purposes. The basic Irish alphabet in modern typeface is shown in Figure 4.1. It contains thirteen consonant symbols and five short vowel symbols, which may be marked by an acute accent (a *sneadh fada* ‘long mark’ or *fada*, as it is called in Irish and in Irish English, respectively).

\[
a, \acute{a}, b, c, d, e, \acute{e}, f, g, h, i, \acute{i}, l, m, n, o, \acute{o}, p, r, s, t, u, \acute{u}
\]

Figure 4.1: Modern Irish alphabet. Note that there are only 18 letters (5 of which can be accented). The letters j, k, q, v, w, x, y, z are not part of traditional alphabet, although they are used in loan words (e.g. *jab* ‘job’).

In modern times up until the 1960s, Irish was commonly written using a style of typeface based on manuscript tradition known collectively as *Cló Gaelach* ‘Gaelic script’. To this day, some older Irish people report difficulty reading Irish printed in Roman letters.\(^{31}\) As seen in Figures 4.2 and 4.3, in Cló Gaelach a dot was placed over a letter to indicate lenition. In Roman script, lenition is now marked with a following ‘h’ (e.g. *a chara* ‘friend’).

---

\(^{31}\)For more on the history of argument and choice between the Roman and Irish character sets see McGuine, 1992.
4.4 The modernisation of Irish spelling

Irish spelling underwent a great deal of change from the Old Irish period (600–900 A.D.) to the Classical Irish period (1200–1600 A.D.). By the start of the Classical period, the spelling rules of the classical modern norm had emerged, for example, conventions for indicating lenition of voiced consonants (a superscript \( h \) or a dot over the letter), for marking nasalisation/eclipsis ("e.g. Old Irish a cat 'their cat' was written a ccat or a gcat."  Ó Murchú, 1977, p. 270), and for marking the palatalised/velarised opposition in consonants (using "glide letters", see § 4.6.2) (Ó Murchú, 1977). Throughout this period, the spelling system remained standardised amongst the educated classes. This strict written standard was set out by the bardic schools of the time, although as Ó Murchú (1977, pp. 270, 271) points out, these spelling rules "were not consistently applied in manuscripts by the majority of scribes until after they had been exemplified in the first printed books and therefore stabilized". These bards were subject to the Gaelic elite of the time and the written standards they adhered to initially reflected the Irish spoken by the upper echelons of Gaelic society. By the time
of the Flight of the Earls in 1608, which saw the Gaelic aristocracy in Ireland gradually replaced by an English-speaking elite, the written standard no longer reflected how Irish was spoken (Ó Murchú, 1985). Unfortunately, due to the lack of an educated class, many of which had fled the country following defeat by the English, no new written standard was set out that might have better reflected the way in which Irish was spoken at the time. Irish thus continued to be written using the Classical Irish spelling system of the bards (McCone et al., 1994).

The Irish spoken by the ordinary people was even further removed from this old written standard. Of course, Irish was not exceptional in this regard. Irish was evolving just as all languages evolve. In Middle English, for example, the word knight was pronounced with an initial [k] and a velar fricative represented in the spelling by the grapheme (gh). Hundreds of years later, English has changed but the spelling of words like knight still reflects earlier pronunciations.

One of the first authors to move towards a change in the way Irish was written was Theobald Stapleton, who in 1639, published a religious catechism (Figure 4.4) using spelling that would have been more accessible to ordinary Irish speakers as it was written using spelling that better reflected how Irish was spoken at the time. Other authors did not however take up Stapleton’s lead, and Irish continued to be written using a spelling system based on the former standard of the bards and Gaelic typeface (Ahlqvist, 1994).
By the late 19th century and the beginning of the Irish language revival movement, led by *Conradh na Gaeilge* 'The Gaelic League', a growing number of books, not only religious but also secular, had been published using the classical modern spelling norms. These norms were reinforced by the publication of Dineen Irish-English dictionary (Dineen, 1904, 1927). Some writers however were modifying the classical spelling system to more closely reflect the characteristics of their variety of the spoken language. At the same time, there was growing...
dissatisfaction with a spelling system that was increasingly viewed as archaic and cumbersome, leading to various efforts to modify, reform or simplify the system.

One of the more organised of these efforts led by a group called the Society for the Simplification of Irish Spelling was pushing radical reform (O’Daly et al., 1911). Their An Letiriú Shimplí ‘Simplified Spelling’ system was well documented and a number of books were edited in simplified spelling and published (see Figure 4.5). The name of this proposed new system, An Letiriú Shimplí, itself illustrates one of the most drastic recommended changes to the old spelling, where phonemically palatalised /c/ was written as ⟨sh⟩. Other suggested changes were the representation of lenited and nasalised/eclipsed consonants with a single consonant letter (e.g. suggested ⟨sh⟩ → ⟨h⟩, ⟨bh⟩ and ⟨mh⟩ → ⟨v⟩). Such changes were undoubtedly Anglicisations rather than simplifications (Ó Murchú, 1977) and were not adopted.

32 Although this suggested change was not adopted into the subsequent Official Standard, it has strangely lived on in the family name of one of the founding members of the Society for the Simplification of Irish Spelling, Shán Ó Cuív, whose name would previously have been spelt ‘Seán Ó Caół mh’. This might not have been significant only for the fact that his nephew Brian Ó Cuív went on to become one of the most renowned Celtic Scholars of the 20th century, and whose son Éamonn Ó Cuív was the longtime Irish Government minister for Community, Rural and Gaeltacht Affairs.
Among the suggested changes that were later adopted were those that brought the spelling more in line with the spoken form in all the modern dialects. For example forms like *diteam hail* 'local' and *oidhche* 'night' which were historically pronounced with intervocalic fricatives or fricative sequences are now pronounced with long vowels, as is indicated by the fada (long mark) over the vowel in the standardised and simplified spelling (*ditiúil, oíche*).

One of the major problems with the Simplified Spelling system was that it was based on only one sub-dialect of Irish, that of West Muskerry, Co. Cork. According to its proponents, books edited into Simplified Spelling were “spelled according to the pronunciation current in West Munster, in the native district of the greatest living writer of the Irish language [An tAthair Peadar Ó Laoghaire]” (O’Daly et al., 1911). Unfortunately, this meant that some of the simplifications

---

**Figure 4.5:** The poem *Dómnall Bán* in classical spelling and in *Letiriú Shímplí ‘Simplified Spelling’* (O’Leary & Bergin, 1911).
in fact came no closer to representing the spoken reality in dialects other than that of West Muskerry. For example, the use of the grapheme ⟨v⟩ instead of ⟨bh⟩ or ⟨mh⟩ reflects Munster pronunciation in cases like mo Dhomhnaill Bháin/mo Ghónuil Váin 'my Fair Domhnall.GEN' (see the first line of the poem in 4.5), but not that of Connaught and Ulster, where the adjective begins with a [w].

The proposals of the spelling reformers did not entirely succeed in gaining widespread support and they were opposed by those who argued for the retention of spellings that reflected the history of the language. These conservatives included the Gaelic League and the lexicographer Father Patrick Dineen (Ó Murchú, 1977). A compromise position was advocated by Thomas S. O’Rahilly "‘[s]uch cumbrous spellings as beirbhitughadh (for beiriú), imthíghthe (for imithe), faghbháil (for fáil), urradhas (for urráis) and filidheacht (for filíocht) would be a serious handicap on any language, and are simply impossible in the case of Irish if we really mean to give it a fair chance of life’" (quoted in Ó Murchú, 1977, p. 275). The establishment of the Irish Free State in 1922 and the resulting demand for Irish language materials for schools and government added to the growing pressure for some sort of modernised written standard to be set out. This culminated in the publishing of an official standard Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil (‘Irish grammar and Irish spelling: The Official Standard’) by state authorities in 1958 (Translation Section, the Houses of the Oireachtas, 1958). This was the first time since the time of the bardic poets that an official written standard existed for Irish.

The new standard, while retaining the traditional features of Irish spelling (e.g. the use of glide letters to indicate consonant quality), resolved the issues of variability by recommending one particular variant (e.g. ‘a’ for the nucleus of
unstressed syllables where previously either ‘o’ or ‘u’ were also used). It also modified cumbersome spellings that no longer reflected pronunciation in any spoken variety (i.e. cuibhdhe > cuí ‘more appropriate’, duilleabhar > duilliúr ‘foliage’, adhbhar > ábhar ‘matter’, and liomhtha > líofa ‘fluent’) (Ó Murchú, 1977). More information about the correspondences between the classical modern spelling system and the modern standard and their pronunciations can be found in the charts in (Ó Murchú, 1977, p. 277, 288). A thorough account of the correspondences between the pre-1958 spelling system (used in the Dineen dictionary, Dineen, 1927) and the spelling system adopted in the new official standard is given in Ahlqvist (1994, pp. 58,59). Ahlqvist cautions “[a]r eagla na mithuisceana, ámh, ní mó r a áitiú nach í stair na bhfoghar Gaeilge atá á ríomh in aon chor, ach stair na litreachá amháin,.i. stair na gcomharthá scriófa agus na caoi a bhfreagraíonn siad dá chéile” (‘to avoid misunderstanding, it is important to specify that it is not at all the history of the sounds of Irish that is being recounted here, but the history of letters [spelling] alone, i.e. the history of written symbols and how they correspond to one another’). That said, even if these spellings do not give us a snapshot of pronunciations at a particular point in the past, they undoubtedly reflect something about historical pronunciations.

In cases where modern dialects had diverged in terms of pronunciation, the spelling of the classical modern norm was retained (e.g. gabhar ‘goat’, /gáùr/ in Connaught and Munster, /go:ùr/ in Ulster). However, some changes resulted in certain dialectal pronunciations being underivable from the new spelling (e.g. cruaidhe > crua but pronounced /krúùaja/ in Ulster). As (Ó Murchú, 1977, p. 278) (and others have noted) “the bisyllabic forms are in no way derivable from the new spelling, whereas the monosyllabic forms are derivable by a simple rule from spellings such as cruaidhe or better * cruaiaghe, etc.”, and Ó Siadhail (1981, 73)
notes that “it must be stressed that examples such as páighe and truaighe are not isolated cases”. As (Ó Murchú, 1977, p. 278, 279) notes, “the new spelling reform represents a compromise between conservatism, on the one hand, and different approaches to reform, on the other, and, as with all compromises, it is not entirely consistent and few were completely satisfied with it.”

Adoption of the new standard was reinforced by the subsequent publication of two significant dictionaries that conformed to the new spelling rules. The first was an English-Irish dictionary, de Bhaldraithe (1959), the second, an Irish-English dictionary, Ó Dónaill (1977). Ó Dónaill (1977) serves as the de facto standard, as it fills many gaps in An Caighdeán Oifigiúil. Though out-of-date (modern terminology included in the national terminology database for Irish available at http://www.focal.ie/ is not to be found in these dictionaries), these two documents still constitute essential spelling and grammar references for Irish language students and scholars.

Over fifty years since the publication of An Caighdeán Oifigiúil, demand for a new, updated, and more comprehensive standard has never been greater as the number of official documents being translated into Irish has increased drastically since the signing into law of the Official Languages Act 2003\(^3\), and since Irish became an official language of the European Union in 2007. This demand for a new standard relates primarily to grammatical and terminological questions, rather than to questions of spelling. It seems therefore, that any new standard is unlikely to change any of the basic principles of Irish spelling and is likely only to suggest changes in how certain words are spelt if the current spelling no longer reflects how the words are pronounced in any of the living dialects. The results of this research would indicate that even if the spelling changes, as long as it changes

in a regular rule-governed way, TTS can be adapted to handle them. And as we have discussed, TTS-based applications, are vital for accessibility. TTS is the main way visually impaired people read and get access to information. They are also a vital communication tool for people with voice disorders.

4.5 Phonetisation and the new standard

The new standard contains a numbers of features which are quite useful to synthesis. For example, the system is morphologically transparent: the root of a word and its internal structure are clear. This can be illustrated by the orthographic conventions used when writing future tense and conditional mood verb forms. A particular class of future and conditional verb endings in Irish begin with the letter 'f'. This 'f' is not typically pronounced as [f] but rather denotes either the devoicing of the verb stem’s final voiced consonant (gearrfaimid [ɾ̪ɣ] ‘we will cut’, or thógfaimis [k] ‘we would take’), or else the insertion of a [h] sound if the verb root ends in a vowel (leadfaird [h] ‘will melt’). If the verb stem ends in a voiceless consonant, the ‘f’ is not pronounced (chloisfinn [c] ‘I would hear’). These words are thus morphologically transparent since the verb stem remains unchanged in the spelling, apart from lenition of lenitable initial consonants in the conditional\(^{35}\), something which is also morphologically transparent (tóg ‘take’ → thógfaimis ‘we would take’). This morphological transparency is useful to human readers and, importantly for our purposes, to TTS systems (as we will see in Chapter 5) in identifying the root word. This morphological transparency is

\(^{34}\) 'f' is pronounced as [f] in some dialects with autonomous verb forms, e.g. gearrfar [f] ‘will be cut’ or gearrfaí [f] ‘would be cut’ (by someone unspecified).

\(^{35}\)There is one exception to this rule: the irregular verb abair ‘say.2SG.IMP.’ has the alternate stem ‘déar’ in the conditional that remains unlenited despite /dʰ/ being lenitable.
helpful to human readers (including learners) as it has the effect of keeping words recognisable after they have undergone a change in form. It is also of benefit when building a phonetisation module for TTS synthesis as it may facilitate effective morphological analysis, another important component of TTS synthesis systems (see § 5.10.1). And it may also be advantageous for multilingual (e.g. English and Irish) synthesis as it might allow a machine to differentiate between English and Irish text (Romsdorfer & Pfister, 2007).

The new standard also retained the convention of using an additional letter to represent a mutated (lenited or nasalised/eclipsed) root consonant, which makes these forms morphologically transparent. For example, the word prionsabal ‘principle’ is pronounced with an initial [pj], but when nasalised/eclipsed, as in i bprionsabal ‘in principle’, it is pronounced with an initial [by]. Rather than the ⟨p⟩ being replaced with a ⟨b⟩, it is prefixed with ‘b’ and becomes ⟨bp⟩. The word-initial ⟨bp⟩ always represents a mutated consonant. Similarly, when the name Sadhbh occurs in the vocative case (A Shadhbh!), the initial consonant is lenited and is pronounced [hoiv'], but the base form, which begins with ‘s’, is still represented orthographically in Shadhbh, despite the absence of an initial ‘s’ sound. Almost all initial mutation two-letter graphemes unambiguously indicate a mutated consonant. It is important to note that not all Celtic languages deal with mutation in the same way in the spelling. For example, in the case of so-called soft mutation in Welsh, the initial consonant is replaced in the spelling by the mutated consonant. Soft mutation occurs in a number of contexts, for example, singular feminine nouns occurring after the definite article y. For example, the word merch ‘girl’ is written with an initial ‘f’ in the phrase y ferch ‘the girl’. This poses a potential difficulty: it may not be obvious to a learner or to a machine

36The only exception to this rule is a small closed class of words, mainly adverbs such as thios ‘down’, thuas ‘up’, etc.
that *ferch* is the mutated form of the word *merch*. Breton, Manx, and Cornish orthography also indicate mutation in this way, by replacing the initial consonant letter with a different letter.

In addition, if the new standard spelling had incorporated the English-influenced graphemes suggested by the Letiriú Shímplí proponents (e.g., ⟨v⟩ for ⟨mh⟩), this may have complicated one of the longterm goals of Irish TTS synthesis – a synthesis system that is able to deal with the code-mixing between Irish and English that is common in Irish speech and writing. Irish TTS systems will eventually need to deal with Irish and English input co-occurring (Ní Chasaide et al., 2006) and the orthographic distance between Irish and English should facilitate this task.

### 4.6 Reading rules for Irish

#### 4.6.1 Consonantal graphemes

As discussed in Chapter 3, consonants in Irish are either phonemically velarised or phonemically palatalised. This phonemic distinction is difficult to represent in a spelling system that uses a Roman-based alphabet with only 21 consonant symbols and no consonant diacritics. Irish has up to 34 consonants that can be organised into velarised-palatalised pairs. Additionally, up to 13 devoiced sonorant segments may also occur in certain grammatical morphemes.

As consonantal diacritics are not used in Irish spelling, the phonemically velarised/phonemically palatalised distinction is indicated by the adjacent vowel symbol. If a consonant symbol occurs next to an ‘a’, ‘á’, ‘o’, ‘ó’, ‘u’, or ‘ú’, the
consonant being represented is phonemically velarised. If a consonant symbol occurs next to an ‘e’, ‘é’, ‘i’, or ‘í’, the consonant being represented is phonemically palatalised. For example, the Irish girl’s name Sorcha starts with the phoneme /sˠ/, whereas the Irish boy’s name Seán starts with the phoneme /c/. These two sibilant consonants are clearly two separate phonemes (/sˠ/~ /c/, údarás ‘authority’ /uːd̪ˠəɾˠaːsˠ/ údarás ‘authorities’ /uːd̪ˠəɾˠaːɾˠ/)

The grouping of consonant phonemes into velarised-palatalised pairs is important when describing how to read Irish aloud since both siblings in each pair are represented using the same grapheme in the spelling.

If consonants appear in clusters, the consonants of the cluster will agree in terms of phonetic quality. Given a system where consonants and consonant clusters are represented using a combination of both consonant and vowel letters, the system of orthographic vowel sequences that encode both vowel and adjacent consonant quality in Irish is complex. For example, the word sneachta ‘snow’ starts with ‘snea’. This is pronounced /p̪ˠənˠa/, where the letter ‘e’ encodes the fact that both ⟨s⟩ and ⟨n⟩ denote phonemically palatalised consonants. The vocalic graphemic system is described in § 4.6.2. The consonantal graphemic system is illustrated in Tables 4.1–4.13 below. The tables are organised alphabetically according to the first letter of the grapheme.

---

37There are some exceptions. Word-initial ⟨s⟩ is always phonemically velarised before ⟨f⟩, ⟨m⟩, or ⟨p⟩ (e.g. sféar ‘sphere’, spéir ‘sky’), and ⟨r⟩ is always phonemically velarised word-initially and before ⟨d⟩, ⟨l⟩, ⟨n⟩, ⟨t⟩ and ⟨s⟩.

38While the letter ‘e’ is treated as being part of the vocalic grapheme ⟨ea⟩, it can also be thought of as a part of a discontinuous consonant grapheme. In a word like sneachta /cp̪ˠ̚æːt̪ˠə/ ‘snow’, the ‘e’ signals that the ‘s’ is phonemically palatalised. However, we use the simplified notation ⟨s⟩ for the grapheme that corresponds to both /sˠ/ and /c/.
### Table 4.1: 'b'-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨b⟩</td>
<td>/b̪/</td>
<td>gob</td>
<td>/gɐb̪/</td>
</tr>
<tr>
<td>⟨b⟩</td>
<td>/ḅ/</td>
<td>forib</td>
<td>/fɐɾi̯ḅ/</td>
</tr>
<tr>
<td>⟨bf⟩</td>
<td>/p̪/</td>
<td>gobfaidh</td>
<td>/gɐp̪aɪd̪/</td>
</tr>
<tr>
<td>⟨bf⟩</td>
<td>/p̣/</td>
<td>foribfadh</td>
<td>/fɐɾi̯bfad̪/</td>
</tr>
<tr>
<td>⟨bh⟩</td>
<td>/w̪/</td>
<td>diobh</td>
<td>/dɪw̪/</td>
</tr>
<tr>
<td>⟨bh⟩</td>
<td>/ẉ/</td>
<td>foribthe</td>
<td>/fɐɾi̯ẉ/</td>
</tr>
<tr>
<td>⟨bhf⟩</td>
<td>/p̪/</td>
<td>diobhfaidh</td>
<td>/gɐp̪aɪd̪/</td>
</tr>
<tr>
<td>⟨bhf⟩</td>
<td>/p̣/</td>
<td>foribfadh</td>
<td>/fɐɾi̯bfad̪/</td>
</tr>
<tr>
<td>⟨bth⟩</td>
<td>/p̪/</td>
<td>foribthe</td>
<td>/fɐɾi̯ẉ/</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨c⟩</td>
<td>/k/</td>
<td>ioc</td>
<td>/ɪk/</td>
</tr>
<tr>
<td>⟨c⟩</td>
<td>/c/</td>
<td>feic</td>
<td>/fɛɪk/</td>
</tr>
<tr>
<td>⟨cf⟩</td>
<td>/k/</td>
<td>iocfaidh</td>
<td>/ɪkəɪd̪/</td>
</tr>
<tr>
<td>⟨cf⟩</td>
<td>/c/</td>
<td>feicfidh</td>
<td>/fɛɪkɪd̪/</td>
</tr>
<tr>
<td>⟨ch⟩</td>
<td>/x/</td>
<td>múch</td>
<td>/mʊx/</td>
</tr>
<tr>
<td>⟨ch⟩</td>
<td>/x̪/</td>
<td>stroich</td>
<td>/sтроi̯x/</td>
</tr>
<tr>
<td>⟨chf⟩</td>
<td>/x/</td>
<td>múchfaidh</td>
<td>/mʊxəɪd̪/</td>
</tr>
<tr>
<td>⟨chf⟩</td>
<td>/x̪/</td>
<td>stroichfadh</td>
<td>/sтрои̯xəɪd̪/</td>
</tr>
<tr>
<td>⟨cth⟩</td>
<td>/k/</td>
<td>ioctha</td>
<td>/ɪkə/</td>
</tr>
<tr>
<td>⟨cth⟩</td>
<td>/c/</td>
<td>feicthe</td>
<td>/fɛɪkə/</td>
</tr>
</tbody>
</table>

### Table 4.2: 'c'-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨d⟩</td>
<td>/d̪/</td>
<td>stad</td>
<td>/sød̪/</td>
</tr>
<tr>
<td>⟨d⟩</td>
<td>/ḍ/</td>
<td>sēid</td>
<td>/céiḍ/</td>
</tr>
<tr>
<td>⟨df⟩</td>
<td>/t̪/</td>
<td>stadaidh</td>
<td>/søt̪aɪd̪/</td>
</tr>
<tr>
<td>⟨df⟩</td>
<td>/ṭ/</td>
<td>sēidfadh</td>
<td>/céɪḍə/</td>
</tr>
<tr>
<td>⟨dh⟩</td>
<td>/y/</td>
<td>dhún</td>
<td>/yʊn̪/</td>
</tr>
<tr>
<td>⟨dh⟩</td>
<td>/j/</td>
<td>dhiúl</td>
<td>/jʊl̪/</td>
</tr>
<tr>
<td>⟨dt⟩</td>
<td>/t̪/</td>
<td>stada</td>
<td>/søt̪aɪd̪/</td>
</tr>
<tr>
<td>⟨dt⟩</td>
<td>/ṭ/</td>
<td>sēidte</td>
<td>/céɪḍə/</td>
</tr>
</tbody>
</table>

### Table 4.3: 'd’-initial consonantal graphemes.
<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨f⟩</td>
<td>/f/</td>
<td>feá</td>
</tr>
<tr>
<td>⟨f⟩</td>
<td>/f/</td>
<td>fáth</td>
</tr>
<tr>
<td>⟨ fh⟩</td>
<td>∅</td>
<td>d'fhág</td>
</tr>
</tbody>
</table>

Table 4.4: 'f'-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨g⟩</td>
<td>/g/</td>
<td>tóg</td>
</tr>
<tr>
<td>⟨g⟩</td>
<td>/g/</td>
<td>treig</td>
</tr>
<tr>
<td>⟨ gf⟩</td>
<td>/k/</td>
<td>tóglaidh</td>
</tr>
<tr>
<td>⟨gf⟩</td>
<td>/c/</td>
<td>tréigfeadh</td>
</tr>
<tr>
<td>⟨ gh⟩</td>
<td>/y/</td>
<td>ghalaigh</td>
</tr>
<tr>
<td>⟨gh⟩</td>
<td>/j/</td>
<td>gheall</td>
</tr>
<tr>
<td>⟨ gth⟩</td>
<td>/k/</td>
<td>tóglaidh</td>
</tr>
<tr>
<td>⟨gth⟩</td>
<td>/c/</td>
<td>tréigfeadh</td>
</tr>
</tbody>
</table>

Table 4.5: 'g'-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨h⟩</td>
<td>/h/</td>
<td>hat</td>
</tr>
</tbody>
</table>

Table 4.6: 'h'-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨l⟩</td>
<td>/l/</td>
<td>diol</td>
</tr>
<tr>
<td>⟨l⟩</td>
<td>/l/</td>
<td>buail</td>
</tr>
<tr>
<td>⟨ lf⟩</td>
<td>/l/</td>
<td>diolaidh</td>
</tr>
<tr>
<td>⟨lf⟩</td>
<td>/l/</td>
<td>bualidh</td>
</tr>
<tr>
<td>⟨ ll⟩</td>
<td>/l/</td>
<td>toll</td>
</tr>
<tr>
<td>⟨ll⟩</td>
<td>/l/</td>
<td>fill</td>
</tr>
<tr>
<td>⟨ llf ⟩</td>
<td>/l/</td>
<td>tollaidh</td>
</tr>
<tr>
<td>⟨llf⟩</td>
<td>/l/</td>
<td>fillidh</td>
</tr>
</tbody>
</table>

Table 4.7: 'l'-initial consonantal graphemes.
<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Table 4.8: ‘m’-initial consonantal graphemes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m)</td>
<td>/m/</td>
<td>crom</td>
<td>‘crouch’ /krʌm/</td>
</tr>
<tr>
<td>(m)</td>
<td>/m/</td>
<td>léim</td>
<td>‘jump’ /lem/</td>
</tr>
<tr>
<td>(mf)</td>
<td>/mf/</td>
<td>cromfaidh</td>
<td>‘will crouch’ /krʌmfɑid/</td>
</tr>
<tr>
<td>(mf)</td>
<td>/mf/</td>
<td>léimfidh</td>
<td>‘will jump’ /lemifɑid/</td>
</tr>
<tr>
<td>(mh)</td>
<td>/w/</td>
<td>snámh</td>
<td>‘swim’ /snaːm/</td>
</tr>
<tr>
<td>(mh)</td>
<td>/v/</td>
<td>nimh</td>
<td>‘poison’ /nɪm/</td>
</tr>
<tr>
<td>(mhf)</td>
<td>/f/</td>
<td>snámhfaidh</td>
<td>‘will swim’ /snaːmfɑid/</td>
</tr>
<tr>
<td>(mth)</td>
<td>/m/</td>
<td>cromtha</td>
<td>‘crouched’ /krʌmfɑid/</td>
</tr>
<tr>
<td>(mth)</td>
<td>/m/</td>
<td>léimthe</td>
<td>‘jumped’ /lemifɑid/</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Table 4.9: ‘n’-initial consonantal graphemes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>/n/</td>
<td>lión</td>
<td>‘fill’ /liːn/</td>
</tr>
<tr>
<td>(n)</td>
<td>/n/</td>
<td>bain</td>
<td>‘remove’ /bəin/</td>
</tr>
<tr>
<td>(nd)</td>
<td>/nd/</td>
<td>á ndúnadh</td>
<td>‘being sold’ /aː nduːnɑd/</td>
</tr>
<tr>
<td>(nd)</td>
<td>/nd/</td>
<td>á ndíol</td>
<td>‘being sold’ /aː n̪diːl/</td>
</tr>
<tr>
<td>(nf)</td>
<td>/nf/</td>
<td>liónfaidh</td>
<td>‘will fill’ /liːnfaid/</td>
</tr>
<tr>
<td>(nf)</td>
<td>/nf/</td>
<td>bainfidh</td>
<td>‘will remove’ /bənfaid/</td>
</tr>
<tr>
<td>(ng)</td>
<td>/ŋ/</td>
<td>ung</td>
<td>‘anoint’ /ʌŋ/</td>
</tr>
<tr>
<td>(ng)</td>
<td>/ŋ/</td>
<td>ling</td>
<td>‘spring’ /lɪŋ/</td>
</tr>
<tr>
<td>(ngf)</td>
<td>/ŋ/</td>
<td>ungfaidh</td>
<td>‘will anoint’ /ʌŋfaid/</td>
</tr>
<tr>
<td>(ngf)</td>
<td>/ŋ/</td>
<td>lingfídh</td>
<td>‘will spring’ /lɪŋfaid/</td>
</tr>
<tr>
<td>(nn)</td>
<td>/nn/</td>
<td>bronn</td>
<td>‘award’ /b迥ɔː/</td>
</tr>
<tr>
<td>(nn)</td>
<td>/nn/</td>
<td>cinn</td>
<td>‘decide’ /kɪn/</td>
</tr>
<tr>
<td>(nnf)</td>
<td>/nn/</td>
<td>bronntaidh</td>
<td>‘will award’ /b迥ɔːntaid/</td>
</tr>
<tr>
<td>(nnf)</td>
<td>/nn/</td>
<td>cinnfídh</td>
<td>‘will decide’ /kɪnfaid/</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Table 4.10: ‘p’-initial consonantal graphemes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p)</td>
<td>/p/</td>
<td>cearp</td>
<td>‘appoint’ /kaːp/</td>
</tr>
<tr>
<td>(p)</td>
<td>/p/</td>
<td>teip</td>
<td>‘fail’ /tɪp/</td>
</tr>
<tr>
<td>(pf)</td>
<td>/p/</td>
<td>ceapfaidh</td>
<td>‘will appoint’ /kaːpfaid/</td>
</tr>
<tr>
<td>(pf)</td>
<td>/p/</td>
<td>teipfídh</td>
<td>‘will fail’ /tɪpfɑid/</td>
</tr>
<tr>
<td>(ph)</td>
<td>/f/</td>
<td>phoc</td>
<td>‘hit’ /fək/</td>
</tr>
<tr>
<td>(ph)</td>
<td>/f/</td>
<td>phioc</td>
<td>‘picked’ /fɪək/</td>
</tr>
<tr>
<td>(pth)</td>
<td>/p/</td>
<td>ceaptha</td>
<td>‘appointed’ /kaːpta/</td>
</tr>
<tr>
<td>(pth)</td>
<td>/p/</td>
<td>teipthe</td>
<td>‘failed’ /tɪpθe/</td>
</tr>
</tbody>
</table>
### Table 4.11: 'r'-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Irish Spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r)</td>
<td>/rɪ/</td>
<td>tuar</td>
<td>/tuəɾɪ/</td>
</tr>
<tr>
<td>(r)</td>
<td>/rɪ/</td>
<td>oir</td>
<td>/ɔɾɪ/</td>
</tr>
<tr>
<td>(rf)</td>
<td>/rɪ/</td>
<td>tuarfaidh</td>
<td>/tuəɾ法人/</td>
</tr>
<tr>
<td>(rf)</td>
<td>/rɪ/</td>
<td>oirfidh</td>
<td>/ɔɾ法人/</td>
</tr>
<tr>
<td>(rr)</td>
<td>/rɪ/</td>
<td>bearr</td>
<td>/bəɾɪ/</td>
</tr>
<tr>
<td>(rr)</td>
<td>/rɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(rrf)</td>
<td>/rɪ/</td>
<td>bearrfaidh</td>
<td>/bɾ法人/</td>
</tr>
<tr>
<td>(rrf)</td>
<td>/rɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(rrth)</td>
<td>/rɪ/</td>
<td>bearrtha</td>
<td>/bɾ法人/</td>
</tr>
<tr>
<td>(rrth)</td>
<td>/rɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(rth)</td>
<td>/rɪ/</td>
<td>tuartha</td>
<td>/tuəɾ法人/</td>
</tr>
<tr>
<td>(rth)</td>
<td>/rɪ/</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

### Table 4.12: ‘s’-initial consonantal graphemes.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Example</th>
<th>Irish Spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s)</td>
<td>/sɪ/</td>
<td>cas</td>
<td>/kæsɪ/</td>
</tr>
<tr>
<td>(s)</td>
<td>/sɪ/</td>
<td>bris</td>
<td>/bɾɪs/</td>
</tr>
<tr>
<td>(sf)</td>
<td>/sɪ/</td>
<td>casfaidh</td>
<td>/kæs法人/</td>
</tr>
<tr>
<td>(sf)</td>
<td>/sɪ/</td>
<td>brisfidh</td>
<td>/bɾ法人/</td>
</tr>
<tr>
<td>(sh)</td>
<td>/ʃɪ/</td>
<td>sheachaid</td>
<td>/ʃɛɾ法人/</td>
</tr>
<tr>
<td>(sh)</td>
<td>/ʃɪ/</td>
<td>shocraigh</td>
<td>/ʃakr法人/</td>
</tr>
<tr>
<td>(sh)</td>
<td>/ʃɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(shl)</td>
<td>/ʃɪ/</td>
<td>shlog</td>
<td>/ʃɬøɡ/</td>
</tr>
<tr>
<td>(shl)</td>
<td>/ʃɪ/</td>
<td>shleamhnaigh</td>
<td>/ʃɬɨɾ法人/</td>
</tr>
<tr>
<td>(shm)</td>
<td>/ʃɪ/</td>
<td>shleamhnaigh</td>
<td>/ʃɬɨɾ法人/</td>
</tr>
<tr>
<td>(shm)</td>
<td>/ʃɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(shr)</td>
<td>/ʃɪ/</td>
<td>shruthlaigh</td>
<td>/ʃɾɬøɡ/</td>
</tr>
<tr>
<td>(shr)</td>
<td>/ʃɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(sh)</td>
<td>/ʃɪ/</td>
<td>shnamh</td>
<td>/ʃɬeiw/</td>
</tr>
<tr>
<td>(sh)</td>
<td>/ʃɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(s)</td>
<td>/ʃɪ/</td>
<td>shnamh</td>
<td>/ʃɬeiw/</td>
</tr>
<tr>
<td>(s)</td>
<td>/ʃɪ/</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(s)</td>
<td>/ʃɪ/</td>
<td>spéir</td>
<td>/ʃɬɬɪɾ/</td>
</tr>
</tbody>
</table>

---

Phonetisation and Irish spelling
Grapheme | Phoneme | Example  
--- | --- | ---  
\(\langle t\rangle\) | /t\|^v/ | at 'swell' /at\|^v/  
\(\langle t\rangle\) | /t\|^i/ | tit 'fall' /t\|^i/  
\(\langle tf\rangle\) | /t\|^y/ | atfaidh 'will swell' /at\|^y/\(\alpha\)/  
\(\langle tf\rangle\) | /t\|^i/ | titfidh 'will fall' /t\|^i/\(\alpha\)/  
\(\langle th\rangle\) | /h/ | thit 'fell' /h/\(t\)/  
\(\langle thf\rangle\) | /h/ | leathfaidh 'will dilate' /h/\(\alpha\)/\(a\)/  
\(\langle thf\rangle\) | /\(\iota\)/ | crithfidh 'will vibrate' /\(\iota\)/\(\iota\)/\(\alpha\)/  
\(\langle thl\rangle\) | /\(\iota\)/ | a thlí 'his tongs' /\(\iota\)/\(\iota\)/\(u\)/  
\(\langle thl\rangle\) | /\(\imath\)/ | — —  
\(\langle thn\rangle\) | /\(\imath\)/ | thnáith 'collapsed' /\(\imath\)/\(\imath\)/\(a\)/  
\(\langle thn\rangle\) | /\(\imath\)/ | — —  
\(\langle thr\rangle\) | /\(\varepsilon\)/ | a thrá 'his beach' /\(\varepsilon\)/\(\varepsilon\)/\(a\)/  
\(\langle thr\rangle\) | /\(\iota\)/ | a théad 'his flock' /\(\iota\)/\(\varepsilon\)/\(\varepsilon\)/\(a\)/  

Table 4.13: ‘t’-initial consonantal graphemes.

### 4.6.2 Vocalic graphemes

As discussed in Chapter 3, there are up to twelve vowel phonemes in Irish, depending on the dialect. Apart from schwa, /\(\alpha\)/ (realised as [\(\alpha\)] and [\(i\)]) and /\(\varepsilon\)/, the vowels fall into pairs based on the phonemic distinction between long and short vowels. The Irish vowel long-short phoneme pairs are shown in Figure 4.14.

```
Front   Mid   Back  
High    i:  i  u:  u  
Mid      e  e   o:  \(\alpha\)  
Low      a:  a
```

Table 4.14: Irish vowel long-short phoneme pairs

In addition to the monophthongs, and as discussed in Chapter 3, there are three or four diphthong phonemes in Irish, depending on the dialect: /\(\alpha\)\(\alpha\)/, /\(\alpha\)\(\alpha\)/, /\(\alpha\)\(\varepsilon\)/, and /\(\varepsilon\)\(\alpha\)/. In the spelling, these diphthongs are written as ‘a(i)dh’, ‘(e)amh(ai)’,
'ia(i)', and 'ua(i)' respectively.

The vowel graphemes of Irish are presented in Tables 4.15–4.28 below, but before presenting these tables, I discuss some aspects of the orthographic representation of vowels. The letters 'a', 'e', 'i', 'o', and 'u' are used to represent the short vowels. The same symbols with acute accents, i.e. 'á, é, í, ó, ú', are used to represent their long vowel counterparts. These vowels can also be represented by more complex orthographic sequences. For example, /et/ is sometimes represented with 'é(i)' (e.g. séid 'blow'), but other times is represented with 'ae(i)' (e.g. Gaeilge 'Irish') sometimes followed. This is also the case with /ii/, which can be represented with either '(u)i(o)' (e.g. cuíosach 'fairly') or 'ao(i)' (e.g. faoi 'under'). In this case, the 'ao' sequence historically corresponded to a different phoneme /uː/, realised as a high mid unrounded vowel, which has today disappeared from almost all dialects, except certain sub-dialects in Donegal. For /oː/, the historically present fada (long mark) (ó) is nowadays conventionally omitted from over the 'o' when this vowel follows a phonemically palatalised consonant or occurs word initially in the sequence 'eo'. For example, words such as ceol 'music' are pronounced with a long /oː/. The only exceptions to this rule are the words seo [ɕo] 'this', anseo [an̪ˠoː] 'here', deoch [d̪wa] 'drink', eochair [a̯kʰaɾ] 'key', in which the 'eo' sequence represents the short vowel phoneme /ʌ/.

As discussed earlier (§ 3.4), the unstressed short vowels in Irish are unstable and are reduced to /a/ ([a] or [ɪ]). This results in a lack of correspondence between the vowel letter and the pronunciation. However, if one bears in mind that for the most part, only the first syllable in Irish is stressed, this reduction rule is easy to apply when reading Irish aloud.  

39 Apart from a closed class of words (mostly adverbs) with an unstressed first syllable such as inniu [a̯nˠuː] 'today', inné [a̯nˠeː] 'yesterday', in compound words, in borrowings, and in certain conditions in Munster Irish, the stress is on the first syllable in Irish.
Another potential difficulty is caused by the vowel being obscured by the *glide letters*, as I will refer to them here (following Ó Murchú, 1977). Glide letters are the vowel letters inserted between vowels and consonants that differ in quality (as in *ceoil* ‘music’, and *sui* ‘sitting’). The front vowels represented by the letters ‘e’, ‘é’, ‘i’, and ‘í’ differ in quality from velarised or ‘back’ consonants (velarisation is a secondary articulation which takes place towards the back of the oral cavity), and likewise, the back vowels represented by the letters ‘a’, ‘á’, ‘o’, ‘ó’, ‘u’, and ‘ú’ differ in quality from palatalised or ‘front’ consonants (palatalisation is a secondary articulation which takes place towards the front of the oral cavity). So when, for example, a back vowel occurs between two palatalised consonants, it is flanked in the spelling by orthographic glide vowels on both sides, as shown in bold in the following example: *ceoil* ‘music.GEN.SG.’, transcribed phonemically as */koːl/\. The function of the orthographic glide vowel is to indicate the quality of the adjacent consonant given that it differs from the quality of the vowel. The pronoun *si* /ciː/ ‘she’, however, consists of a palatalised or ‘front’ ‘s’ followed by the front vowel /iː/. As they agree in place, there is no requirement for the insertion of an orthographic glide vowel. In the case of the verbal noun *sui* /ʃiː/ ‘sitting’, though, the initial ‘s’ is velarised or ‘back’. As this does differ in quality from the following vowel, an orthographic glide vowel (‘u’), is required to indicate consonant quality (phonemically velarised).

In this chapter on spelling, I focus primarily on the correspondences between the orthography and the phonology. It is important to note at this point however that the glide letters are not always simply orthographic markers of consonant quality (phonemically velarised vs. phonemically palatalised). In some cases they correspond to audible glides from a front consonant to a back vowel or vice versa (e.g. *beo* ‘alive’ [bʲjoː]), or from a back consonant to a front vowel or vice versa.
(e.g. *súi* ‘sit’ [sˠuːiː]).

This situation is further complicated in the case of the short low vowel /a/. This vowel is unstable and its quality changes to mirror the quality of the adjacent consonant (see § 3.4). In this case, therefore, a glide letter indicates consonantal quality but also serves to indicate a change in vowel quality rather than an actual glide from consonant to vowel. For example, in a word like *gal* /gaɭ/ ‘steam’, both the initial consonant and the vowel are articulated towards the back of the mouth, and there is therefore no need for a glide vowel: *gal* [gaɭ]. However, the word *geal* /jaɭ/ ‘bright’ is pronounced [jaɭ]: here there is no audible glide vowel from the front consonant to the back vowel, but rather a vowel shift from back to front.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(eal)</td>
<td>/a/</td>
<td>CVC˘</td>
<td><em>leíc</em> ‘lac’ /lək/</td>
</tr>
<tr>
<td>(ea)</td>
<td>/a/</td>
<td>CVC˘</td>
<td><em>leac</em> ‘flagstone’ /lək/</td>
</tr>
<tr>
<td>(ai)</td>
<td>/a/</td>
<td>CVC˘</td>
<td><em>baíc</em> ‘nape’ /bək/</td>
</tr>
<tr>
<td>(a)</td>
<td>/a/</td>
<td>CVC˘</td>
<td><em>bac</em> ‘hindrance’ /bək/</td>
</tr>
</tbody>
</table>

Table 4.15: Graphemes corresponding to the phoneme /a/. Here and in following tables, the third column in each table pertains to the consonantal context in which the vowel is occurring. ‘C˘’ stands for any velarised consonant, and C˘ stands for any palatalised consonant. ‘V’ represents the vowel in context.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(e(i))</td>
<td>/ɛ/</td>
<td>CVC˘</td>
<td><em>deís</em> ‘opportunity’ /dəɛ/</td>
</tr>
<tr>
<td>(e)</td>
<td>/ɛ/</td>
<td>CVC˘</td>
<td>—</td>
</tr>
<tr>
<td>(ue)</td>
<td>/ɛ/</td>
<td>CVC˘</td>
<td><em>bhuel</em> ‘well!’ /vəl/</td>
</tr>
<tr>
<td>—</td>
<td>/ɛ/</td>
<td>CVC˘</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 4.16: Graphemes corresponding to the phoneme /e/.
<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>/i/</td>
<td>C'VCJ</td>
<td>criticiúil</td>
</tr>
<tr>
<td>(io)</td>
<td>/i/</td>
<td>C'VCv</td>
<td>crios</td>
</tr>
<tr>
<td>(ui)</td>
<td>/i/</td>
<td>CVVCi</td>
<td>crudín</td>
</tr>
</tbody>
</table>

Table 4.17: Graphemes corresponding to the phoneme /i/.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>/a/</td>
<td>C'VCJ</td>
<td></td>
</tr>
<tr>
<td>(eo)</td>
<td>/a/</td>
<td>C'VCv</td>
<td>deoch</td>
</tr>
<tr>
<td>(oi)</td>
<td>/a/</td>
<td>CVVCi</td>
<td>dois</td>
</tr>
<tr>
<td>(o)</td>
<td>/a/</td>
<td>CVVCv</td>
<td>dos</td>
</tr>
</tbody>
</table>

Table 4.18: Graphemes corresponding to the phoneme /o/.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>/u/</td>
<td>C'VCJ</td>
<td></td>
</tr>
<tr>
<td>(iu)</td>
<td>/u/</td>
<td>C'VCv</td>
<td>fiuch</td>
</tr>
<tr>
<td>—</td>
<td>/y/</td>
<td>CVVCi</td>
<td></td>
</tr>
<tr>
<td>(u)</td>
<td>/y/</td>
<td>CVVCv</td>
<td>bun</td>
</tr>
</tbody>
</table>

Table 4.19: Graphemes corresponding to the phoneme /u/.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(eá)</td>
<td>/a:/</td>
<td>C'VCv</td>
<td>meán</td>
</tr>
<tr>
<td>(eá)</td>
<td>/a:/</td>
<td>C'VCv</td>
<td>meán</td>
</tr>
<tr>
<td>(áí)</td>
<td>/a:/</td>
<td>CVVCi</td>
<td>cáig</td>
</tr>
<tr>
<td>(á)</td>
<td>/a:/</td>
<td>CVVCv</td>
<td>cáig</td>
</tr>
</tbody>
</table>

Table 4.20: Graphemes corresponding to the phoneme /a:/.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(é(i))</td>
<td>/e:/</td>
<td>C'VCJ</td>
<td>béil</td>
</tr>
<tr>
<td>(éa)</td>
<td>/e:/</td>
<td>C'VCv</td>
<td>béal</td>
</tr>
<tr>
<td>(æei)</td>
<td>/e:/</td>
<td>CVVCj</td>
<td>Gaeil</td>
</tr>
<tr>
<td>(æe)</td>
<td>/e:/</td>
<td>CVVCv</td>
<td>Gaeil</td>
</tr>
</tbody>
</table>

Table 4.21: Graphemes corresponding to the phoneme /e/.
<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>{i}</td>
<td>[iː]</td>
<td>CVC</td>
<td>'of net'</td>
</tr>
<tr>
<td>{io}</td>
<td>[iː]</td>
<td>CVC</td>
<td>'net'</td>
</tr>
<tr>
<td>{af}, {uf}, {aói}</td>
<td>[iː]</td>
<td>CVC</td>
<td>'of site'</td>
</tr>
<tr>
<td>{aío}, {úio}, {ao}</td>
<td>[iː]</td>
<td>CVC</td>
<td>'site'</td>
</tr>
</tbody>
</table>

Table 4.22: Graphemes corresponding to the phoneme [iː].

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>{eoi}</td>
<td>[oː]</td>
<td>CVC</td>
<td>'of music'</td>
</tr>
<tr>
<td>{eo}, {ió}</td>
<td>[oː]</td>
<td>CVC</td>
<td>'music'</td>
</tr>
<tr>
<td>{ói}</td>
<td>[oː]</td>
<td>CVC</td>
<td>'of sod'</td>
</tr>
<tr>
<td>{ó}</td>
<td>[oː]</td>
<td>CVC</td>
<td>'sod'</td>
</tr>
</tbody>
</table>

Table 4.23: Graphemes corresponding to the phoneme [oː].

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>{iúi}</td>
<td>[uː]</td>
<td>CVC</td>
<td>'of thickness'</td>
</tr>
<tr>
<td>{iú}</td>
<td>[uː]</td>
<td>CVC</td>
<td>'thickness'</td>
</tr>
<tr>
<td>{uí}</td>
<td>[uː]</td>
<td>CVC</td>
<td>'resolutions'</td>
</tr>
<tr>
<td>{ú}</td>
<td>[uː]</td>
<td>CVC</td>
<td>'resolution'</td>
</tr>
</tbody>
</table>

Table 4.24: Graphemes corresponding to the phoneme [uː].

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>{eadhai}</td>
<td>[aː]</td>
<td>CVC</td>
<td>'troop'</td>
</tr>
<tr>
<td>{eadh(a)}</td>
<td>[aː]</td>
<td>CVC</td>
<td>'whey'</td>
</tr>
<tr>
<td>{adhai}</td>
<td>[aː]</td>
<td>CVC</td>
<td>'of sight'</td>
</tr>
<tr>
<td>{adh(a)}</td>
<td>[aː]</td>
<td>CVC</td>
<td>'sight'</td>
</tr>
</tbody>
</table>

Table 4.25: Graphemes corresponding to the phoneme [aː].

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>{eamhai}</td>
<td>[əuː]</td>
<td>CVC</td>
<td>'demons'</td>
</tr>
<tr>
<td>{eamh(a)}</td>
<td>[əuː]</td>
<td>CVC</td>
<td>'demon'</td>
</tr>
<tr>
<td>{amhai}</td>
<td>[əuː]</td>
<td>CVC</td>
<td>'trunks'</td>
</tr>
<tr>
<td>{amh(a)}</td>
<td>[əuː]</td>
<td>CVC</td>
<td>'trunk'</td>
</tr>
</tbody>
</table>

Table 4.26: Graphemes corresponding to the phoneme [əuː].
Phonetisation and Irish spelling

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨iai⟩</td>
<td>/iə/</td>
<td>CV:CJ</td>
<td>rian</td>
</tr>
<tr>
<td>⟨ia⟩</td>
<td>/iə/</td>
<td>CV:CV'</td>
<td>rian</td>
</tr>
<tr>
<td>—</td>
<td>/iə/</td>
<td>CV:CV'</td>
<td>—</td>
</tr>
<tr>
<td>—</td>
<td>/iə/</td>
<td>CV:CV'</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 4.27: Graphemes corresponding to the phoneme /iə/.  

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Context</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>⟨uai⟩</td>
<td>/uə/</td>
<td>CV:CV'</td>
<td>cuain</td>
</tr>
<tr>
<td>⟨ua⟩</td>
<td>/uə/</td>
<td>CV:CV'</td>
<td>cuan</td>
</tr>
</tbody>
</table>

Table 4.28: Graphemes corresponding to the phoneme /uə/.  

As we have seen, the reading rules of Irish are quite complex. Adjacent vowel symbols must be taken into account when pronouncing consonants, and vowel graphemes encode information about the quality of the vowel and the adjacent consonant(s). Despite all this however, the system is fairly regular and governed by a system of rules with few exceptions. We exploit this transparency in our development of a phonetiser for Irish described in the next chapter.
Chapter 5

A modular multi-dialect phonetiser for Irish

5.1 A standard approach to phonetisation

As described in detail in Chapter 2, phonetisation is the process in which normalised text (see § 2.4.1) is converted to a phonemic representation, as shown in Figure 5.1. In this example, the phonetised output shown is the result of the input text *an 3ú mórsíul* ('the 3rd procession') being passed through a normaliser and then through a phonetiser.

The phonetised output in Figure 5.1 consists of a phone sequence with syllable boundaries marked with a full stop ('). Stress/no stress is marked with a number 0 (no stress), 1 (primary stress), or 2 (secondary stress) preceding the syllable. This sequence, or transcription, constitutes what will be searched for in the annotated speech corpus in corpus-based speech synthesis.
A modular multi-dialect phonetiser for Irish

INPUT TEXT: an 3ú mórsiuil ('the 3rd procession')

NORMALISER

NORMALISED TEXT: an trí mórsiuil

PHONETISER

PHONETISED TEXT: /O @ nn . 1 tj rj ii . 0 uu . 1 m oo r . 2 xj uu ll/

Figure 5.1: An schematic illustration of text being passed through a normaliser and a phonetiser in turn.

The reader will note that this transcription is not given using the International Phonetic Alphabet (IPA). The reason for this is that the computer software that performs the phonetisation needs to be able to understand the input, search for matching patterns in rule and dictionary files to work out the correct transcription of the input, and then match that transcription with annotations in a speech corpus. Although IPA symbols and diacritics can nowadays be machine-encoded using an encoding known as Unicode,\textsuperscript{40} Unicode compatibility is not widespread across computer software, programming languages, code editors, and operating systems. As this lack of compatibility has the potential to cause problems for the synthesis software, it is safer to work entirely with plain text or machine-readable (computer-readable) rule and dictionary files in the case of the phonetisation modules (as is the case with other modules in the Irish TTS system).

To deal with this issue, the phonetiser actually produces output using a phoneset that I call the Machine Readable Phonetic Alphabet for Irish (MRPAI). The name and approach is based on a similar phoneset developed for Welsh phonetisation (Williams, 1993, 1994a). The main difference between the MRPAI and

\textsuperscript{40}Unicode is a computing industry standard that allows computers to consistently represent and manipulate text expressed in most (and eventually all) of the world's writing systems, including the IPA.
the IPA is that no phonetic symbols, phonetic diacritics, accents, punctuation, or superscript characters are used in MRPAI. In the case of Irish, this amounts to palatalisation being marked as a ‘j’ as opposed to a ‘i’, ‘í’ or an apostrophe, and vowel length is marked by doubling the vowel symbol instead of using a ‘ː’ or a colon. Other symbols and diacritics such as [y], [i], and [ŋ] are also replaced. A full comparison between MRPAI and the IPA-based phonesets, as well the more traditional phonesets used by Celticists, can be found in Appendix A.41

Phonetisation is normally done in one or two steps. In the first step, the words in the input text are searched for in a pronunciation dictionary (or lexicon). An excerpt from a lexicon commonly used for American English is shown in Figure 5.2. This is a machine-readable alphabetised lexicon in which each line contains exactly one word and its pronunciation. The word and its pronunciation are separated by a tab space, and the pronunciation is a space-separated sequence of phone labels. These phone labels are from a commonly used American English machine-readable phoneset.42 After each vowel, there is a number 0, 1, or 2. An unstressed vowel is followed by a 0, a vowel with primary stress by a 1, and a vowel with secondary stress by a 2. Given this lexicon, were the proper name Liffey, as in the River Liffey, which runs through the centre of Dublin, to be synthesised, it would not be found in the lexicon. If this is the case, and the word being searched for is not found in the lexicon, the second step is applied.

41I attempted to develop an X-SAMPA (see http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm) phoneset for Irish, but it proved to be incompatible with the speech synthesis software we were using. It is included in Appendix A as a reference.
42This phoneset has 39 phonemes and is listed here: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
A modular multi-dialect phonetiser for Irish

LIFESAVING L AY1 F S EY2 V IH0 NG
LIFESPAN L AY1 F S P AE2 N
LIFESPANS L AY1 F S P AE2 N Z
LIFESTYLE L AY1 F S T AY2 L
LIFESTYLES L AY1 F S T AY2 L Z
LIFETIME L AY1 F T AY2 M
LIFETIME'S L AY1 F T AY2 M Z
LIFETIMES L AY1 F T AY2 M Z
LIFF L IH1 F
LIFFE L IH1 F
LIFFE'S L IH1 F S
LIFLAND L IH1 F L AH0 N D
LIPO L IY1 F OWO
LIFORD L IH1 F ERO D
LIFSEY L IH1 F S IY0
LIFSHITZ L IH1 F SH IH0 T S
LIFSON L IH1 F S AA2 N
LIFSON'S L IH1 F S AA2 N Z
LIFT L IH1 F T
LIFTED L IH1 F T AH0 D

Figure 5.2: An excerpt from the Carnegie Mellon University (CMU) Pronouncing Dictionary for North American English (presented earlier as Figure 2.4).

The second step involves the application of a set of letter-to-sound (LTS) rules to derive the pronunciation of a word not in the lexicon. English LTS rules might have a rule such as: ey/._# → [i] (# stands for a word boundary), which would convert the word-final grapheme (ey) in Liffey, and so on. In many text-to-speech (TTS) systems, LTS rules have only a limited application as the lexicon is usually quite large. For example, the excerpt in Figure 5.2 is from the Carnegie Mellon University (CMU) Pronouncing Dictionary for North American English, which is constantly being enlarged: the current version contains roughly 133,700 words and their pronunciations.43

43The current CMU dictionary is available to download from: https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict/
5.2 Methods for developing LTS rules

There are two methods for developing LTS rules: automatically producing rules using automatic machine learning methods and writing rules by hand. Hand-written re-write (mapping) rules convert graphemes to phones. The orthographic sequence on the left-hand side of a rule is mapped to a phone sequence on the right-hand side of a rule. Context is taken into account when matching with the left-hand side of a rule, and therefore the rules are said to be "context-dependent". An example of a rule from the current system that says "if the orthographic sequence 'gc' occurs word initially, it is pronounced as a /g/ (G)"., is shown in (1). The ordering of the rules is critical with more specific rules coming before more general rules. Hand-written rules work well for languages with transparent orthographies, that is, regular letter-to-sound correspondence. For example, Spanish has one-to-one letter-to-sound correspondences for vowels and simple correspondences for consonants (e.g., plátano ['platano] 'banana'; juntos ['xuntos] 'together'; dañino [da'ñino] 'harmful'). A basic LTS rule set for Spanish, consisting of just 50 rules, is given in Black et al. (2002).

(1) \# [gc] = G

For example: a gcótaí 'their coats', pronounced [a go:ta:i]

In fact this rule is preceded by a more specific rule that says "if the orthographic sequence 'gc' occurs word initially and is followed by zero or more consonant letters which are followed by either an 'e' or an 'i', it is pronounced /j/ (Gj)". For example, the 'gc' sequence in the phrase a gcleachtaí 'their exercises' is pronounced [j]. See § 3.3 for a discussion on the phonemically velarised and phonemically palatalised consonant distinction in Irish and § 4.6.1 for a explana-
There are also a variety of machine learning techniques that have been used to build LTS rules. These methods have been used for languages like English that have an opaque (non-transparent) orthographic system, with complicated and unpredictable letter-to-sound correspondences (Black et al., 2002; Taylor, 2009). The successful use of automatic methods depends crucially on the availability of a lexicon (corpus) fit for purpose. In particular, the lexicon (corpus) must be very large, machine-readable, phonetically transcribed according to the dialect of the language to be phonetised, and should contain inflected as well as base forms.

Sproat et al. (1998) discuss the advantages and disadvantages of the two methods. On the one hand, developing LTS rules by hand is labour-intensive, and the interaction and ordering of rules can be difficult to manage. Using automatic methods to build LTS rules, on the other hand, has often been claimed to be cheaper and faster and to give better, more predictable results. Sproat et al. (1998, p. 75) dispute this claim, writing: “learning an accurate model may ultimately not be much cheaper than developing a model by hand, and there is certainly no reason to believe that the results of training a decision tree, a decision list, an inferable stochastic transducer or a neural net will necessarily result in a system that is more robust or predictable on unseen data than a carefully developed set of rules”. Sproat and colleagues also emphasize that appropriate corpora on which to train models are often not available, except for widely spoken Western European languages. This is the case for Irish. As discussed earlier, the only available resource for Irish with any kind of phonetic transcription is the *Foclóir Póca* Irish-English dictionary. This pronunciation dictionary is available in a machine-readable format, but does not meet the three other requirements.
for successful use in the automatic approach to developing LTS rules: it is small (14,000 words vs. 133,000 words in the CMU American English lexicon), it is transcribed in an artificial dialect rather than any actual spoken dialect of the language, and like most pronunciation dictionaries, it contains mostly headwords with only a few inflected forms (and as we have seen, Irish has a fairly rich system of inflections). Fortunately, as discussed in the previous chapter, Irish falls into the category of languages with a reasonably transparent orthography and is thus well-suited to an approach that uses hand-written LTS rules.

5.3 Multi-dialect phonetisation

Of course, all languages have dialects. However, depending on a variety of factors, the demand for multi-dialect synthesis may be lesser or greater. If a language has a standard variety, it may be acceptable to develop synthesis only for the standard variety, at least as a starting point. As described in § 2.6.2, Irish does not have a standard spoken variety, but rather three main dialects. Even if a language has a standard variety (e.g. English, which has at least two major standards: General American and Received Pronunciation), there are other factors that might motivate multi-dialect synthesis. For example, vocally-impaired people that use synthesisers “might prefer an accent closer to the one they are familiar with...since the voice becomes their persona” (Williams & Isard, 1997); an Irish person from Cork will not want to sound like an American from Indiana (or like a Dubliner, for that matter!).

Whatever the motivation for developing multi-dialect synthesis, there are two basic approaches to doing this type of synthesis. In the first, the various com-
ponents of a system built for a standard variety can be built upon or modified. This is a common situation where the components required to develop synthesis for a new dialect already exist for a standard dialect of the language and may be re-used and adapted. An example of this approach is described in Neubarth et al. (2008), where multiple synthetic voices representing various dialects (or sociolects) of Austrian German are developed. Existing resources from a TTS system for standard German (the standard German of Germany) are bootstrapped onto both standard Austrian German and the Viennese dialect. For example, they modified the standard German phoneset, added rules to handle phonological processes in the dialects, and modified the lexicon. They note, as we have noted for Irish, that dealing with some dialect differences is "rather unproblematic, since they can be covered by the local context and are fully represented in the acoustic data" (Neubarth et al., 2008, p. 1878), for example, lenition of intervocalic stops.

In the second approach, multi-dialect synthesis is envisaged and planned for from the outset. In this approach, TTS systems are designed in such a way as to allow new dialects to be added easily by re-using existing resources (see, for example, Hertz, 1990; Hertz et al., 1994; Williams & Isard, 1997; Sproat, 1998). As Hertz (1990, p. 225) notes "[b]esides the theoretical significance of such an approach, it results in more cost-effective development". She describes a system that uses "a modular approach to rule-based speech synthesis... [that is] divided into language-universal, language-specific/dialect-universal, and dialect-specific rule modules" (p. 225). These rule modules include LTS rule module that comprises dialect-universal and dialect-specific modules, and employs the concept of the diaphoneme, described as "underlying 'phonemes' common to all dialects" (Hertz, 1990, p. 226). The dialect-universal module maps text to diaphonemes, and the dialect-specific module maps diaphonemes to phones particular to the
A modular multi-dialect phonetiser for Irish

dialect being synthesised. In this system, for example, the diaphoneme /ai/ as in *pie*, would map to either [əi] in General American, or to [a] in African American English. Something similar is described in Williams & Isard (1997), who use a *keyvowel* approach based the concept of lexical sets (Wells, 1982). These researchers use keyvowels to specify all the vowel phonemes in the lexicon and LTS rules. For example, the GOOSE vowel is an abstract unit (a keyvowel) that represents a set of words (*loop, group, move, etc.*); words in this set contain the same vowel, although the phonetics of the vowel are dialect-dependent (e.g., [u] in RP, but [ũ] in Scottish).

5.4 A phonetiser suited to Irish: a global-local modular approach

Our approach to development of TTS and associated resources for Irish is in line with the development philosophy of Hertz (1990) and other researchers in that it promotes re-use of resources within a modular framework, and allows for straightforward extension to new dialects. As we wrote in Ní Chasaide et al. (2006, p. 181) “while working on [the Gaoth Dobhair] dialect we were from the outset conscious of the need to adopt strategies that will maximally facilitate similar developments for further dialects”. For phonetisation, I use an intermediate, dialect-neutral (global) representation,

To facilitate multi-dialect phonetisation for Irish, I adopted a *global-local modular approach*, which builds on what the dialects have in common, while dealing with the challenges caused by the multi-dialect dimension. This approach, which is inspired by the work in Ó Murchú (1969) and subsequent work in Ó Baoill
A modular multi-dialect phonetiser for Irish

(1986a) and Ó Baoill (1986b) (see § 3.2), derives an abstract dialect-neutral intermediate representation from the spelling, which can in turn be converted into a dialect-specific representation by the application of dialect-specific rules. It is in some ways similar to the approaches of Hertz (1990) and Williams & Isard (1997), although we do not use the concepts of diaphonemes and keyvowels. The intermediate representation used by the phonetiser is more abstract than a Láirchamúint transcription (see § 3.2). Where an orthographic sequence maps to different phones in different dialects, the Láirchamúint makes a choice and prescribes only one of the potential phones. For example, while the graphemes ⟨aoi⟩ and ⟨ao⟩ are pronounced as /iː/ in some dialects but as /eː/ in others, the Láirchamúint transcription is /iː/. The global-local modular phonetiser described here does not force such a choice, but rather allows the application of the appropriate dialect-specific rules.

For the remainder of this thesis, I will refer to the part of the phonetiser that converts ordinary text to this intermediate representation as the global module, with its output referred to as global transcription. It is global in the sense that it holds for the Irish language as a whole, regardless of dialect. The part of the phonetiser that converts this global transcription to a local (dialectal) transcription will be referred to as the local module, and the dialectal transcription it produces as local transcription.

5.5 Abstract transcription vs. Láirchamúint

The fundamental difference between the global transcription output by the phonetiser’s global module, and a Láirchamúint transcription, is that the former is map-
pable to local (dialectal) representations, while the latter is not. Láirchánúint transcriptions constitute abstract representations of Láirchánúint pronunciations, in the same way as if Láirchánúint were an actual regional dialect (see § 7.3). However, despite being designed to be in principle equally related to all dialects of Irish, it is not possible to map from Láirchánúint transcriptions to local dialects. Global transcriptions on the other hand, are not only mappable to the living dialects, but can also be mapped to Láirchánúint transcriptions, and a Láirchánúint local module is provided as part of this implementation.

This distinction is crucial and is central to this thesis. While Láirchánúint provides a suggested pronunciation, the global transcriptions have to be mapped to a particular dialect. For this to be possible, the global system has to be able to handle the many ways in which the dialects diverge in terms of pronunciation. It does so by remaining abstract from all dialects. In addition to this, since the phonetiser consists of both a system of LTS rules as well as a lexicon, it can handle dialectal word forms as well.

The global inventory, while comparable to the Láirchánúint phoneset, is fundamentally different. Where "phonemes" are excluded from the Láirchánúint because they are not common to all dialects, they are represented in the global set. For example, the Láirchánúint only allows for a two-way distinction among laterals, which is suitable for Munster Irish, but not for Connaught nor for Ulster Irish where there exists a robust three-way, possibly four-way distinction (Ní Chasaide, 1977), and the global inventory is equipped to deal with this. It is also important to note that members of the global inventory are not equivalent to phonemes, and should not be referred to as such. The schematic illustrations in Figure 5.3 serve to highlight the difference between Láirchánúint and global-local
coverage.

![Figure 5.3: A schematic representation of the difference between the (theoretic) coverage of the Láráinínt artificial dialect (A), the phonetiser’s global module (B), and the phonetiser’s local module (C). In each diagram, the circles represent regional dialects of Irish. The shaded area in A represents how the Láráinínt in theory relates to regional dialects. The shaded area in diagram B represents the way in which the global module covers all regional dialects, and the shaded area in C shows how local modules cover only features of the regional dialects.](image)

Note that while it might be possible to force align Láráinínt transcription (LC) with the speech corpora for the various dialects, such an approach would yield poor results and would necessarily lose dialect-specific information. This would undermine the fundamental goal of our research. For example, in the Láráinínt transcriptions in *Focloir Póca*, words like *póca* ‘pocket’ LC: /po:kə/ and *piocadh* ‘pick.vn’ LC: /pˈokə/ are both transcribed with final schwa. But in Ulster dialects, these words end in different vowels. This forced alignment would result in units being selected that contain incorrect segments and would result in synthesised speech that is incompatible with the target dialect. These types of examples, where a single Láráinínt segment would map to more than one phoneme in the target dialect, are not uncommon. Another type of problem that such an approach would result in is the loss of phonemic distinctions crucial in some dialects. For example, Ulster Irish contains the phoneme /ɔː/ (*go* *leor* /lɔːr/ ‘enough’ vs. *leabhar* /lɔːr/ ‘book’), but Láráinínt does not.
Similarly, as discussed in § 3.3, Ulster and Connaught Irish have more nasal and lateral phonemes than does the Lárchanúint.

5.6 Global transcriptions

The global system of transcription is used as an intermediate abstract representation of the sounds in the text that can be converted to a dialectal (local) realisation. The global inventory consists of a set of abstract items being either interdialectal abstract segments, grapheme that diverge to different phones at a local level (e.g. ⟨ao⟩ is pronounced as /iː/ in some dialects but as /eː/ in others), and interdialectal grammatical morphemes that are pronounced differently depending on dialect (e.g. word-final ‘-adh’ is pronounced as /ax/ in some dialects but as /uː/ in others). These items are mapped to phones by the phonetiser’s local module.

The items in this intermediate system, which are derived from text and map to local (dialectal-specific) transcriptions, consist of the superset of all dialectal phonesets, a small set of grapheme-based sequences, and a set of word-final grammatical morphemes, mostly verbal endings. Conventionally, these global units are written in uppercase plain text (e.g. PJ), and are comparable to their lowercase counterparts in the MRPAI (see § 5.1). The grapheme-based sequences are written as is but in uppercase (e.g. IO, representing the grapheme ⟨io⟩), and the word-final grammatical morphemes are written in uppercase prefixed with a hyphen (‘-’) and with a hash mark (#) suffixed to indicate a word boundary. These indicate that it is word-final (e.g. -IDH#, representing the verbal ending

44The pronunciation also depends on context, which could be handled by post-lexical rules, but this is a separate question.
5.7 Multi-dialect phonetisation step-by-step

What follows is an overview of the steps taken by the Irish multi-dialect phonetiser to convert ordinary text into a local transcription. The process involves a number of steps which are outlined below and presented schematically in Figure 5.4:

1. Input word.

2. Check if word is contained in a list of cross-dialect exceptions, i.e. a global exceptions lexicon. If so, convert to global transcription and skip to step 4. If not, proceed to step 3. This component currently interacts with a simple morphological analysis component.

3. Convert word to global transcription using global letter-to-sound (LTS) rules, and proceed to step 4.

4. Check if word is contained in a list of dialect-specific exceptions, i.e. a local exceptions lexicon. If so, convert to local transcription and skip to step 7. If not, proceed to step 5. This component currently interacts with a simple morphological analysis component.

5. Mark vowel for stress if a stress mark has not already been applied by the local exceptions lexicon, and proceed to step 6.

6. Convert global transcription to local transcription using local LTS rules and proceed to step 7.

7. Syllabify, and proceed to step 8.
8. Output phonetised word.

Figure 5.4: Phonetisation step-by-step. Separate components that interact with the phonetiser are shaded in grey.
5.8 The structure of the lexicon

As a lexicon (pronunciation dictionary) normally forms part of any phonetisation module, multiple lexicons are required for a multi-dialect phonetiser. Often, TTS systems use very large lexicons in parallel with letter-to-sound rules to perform phonetisation. As no comprehensive dialectal lexicons are available for Irish however, multiple lexicons would have to be created from scratch. Creating and maintaining multiple large lexicons was not possible within the scope of this project, nor was it desirable. The alternative to this approach is to rely more heavily upon the LTS rules and use a small exceptions lexicon as opposed to a large lexicon for each dialect. This approach is more economical in terms of production and maintenance costs, if the language in question has a more regular letter-to-sound correspondence. This is the case with Irish, and hence, this is the approach we have adopted in developing a multi-dialect phonetiser for Irish.

Given the multi-dialect nature of the phonetiser, an exceptions lexicon is required for each dialect. It became apparent that in the case of Irish, certain exceptions to regular spelling rules are not dialect-specific but rather cross-dialectal. For example, in (2) the word *deoch* is pronounced with a short /o/ vowel in all dialects.

(2) Tá deoch de dhíth orm.

‘I need a drink.’

This is an exception because the grapheme ⟨eo⟩ ordinarily corresponds to a long /ɔː/ in all dialects (e.g. *beo* ‘live’, *ceo* ‘fog’, etc.). This being the case, redundancy is likely in dialect-specific exceptions lexicons. To avoid this, a global-local approach was adopted in the construction of the exceptions lexicons that mirrored
the global-local approach of the LTS rules.

This approach meant that there would be a global exceptions lexicon that would be used by the phonetiser regardless of dialect, and a local dialect-specific exceptions lexicon for each dialect covered. The two types of exceptions lexicons (global and local) differ fundamentally in that the global lists convert ordinary text to global transcription and the local lists take global output from either the global exceptions lexicon or the global LTS rules, and output local transcriptions. Both the global exceptions lexicon and the local exceptions lexicons are described in § 5.8.1 and § 5.8.2.

These exceptions lexicons are functionally the same as the CMU pronunciation dictionary described in § 5.1, although they differ slight in format. The main difference is that word-pronunciation pairs are spread across two lines, allowing for multiple pronunciations. This change was made simply to enhance readability. The syntax, followed by an example, is given in (3).

(3) \texttt{<orthography>}
\texttt{<tab><transcription>}
\ldots

An example from the global exceptions lexicon (see Appendix B.4):

\begin{verbatim}
abhus
  0 @@ . 1 V U S

abú
  0 @@ . 1 B UU

acu
  1 A . 0 K U
\end{verbatim}
5.8.1 Global exceptions lexicons

The global exceptions lexicon included as part of the phonetiser is split into four subsets in this implementation, but the separate lists are compiled into one list (the global exceptions lexicon) by the rule interpreter. The lists are kept separate according to type of exception in order to simplify the maintenance of the lexicon as a whole. The first subset contains global exceptions (e.g. deoch ‘drink’, pronounced with a short /ʌ/ vowel in all dialects). The second subset contains a closed class of words that break normal spelling rules, as described in Chapter 4, across all dialects. This set of words consists mostly of adverbs as well as some other grammatical words such as prepositions, pronouns, and forms of the copula. The ways in which the different categories of words in this second subset break normal spelling rules are as follows:

**Category 1:** Adverbs with an unstressed first syllable. The vowel in this first syllable is therefore reduced, and the second syllable will have primary stress. This subgroup contains words like inné ‘yesterday’, inniu ‘today’, and amárach ‘tomorrow’.

Category 3: Various forms of the copula (is, ba) and some simple prepositions (le, ‘with’, i ‘in’, etc.) and possessive pronouns (mo ‘my’, do ‘your’, a ‘his/her/its/their’, etc.) that are unstressed and therefore reduced.

The third subset contains borrowings that retain their native stress patterns, usually from Irish English, across all dialects. This change in stress pattern also has an effect on vowel reduction. Many of the words on this list are scientific words such as séimeantaic, ‘semantics’, pronounced /2 SJ EE . 1 MJ A NN . 0 T  @ @ KJ/, in which the second syllable is stressed and its vowel is unreduced, and the first syllable has secondary stress. Other examples include words like coinbhinsiún ‘convention’ pronounced /2 K O NNJ . 1 V J I NNJ . 0 SJ UU NN/, sioráf ‘giraffe’, pronounced /2 SJ I . 1 R AA F/, or tobac ‘tobacco’ pronounced /0 T  @ @ . 1 B A K/. This list was easily extractable from Foclóir Póca, since stress is only marked in the dictionary if irregular (i.e. not on first syllable). This list is open-ended as the words in it do not form a closed class and more words may be added to it.

The fourth subset is a list of compound words. If the phonetiser tries to use LTS rules to phonetise a compound word, vowel reduction will be over-applied. In the absence of a tool to automatically recognise and split compounds, a list of compounds has been compiled. When such a tool is developed and integrated into the system, this subset can be removed from the global exceptions list.

5.8.2 Dialect-specific local exceptions lexicons

For each dialect, there will be one dialect-specific or local exceptions lexicon. Exceptions lexicons are provided as part of the phonetiser for three major dialects, Gaoth Dobhair, Conamara, and Corca Dhuibhne, but these lists are meant to be
open-ended, and can be edited or expanded at any time. An example entry from the Gaath Dobhair list is the word *madra* pronounced /1 m a . 0 d uu/. Normal letter-to-sound rules would predict /1 m a . 0 d r @/. These lists were initially based to some extent on the word lists contained as appendices to the series of short dialect descriptions called *An Teanga Bheo* 'the living language' (Ó Sé, 1995; Ó Baoill, 1996; Ó Murchú, 1998; Ó Buachalla, 2003). The lists should ideally be edited in consultation with native speaker consultants. Separating the lists in this way makes it easier for developers to understand and modify for the current dialect, and to extend to other dialects.

5.9 The letter-to-sound (LTS) rules

The global-local architecture of the phonetiser is the same with regard to the application of letter-to-sound rules as it is for the lexicon. Words that are not exceptions are phonetised by rule, firstly by being passed through the global LTS rules, with the result being passed through the appropriate set of local LTS rules.

As already outlined, all dialects of Irish are closely related at a structural level, and one of the goals of this research to harness this in the production of the global and local LTS rule sets. The goal therefore was to ensure that only rules pertaining to surface realisations reside in the local rule sets. As a result, most of the work of the LTS rule component of the phonetiser is carried out by the global rule set. Each local rule set is lightweight relative to this.

As the implementation of the global and local rule set constitute a central part of this work, they are described in detail in Chapter 6.
5.10 Interaction between the phonetisation and linguistic analysis components of TTS synthesis

There are a number of other components that have not yet been developed and integrated into the TCD Irish TTS synthesis system, of which the phonetisation developed here is a component. These include linguistic analyses such as morphological analysis (lemmatisation, compound recognition), syntactic analysis (e.g. part-of-speech tagging), and syllabification, as well as post-lexical rules (see § 5.10.1, § 5.10.2, and § 5.10.3 for further discussion).

5.10.1 Morphological analysis and part-of-speech tagging

Irish has an extensive system of initial mutations and is also fairly heavily inflected. This means that for any word listed in an exceptions lexicon, there may well be a number of possible variations in the form of that word. For the phonetiser to produce correct pronunciations for all of these forms, either they each have to be listed in the exceptions lexicon, or the phonetiser needs to be able to recognise if a word is a mutated or inflected form of a word contained in the exceptions lexicon. For example, *mo mhadra* 'my dogs' /0 m @ . 1 v a . 0 d uu/, contains a mutated form of *madra* 'dog' /1 m a . 0 d uu/ contained in the Gaoth Dobhair exceptions lexicon.

It is desirable for the phonetiser to have access to information about whether a word is mutated or inflected. However, the morphological analysis required to correctly identify relationships between inflected and base forms (lemmatisation
is a separate component of the TTS system, part of the linguistic analysis component. Some work on morphological analysis for Irish has been done and more is in progress (Uí Dhonnchadha, 2003). Until we are able to integrate a more sophisticated, elegant approach into the system, we used the following temporary solution. A simple set of rules to strip mutations and inflections from words before they are looked up in the exceptions lists.

Homograph disambiguation, which depends in part on the syntactic analysis component, is also not currently handled by the system that uses the phonetiser described here. For example, certain homographs (orthographically identical words) may be pronounced differently depending on part-of-speech. For example, when the sequence of letters ‘ag’ denotes a preposition meaning ‘at’, the pronunciation is [æj]. However, if it occurs as a verbal noun particle, its pronunciation varies. For example, when followed by a vowel as in _ag ól_ ‘drinking’ and _ag ithe_ ‘eating’, it is pronounced as either [æg] or [aj], depending on the quality of the following vowel (a post-lexical constraint). When it is followed by a consonant, as in _ag scríobh tráchtais_ ‘writing a thesis’, it is pronounced [a], except in carefully read speech.

### 5.10.2 Stress assignment in compound words and borrowings

Irish words generally have stress on the first syllable. The issue regarding stress patterns of borrowings is partially dealt with by the phonetiser’s global exceptions lexicon. Compound words in Irish can also have a variety of stress patterns: primary-primary (‘príomhchathair ‘capital city’), primary-secondary (‘lárchanúint), secondary-primary (‘do’thuigthe ‘incomprehensible’). As already
explained in § 3.6, correct phonetisation depends in part on a knowledge of where stress falls. The reason for this is that in general, short vowels in unstressed syllables are reduced. Given the varying stress patterns of compounds therefore, and the fact that it cannot be assumed that syllables other than the first are stressed, this information must be available to the phonetiser for optimal phonetisation.

The linguistic analysis component of TTS synthesis includes identifying and marking compound boundaries. In Irish, these boundaries are often recognisable when there is lenition in the middle of a word (e.g. ‘bog.shodar ‘jogging’), or when two consonants that differ in quality (phonemically velarised and phonemically palatalised) come together (e.g. ‘mór.shiúil ‘procession’). However, as compounds in Irish are not generally orthographically marked (e.g., by hyphenation), and many non-compound words, or words that historically were compounds, are indistinguishable from compounds, for example ‘comh’ainm ‘homonym’ (comh ‘co-’ + ainm ‘name’) vs. ‘comhad ‘file’, a sophisticated solution is required. As a linguistic analysis component is not yet implemented for the Irish TTS system, an alternate interim solution is employed. This involves first identifying compound words and then separating them into three categories based on stress pattern and characteristics as follows:

Category 1: Compounds with a primary-primary stress pattern mostly consist of an intensive prefix and a simple word. The intensive prefixes accounted for are the following: an- ‘very’, ard(-) ‘arch-, high, chief’, barr(-) ‘top’, có(i)(-) / comh(-) ‘mutual, joint, common; co-, fellow-’, dea- ‘good; well-’, dearg(-) ‘red; real; intense, utter’, dian(-) ‘intense, strong; hard, severe’, droch(-) ‘bad; poor, evil; ill-, un-’, dubh(-) ‘black, dark; great, intense; evil; unknown’, fíor(-) ‘true, real; intense, very’, glan(-) ‘clean,
clear; pure, bright, clear-cut', \textit{gnáth(-)} 'customary; vulgar, common; constant', \textit{iar-} 'after-, post-; late, ex-; west, western', \textit{lán(-)} 'full', \textit{oir(-)} 'east, eastern', \textit{os(-)} 'over, above; super-, supra-', \textit{priomh(-)} 'prime, principal, major, cardinal', \textit{rí(-)} 'royal, kingly; exceedingly, very, ultra-', \textit{ró(-)} 'too, most, very; over-; excessive', \textit{sáร(-)} 'exceeding, surpassing; excellent; ultra-, most', \textit{síor(-)} 'perpetual, continual; ever-', \textit{uile(-)} 'all'.

\textbf{Category 2}: Compounds with a primary-secondary stress pattern, for example \textit{'aer-phórsa} 'air force', \textit{'béal-oideas} 'folklore', etc.

\textbf{Category 3}: Compounds with a secondary-primary stress pattern, for example \textit{'dothuigthe} 'incomprehensible', \textit{'in-fheicthe} 'visible', etc.

The interim solution uses a list of intensive prefixes to find the Category 1 compounds. A similar solution was adopted for Category 3 compounds. However, this approach does not work for Category 2 compounds as the set of possible first elements is infinite, since any word can be prepended. As mentioned in § 5.8.1, the interim solution involved compiling a list of these words (about 2,000), based on an initial extraction from \textit{Foclóir Póca}. As these stress rules are global, this list will form part of the global exceptions list, but may be removed when a tool to recognise and split compounds is added to the phonetiser. Another reason to develop such a tool is to be able to deal with nested compounds such as \textit{ródhráchbhéasach} 'too impolite' (\textit{ró} 'too' + \textit{droch} 'bad' + \textit{béasach} 'mannerly').

\textbf{5.10.3 Syllabification}

Syllabification is another important component of any TTS system is a syllabifier. Research is underway to develop such a tool for Irish. Syllable boundaries
often (though not always) respect morpheme boundaries in compound words, so the development of the linguistic analysis component of the TTS system is also relevant here. Until a more sophisticated solution is available, we use a fairly simple interim syllabification strategy.

We put the syllable boundary before the intervocalic consonant in all such sequences (V.CV), regardless of vowel length and stress, assuming an obligatory onset principle, which requires syllables to have onsets if possible. We are, however, mindful that there are conflicting accounts of the syllable affiliations of intervocalic consonants in Irish, for example, some would argue that a word like *buile* 'madness' (with stress on the first syllable and two short vowels) should be syllabified as /b̪i.ə/, with the intervocalic /l/ syllabified as the coda of the first syllable (Ó Searcaigh, 1925; Ó Cuív, 1944; de Búrca, 1958; Mhac an Fhailigh, 1968). Sjoestedt (1931) argued for the syllabification /b̪i.ə/, with intervocalic /l/ syllabified as the onset of the second syllable, and other scholars note that the syllabic affiliation of such intervocalic consonants is unclear (Quiggin, 1906; de Bhaldraithe, 1945; Breatnach, 1947). The vast majority of scholars assume that words like *cóta* 'coat' (with stress on the first syllable, a long first vowel and a short second vowel) are syllabified as /ko.ə/, with the intervocalic consonant syllabified as the onset of the second syllable. Ó Cuív (1944), however, treats it as the coda of the first syllable. The intervocalic consonant in words like *cosán* 'path', with long second syllables, are consistently treated as the onset of second syllable.

We assume a maximum onset principle, constraining to disallow nonpermissible clusters. For example, a word like *stiúcra* 'sugar' will be syllabified /ʃu.krə/, since /kr/ is a possible onset in the language. A word like *ordóg* 'thumb', however,
will be syllabified /or\d^d\vog\d/, since /rd/ is not a possible onset in Irish. Scholars of Irish have consistently treated intervocalic consonant sequences, as in taisce ‘treasure’, siúcra ‘sugar’, and coláiste ‘college’, as belonging to different syllables (taisce /\t^ty\af\ca/). We do not follow this structure, though. This decision is motivated in large part by the observation that when voiceless stops appear as simple syllable onsets in Irish, they are aspirated (in cead [k^h\af\d\v] ‘permission’, for example), but when they appear intervocally in s + stop sequences, they are unaspirated (in taisce, for example).

Our LTS rules also insert an epenthetic (svarabhakti) schwa /a/ to break up impermissible coda consonant clusters in words like ainm ‘name’, borb ‘abrupt’, and dearg ‘red’ (for a phonological account, see Ní Chiosáin, 1991, 1999).

The results of a recent experimental study of Irish suggest that Irish is not a typological exception with respect to syllabification (Ní Chiosáin & Welby, 2010). The same types of factors that have been shown to affect syllabification boundary judgements in over a dozen other languages also affect these judgements in Irish (e.g., the shorter a vowel, the more likely it is to attract a coda consonant). We therefore have no reason to believe that we should not use an approach to syllabification for Irish similar to those successfully used for other languages for which TTS systems have been developed. Of course, if any of our assumptions turn out to be incorrect the algorithm can be easily modified.
Chapter 6

A detailed description of the global and local rules

6.1 Introduction

In this chapter, I will describe in detail the letter-to-sound (LTS) rules listed in Appendix B that, in conjunction with the lexicon described in § 5.8, constitute an implementation of the multi-dialect phonetiser for Irish. A global-local modular approach has been adopted (see § 5.4), and this architecture applies to both the LTS rules and the lexicon in exactly the same way.

As discussed, this approach was inspired by the common core approach suggested in Ó Murchú (1969) and the Lárchanúint 'core dialect' developed in Ó Baoill (1986b), but achieves multi-dialect applicability by virtue of the global-local modular approach adopted. The approach uses global rules and exceptions lists to derive an abstract dialect-neutral global transcription from the spelling,
followed by local rules and exceptions lexicons to produce a dialect-specific local transcription suitable for use in text-to-speech (TTS) synthesis systems. The implementation of the LTS rules is described below.

6.2 The structure of the phonetiser

The phonetiser consists of a global module and a local module for each dialect implemented. The purpose of the modular approach adopted was to facilitate the development of a multi-dialect phonetiser. Each module consists of a set of LTS rules, and an exceptions list. The implementation of the global and local rules is described in §6.5 and §6.6 respectively. Firstly, however, the rule syntax and interpretation is described in §6.3.

6.3 Rule syntax and interpretation

The phonetiser’s LTS rules are rewrite rules. Each rule has a left-hand side (LHS) and a right-hand side (RHS), which are separated by an equals sign (‘=’). The LHS is a search pattern and the RHS is a replacement pattern. The rules are fed input and when search patterns match portions of the input, the input is rewritten as the corresponding replacement patterns in the output. To give a very simple example in (4), an orthographic ⟨i⟩ in the input is written to the phonetic symbol for a long high front vowel in the output, reflecting its pronunciation in words like sí ‘she’, or the girl’s name Róisín.

(4) \( i = [i:] \)
6.3.1 Regular expressions

The left-hand side (LHS) of the rewrite rules is a type of regular expression, a tool that is used to represent text patterns. The most common use of a regular expression is for text searching. For example, in searching a passage of text for all references to the Gaelic game of hurling, one should search for both *hurling* and related forms including *hurl*, *hurls*, *hurler*, *hurlers*, *hurley*, *hurleys*, etc. Since all these words start with ‘hurl’, essentially one is searching for the pattern of letters ‘hurl’, with or without something else suffixed. A regular expression for this is given in (5).

\[(5) \text{hurl}.*\]

Regular expressions provide a standardised way of succinctly expressing such a pattern, as well as other more complex patterns.

Although the syntax of regular expressions can vary depending on the interpreter used, that is, the computer program used to parse the regular expression, variations do not tend to depart drastically from the various well-defined standards. The most important features of standard regular expression syntax therefore are as follows:

- Alternates: alternate patterns are expressed using a vertical bar binary operator (the *pipe* symbol). For example, *palatalised*|*palatalized*, equivalent to ‘palatalised’ or ‘palatalized’, will match with either spelling of the word. Square brackets can be used to express a group of single character alternatives. For example, *[bcdfglnprst]* will match with any lowercase consonant letter from the traditional Irish alphabet (see Figure 4.1).
• Groups: operators can be applied to groups of tokens if the tokens are enclosed in parentheses. For example, palatali(s|z)ed, will also match either spelling of the word. Grouped expressions can also be referred back to. This is a feature that can be used in replacement rules to move elements from the search pattern around in the replacement pattern.

• Wild card: a full stop (period) '.' will match with any character.

• Quantities: a required quantity can be specified by using one of the following operators that apply to the element that precedes them. These operators are called unary operators because they apply to a single item (as opposed to binary operators that apply to two items):

  ? “zero or one”. For example, whiske?y would match with whisky or whiskey, and ch?eann would match with ceann ‘head’, or cheann ‘head’ (lenited form).

  * “zero or more”. For example, ch?eann.* would match with ceann, cheann, ceannóidh ‘will buy’, cheannaigh ‘bought’, etc.

  + “one or more”. For example, pr(e|o)ce+ding would match with ‘preceding’ or ‘proceeding’.

As an example of how various features and operators can be combined to search for more complex patterns, the regular expression palatali(s|z).*”, which uses grouping, the wild card, and the unary operator *, would find words like palatalised, palatalized, palatalisation, etc., if these words are present in the file being searched.
6.4 Linguist-friendly letter-to-sound rules

In general terms, the rules being discussed are letter-to-sound (LTS) rules. However, given the global-local modular approach adopted, global and local LTS rules (and similarly global and local exceptions lists) differ in terms of the type of input they are designed to deal with and the type of output they are designed to produce. When we talk in terms of input and output, the input is what the left-hand side (LHS) of the rewrite rules is designed to match with and the output is what is on the right-hand side (RHS) of the rewrite rules. The global rewrite rules take text patterns as input and replace these text patterns with an abstract global transcription. The local rules take the output of the global rules as input and produce a local transcription as output.

As mentioned in § 6.3, the LHS of the rewrite rules is a type of regular expression, a pattern of sorts. The LHS syntax differs slightly from strict regular expression syntax, for historical reasons, which will be described below.

The rule file format is based on an existing format referred to as “Linguist-friendly format” (LFF), hence the unusual .lff file extensions (see Appendix B). This format was adopted for the development of Welsh LTS rules (Williams, 1993, 1994a), and an LFF interpreter was added to the Festival speech synthesis system to handle rules in this format.

This Festival LFF format has been adapted slightly for this project, and the LTS rule interpreter of the Irish TTS system has been adapted accordingly. The main difference is that the Festival LFF format requires both spacing between items on the LHS and spacing between items on the RHS. Note that there are spaces either side of items on the LHS in the Festival format but not in the version
of the LFF format used in the examples in (6).

(6) \(\# [s] p = s\) (Festival LFF syntax)
\[\#[s]p = s\] (LFF syntax used by Irish TTS system)

The Festival format also requires additional parentheses. The spaces are there to allow the rules to deal with orthographic or phonetic input, and to allow for computer-readable phone representations that consist of more than one character. The LFF variant devised for the Irish rules is different in that orthographic input can be unspaced. This helps with readability as normal orthography is unspaced. It is for this reason that a type of regular expression syntax was devised for the LHS of the LFF rules used here.

The main difference between the syntax of the LHS of the global rules and normal regular expressions, is the use of variables, and the syntax of the variable declarations. Someone familiar with standard regular expression syntax would find it odd that when variables are declared at the beginning of the rule file (see Appendix B.1), the variable assignment is enclosed in square brackets, rather than parentheses. This odd syntax reflects earlier LFF formats.

Another difference is that square brackets are used to enclose the text pattern being searched for, separating it from the context, if any.

### 6.5 The global rule set

The global rule set is a set of rewrite rules that take orthographic input (ordinary spelling) and produce an abstract dialect-neutral *global* transcription as output.
A detailed description of the global and local rules

This global transcription can be converted to a dialect-specific local transcription by applying the dialect-specific rules contained in one of the local modules.

One of the goals of this research was to use to our advantage the structural similarities between the dialects of Irish. Assuming great structural similarity, it was hoped that most of the work involved in converting from orthography to a transcription of sounds would be done by the global module. In addition to this, given the approach adopted of relying more heavily on LTS rules than on a large lexicon, it was envisaged from the start that the global rule set would be the engine room of the Irish phonetiser as a whole.

It is worth emphasising at this point that this goal was achieved, and as a result, the global rule set is substantial, and requires a detailed description. I will now go through how the rules work in relation to the following salient aspects of Irish phonology and grammar that are encoded in the spelling:

- The phonemically velarised vs. phonemically palatalised consonant distinction.
  

- The system of initial mutation, characteristic of the Celtic languages.
  
  contaé ‘county’ [k], a chontae ‘his county’ [x], a gcontae ‘their county’ [g].

- Epenthesis: the insertion of an extra vowel into certain consonant clusters.
  
  dorn ‘fist’ [d[缺陷]or[缺陷]n].

- The system of grammatical inflection.
  
  tógaídh ‘will take’ [t[缺陷]oika] (M)
As was described in detail in Chapter 3, Irish has a series of phonemically velarised and a series of phonemically palatalised consonants, which are phonemically contrastive. Consonants in either series will have a counterpart in the other series whose phonetic characteristics are similar, except for the presence of a secondary articulation (This contrast is not always conveyed phonetically by velarisation/palatalisation. For details see Chapter 3). For example, a t-sound will have a phonemically velarised and phonemically palatalised counterpart, the former being coloured by velarisation, the latter being coloured by palatalisation. So while there are two t-sounds in Irish, there is only one consonant symbol to denote a t-sound in the Roman alphabet used to write Irish (see Figure 4.1). A device is therefore required to depict this phonological distinction in the spelling. The historic solution to this problem in Irish spelling has been to use adjacent vowels letters to denote consonant quality (phonemically velarised or phonemically palatalised).

In Irish orthography, a consonant symbol denoting a phonemically velarised consonant will be flanked by the letters ‘a’, ‘á’, ‘o’, ‘ó’, ‘u’, or ‘ú’,45 and a phonemically palatalised consonant will be flanked by the letters ‘e’, ‘é’, ‘i’, or ‘í’. If we denote all orthographic consonant letters as $C_i$, all phonemically velarised consonants as $C$, and all phonemically palatalised consonants as $C'$, these two spelling rules for Irish consonants, using regular expression patterns on the LHS, could be written as is shown in (7).

45The situation is actually a bit more complicated. For example, the graphemes (ae) counts as phonemically velarised (hence forms like tráenach ‘train.GEN.SG’).
(7) \( C_i/[aáóúú]\_ [aáóúú] \rightarrow C \)

For example, the letter ‘s’ in the word *ásal* ‘donkey’ is pronounced with a phonemically velarised [s\(^\nu\)]

\( C_i/[eéíí]\_ [eéíí] \rightarrow C^j \)

For example, the letter ‘s’ in the context *áséal* ‘low’ is pronounced with a phonemically palatalised [c]

The orthographic context of the grapheme is given to the right of the slash on the LHS of the rule.

However, these two rules will not deal with the fact that consonants in Irish can occur in sequences. A vowel letter may therefore denote the quality of more than one consonant, only one of which will be adjacent to the vowel symbol (e.g. *Ástráil* /as\(^\nu\)t\(^\nu\)l/ ‘Australia’, *aistriu* /a\(^\nu\)\(^t\(_t\)\(^t\_u\)/ ‘switching’). The two rules in (8) deal with this, and say that zero or more (*) consonant letters can occur between the consonant letter in question and the flanking vowels.

(8) \( C_i/[aáóúú]C_i^*C_i'[aáóúú] \rightarrow C \)

\( C_i/[eéíí]C_i^*C_i'[eéíí] \rightarrow C^j \)

These rules do not however deal with the fact that if a consonant or consonant cluster occurs next to a word boundary, there will only be vowel context on one side (e.g. *coar\(_t\)t\(^\nu\)l* ‘correct’, *cr\(_a\)t\(^\nu\)j* ‘framework’). To deal with this the rules need to include the option of either a left or a right context. This entails two separate patterns on the left-hand side and is therefore best divided into two separate rules as is shown in (9), in which # denotes a word boundary.

(9) \( C_i/[aáóúú]C_i^*C_i'[# \rightarrow C \)
The rules shown in examples (7)–(9) are not given in any particular order. However, as LTS rules are applied in sequence, they must be ordered, with more specific cases coming before more general ones. In addition to this, once a matching rule is found, the right-hand side of that rule is output, and the input cursor moves forward. This allows us to reduce the number of rules needed to deal with the phonemically velarised–phonemically palatalised distinction.

The approach adopted here is to treat phonemically velarised consonants as the default case. Doing it this way allows us to cover the phonemically palatalised case using two rules, and the phonemically palatalised case using just one general rule. This approach is illustrated by the rules shown in (10).

\[
\begin{align*}
C_l/^/C_l^*\cdot C_l^*[\text{æoøuû}] & \rightarrow C \\
C_l//[\text{eéíí}]C_l^*\cdot C_l^*# & \rightarrow C_l^\\nC_l/^/C_l^*\cdot C_l^*[\text{eéíí}] & \rightarrow C_l^\\
\end{align*}
\]

The phonemically velarised–phonemically palatalised rules in (10) are general and would match with any consonant. In the actual rule file, a set of phonemically velarised–phonemically palatalised rules are needed for all consonants, as each set will have a different right-hand side. The phonemically velarised–phonemically palatalised rules for the grapheme (b) and the sounds /b\textsuperscript{v}/ and /b\textsuperscript{j}/, denoted B and BJ respectively (using MRPAI, see 5.1) in the rule files (see § 5.1), are shown in (11). These rules would each be required to correctly convert the letter ‘b’ in the words *bris* ‘break’, *clib* ‘tag’, and *clab* ‘mouth’, in that order.
A detailed description of the global and local rules

(11) \[ \text{CS}_* \text{RCSV}_\text{NONSYLLABIC}_ = \text{BJ} \]
\[ \text{NONSYLLABIC}_\text{LCSV}_\text{CS}_*[b] = \text{BJ} \]
\[ [b] = B \]

The context is specified using variable symbols that represent various letters or groups of letters. A convention of suffixing variable names with an underscore (_) was adopted to make the rules more readable. Readability is important because the system is designed to be extendable to other dialects and it is important that other users and developers be able to understand the system when working on it in the future. All variables are declared at the beginning of the rule file (see Appendix B.1).

A number of other points need to be made about the rules shown in (11), which are part of the actual implementation.

There are four variables used in the phonemically velarised–phonemically palatalised rules. The variable \( \text{CS}_* \) stands for all Consonant Symbols. The variable \( \text{RCSV}_* \) stands for Right Context Slender Vowel,\(^{46}\) and similarly, the variable \( \text{LCSV}_* \) stands for Left Context Slender Vowel. The variable \( \text{NONSYLLABIC}_* \) stands for anything that does not constitute a syllable nucleus, including consonants and word boundaries. How these variables actually resolve is shown in (12). (Recall that the pipe symbol '|' pipe means 'or'.)

(12) \[ \text{CS}_* \ [b|c|d|f|g|h|i|j|k|l|m|n|p|q|r|s|t|v|w|x|y|z] \]
\[ \text{NONSYLLABIC}_* \ [b|c|d|f|g|h|i|j|k|l|m|n|p|q|r|s|t|v|w|x|y|z|\#] \]

\(^{46}\)Note the use of the term Slender in the formation of the acronym. This is a Celticist term for what is referred to as phonemically palatalised throughout this thesis. See § 3.3 for further details.
A detailed description of the global and local rules

It is important to note that the vowel context matched with in RCSV_ and LCSV_ is the entire orthographic vowel sequence and covers both vowels and diphthongs. This allows for more precise control of rule application and makes rule application more transparent.

6.5.2 Initial mutations

Initial consonants in Irish are subject to various types of mutation under certain conditions. There are two main types of initial mutation: lenition and nasalisation/eclipsis. Consonants change as a result of lenition or nasalisation/eclipsis, however, these initial mutations are dealt with in the spelling in a way that is morphologically transparent (see § 4.5). As a result, certain word-initial two- and three-letter graphemes map to single sounds. For example, the grapheme ⟨bp⟩ (i bPort Mearnóg ‘in Portmarnock’), which represents an nasalised/eclipsed ‘p’, maps to the sounds /bˠ/ and /bʲ/. The grapheme ⟨ph⟩ (ó Phort Mearnóg ‘from Portmarnock’), which represents a lenited ‘p’, maps to the sounds /fˠ/ and /fʲ/.

Rules, similar to those for the ⟨bp⟩ grapheme shown in (13), are required for each mutated sequence.

(13) #[bp]CS_*RCSV_NONSYLLABIC_ = BJ
    #[bp] = B
6.5.3 Epenthesis

Epenthesis occurs in Irish when an additional unstressed vowel is inserted into certain consonant clusters in Irish. The epenthetic vowel (also called *svarabhakti*), is inserted after a voiced sonorant that follows a stressed short vowel and precedes a non-dental consonant other than /p^v/, /p^j/ or /k/, /c/ (e.g. *dorn* ‘fist’, *gorm* ‘blue’, etc.). It is the distribution of this epenthetic vowel that is important for our purposes, but for more details, and a phonological treatment, see Ní Chiosáin (1990, 1991), and references therein. Although the insertion of this vowel is systematic, it is not reflected in the spelling system, and thus extra rules were needed in the global rule set to deal with all the clusters into which this extra vowel is inserted. An example of these rules is shown in (14).

\[
\begin{align*}
\text{NONSYL}_{\text{ABIC}}_{\text{LCSBV}}_{\text{LNR}}_{\pm} + [b] &= \text{BB} \\
\text{NONSYL}_{\text{ABIC}}_{\text{LCSSV}}_{\text{LNR}}_{\pm} + [b] &= \text{BJ}
\end{align*}
\]

The effect of these rules is to rewrite the letter ‘b’ as /ə b^v/ or /ə b^j/, depending on whether it is phonemically velarised or phonemically palatalised as always, when it follows a voiced sonorant, represented by the variable LNR., which resolves to the letters ‘l’, ‘n’, or ‘r’.

6.5.4 Future, conditional, and verbal adjective endings

As was described in § 4.5, future and conditional verb endings in Irish are morphologically transparent. They often begin with the letter ‘f’. As a result, rules can be included in the global rule set to apply the devoicing effect of these verb endings on voiced final consonants of verb stems. The rules to deal with verb
stems that end with the letter ‘b’ are shown in (15). The variables BFCE_ and SFCE_, stand for Broad Future and Conditional Endings and Slender Future and Conditional Endings respectively.

(15) \[ \text{BFCE}_# = P \]
    \[ \text{SFCE}_# = PJ \]

Similarly, verbal adjective endings that begin with the grapheme (th) cause devoicing of voiced final consonants of verb stems, and an example of rules to deal with this are shown in (16).

(16) \[ \text{LCBV.CS}_*[\text{bth}] = P \]
    \[ \text{LCSV.CS}_*[\text{bth}] = PJ \]

6.5.5 Rule ordering and grouping

In keeping with the goal of having linguist-friendly readable rules, the rules in the global rule file have been grouped and commented. They are grouped according to the first letter of the sequence of letters being replaced, i.e. the sequence of letters between the square brackets on the left-hand side. These groups are then ordered alphabetically. Within each group of rules, the ordering is critical with more specific cases coming before less specific cases. To illustrate this, a complete group of global rules is shown in (17). This group deals with all orthographic sequences that begin with the letter ‘b’.

(17) 409 // future & conditional ‘b’
    410 \[ \text{BFCE}_# = P \]
A detailed description of the global and local rules

411  [bf]SFCE_# = PJ

412  // eclipsis 'bhf'

413  #[bhf]CS_\*RCSV_NONSYLLABIC_ = VJ

414  #[bhf] = V

415  // future & conditional 'bh'

416  [bhf]BFCE_# = F

417  [bhf]SFCE_# = FJ

418  // epenthesis 'bh'

419  NONSYLLABIC_LCSBV_LNR_+[bh] = @@@ V

420  NONSYLLABIC_LCSSV_LNR_+[bh] = @@@ VJ

421  // lenition 'bh'

422  #[bh]CS_\*RCSV_NONSYLLABIC_ = VJ

423  #[bh] = V

424  // base case 'bh'

425  [bh]CS_\*RCSV_NONSYLLABIC_ = VJ

426  //NONSYLLABIC_LCSV_CS_+[bh] = VJ

427  NONSYLLABIC_LCSV_CS_+[bh] = VJ

428  NONSYLLABIC_LCSV_[bh] = VJ

429  [bh] = V

430  // eclipsis 'bp'

431  #[bp]CS_\*RCSV_NONSYLLABIC_ = BJ

432  #[bp] = B

433  // verbal adjective 'b'

434  LCBV_CS_+[bth] = P

435  LCSV_CS_+[bth] = PJ

436  // epenthesis 'b'
A detailed description of the global and local rules

There is one local module for each dialect implemented. Currently, the dialects of Gaoth Dobhair (Ulster) (see Appendices B.10–B.12), Conamara (Connaught) (see Appendices B.13–B.15), and Corca Dhuibhne (Munster) (see Appendices B.16–B.18) have been implemented and are included in the appendices. There is also an implementation for the Lácharumúint, that treats this artificial dialect as if it were a regional dialect (see Appendices B.7–B.9).

As explained earlier, each local module consists of a letter-to-sound (LTS) rule set, an exceptions lexicon, and a test set. With local modules, the LTS rule set and exceptions lexicon convert the abstract linguistic output of the global module to a local phone-level transcription.

The LTS rules have to contain at least enough rules to deal with all possible units that can be output by the global module. They should also contain enough rules to produce phonetic output for the regional dialect in question. The idea is that this output could be used in dialect-specific speech synthesis. Testing of the
local LTS rules is done in the same way as testing of the global LTS rules with tests interspersed within the rules. Testing is described in § 7.2.

The local exceptions lexicon should only contain words and their pronunciations where the word-pronunciation correspondence is an exception in that dialect (e.g. Gaath Dobhair: *madra* /m\^ad\^\textsuperscript{u}/ ‘dog’, other dialects /m\^ad\^\textsuperscript{r}\^\textsuperscript{a}/; Conamara: *bolgam* /b\^\textsuperscript{\textalpha}\^\textgamma\textalpha\textomega\textgamma/ ‘mouthful’, other dialects /b\^\textgamma\textalpha\textomega\textgamma/; Corca Duibhne: /b\^\textgamma\texteta\textgamma\textalpha\textomega\textgamma/ *brad\'an* ‘salmon’, other dialects /b\^\textgamma\textalpha\textomega\textgamma/), but not an exception in all dialects. In cases where a word-pronunciation correspondence is an exception in all dialects, it should be included in the global exceptions lexicon (e.g. *deoch* /dK\^\textalpha\textlambda\texteta/ ‘drink’ where the (eo) is pronounced with a short /\^\textalpha/, *is* /is\^\texteta/ (the copula) where the /s\^\texteta/ is phonemically velarised, etc.). The syntax of entries is the same as with the global exceptions lexicon (see § 3).

The test file can be used to test the dialectal phonemic output of a particular local module given orthographic input. This test file, if empty, is simply overlooked by the interpreter.

### 6.7 Adding a new dialect

To add a new dialect, a new local module is required. As described in § 6.6, a local module includes a set of LTS rules, and an exceptions list. Creating a new set of LTS rules involves mapping all global units (see § 5.4) to suitable dialectal phonetic transcriptions. To do this, a developer would take the template given in Appendix B and, using the comments and examples included, add the appropriate dialectal transcriptions. As a starting point, it may be possible to use information from the many traditional dialect descriptions that have been
published for the various Irish dialects, such as the dialect monographs published by the Dublin Institute for Advanced Studies. These and other works were useful to me in developing local modules for Gaoth Dobhair (Ní Chasaide, 1999), Conamara (de Bhrdalraithe, 1945), and Corca Dhuibhne (Ó Sé, 2000). There are also published descriptions on other dialects including dialects of Cork (e.g. Ó Cuív, 1944), Waterford (e.g. Breathnach, 1947) and Mayo (e.g. de Bhrca, 1958; Mhac an Fhailigh, 1968; Stockman, 1974). As we point out in Ní Chasaide et al. (2006, p.182), though, “these are an excellent resource, but they are not always ideal for the task at hand, some of them being quite dated, and directed at the speech of the oldest members of the community”, so developers will need to work with contemporary native speaker consultants (if they are not native speakers themselves) to check the pronunciations. Exceptions can then be added to the local exceptions list. It is envisaged that the exceptions list can remain open-ended, and be added to over time. Again, many printed resources are available. The An Teanga Bheo series on contemporary Irish dialects was particularly helpful in the development of the exceptions lists for Gaoth Dobhair (Ó Baoill, 1996), Conamara (Ó Murchú, 1998), and Corca Dhuibhne (Ó Sé, 1995), as each book in the series includes a glossary of dialectal exceptions. There is another book in this series for the dialect of Cléire, Co. Cork (Ó Buachalla, 2003) which would be of use if synthesis for that dialect were developed.

Of course, catering for a new dialect for TTS synthesis involves more than just phonetisation. Normalisation rules and post-lexical rules, for example, will need to be added. For example, normalisation rules for telling the time are dialect-specific, and a Gaoth Dobhair normalisation rule might convert 9:15 to ceathrú i ndiaidh a naoi while a Conamara rule would convert it to ceathrú tar éis a naoi.
6.8 Tracing the application of the phonetiser: a detailed example

To illustrate how the process of phonetisation works from start to finish, I included here a trace or a log file of how an actual word is processed by the phonetiser. This log file is what is actually printed out by the phonetiser if required, and allows developers to identify at which point a problem is encountered and make necessary changes. For the purposes of this illustration, I have chosen to use a word that will not be contained in the lexicon so as to highlight in particular LTS rule application. The word chosen is the Irish boy's name *Tadhg*, for the Gaath Dobhair local module.

GLOBAL INPUT: "tadhg"

SEARCHING GLOBAL EXCEPTIONS LEXICON...
--> No matching entry

APPLYING GLOBAL RULES...

Applying rules to [#tadhg#]

Matching focus in rule [t]CS_*RCSV_NONSYLLABIC_ = TJ
LEFT CONTEXT: #
RIGHT CONTEXT: a d h g #

Trying rule: [t]CS_*RCSV_NONSYLLABIC_ = TJ
--> Left Context: PASS
--> RIGHT Context: FAIL

Matching focus in rule NONSYLLABIC_LCSV_CS_+[t] = TJ
LEFT CONTEXT: #
RIGHT CONTEXT: a d h g #

Trying rule: NONSYLLABIC_LCSV_CS_+[t] = TJ
A detailed description of the global and local rules

--> Left Context: FAIL

Matching focus in rule NONSYLLABIC_LCSV_[t] = TJ
LEFT CONTEXT: #
RIGHT CONTEXT: a d h g #

Trying rule: NONSYLLABIC_LCSV_[t] = TJ
--> Left Context: FAIL

Matching focus in rule [t] = T
LEFT CONTEXT: #
RIGHT CONTEXT: a d h g #

Trying rule: [t] = T
--> Left Context: PASS
--> RIGHT Context: PASS
RULE APPLIED: [t] = T
OUTPUT: "T"
REST: "a d h g #"

Matching focus in rule [adh]# = ADH#
LEFT CONTEXT: # t
RIGHT CONTEXT: g #

Trying rule: [adh]# = ADH#
--> Left Context: PASS
--> RIGHT Context: PASS
RULE APPLIED: [adh] = AI
OUTPUT: "T AI"
REST: "g #"

Matching focus in rule NONSYLLABIC_LCSVV_LNR_+[g] = @@ G
LEFT CONTEXT: # t a d h
RIGHT CONTEXT: #
A detailed description of the global and local rules

Trying rule: NONSYLLABIC_LCSBV_LNR_+[g] = @@ G
--> Left Context: FAIL

Matching focus in rule NONSYLLABIC_LCSV_LNR_+[g] = @@ GJ
LEFT CONTEXT: # t a d h
RIGHT CONTEXT: #

Trying rule: NONSYLLABIC_LCSSV_LNR_+[g] = @@ GJ
--> Left Context: FAIL

Matching focus in rule [g]CS_*RCSV_NONSYLLABIC_ = GJ
LEFT CONTEXT: # t a d h
RIGHT CONTEXT: #

Trying rule: [g]CS_*RCSV_NONSYLLABIC_ = GJ
--> Left Context: PASS
--> RIGHT Context: FAIL

Matching focus in rule NONSYLLABIC_LCSV_CS_+[g] = GJ
LEFT CONTEXT: # t a d h
RIGHT CONTEXT: #

Trying rule: NONSYLLABIC_LCSV_CS_+[g] = GJ
--> Left Context: FAIL

Matching focus in rule NONSYLLABIC_LCSV_[g] = GJ
LEFT CONTEXT: # t a d h
RIGHT CONTEXT: #

Trying rule: NONSYLLABIC_LCSV_[g] = GJ
--> Left Context: FAIL

Matching focus in rule [g] = G
LEFT CONTEXT: # t a d h
RIGHT CONTEXT: #

Trying rule: [g] = G
--> Left Context: PASS
--> RIGHT Context: PASS
RULE APPLIED: [g] = G
OUTPUT: "T AI G"
REST: "#"
GLOBAL LOOKUP RETURNING: "T AI G"
ADDING STRESS...
OUTPUT: "T 1 AI G"

LOCAL INPUT: "tadhg"

DIALECT: Gaoth Dobhair

SEARCHING LOCAL EXCEPTIONS LEXICON...
--> No matching entry

APPLYING LOCAL RULES
Applying rules to [#T 1 AI G#]

Matching focus in rule [T] = t
LEFT CONTEXT: #
RIGHT CONTEXT: 1 AI G #

Trying rule: [T] = t
--> Left Context: PASS
--> RIGHT Context: PASS
RULE APPLIED: [T] = t
OUTPUT: "t"
REST: "1 AI G #"

Matching focus in rule [1] = 1
LEFT CONTEXT: # T
RIGHT CONTEXT: AI G #

Trying rule: [1] = 1
--> Left Context: PASS
--> RIGHT Context: PASS
RULE APPLIED: [1] = 1
OUTPUT: "t 1"
REST: "AI G #"

Matching focus in rule [AI] DNST_ = ai
LEFT CONTEXT: # T 1
RIGHT CONTEXT: G #

Trying rule: [AI] DNST_ = ai
--> Left Context: PASS
A detailed description of the global and local rules

---> RIGHT Context: FAIL

Matching focus in rule [AI] = ee
LEFT CONTEXT: # T 1
RIGHT CONTEXT: G #

Trying rule: [AI] = ee
---> Left Context: PASS
---> RIGHT Context: PASS
RULE APPLIED: [AI] = ee
OUTPUT: "t 1 ee"
REST: "G #"

Matching focus in rule [G] = g
LEFT CONTEXT: # T 1 AI
RIGHT CONTEXT: #

Trying rule: [G] = g
---> Left Context: PASS
---> RIGHT Context: PASS
RULE APPLIED: [G] = g
OUTPUT: "t 1 ee g"
REST: "#"

LOCAL LOOKUP RETURNING: tadhg t 1 ee g
SYLLABIFYING...
OUTPUT: "1 t ee g"

PHONETISED OUTPUT: "1 t ee g"
Chapter 7

Testing and evaluating the phonetiser

7.1 Testing that the rules work together: internal rule compatibility

It goes without saying that a system such as the phonetiser described here is only useful if it produces good results. With a complex rule system such as this, it is difficult to predict how the rules will interact. It was common during the development of the Irish phonetiser, for example, for rules added to be too general and thus override already existing rules.

In order to deal with this challenge, a system was devised to test the rules during development. The goal was to be able to test the functionality of particular rules in a given module (global or local) so that if a new rule disrupted the functionality of an existing rule, the problem would be identified immediately as the test that
had previously passed would fail. This system for testing the rules is described in more detail in § 7.2.

Once rules are developed that pass all tests and in principle, produce the desired output, it is then necessary to evaluate the performance of the phonetiser in the real world. By this I mean to compare the transcriptions produced by the phonetiser to existing local transcriptions known to be correct. Ideally, a dialect-specific phonetically transcribed corpus would be used as a benchmark. We explored the possibility of using a dialect-specific lexicon that was created using a version of the *Foclóir Póca*. During the development of the Ulster Irish speech corpus created during the WISPR project (see § 2.6 and Ni Bhriain, 2008), this dialect-specific lexicon was semi-automatically modified for Gaoth Dobhair Irish and expanded to include all words recorded in the speech corpus. Unfortunately, given the semi-automatic techniques used to develop this lexicon, it cannot be treated as a gold standard, and tests to compare the output of the phonetiser and the transcriptions in this lexicon proved this to be the case.47

As discussed earlier, no suitable dialect-specific phonetically transcribed corpora or pronunciation dictionaries were available (see § 1.4). Given the absence of appropriate corpora, we decided to use the Lárchanúint artificial dialect transcriptions in the *Foclóir Póca* pronunciation dictionary (see § 3.2 for a description of the Lárchanúint).

---

47 The phonetiser only produced the same transcription as was stored in the lexicon in 54.0% of all entries tested. This result can on the one hand be attributed to the phonetiser's lack of interaction with fully developed linguistic analysis components (i.e. morphological and syntactic analysis), and on the other, to inconsistencies in the lexicon.
7.2 Syntax of the tests

A system for testing was devised to allow testing of individual rules in particular modules (global or local), as well as testing of the process of converting ordinary Irish text to a dialect-specific (local) representation as a whole. Tests for individual rules can be applied to both global and local modules. These tests fail if the rule being tested is either not written correctly, or does not interact correctly with all the other rules. This is useful because, if a test for an existing rule passes prior to the addition of a new rule, but fails following the addition of a new rule, the lack of correct interaction between particular rules is flagged immediately.

The tests themselves are a type of rule with a left-hand side (LHS) and right-hand side (RHS), which are separated by an arrow (\(--\rightarrow\)). The LHS is the test input: an orthographic word in the case of the global rule file, or a global transcription in the case of the local rule files. The RHS of the test is the desired output. The way the tests are applied is that the RHS is compared with the output of the stage of phonetisation in question, either global or local. If the output, which is the result of the application of the rule file as a whole, matches the desired output as transcribed on the RHS of the test, then the test passes. If not, then the test fails. The software is designed so that testing can be switched on or off, and if switched on (test mode), the results of the tests are outputted at runtime.

The tests are interspersed within the rule files, both global and local. In keeping with the goal to have linguist-friendly rule files and exceptions lexicons that are human-readable, the tests are clearly distinguishable from the rules and have the syntax shown in (18).
An example of a test from the global rule file is shown in (18a). This rule tests if the orthographic sequence ‘adh’ is correctly mapped to the global unit AI. An example of a test from the Gaoth Dobhair local rule file is shown in (18b). This rule tests if the global unit AI is correctly mapped to the local phone /ee/, pronounced [eː].

In addition to the module-specific testing, the phonetisation process from start to finish, i.e. ordinary spelling to local transcription, can be tested as a whole. These tests are included in a separate file as part of the local modules. In the current implementation, a file called test.txt is included with each local module. The difference in the format of these tests from the module-specific tests already described is that the left-hand side (input) of these tests is orthographic, and the right-hand side (output) is local or dialectal, as in (19), where the Gaoth Dobhair pronunciation is outputted.

(19) abhainn 1 oo nnj

The reader should note from this example that the syntax of these tests differs from the module-specific tests in that they more closely resemble an entry in a pronunciation dictionary. The reason for this was to allow the inclusion as test sets of any dialect-specific pronunciation dictionaries that may become available in the future.
7.3 Evaluating the output

7.3.1 A point of comparison: dictionaries vs. phonetised corpora

Despite the rigorous testing during the development stage, it is desirable to evaluate the output of such a rule system against some real world benchmark. As mentioned in the introduction to this chapter, the best benchmark against which to evaluate the phonetiser’s output would be a corpus with dialect-specific phonetic transcriptions. A corpus of news articles, blog entries, e-mails, and other types of documents would contain inflected and mutated forms as well as base forms of words. A pronunciation dictionary, on the other hand, typically contains uninflected base forms as headwords. Consider, for example, the excerpt in (20) taken from the online magazine beo.ie, where the limited number of unique dictionary headwords found in Foclóir Póca are underlined.

(20) Bhi Humptáí Dumptaí thuas ar an mballa, shéid an ghaoth is thit sé ar an talamh, tháinig amach an tseanbhean bhuf, is nuair chonaic sí HD ina lui, “anois” a deir sí, “ní fhéadfadh an rí, ná an méid capall is fear atá faoi, thú a chur ar aí arís mar a bhí!” (Nic Gabhann, 2009)

Note that the inflected and mutated forms are not present in the dictionary (shéid, ghaoth, thit, etc.). A better test of a phonetiser would be how it handles the more common inflected and mutated forms. In addition, it is possible to calculate lexical frequency (an estimate of how common a word is) from corpora. This is an advantage because it is particularly important for the phonetiser to deal with words that occur most frequently in the language.
The issue of certain word forms not being represented in a pronunciation dictionary is significant for a fairly heavily inflected language such as Irish (see § 4.5), where words can have many different grammatical forms (see Figure 2.3) as well as multiple mutated forms (e.g. bris, mbrise, bhris, etc.). Most of these forms will not appear in a dictionary at all, since dictionaries contain primarily uninflected base forms as headwords. Therefore if evaluating against a dictionary as opposed to a corpus, the results will not inform us as to how well the phonetiser deals with mutations and inflectional endings.

In addition to this, dialectal word forms are more likely to occur in a corpus than in a dictionary. By this I mean words that are pronounced so differently from the standard spelling that local writers actually replace the standard spelling with a form that is more closely related to the local dialect. For example, the word form nuacht /\ ny\uaxt/ 'news' is replaced with nuaíocht /\ ny\u:ja\axt/ by writers of Donegal Irish, reflecting the pronunciation in those dialects. That dialectal words be represented is particularly important to our goals in multi-dialect synthesis.

The issue of word frequency information being absent from a pronunciation dictionary also has the potential to skew the results of an evaluation performed against a dictionary. The reason for this is that it is much more important that the phonetiser produce accurate results on the more common words and sequences of letters. Whether or not the phonetiser actually does this will not be apparent following an evaluation against a dictionary.
7.3.2 Test dialect: the Lárchanúint

Given the lack of a dialect-specific phonetically transcribed corpus, we explored the only option currently available to us, and used the Foclóir Póca (FP) pronunciation dictionary in the evaluation. As explained earlier, no gold-standard pronunciation dictionary was available, and the transcriptions in the FP, the only pronunciation dictionary available for Irish, are given in the Lárchanúint artificial dialect. However, since Lárchanúint suggests actual pronunciations, it can be treated in the context of multi-dialect phonetisation in the same way as a naturally occurring living dialect. A decision was made therefore to add a Lárchanúint local module to the phonetiser and add FP as a test set for this new module, thus evaluating the process of multi-dialect phonetisation.

The basic method for evaluation was to include an electronic version of the Irish headwords and their pronunciations from FP as a test set in the Lárchanúint local module. Additionally, the pronunciations were converted to MRPAI (see § 5.1). And since the phonetiser cannot currently handle words such as borrowings and compounds with marked stress patterns (e.g. to'bac ‘tobac’, ‘aer;fhórsa ‘air force’), all such words were removed from the test set. Since the vast majority of words have first syllable stress, only words with marked stress patterns have stress indicated in FP, therefore these words were easy to find and remove\(^48\). The remaining test set consisted of 12,191 entries. As was described in § 7.2, each entry in the test set is then treated as a test. Whether or not each test passes or fails is outputted when the phonetiser is run in test mode.

This evaluation of Lárchanúint local pronunciations produced by the phonetiser against the Lárchanúint transcriptions in FP showed reasonably good results. The

\(^{48}\text{2,462 (16.8\%) of the 14,653 entries in Foclóir Póca were removed before testing.}\)
phonetiser output matched the FP pronunciation in 10,968 out of 12,191 tests. This constitutes a success rate of 90.0%.

These test results merit investigation. A closer inspection of the results uncovered a number of reasons where the phonetiser failed repeatedly despite producing a desired result. The reasons for this relate to the transcriptions in FP. For example, while the phonetiser treats devoiced sonorants as single segments, Lárchanúint treats them as two segments. The difficulty is that the two segments can come in either order in Lárchanúint. That is to say, the global unit $R_D$ has to map to both /r h/ and /h r/ in the Lárchanúint. The same problem occurs with the global units $RJ_D$, $NJ_D$, $N_D$, $L_D$, and $LJ_D$ ($D$ means devoiced). For example, the letters ‘thr’ in wordPEOÁTHRAI ‘supplier’, are transcribed /hr/ in the Lárchanúint, whereas the letters ‘rth’, in the word SCOLÁRTHA ‘scholarly’, are transcribed /rh/. This resulted in 218 false test fails. In addition, while the phonetiser applies vowel reduction to all unstressed short vowels, reduction is only applied sporadically in FP to the short vowel /i/ when in an unstressed syllable. This resulted in 64 false test fails.

Taking these types of issues into account, the success rate can be revised up to 92.3%. It seems likely that were the phonetiser to be evaluated against a corpus rather than a dictionary, that the results would be even more satisfactory, in particular because of the lexical frequency issue mentioned in § 7.3.1.

We examined our evaluation results to see whether there was any effect of lexical frequency. We obtained token frequencies for the 12,191 entries of Foclóir Póca (FP, our test set) from Nua-Chorpas na hÉireann (The New Corpus for Ireland),\(^{49}\) which includes a very large corpus of Irish (30 million words).\(^{50}\) An

\(^{49}\)http://www.focloir.ie/corpus/
\(^{50}\)These frequencies are not publicly available. I thank Michal Boleslav Měchura for extract-
examination of this lexical frequency information shows that Foclóir Póca includes a large percentage of relatively rare words. About 10% of the words in FP (1,147 tokens) do not appear in Nua-Chorpas na hÉireann, and 30% (3,598 tokens) appear 6 times or fewer. And of course, as we have noted, very few inflected and mutated forms are included (some irregular inflected forms are included as headwords). Bearing in mind these limitations, we used the available lexical frequency information to examine success rates for words in different frequency ranges, dividing the corpus into 5 or 10 sets, depending on lexical frequency (e.g., the first 10%, the second 10%, etc.). These analyses revealed no frequency effects: the results for the different groups were stable, varying from the overall success rate by a couple of percentage points at most.

7.3.3 Hand-written rules versus automatically trained rules

The phonetiser was also evaluated against automatically trained LTS rules (see § 5.2 for a discussion of the technical requirements as well as advantages and disadvantages of automatic or self-organising techniques). Once again, given that Foclóir Póca (FP), a dictionary containing pronunciations for an artificial dialect, was the only reliable data available; it was therefore used as a training set. Using the CART building programme Wagon, which Festival offers (see § 2.2), LTS rules were both trained and evaluated against FP. This test produced 87.6% accuracy. One of the arguments put forward in this thesis has been that hand-written rules are better suited to the task of phonetisation for Irish (see § 5.2). Hand-written LTS rules produce results that are slightly better than automatically trained LTS rules (92.3% vs. 87.6%). Also, the automatically trained rules are based on the

ing this information from the corpus. Note that this is not a phonetised corpus.
artificial dialect in FP and thus would not work for actual dialects. Additionally, given the lack of dialect-specific data upon which to train LTS rules, producing dialect-specific LTS rules in this way is currently not technically possible. Producing sufficiently large databases for the three dialects treated here would be much more costly (in time and in person-hours) than the approach adopted here. Sproat (1998, p. 84) make a similar observation about phonetisation for Russian:

Under that scenario, a self-organizing technique would not afford much of a savings in effort. Add to this the fact that such techniques depend upon having a large database that gives the correspondence between the spelling and the pronunciation of words, and that such databases are often lacking for languages other than a few of the more intensively studied Western European ones. For example, we know of no database that gives orthography/phonology mappings for Russian inflected words: there are pronunciation dictionaries, but these almost exclusively list base forms. Yet such a database for inflected words would be critical for training a self-organizing method for the pronunciation of unrestricted Russian text. (p. 84)

Note that for Irish, unlike Russian, there are not even any dialect-appropriate pronunciation dictionaries.

7.4 Improving the results

In addition to the issues with *Focloir Póca* transcriptions, there are a number of other reasons that contribute to problems with the phonetisation. Some of
these issues are general issues that have already been flagged as resolvable follow­
ing future work, in particular, the development of syntactic and morphological analysis components which could supply grammatical and morphological infor­
mation to the phonetisation, for example in disambiguating the preposition \textit{ag}
‘at’ and the verbal noun particle \textit{ag} (\textit{ag òl} ‘drinking’, \textit{ag ithe} ‘eating’), which have
different pronunciations (see § 5.10). Others relate to ongoing tweaking of the
rules. It is envisaged that if suitably transcribed and verified dialectal corpora
become available in the future, they will be used to improve the performance of
the phonetiser.

In addition, as described in § 2.6.1, the TTS speech synthesiser available at
\textit{abair.ie} includes a feature that allows users to give feedback on the quality of
the synthesis, a feature inspired by the feedback feature of the AT&T speech
synthesis demo web page.\footnote{http://www2.research.att.com/~ttsweb/tts/demo.php, Pauline Welby, personal com­
munication.} The \textit{abair.ie} feedback feature stores the text inputted
by the user, and the phone string and sound file generated, as well as optional
user comments entered in a simple feedback form. Of course, problems with
synthesis output can have many sources (errors in the normalisation or the corpus
annotation, bad joins, etc.), but some are phonetisation problems; identifying
such problems can help us to improve the phonetiser.
Chapter 8

Irish phonetisation: applications beyond synthesis

8.1 Introduction

The main contribution of phonetisation for Irish is in multi-dialect text-to-speech (TTS) synthesis, that is, transforming written Irish to phone strings that can be used by a synthesis system to produce dialect-appropriate speech. Text-to-speech synthesis, in turn, has a number important applications, including addressing the access needs of the visually-impaired and the communication needs of the vocally impaired.

But there are applications beyond synthesis for Irish phonetisation. I will address a number of these potential applications in this chapter, discussing in some detail the use of phonetisation in the development of dialect-specific pronunciation dictionaries and other computer-assisted language learning (CALL) technologies
and in psycholinguistic research.

8.2 Pronunciation guides

While a broad level of transcription is suitable for phonetisation for unit-selection TTS synthesis (see § 2.3.1), it would not be suitable for some of the other applications envisaged. The first of these relates to education. Learners may require a finer level of detail than native speakers of a language. Adult learners of Irish, students of Irish interested in the dialects and the differences between them, and school children in Ireland, could benefit from pronunciation guides such as dialect-specific pronunciation dictionaries that provided a narrower level of pronunciation transcription than is found in *Focloir Póca*. As we have seen, the Lárchanúint pronunciations in *Focloir Póca* are by design, not consistent with any single living dialect (see § 3.2) and may be difficult for learners to understand because of their somewhat abstract nature. For example, the word *cad* ‘what’ is transcribed */kad/ in *Focloir Póca* and the word *cead* ‘permission’ is transcribed */cad/*, with the same vowel symbol */a/*. But as actually pronounced by native speakers, the vowel in these two words is quite different. Following the initial phonemically velarised consonant in *cad*, the vowel is pronounced as */æ/*, but following the initial phonemically palatalised consonant in *cead*, it is pronounced */æ/* in all dialects. This letter-to-sound correspondence is quite regular in the language (for example, *gar* */gær/ ‘near’ [gær] vs. *gearr* */jær/ ‘short’ [jar], see § 3.4 on the realisation of the low vowel phoneme), but it is not present in the Lárchanúint transcription, and to correctly produce it, learners need somehow to remember this and then apply this transformation to the Lárchanúint transcriptions. Another reason the *Focloir Póca* pronunciations are limited in usefulness to learners is the
fact that they do not distinguish between the dialects. For example, in the word 
dearbhán ‘voucher’, transcribed as /dˈarəvən/ in Foclóir Póca, the /v/ conso-
nant is pronounced as [v^] in Corca Dhuibhne (Munster) but as [w] in Conamara
(Connaught) and Gaoth Dobhair (Ulster). In addition to this, the initial conso-
nant is pronounced as an alveolar affricate [tʃ^] in Conamara but is pronounced
as an postalveolar affricate [tʃj] in Gaoth Dobhair (see § 3.3). This difference is
not reflected in the transcription used for synthesis—it does not need to be—but
it is a striking difference between dialects.

Dictionaries containing pronunciations are common for other languages, par-
ticularly for languages without one-to-one letter-to-sound correspondences, in
both print and electronic formats (for example, the American Heritage Diction-
ary (Pickett, 2000), the Oxford English Dictionary (Simpson & Weiner, 1989),
and the Oxford-Hachette French Dictionary, Corréard et al., 2007). Adding pro-
nunciations to Irish language resources such as the new English-Irish dictionary,
currently being written,52 the online phonetiser tool on abair.ie,53 the online
terminology database focal.ie54 (see Figure 8.1), and the online dictionary for
learners potafocal.com55 could be highly beneficial to users of these resources.

Entries in online electronic dictionaries could be enhanced to include not only
written words and glosses or definitions, but also a phonetic transcription for each
major dialect and a synthesis-generated sound file. And since this information
would be available online, learners are more likely to use it: research shows that
learners look up words more frequently when they are reading texts linked to elec-
tronic dictionaries or online glosses than when they are reading traditional texts

52http://www.focloir.ie/
53http://www.abair.ie/
54http://www.focal.ie/
55http://www.potafocal.com
and using printed dictionaries (Goyette, 1995). CALL applications integrating TTS synthesis have been used in the teaching of another Celtic language, Breton: a bilingual Breton/French and courseware for learning spelling-pronunciation correspondences and intonation patterns (Auclerc et al., 2000; Mercier et al., 2000; Messager et al., 1998).

The synthesis voice used for the Gaeth Dobhair voice, which uses the phonetiser developed for this thesis, is already used in this type of application. For example, the English-Irish translation application for mobile phones called Get the Focal\footnote{http://www.getthefocal.com/} uses to Gaeth Dobhair voice provided by abair.ie to synthesise pronunciations for the Irish entries.

With the click of a mouse, pronunciations could be provided not only for headwords, but also for all inflected and mutated forms. This would be particularly useful given that learners are often quite sure about the pronunciations of base forms that appear as headwords in dictionaries, but may have difficulty with other forms of those words. As was shown in Figure 2.3, Irish words, such as the verb *bris* 'to break', may have many inflected and mutated forms, for which pronunciations may not be known. For example, the learner may not know that the ‘f’ in the conditional form *brisfeadh* is not pronounced in any dialect and that this form of the verb is pronounced [b\textsuperscript{h}ri\textsuperscript{f}a\textsuperscript{x}] in Conamara, but [b\textsuperscript{h}ri\textsuperscript{f}u\textsuperscript{i}z] in Gaeth Dobhair (or [b\textsuperscript{h}ri\textsuperscript{f}u\textsuperscript{i}z] when there is a pronominal subject). Other forms that might cause difficulty include mutated forms including *mbrisfeadh*, *bhrisfeadh*, and so on. Initial mutations are pervasive in the language but are difficult for learners to decode (Ó Direáin, 2000; Scott & Kenny, 2000).

To accomplish this, another layer of rules must be written, which could then
Irish phonetisation: applications beyond synthesis

be used to automatically produce more user-friendly dialectal pronunciations. If the level of transcription and phoneset are well defined, a good deal should be achievable through automatic methods, although some manual checking will of course be required.

Although full development of this extra layer is beyond the scope of this thesis, it is envisaged that any such future development should follow certain basic principles, in particular those of the International Phonetic Alphabet, devised in the 1800s and now established best practice. One of these principles states that "[w]hen any sound is found in several languages, the same sign should be used in all. This applies also to very similar shades of sound". Another states that "[i]n assigning values to the Roman letters, international usage should decide" (International Phonetic Association, 1888) (quoted in Albright, 1958, p. 66). These
basic principles entail, for example, that a v-symbol should always be used for a v-sound, a w-symbol should always be used for a w-sound, and so on. So in the case of Irish dialects, when a v-sound is pronounced, it should be transcribed with a v-symbol, and likewise when a w-sound is pronounced, it should be transcribed with a w-symbol, even if this distinction is to do only with dialectal surface realisations. Intuitively this makes sense. Learners of Irish grow up reading other languages (particularly English) where it is standard for a v-symbol to represent a v-sound. The approach of the developers of the IPA is supported by research showing interference of L1 orthography on the learning of L2 (see § 4.2). Transcribing a v-sound with anything other than a v-symbol is likely to lead to confusion. For example, using (mo) mhála /və:lə/ to represent the Donegal pronunciation [wæ:ə] is not likely to help learners learn Donegal Irish.

It is important to develop and test any such online pronunciation aids in conjunction with users. Non-linguists are not typically familiar with IPA transcription and its use may even be counterproductive, “partly owing to the deterrent effect which a free use of such [phonetic] symbols has upon many readers” (O’Rahilly, 1972, p. XI). Furthermore, there is at least some research that shows that learners cannot take full advantage of pronunciations in dictionaries without training (Fraser, 1997). A complementary guide to the transcription system used should therefore also be developed. Such a guide could, for example, give instructions such as “this sound should be produced with rounded lips” (e.g. for the ⟨p⟩ in páiste ‘child’ in Donegal Irish) or “in this dialect this sound is pronounced like the Irish English j-sound” (e.g. for the ⟨d⟩ of dearbhán ‘voucher’ in Donegal) as to how to pronounce a particular symbol. A tool that would give similar assistance to users was proposed by Ó Néill & Mac Lochlainn (2007).
8.3 Theoretical and applied psycholinguistics

These rules could also be applied to the development of a dialect-specific phonetised corpus which could be used to investigate the influence of phonological pattern frequency on different aspects of language processing. A growing body of research on a variety of languages has used phonetised databases to show that listeners are sensitive to the frequency of phonetic patterns in their native language. These databases include phonetic transcriptions and other information, such as part of speech, syllable structure, token frequency, lemma frequency, neighbourhood density and uniqueness point (CELEX: Baayen et al., 1995; Hoosier Mental Lexicon for English: Nusbaum et al., 1984; Lexique for French: New, 2006; and the Japanese Mental Lexicon: Kess & Miyamoto, 1999, among others). A host of findings in psycholinguistic, phonetic, phonological, and reading research have used research based on these kinds of databases. As Davis (2005, p.66) notes, "[t]he fact that high-frequency words are responded to more rapidly than low-frequency words are in word-reading tasks is possibly the most reliable finding in experimental psycholinguistics". Phonological pattern frequency has also been used to explain differing patterns of acquisition (e.g. Beckman et al., 2003 for English vs. Japanese), to explain word segmentation preferences (e.g. Welby, 2003 for French), and to examine many other research questions.

One could imagine a number of questions about Irish that could explored. For example, what are the relative frequencies of different types of phonemes, consonant sequences, syllables, and syllable onsets or codas in Irish and do children acquire more frequent types earlier? Some of this information is not available in a phonetised lexicon (dictionary). For example, some onset clusters are only found in nasalisation/eclipsis contexts (/nr/ ár ndráma ‘our drama’, /ml/ i mbliana
Irish phonetisation: applications beyond synthesis

'this year') and would not be present in a lexicon. And some phones have a low type frequency, but a high token frequency. For example, in English the voiced interdental fricative [ð] is present only in a few words (low type frequency), but one of them, the definite article the, is one of the most common words in the language. In addition, dialect-specific phonetisations are important for addressing certain questions. Researchers examining the acquisition of fricatives would want, for example, to distinguish between the intervocalic [h] in Donegal Irish and the intervocalic [x] in other varieties in words like clocha 'stones'. In addition, Irish is characterised by a number of typologically interesting characteristics, including its system of initial mutations and its phonological opposition between velarised and palatalised segments, the production and perception of which cannot be adequately studied without a database containing the types of information described above.

The development for similar databases for other languages is well-documented, and resources that could be used to develop a database for Irish are publicly available (N-Watch Davis, 2005, E-Hitz Perea et al., 2006). Existing corpora of written Irish (The National Corpus for Ireland, for example) could be passed through a phonetiser to build a phonetised database for Irish. Additional processing would extract word frequency (how often a given word appears, usually expressed per million words) and other types of information. Of course, writing differs in important ways from speaking—different vocabulary is used and certain themes are over- or under-represented. Most phonetised corpora are based on written sources, though, and these corpora share similar drawbacks. For example, the Irish word ach 'but' and therefore the phone [x] might be much more common in speech than in writing, where words associated with a more formal register might be used. Similarly, English loan words and code switching are very common in
spoken Irish, but less common (though present) in writing. Also, dialect-specific words like *bomaite 'minute' (Ulster) are likely to be under-represented in a written corpus. But even though a phonetised corpus based on written sources would not be perfect, it would still be an important resource.

8.4 Conclusion

In this chapter I have looked at two areas where the capabilities of the Irish phonetiser could be applied: the area of dialect-specific pronunciation dictionaries, and the area of theoretical and applied psycholinguistics. Although we focused on these two area in particular, there may of course be other applications. For example, the tool that converts older spelling to modern spelling\(^57\) could be interfaced with the phonetiser to read aloud older texts such as *An Béal Bocht* (see Figure 4.2).

\(^{57}\)See http://borel.slu.edu/
Chapter 9

Conclusion

9.1 Goals and results

The goal of this thesis, as set out in the introduction, was to develop an Irish-language phonetiser, a tool to convert orthographic transcription to phonemic transcription, letters to sounds. Furthermore, given the lack of a standard spoken variety of Irish, and the prominence of several dialects, a phonetiser that produces dialectal (local) phonemic output was required for at least the three main dialect groups of Irish (Ulster, Connaught, and Munster). Previous research on the sound system of Irish (Ó Murchú, 1969; Ó Baoill, 1986b) argued that the various dialects of Irish have much in common at a structural level. Taking this into account, it was envisaged that what the dialects had in common could be harnessed to facilitate the development of a multi-dialect phonetisation framework for Irish dialects, rather than developing a separate phonetiser for each dialect.

To achieve the goal of multi-dialect phonetisation, a global-local modular ap-
proach was adopted. This approach distinguishes between what is encompassed globally by all dialects, and what is purely local to a particular dialect (hence global-local). The modularity of the approach is crucial to facilitating mapping from a more abstract global level to the local level. The modularity means that the framework now in place and already at least partially implemented for a number of dialects allows phonetisation to be extended to other dialects and sub-dialects. For example, while the current work includes implementations for Gaoth Dobhair (Ulster), Conamara (Connaught), and Corca Dhuibhne (Munster) Irish, researchers and users may be interested in the development of local modules for some of the smaller dialects in other counties of these provinces (Co. Mayo and Co. Cork, for example). Because the bulk of the conversion is done by the global module, which does not have to be modified, it should be possible to add new dialects with relative ease.

As of writing, the phonetisation work described here has directly contributed to a number of resources integrating synthesis. The research team at TCD has used the phonetiser to develop two unit selection synthesis voices: one for Gaoth Dobhair (Co. Donegal, Ulster) Irish and one for Conamara (Co. Galway, Connaught) Irish, as well as a HTS voice for Conamara Irish. The Gaoth Dobhair voice is currently available at abair.ie, and the two Conamara voices will be soon added to the site. The HTS voice has been used to produce a beta version of a screen reader. This tool is not yet publicly available, but was pilot-tested by a blind user just last month. An English-Irish translation application for mobile phones, called “Get the Focal” \textsuperscript{58} uses the Gaoth Dobhair unit selection voice to synthesis pronunciations for the Irish entries.

\textsuperscript{58}http://www.getthefocal.com/
9.2 Future work

There are a number of ways in which the performance of the phonetiser might be improved. In discussing these potential improvements, it is important to distinguish between improvements to the architecture of the phonetiser itself (i.e., the letter-to-sound (LTS) rules and the exceptions lexicons) and the development and inclusion of other components which will result in better phonetisation. In the first case, it is possible that the division of labour between global and local rules could be further optimised. It is also essential that development time and resources be devoted to soliciting and using native speaker feedback to tweak the LTS rules and exceptions lexicons, particularly for the Conamara and Corca Dhuibhne modules, which has not yet been incorporated into the TTS system.

In the second case, we have noted a number of places where temporary solutions were used (to strip initial mutations, for example; see § 5.10), in the absence of a specially designed component. The performance of the phonetiser would be improved by the development and inclusion of additional linguistic analysis components as discussed below.

For example, if a morphological analysis module is developed and incorporated, the phonetiser, which contains an entry for *eochair* ‘key’ in the global exceptions lexicon because of the short /o/ vowel in the word (/eο/ usually pronounced with a long /o:/), would produce the correct transcription (with short /o/) for *heochair, n-eochair, eochrach, heochrach, eochracha, heochracha*, and *n-eochracha*, the various mutated and inflected forms of the word.

If syntactic analysis, and so part-of-speech information is incorporated into the phonetiser, it could be used to distinguish between, for example, the prepositional
Conclusion

163

Ag ‘at’ [æ] and the verbal noun particle ag [æ], [æɡ], or [æj] (ag rith ‘running’, ag ól ‘drinking’, ag ithe ‘eating’). Distinguishing between homographs in this way is important because in this case, and in many other cases, homographs are pronounced differently. For the phonetiser to be able to make the distinction, first the grammatical information would have first to be obtained, either from a dictionary or from a part-of-speech tagger (Úi Dhonnchadha & Van Genabith, 2006), and then used to distinguish between homographs. After this, the addition of a post-lexical module, a module that deals with cross-word coarticulation and other phenomena, would enable the phonetiser to select the correct one of the three possible pronunciations of the verbal noun particle ag ([æ], [æɡ], or [æj]).

Deciding what pronunciation is correct in this case depends on whether or not the verbal noun begins with a consonant, a back vowel, or a front vowel. The addition of such modules would constitute a significant enhancement to the phonetiser.

One other possible enhancement would be the full integration of phonetisation with spell checking. This way, spelling errors could be removed before phonetisation, thus improving the quality of phonetisation, which depends crucially on the quality of the spelling.

Finally, the extent to which the phonetiser meets some of its stated goals remains to be seen. For example, while the architecture of the system should allow a fairly straightforward extension to new dialects by other researchers and developers, this has not yet been demonstrated. And while the phonetiser has so far been used to build three synthetic voices that have been used in a number of applications, many other applications, such as the use of an extended version of the phonetiser in electronic dictionaries (see Chapter 8), remain to be explored.

I look forward to working on some of these research questions and also hope
that this thesis provides a basis for work in this area by other researchers.
Bibliography


Corréard, Marie-Hélène, Valerie Grundy, Jean-Benoît Ormal-Grenon & Natalie
Bibliography


Ní Chiosáin, Máire. 1999. Syllables and phonotactics in Irish. In Harry van der


Translation Section, the Houses of the Oireachtas. 1958. {	extit{Gramadach na Gaeilge agus litriú na Gaeilge: An caighdeán oifigiúil}. Dublin: Stationery Office.}


Ui Dhonnchadha, Elaine. 2009. {	extit{Part-of-speech tagging and partial parsing for irish using finite-state transducers and constraint grammar}. Dublin: Dublin City University Ph.D. thesis.}


Wagner, Heinrich. 1964. {	extit{The linguistic atlas and survey of Irish dialects}}, vol. 2: The dialects of Munster. Dublin: Dublin Institute of Advanced Studies.


Wagner, Heinrich. 1969. {	extit{The linguistic atlas and survey of Irish dialects}}, vol. 4: The dialects of Ulster and the Isle of Man. Specimens of Scottish Gaelic dialects.
Bibliography


Welby, Pauline. 2003. The slaying of Lady Mondegreen, being a study of French tonal association and alignment and their role in speech segmentation. Columbus: The Ohio State University dissertation.


Appendix A

Irish phoneset comparison tables
### Table A.1: Irish labial phones.

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Lárchanúint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>/p′/</td>
<td>/p′/</td>
<td>/p′/</td>
<td>/pj/</td>
</tr>
<tr>
<td>/p’/</td>
<td>/p/</td>
<td>/p′/</td>
<td>/p′/</td>
<td>/p/</td>
</tr>
<tr>
<td>/b/</td>
<td>/b′/</td>
<td>/b′/</td>
<td>/b′/</td>
<td>/bj/</td>
</tr>
<tr>
<td>/b’/</td>
<td>/b/</td>
<td>/b′/</td>
<td>/b′/</td>
<td>/b/</td>
</tr>
<tr>
<td>/θ/</td>
<td>/θ′/</td>
<td>/θ′/</td>
<td>/θ′/</td>
<td>/θj/</td>
</tr>
<tr>
<td>/ρ/</td>
<td>/ρ′/</td>
<td>/ρ′/</td>
<td>/ρ′/</td>
<td>/ρ/</td>
</tr>
<tr>
<td>/v/</td>
<td>/v/</td>
<td>/v/</td>
<td>/v/</td>
<td>/v/</td>
</tr>
<tr>
<td>/m/</td>
<td>/m′/</td>
<td>/m′/</td>
<td>/m′/</td>
<td>/mj/</td>
</tr>
<tr>
<td>/m’/</td>
<td>/m/</td>
<td>/m′/</td>
<td>/m′/</td>
<td>/m/</td>
</tr>
</tbody>
</table>

### Table A.2: Irish coronal phones (Part I).

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Lárchanúint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ʃ/</td>
<td>/ʃ  t′s′/</td>
<td>/t′s′/</td>
<td>/t′s′/</td>
<td>/tjsj/</td>
</tr>
<tr>
<td>/ç/</td>
<td>/ç  d′z′/</td>
<td>/d′z′/</td>
<td>/d′z′/</td>
<td>/djzj/</td>
</tr>
<tr>
<td>/t/</td>
<td>/t/</td>
<td>–</td>
<td>–</td>
<td>/t/</td>
</tr>
<tr>
<td>/t’/</td>
<td>/t’/</td>
<td>/t’/</td>
<td>/t’/</td>
<td>/t/</td>
</tr>
<tr>
<td>/t’/</td>
<td>/t/</td>
<td>/t/</td>
<td>/t/</td>
<td>/t/</td>
</tr>
<tr>
<td>/d/</td>
<td>/d/</td>
<td>–</td>
<td>–</td>
<td>/d/</td>
</tr>
<tr>
<td>/d’/</td>
<td>/d’/</td>
<td>/d’/</td>
<td>/d’/</td>
<td>/d/</td>
</tr>
<tr>
<td>/d’/</td>
<td>/d/</td>
<td>/d/</td>
<td>/d/</td>
<td>/d/</td>
</tr>
<tr>
<td>/l’/</td>
<td>/l’  l′/</td>
<td>/l′/</td>
<td>/l′/</td>
<td>/lj/</td>
</tr>
<tr>
<td>/l’/</td>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
</tr>
<tr>
<td>/l’/</td>
<td>/l’  hl′/</td>
<td>–</td>
<td>–</td>
<td>/lj_d/</td>
</tr>
<tr>
<td>/l/</td>
<td>/l/</td>
<td>–</td>
<td>–</td>
<td>/l_d/</td>
</tr>
<tr>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
</tr>
<tr>
<td>–</td>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
<td>/l/</td>
</tr>
</tbody>
</table>
### Table A.3: Irish coronal phones (Part II).

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Láirchenúint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/u/</td>
<td>/u/</td>
<td>-</td>
<td>-</td>
<td>/nnj.d/</td>
</tr>
<tr>
<td>/ũ/</td>
<td>/ũ/</td>
<td>-</td>
<td>-</td>
<td>/nn.d/</td>
</tr>
<tr>
<td>/ũ/</td>
<td>/ũ/</td>
<td>/n/</td>
<td>/n/</td>
<td>/nnj/</td>
</tr>
<tr>
<td>/ũ̃/</td>
<td>/ũ̃/</td>
<td>/n/</td>
<td>/n/</td>
<td>/nn/</td>
</tr>
<tr>
<td>/ɲ/</td>
<td>/ɲ/</td>
<td>-</td>
<td>-</td>
<td>/nj.d/</td>
</tr>
<tr>
<td>-</td>
<td>/ɲ/</td>
<td>-</td>
<td>-</td>
<td>/n.d/</td>
</tr>
<tr>
<td>/n/</td>
<td>/n/</td>
<td>/n/</td>
<td>/n/</td>
<td>/nj/</td>
</tr>
<tr>
<td>-</td>
<td>/n/</td>
<td>/n/</td>
<td>/n/</td>
<td>/n/</td>
</tr>
<tr>
<td>/ɛ̃/</td>
<td>/ɛ̃/</td>
<td>-</td>
<td>-</td>
<td>/rj.d/</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>/ɛ/</td>
<td>/r/</td>
<td>/r/</td>
<td>/rj/</td>
</tr>
<tr>
<td>-</td>
<td>/ɛ/</td>
<td>/r/</td>
<td>/r/</td>
<td>/r/</td>
</tr>
<tr>
<td>/ɛ̃/</td>
<td>/ɛ̃/</td>
<td>/s/</td>
<td>/s/</td>
<td>/sj/</td>
</tr>
<tr>
<td>/s̃/</td>
<td>/s̃/</td>
<td>/s/</td>
<td>/s/</td>
<td>/s/</td>
</tr>
<tr>
<td>/z/</td>
<td>-</td>
<td>/z/</td>
<td>/z/</td>
<td>/z/</td>
</tr>
</tbody>
</table>

### Table A.4: Irish dorsal phones.

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Láirchenúint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/k/</td>
<td>/k/</td>
<td>/k/</td>
<td>/k/</td>
<td>/k/</td>
</tr>
<tr>
<td>/k̃/</td>
<td>/k̃/</td>
<td>/g/</td>
<td>/g/</td>
<td>/g/</td>
</tr>
<tr>
<td>/g/</td>
<td>/g/</td>
<td>/g/</td>
<td>/g/</td>
<td>/g/</td>
</tr>
<tr>
<td>/x/</td>
<td>/x/</td>
<td>/x/</td>
<td>/x/</td>
<td>/x/</td>
</tr>
<tr>
<td>/j/</td>
<td>/j/</td>
<td>/y/</td>
<td>/y/</td>
<td>/g/,j/</td>
</tr>
<tr>
<td>/y/</td>
<td>/y/</td>
<td>/y/</td>
<td>/y/</td>
<td>/g/</td>
</tr>
<tr>
<td>/ɛ̃/</td>
<td>/ɛ̃/</td>
<td>/ũ/</td>
<td>/ũ/</td>
<td>/ngj/.d/</td>
</tr>
<tr>
<td>/ŋ̃/</td>
<td>/ŋ̃/</td>
<td>/ŋ̃/</td>
<td>/ŋ̃/</td>
<td>/ngj/</td>
</tr>
<tr>
<td>/ŋ/</td>
<td>/ŋ/</td>
<td>/ŋ/</td>
<td>/ŋ/</td>
<td>/ŋ/</td>
</tr>
</tbody>
</table>

### Table A.5: Irish glottal phones.

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Láirchenúint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/h/</td>
<td>/h/</td>
<td>-</td>
<td>/h/</td>
<td>/h/</td>
</tr>
<tr>
<td>/h̃/</td>
<td>/h̃/</td>
<td>-</td>
<td>/h̃/</td>
<td>/h̃/</td>
</tr>
</tbody>
</table>

Table A.5: Irish glottal phones.
<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Lárnachuint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i:/</td>
<td>/i:/</td>
<td>/i:/</td>
<td>/i:/</td>
<td>/ii:/</td>
</tr>
<tr>
<td>/i/</td>
<td>/i/</td>
<td>/i/</td>
<td>/i/</td>
<td>/i/</td>
</tr>
<tr>
<td>/e:/</td>
<td>/e:/</td>
<td>/e:/</td>
<td>/e:/</td>
<td>/ee:/</td>
</tr>
<tr>
<td>/e/</td>
<td>/e/</td>
<td>/e/</td>
<td>/e/</td>
<td>/e/</td>
</tr>
</tbody>
</table>

Table A.6: Irish front vowel phones.

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Lárnachuint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a:/</td>
<td>/a:/</td>
<td>/a:/</td>
<td>/a:/</td>
<td>/aa:/</td>
</tr>
<tr>
<td>/a/</td>
<td>/a/</td>
<td>/a/</td>
<td>/a/</td>
<td>/a/</td>
</tr>
</tbody>
</table>

Table A.7: Irish low vowel phones.

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Lárnachuint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/u:/</td>
<td>/u:/</td>
<td>/u:/</td>
<td>/u:/</td>
<td>/uu:/</td>
</tr>
<tr>
<td>/x/</td>
<td>/u/</td>
<td>/u/</td>
<td>/u/</td>
<td>/u/</td>
</tr>
<tr>
<td>/o:/</td>
<td>/o:/</td>
<td>/o:/</td>
<td>/o:/</td>
<td>/oo:/</td>
</tr>
<tr>
<td>/ɔ:/</td>
<td>/ɔ:/</td>
<td>–</td>
<td>/o:/</td>
<td>/oo_o/</td>
</tr>
<tr>
<td>/ʌ/</td>
<td>/o/</td>
<td>/o/</td>
<td>/o/</td>
<td>/o/</td>
</tr>
</tbody>
</table>

Table A.8: Irish back vowel phones.

<table>
<thead>
<tr>
<th>IPA</th>
<th>CELTIC</th>
<th>Lárnachuint</th>
<th>X-SAMPA</th>
<th>MRPAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ə/</td>
<td>/ə/</td>
<td>/ə/</td>
<td>/ @/</td>
<td>/ @/</td>
</tr>
</tbody>
</table>

Table A.9: Irish unstressed vowel phone.
Appendix B

Letter-to-sound (LTS) rule files
B.1 Global/lts.lff

1 // This file is a component of a multi-dialect modular phonetiser for Irish.
2 // A phonetiser converts orthographic text to dialectal phonemic transcription.
3 // This phonetiser has two layers: one global, one local.
4 // At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
5 // At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
6 // The global layer consists of one global module.
7 // The local layer consists of one local module for each dialect implemented.
8 // Each module contains the following:
9 // (1) A set of rewrite (letter-to-sound) rules.
10 // (2) An exceptions lexicon.
11 // (3) A set of tests.
12
13 // This file contains the global rewrite rules.
14
15 // Variable syntax: <VARIABLE NAME><TAB>[
16 // (1) A set of rewrite (letter-to-sound) rules.
17 // (2) An exceptions lexicon.
18 // (3) A set of tests.
19
20 // This file contains the global rewrite rules.
21
22 // Variable syntax: <VARIABLE NAME><TAB>[
23 // Variable syntax: <VARIABLE NAME><TAB>[<VALUE 1>|<VALUE 2>|...|<VALUE N>]
24 // Rule syntax: <INPUT> = <OUTPUT>
25 // Test syntax: TEST <INPUT> --> <DESIRED OUTPUT>
26
27 // Author: Brian Ó Raghallaigh
28 // oraghalb AG tcd PONC ie
29
30 // VARIABLES
31
32 // Any Letter
33 AZ_ [a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z]
34
35 // Consonant Symbol
36 CS_ [b|c|d|f|g|h|j|k|l|m|n|o|p|q|r|s|t|v|w|x|y|z]
Letter-to-sound (LTS) rule files

29 // Consonant Symbol, 'á', 'ó', 'ú', or word start (syllable boundary)
30 NONSYLLABIC_ [á|b|c|d|f|g|h|j|k|l|m|n|ó|p|q|r|s|t|ú|v|w|x|y|z|#]

32 // Broad Future and Conditional Endings
33 BFCE_ [á|adh|aidh|aidís|aimid|aimis|ainn|as]
35 // Slender Future and Conditional Endings
36 SFCE_ [ea|eadh|eas|idh|idís|imid|imis|inn]
38 L_ [l|l|l]
40 MN_ [m|n]
42 FMP_ [f|m|p]
44 LNR_ [l|n|r]
46 LRST_ [l|r|s|t]
48 DNLST_ [d|n|l|s|t]
50 DNST_ [d|n|s|t]
52 Rdnlr_ [rd|rn|rl|rr]
54 // Vowel Symbol
56 VS_ [a|a|e|é|ê|i|i|ó|u|ú]
58 // Left Context Short Broad Vowel
60 LCSBV_ [ea|io|iu|a|o|u]
62 // Left Context Short Slender Vowel
64 LCSSV_ [ai|ei|ê|iu|i|oi|ui]
66 // Left Context Broad Vowel
68 LCBV_ [adh|ae|ao|á|ê|a|e|ó|e|i|io|iú|ó|o|ó|u|f|ó|u|ú]
70 // Right Context Broad Vowel
72 RCBV_ [aei|æi|ai|ao|a|ái|â|e|o|í|ó|ú|u|f|ó|ú|ú|ú]
74 // Left Context Slender Vowel (incl. diphthongs)
188

L etter-to-sou n d (L T S ) rule files

LCSV_

[aei|aidh|ai|al|aoi|ai|eai|eai|ei|eoi|e|ei|e|iai|iui|

iui|i|i|oi|6i|uai|ui|ui|ui]

//

Right

RCSV_

Context

Slender

Vowel

(incl.

diphthongs)

[eai|ea|eai|ea|ei|eoi|eo|e|ea|ei|e|iai|ia|io|iui|iu|

iiii Iiii Ii I1 o I1 ]
68
69

//

Left

70

L CL V _

Context

Long

Vowel

[aei|ae|aoi|ao|ai|a|eailea|eoi|e6|eo|ei|e|iui|iu|io|i

I6i I6 Iuio Iui I lii Ili]
71
72

//

Left

73

LCSLV_

Context

Slender

Long

Vowel

//

RULES

//

VOWELS

AND

TESTS

76
77

and

DIPHTHONGS

78
79

[Idh]

80

[ai]

81

[a]

= AA
= AA

= AA

82
83

//

star t

84

T ES T

adh

85

T ES T

aiseanna

86

TEST

at ha s

87

//

end

diphthongs)

[aei Ia i d h |aoi Iai |eai |eoi |ei Ie |iiii |i |6i Iuai |ui |lii]

74
75

(incl.

of

testset

-->

of

-->

-->
AA

88

[abh]

90

[adh]#

91

[adh]

= AI

92

[agh]

= AI

93

[aei]

= EE

94

[ae]

95

[alodh]#

= ABH

96

[aio]

97

[ai]

= ADH#

= EE
=

= AlO
= II

‘a ’

AA

testset

89

for

l OD H#

AA

SJ

H @@
for

NN
S

‘a ’

@@


Letter-to-sound (LTS) rule files

98  [aidh]\# = IDH#
99  [aidh] = AI
100 [aigh]\# = IGH#
101 [aigh] = AI
102 [aith]\# = ITHE#
103  #CS_*VS_*abh[ai]CS_+ =
104  #CS_*VS_*adh[ai]CS_+ =
105  #CS_*VS_*agh[ai]CS_+ =
106  #CS_*VS_*amh[ai]CS_+ =
107  #CS_*obh[ai]CS_+ =
108  #CS_*VS_*ðdh[ai]CS_+ =
109  #CS_*odh[ai]CS_+ =
110  #CS_*VS_*ogh[ai]CS_+ =
111  #CS_*VS_*omh[ai]CS_+ =
112  #CS_*VS_*umh[ai]CS_+ =
113 // check rr
114  #CS_*[ai]Rdnlr_ = AA
115  #CS_*[ai] = A
116 // unstressed short vowel reduction
117  VS_+CS_+[ai] = @@
118 [ai] = A
119 [amh]\# = AMH#
120 [amh] = AU
121 [aoi] = AO
122 [ao] = AO
123  #CS_*VS_*abh[a]CS_+ =
124  #CS_*VS_*adh[a]CS_+ =
125  #CS_*VS_*agh[a]CS_+ =
126  #CS_*VS_*amh[a]CS_+ =
127  #CS_*obh[a]CS_* =
128  #CS_*VS_*ðdh[a]CS_+ =
129  #CS_*odh[a]CS_* =
130  #CS_*VS_*ogh[a]CS_* =
131  #CS_*VS_*omh[a]CS_* =
132  #CS_*VS_*umh[a]CS_+ =
133 // check rr
134  #CS_*[a]Rdnlr_ = AA
//CS_*[a]L_NONSYLLABIC_ = 0
#CS_*[a] = A
// unstressed short vowel reduction
VS_+CS_+[a] = @@
[a] = A

// start of testset for 'a'
TEST abhainn --> ABH NNJ
TEST bualadh --> B UQ LL ADH#
TEST sadhbh --> S AI V
TEST saghas --> S AI S
TEST gaelige --> G EE LJ GJ @@
TEST saolaíodh --> S AO LL ÍODH#
TEST gardaí --> G AA R D II
TEST dúnfaidh --> D UU NN_D IDH#
TEST aidhm --> AI MJ
TEST cheadaigh --> XJ A D IGH#
TEST aighneas --> AI NJ @@ S
TEST diúltaithe --> DJ UU LL T ITHE#
TEST seabhaic --> SJ ABH KJ
TEST feadhain --> FJ AU NJ
TEST teaghas --> TJ AI SJ
TEST eamhain --> AU NJ
TEST lobhair --> LL OBH RJ
TEST leóthaí --> LJ OO SJ
TEST bodhair --> B ODH RJ
TEST eoghain --> 00 NJ
TEST broghais --> B R OGH SJ
TEST comhair --> K OO RJ
TEST ciúmhais --> KJ UU SJ
TEST airde --> AA RJ DJ @@
TEST cait --> K A TJ
TEST sodair --> S O D @@ RJ
TEST ait --> A TJ
TEST déanamh --> DJ EE NN AMH#
TEST amharc --> AU R K
TEST gaoil --> G AO LJ
TEST gaol --> G AO LL
TEST seabhac --> SJ ABH K
TEST ceadharaclach --> KJ AU R LL @0 X
TEST teaghasán --> TJ AI S AA NN
TEST lobhar --> LL OBH R
TEST leódhas --> LJ O0 S
TEST bodhar --> B ODH R
TEST eoghan --> OO NN
TEST bogha --> B OGH
TEST comhar --> K O0 R
TEST dumhach --> D UU X
TEST ard --> AA R D
TEST REQUIRED
TEST cat --> K A T
TEST sodar --> S O D @0 R
TEST at --> A T
// end of testset for 'a'

[éa] = EE
[éi] = EE
[é] = EE

// start of testset for 'é'
TEST éan --> EE NN
TEST éinini --> EE NJ II NJ II
TEST ée --> EE
// end of testset for 'é'

[eái] = AA
[eá] = AA
[eabh] = ABH
[eadh]# = ADH#
[eadh] = AU
[eaigh] = AI
[eai] = A
[eamh]# = AMH#
#CS_*[eamh] = AU
Letter-to-sound (LTS) rule files

#CS_*VS_igh[ea]CS_+ =
// check rr
#CS_*[ea]Rdnlr_ = AA
#CS_*[ea] = A
// unstressed short vowel reduction
VS_+CS_+[ea] = @0
[ea] = A
[eidh] = EIDH
[eigh] = EIGH
#CS_*[ei]Rdnlr_NONSYLLABIC_ = EE
#CS_*[ei] = E
// unstressed short vowel reduction
// VS_+CS_+[ei] = @0
[ei] = E
[eódh] = 00
[eoI] = 00
[eó] = 00
[eo] = 00
// borrowings only
#CS_*[e] = E
VS_+CS_+[e] = @0
// borrowings only
[e] = E

// start of testset for 'e'
TEST sheáin --> XJ AA NJ
TEST seán --> SJ AA NN
TEST seabhac --> SJ ABH K
TEST seinneadh --> SJ E NNJ ADH#
TEST ceadharlach --> KJ AU R LL @0 X
TEST teaghlach --> TJ AI LL @0 X
TEST beairic --> BJ A RJ @0 KJ
TEST áireamh --> AA RJ AMH#
TEST sleamhán --> SJ LJ AU NN AA NN
TEST oighear --> OIGH R
TEST ceard --> KJ AA R D
TEST cead --> KJ A D
246 TEST áireamhán --> AA RJ @0 V AA NN
247 TEST eas --> A S
248 TEST feidhm --> FJ EIDH MJ
249 TEST leigheas --> LJ EIGH S
250 TEST ceird --> KJ EE RJ DJ
251 TEST deis --> DJ E SJ
252 TEST REQUIRED
253 TEST eitpheil --> E TJ FJ E LJ
254 TEST ceoil --> KJ OO LJ
255 TEST bainseó --> B A NJ SJ OO
256 TEST ceol --> KJ OO LL
257 TEST REQUIRED
258 TEST uile --> I LJ @0
259 TEST REQUIRED
260 // end of testset for 'e'
261 262 [iodh]# = ÍODH#
263 [ío] = II
264 [í] = II
265 // start of testset for 'í'
266 TEST ceannáiodh --> KJ A NN ÍODH#
267 TEST síos --> SJ II S
268 TEST sí --> SJ II
269 // end of testset for 'í'
270 271 [iadh] = ÍQ
272 [iadh] = ÍQ
273 [iath] = ÍQ
274 [iai] = ÍQ
275 [ia] = ÍQ
276 [idh] = IDH#
277 [igh] = IGH#
278 [io] = IO
279 [ithe]# = ITHE#
280 [iúi] = UU
281 [iú] = UU
282 [iubh] = UBH
Letter-to-sound (LTS) rule files

\[\text{[iumh]} = \text{UU}\]
\[\text{[iui]} = \text{UU}\]
\[\text{[iu]} = \text{U}\]
\[\#\text{CS}_*\text{VS}_\text{idh}[i]\text{CS}_+\# = \]
\[\#\text{CS}_*\text{VS}_\text{igh}[i]\text{CS}_ = \]
// unstressed short vowel reduction
\[\text{VS}_+\text{CS}_+[i] = @@\]
\[\text{[i]} = \text{I}\]

// start of testset for ‘i’

TEST siadhail --> SJ IO LJ
TEST sciath --> SJ KJ IO
TEST riail --> RJ IO LJ
TEST siad --> SJ IO D
TEST seizedh --> SJ E NNJ_D IDH#
TEST cheannaigh --> XJ A NN IGH#
TEST fios --> FJ IO S
TEST imithe --> I MJ ITHE#
TEST siúil --> SJ UU LJ
TEST siúl --> SJ UU LL
TEST tiubh --> TJ UBH
TEST ciumhais --> KJ UU SJ
TEST giuirléid --> GJ UU RJ LJ EE DJ
TEST fiuch --> FJ U X
TEST leighis --> LJ EIGH SJ
TEST foighid --> F OIGH DJ
TEST aithris --> A RJ_D @@ SJ
TEST sin --> SJ I NJ
// end of testset for ‘i’

\[\text{[ódh]}# = \text{ÓDH}\#
\[\text{[ói]} = \text{OO}\]
\[\text{[ó]} = \text{OO}\]

// start of testset for ‘ó’

TEST cheannódh --> XJ A NN ÓDH#
TEST óil --> OO LJ
TEST 6l --> 00 LL
// end of testset for 'ó'
[obh] = OBH
[odh] = ODH
[ogh] = OGH
[oi] = II
[oidh] = OIDH
[oigh] = OIGH
// check rr
#CS_*[oi]Rdnlr_ = 00
// unstressed short vowel reduction
// borrowings
VS_+CS_+[oi] = @@
[oi] = 0I
[omh] = OMH
#CS_*[o]Rdnlr_NONSYLLABIC_ = 00
// unstressed short vowel reduction
// borrowings
VS_+CS_+[o] = @@
[o] = 0
// start of testset for 'o'
TEST lobhadh --> LL OBH ADH#
TEST todhchái --> T ODH X II
TEST toghadh --> T OGH ADH#
TEST oíche --> II XJ @@
TEST oidhe --> OIDH @@
TEST oighhear --> OIGH R
TEST boird --> B 00 RJ DJ
TEST bantaboic --> B A NN T @@ B @@ KJ
TEST soir --> S 0I RJ
TEST comhar --> K 00 R
TEST bord --> B 00 R D
TEST bantaboc --> B A NN T @@ B @@ K
TEST bos --> B 0 S
// end of testset for 'o'
Letter-to-sound (LTS) rule files

357 [úi] = UU
358 [ú] = UU
359
360 // start of testset for ‘ú’
361 TEST súil --> S UU LJ
362 TEST súl --> S UU LL
363 // end of testset for ‘ú’
364
365 [uauth] = UQ
366 [uai] = UQ
367 [ua] = UQ
368 [ubh] = UBH
369 [ue] = E
370 #CS_*[ui]Rdnlr_NONSYLLABIC_ = UU
371 #CS_*[ui] = I
372 // unstressed short vowel reduction
373 // borrowings
374 VS_+CS_+[ui] = @@
375 [ui] = UI
376 [uio] = II
377 [ui] = II
378 [umh] = UU
379 // check rr
380 #CS_*[u]Rdnlr_ = UU
381 #CS_*[u] = U
382 // unstressed short vowel reduction
383 // borrowings
384 VS_+CS_+[u] = @@
385 [u] = U
386
387 // start of testset for ‘u’
388 TEST uathúil --> UQ UU LJ
389 TEST uaine --> UQ NJ @@
390 TEST uan --> UQ NN
391 TEST subh --> S UBH
392 TEST bhuel --> V E LL
TEST guird --> G UU RJ DJ
TEST cuid --> K I DJ
TEST REQUIRED
TEST uile --> I LJ @Q
TEST bruón --> B R II NN
TEST bruíne --> B R II NJ @Q
TEST cumhacht --> K UU X T
TEST burdún --> B UU R D UU NN
TEST cur --> K U R
TEST REQUIRED
TEST bus --> B U S
// end of testset for 'u'

// CONSONANTS

// future & conditional 'b'
[bf]BFCE_# = P
[bf]SFCE_# = PJ
// eclipsis 'bhf'
#[bhf]CS_*RCSV_NONSYLLABIC_ = VJ
#[bhf] = V
// future & conditional 'bh'
[bhf]BFCE_# = F
[bhf]SFCE_# = FJ
// epenthesis 'bh'
NONSYLLABIC_LCSBV_LNR_+[bh] = @Q V
NONSYLLABIC_LCSSV_LNR_+[bh] = @Q VJ
// lenition 'bh'
#[bh]CS_*RCSV_NONSYLLABIC_ = VJ
#[bh] = V
// base case 'bh'
[bh]CS_*RCSV_NONSYLLABIC_ = VJ
//NONSYLLABIC_LCSV_CS_*[bh] = VJ
NONSYLLABIC_LCSV_CS_+[bh] = VJ
NONSYLLABIC_LCSV_[bh] = VJ
[bh] = V
// eclipsis 'bp'
Letter-to-sound (LTS) rule files

431 #[bp]CS_*RCSV_NONSYLLABIC_ = BJ
432 #[bp] = B
433 // verbal adjective 'b'
434 LCBV_CS_*[bth] = P
435 LCSV_CS_*[bth] = PJ
436 // epenthesis 'b'
437 NONSYLLABIC_LCSBV_LNR_+[b] = @@ B
438 NONSYLLABIC_LCSSV_LNR_+[b] = @@ BJ
439 // base case 'b'
440 [b]CS_*RCSV_NONSYLLABIC_ = BJ
441 //NONSYLLABIC_LCSV_CS_*[b] = BJ
442 NONSYLLABIC_LCSV_CS_*+[b] = BJ
443 NONSYLLABIC_LCSV_[b] = BJ
444 [b] = B
445
446 // start of testset for 'b' and 'bh'
447 // future & conditional 'b'
448 TEST scuabfaidh --> S K U@ P IDH#
449 TEST clibfidh --> KJ LJ I PJ IDH#
450 TEST scuabfadh --> S K U@ P ADH#
451 TEST clibfeadh --> KJ LJ I PJ ADH#
452 // eclipsis 'bhf'
453 TEST bhfainne --> V AA NNJ @Q
454 TEST bhfianaise --> VJ I@ NN @Q SJ @Q
455 // future & conditional 'bh'
456 TEST scríobhfaidh --> SJ KJ RJ II F IDH#
457 TEST díbhfidh --> DJ II FJ IDH#
458 TEST scríobfadh --> SJ KJ RJ II F ADH#
459 TEST díbhfeadh --> DJ II FJ ADH#
460 // epenthesis 'bh'
461 TEST searbh --> SJ A R @Q V
462 TEST seirbhís --> SJ E RJ @Q VJ II SJ
463 // lenition 'bh'
464 TEST bhrostaigh --> V R O S T IGH#
465 TEST bhris --> VJ RJ I SJ
466 // base case 'bh'
467 TEST leathbhosca --> LJ A V @Q S K @Q
TEST coibhín --> K OJ VJ II NJ
  // eclipsis 'bp'

TEST bpáistí --> B AA SJ TJ II
TEST bpéisteanna --> BJ EE SJ TJ OO NN OO
  // verbal adjective 'b'

TEST scuabtha --> S K UO P OO
TEST clibthe --> KJ LJ I PJ OO
  // epenthesis 'b'

TEST borb --> B O R OO B
TEST seirbiach --> SJ E RJ OO BJ IO X
  // base case 'b'

TEST bróna --> B R OO NN OO
TEST brian --> BJ RJ IO NN
  // end of testset for 'b' and 'bh'

  // future & conditional
[cf]BFCE_# = K
[cf]SFCE_# = KJ
  // future & conditional
[chf]BFCE_# = X
[chf]SFCE_# = XJ
  // epenthesis
NONSYLLABIC_LCSBV_LNR_+[ch] = OO X
NONSYLLABIC_LCSSV_LNR_+[ch] = OO XJ
  // lenition
# [ch]CS_*RCSV_NONSYLLABIC_ = XJ
# [ch] = X
  // base case
// NONSYLLABIC_LCBV_[ch]t = R
// NONSYLLABIC_LCSV_[ch]RCSV_NONSYLLABIC_ = HJ
[ch]CS_*RCSV_NONSYLLABIC_ = XJ
  // epenthesis
// NONSYLLABIC_LCSV_CS_*[ch] = XJ
NONSYLLABIC_LCSV_CS_+[ch] = XJ
NONSYLLABIC_LCSV_[ch] = XJ
[ch] = X
  // verbal adjective
LCBV_CS_*[cth] = K
LCSV_CS_*[cth] = KJ
// base case
[ct]CS_*RCSV_NONSYLLABIC_ = KJ
// NONSYLLABIC_LCSV_CS_*[ct] = KJ
NONSYLLABIC_LCSV_CS_+[ct] = KJ
NONSYLLABIC_LCSV_[ct] = KJ
[ct] = K

// future & conditional
[df]BFCE_# = T
[df]SFCE_# = TJ
NONSYLLABIC_LCLV_[dha] =
NONSYLLABIC_LCLSL_[dh] = GFJ
#CS_*LCSV_*[dh]# = GFJ
NONSYLLABIC_LCLLV_[dh] =
// lenition
#[dh]CS_*RCSV_NONSYLLABIC_ = GFJ
#[dh] = GF
// deletion
NONSYLLABIC_LCLLV_[dh] =
TEST leódhas --> LJ 00 S
// base case
[dh]CS_*RCSV_NONSYLLABIC_ = GFJ
// NONSYLLABIC_LCSV_CS_*[dh] = GFJ
NONSYLLABIC_LCSV_CS_+[dh] = GFJ
NONSYLLABIC_LCSV_[dh] = GFJ
[dh] = GF
// eclipsis
#[dt]CS_*RCSV_NONSYLLABIC_ = DJ
#[dt] = D
// verbal adjective
LCBV_CS_*[dt] = T
LCSV_CS_*[dt] = TJ
// base case
[d]CS_*RCSV_NONSYLLABIC_ = DJ
// NONSYLLABIC_LCSV_CS_*[d] = DJ
NONSYLLABIC_LCSV_CS_+[d] = DJ
Letter-to-sound (LTS) rule files

542 NONSYLLABIC_LCSV_[d] = DJ
543 [d] = D
544
545 // base case
546 [fh] =
547 // future & conditional
548 VS_[f]BFCE_# = H
549 VS_[f]SFCE_# = HJ
550 // epenthesis
551 NONSYLLABIC_LCSV_LNR_+[f] = @@ F
552 NONSYLLABIC_LCSV_LNR_+[f] = @@ FJ
553 // base case
554 [f]CS_*RCSV_NONSYLLABIC_ = FJ
555 // NONSYLLABIC_LCSV_CS_*[f] = FJ
556 NONSYLLABIC_LCSV_CS_[f] = FJ
557 NONSYLLABIC_LCSV_[f] = FJ
558 [f] = F
559
560 // eclipsis
561 #[gc]CS_*RCSV_NONSYLLABIC_ = GJ
562 #[gc] = G
563 // future & conditional
564 [gf]BFCE_# = K
565 [gf]SFCE_# = KJ
566 // lenition
567 #[gh]CS_*RCSV_NONSYLLABIC_ = GFJ
568 #[gh] = GF
569 NONSYLLABIC_LCSV_[gh] = GFJ
570 #CS_*LCSV_*[gh]# = GFJ
571 // deletion
572 NONSYLLABIC_LCSV_[gh] =
573 // base case
574 [gh]CS_*RCSV_NONSYLLABIC_ = GFJ
575 // NONSYLLABIC_LCSV_CS_*[gh] = GFJ
576 NONSYLLABIC_LCSV_CS_[gh] = GFJ
577 NONSYLLABIC_LCSV_[gh] = GFJ
578 [gh] = GF
// verbal adjective
LCBV_CS_*[gth] = K
LCSV_CS_*[gth] = KJ

// epenthesis
NONSYLLABIC_LCSBV_LNR_+[g] = @ @ G
NONSYLLABIC_LCSSV_LNR_+[g] = @ @ GJ

// base case
[g]CS_*RPCSV_NONSYLLABIC_ = GJ
//NONSYLLABIC_LCSV_CS_*[g] = GJ
NONSYLLABIC_LCSV_CS_+[g] = GJ
NONSYLLABIC_LCSV_[g] = GJ
[g] = G

// base case
[h]CS_*RPCSV_NONSYLLABIC_ = HJ
//NONSYLLABIC_LCSV_CS_*[h] = HJ
NONSYLLABIC_LCSV_CS_+[h] = HJ
NONSYLLABIC_LCSV_[h] = HJ
[h] = H

// base case
[j] = DJZJ

// base case
[k]CS_*RPCSV_NONSYLLABIC_ = KJ
//NONSYLLABIC_LCSV_CS_*[k] = KJ
NONSYLLABIC_LCSV_CS_+[k] = KJ
NONSYLLABIC_LCSV_[k] = KJ
[k] = K

// future & conditional
[llf]BFCE_# = LL_D
[llf]SFCE_# = LLJ_D

// verbal adjective
LCBV_CS_*[llth] = LL_D
LCSV_CS_*[llth] = LLJ_D

// base case
Letter-to-sound (LTS) rule files


// NONSYLLABIC_LCSV_CS_*[11] = LLJ
NONSYLLABIC_LCSV_CS_*[11] = LLJ
NONSYLLABIC_LCSV_[11] = LLJ

// future & conditional
[lf]BFCE_*# = LL_D
[lf]SFCE_*# = LJ_D

// verbal adjective
LCBV_CS_*[lth] = LL_D
LCSV_CS_*[lth] = LJ_D

// base case
[1]CS_*RCSV_NONSYLLABIC_ = LJ
// NONSYLLABIC_LCSV_CS_*[1] = LJ
NONSYLLABIC_LCSV_CS_*[1] = LJ
NONSYLLABIC_LCSV_[1] = LJ
[1] = LL

// eclipsis
#[mb]CS_*RCSV_NONSYLLABIC_ = MJ
#[mb] = M

// future & conditional
[mf]BFCE_*# = M_D
[mf]SFCE_*# = MJ_D

// future & conditional
[mhf]BFCE_*# = F
[mhf]SFCE_*# = FJ

// lenition
#[mh]CS_*RCSV_NONSYLLABIC_ = VJ
#[mh] = V

// epenthesis
NONSYLLABIC_LCSV_LNBV_LNR_+[mh] = @@ V
NONSYLLABIC_LCSV_LBSV_LNR_+[mh] = @@ VJ

// base case
[mh]CS_*RCSV_NONSYLLABIC_ = VJ
// NONSYLLABIC_LCSV_CS_*[mh] = VJ
NONSYLLABIC_LCSV_CS_*[mh] = VJ
NONSYLLABIC_LCSV_[mh] = VJ
[mh] = V

// verbal adjective
LCBV_CS_*[mth] = M_D
LCSV_CS_*[mth] = MJ_D

// epenthesis
NONSYLLABIC_LCSBV_LNR_+[m] = @0 M
NONSYLLABIC_LCSVV_LNR_+[m] = @0 MJ

// base case
[m]CS_*RCSV_NONSYLLABIC_ = MJ
// NONSYLLABIC_LCSV_CS_*[m] = MJ
NONSYLLABIC_LCSV_CS_+[m] = MJ
NONSYLLABIC_LCSV_[m] = MJ
[m] = M

// future & conditional
[nnf]BFCE_# = NN_D
[nnf]SFCE_# = NNJ_D

// verbal adjective
LCBV_CS_*[nnth] = NN_D
LCSV_CS_*[nnth] = NNJ_D

// base case
[nn]CS_*RCSV_NONSYLLABIC_ = NNJ
// NONSYLLABIC_LCSV_CS_*[nn] = NNJ
NONSYLLABIC_LCSV_CS_+[nn] = NNJ
NONSYLLABIC_LCSV_[nn] = NNJ
[nn] = NN

// eclipsis
[#n-]RCSV_NONSYLLABIC_ = NJ
[#n-] = NN

// eclipsis
[#nd]CS_*RCSV_NONSYLLABIC_ = NNJ
[#nd] = NN

// future & conditional
[nf]BFCE_# = NN_D
[nf]SFCE_# = NJ_D

// future & conditional
Letter-to-sound (LTS) rule files

690 [ngf]BFCE_# = NG_D
691 [ngf]SFCE_# = NGJ_D
692 // verbal adjective
693 LCBV_CS_*[ngth] = NG_D
694 LCSV_CS_*[ngth] = NGJ_D
695 // eclipsis
696 #[ng]CS_*RCSV_NONSYLLABIC_ = NGJ
697 #[ng] = NG
698 // ng --> 2 syllables
699 #CS_*LCBV_[ng]RCBV_ = NG G
700 #CS_*LCSV_[ng]RCSV_NONSYLLABIC_ = NGJ GJ
701 NONSYLLABIC_LCSV_[ng]t# = NJ
702 // base case
703 [ng]CS_*RCSV_NONSYLLABIC_ = NGJ
704 // NONSYLLABIC_LCSV_CS_*[ng] = NGJ
705 NONSYLLABIC_LCSV_CS_+[ng] = NGJ
706 NONSYLLABIC_LCSV_[ng] = NGJ
707 [ng] = NG
708 // verbal adjective
709 LCBV_CS_*[nth] = NN_D
710 LCSV_CS_*[nth] = NJ_D
711 // 'tns'
712 // #ts[n]RCSV_NONSYLLABIC_ = RJ
713 // #ts[n] = R
714 // 'nc'
715 NONSYLLABIC_LCSV_[n]c = NGJ
716 [n]c = NG
717 // base case
718 [n]CS_*RCSV_NONSYLLABIC_ = NJ
719 // NONSYLLABIC_LCSV_CS_*[n] = NJ
720 NONSYLLABIC_LCSV_CS_+[n] = NJ
721 NONSYLLABIC_LCSV_[n] = NJ
722 [n] = NN
723
724 // future & conditional
725 [pf]BFCE_# = P
726 [pf]SFCE_# = PJ
// lenition

// base case

// verbal adjective

// future & conditional

// verbal adjective

// base case

// future & conditional
LCBV_CS_*CS_*[rth] = R_D
LCSV_CS_*CS_*[rth] = RJ_D

// base case
#s[r] = R
#[r] = R
[r]DNLST_ = R

[r]CS_*RCSV_NONSYLLABIC_ = RJ

// NONSYLLABIC_LCSV_CS_*[r] = RJ
NONSYLLABIC_LCSV_CS_*+[r] = RJ
NONSYLLABIC_LCSV_[r] = RJ

[r] = R

// future & conditional
[sf]BFCE_# = S
[sf]SFCE_# = SJ

// devoicing
[shl]CS_*RCSV_NONSYLLABIC_ = LJ_D
[shl] = LL_D
[shm]CS_*RCSV_NONSYLLABIC_ = MJ_D
[shm] = M_D
[shn]CS_*RCSV_NONSYLLABIC_ = NJ_D
[shn] = NN_D
[shr]CS_*RCSV_NONSYLLABIC_ = RJ_D
[shr] = R_D

// lenition
#[sh]CS_*RCSV_NONSYLLABIC_ = XJ
#[sh] = H

// base case
[sh]CS_*RCSV_NONSYLLABIC_ = XJ

// NONSYLLABIC_LCSV_CS_*[sh] = XJ
NONSYLLABIC_LCSV_CS_*+[sh] = XJ
NONSYLLABIC_LCSV_CS_*[sh] = XJ

[sh] = H

#s[r] = S

#s]FMP_CS_*RCSV_NONSYLLABIC_ = S

// base case
[s]CS_*RCSV_NONSYLLABIC_ = SJ
// NONSYLLABIC_LCSV_CS_*[s] = SJ
NONSYLLABIC_LCSV_CS_*+[s] = SJ
NONSYLLABIC_LCSV_+[s] = SJ
[s] = S

// prefixed 't'
#*[t-]RCSV_NONSYLLABIC_ = TJ
#*[t-] = T

// future & conditional
[tf]BFCE_# = T
[tf]SFCE_# = TJ

// hack for compound boundaries
[th]CS_h =

// devoicing
[thb]CS_*RCSV_NONSYLLABIC_ = PJ
[thb] = P
[thc]CS_*RCSV_NONSYLLABIC_ = KJ
[thc] = K
[thd]CS_*RCSV_NONSYLLABIC_ = TJ
[thd] = T

// future & conditional
[thf]BFCE_# = H
[thf]SFCE_# = XJ

// devoicing
[thi]CS_*RCSV_NONSYLLABIC_ = LJ_D
[thi] = LL_D
[thm]CS_*RCSV_NONSYLLABIC_ = MJ_D
[thm] = M_D
[thn]CS_*RCSV_NONSYLLABIC_ = NJ_D
[thn] = NN_D
[thp]CS_*RCSV_NONSYLLABIC_ = PJ
[thp] = P
[thr]CS_*RCSV_NONSYLLABIC_ = RJ_D
[thr] = R_D
[ths]CS_*RCSV_NONSYLLABIC_ = SJ
[ths] = S
[tht]CS_*RCSV_NONSYLLABIC_ = TJ
Letter-to-sound (LTS) rule files

\[
[t\tht] = T
\]

// lenition

\# \[\tht\]CS_*RCSV_NONSYLLABIC_ = HJ
\# \[\tht\] = H

// VS_+i[\tht] = XJ

// base case

\[\tht\]CS_*RCSV_NONSYLLABIC_ = HJ
// NONSYLLABIC_LCSV_CS_*[\tht] = HJ
NONSYLLABIC_LCSV_CS_+[\tht] = HJ
NONSYLLABIC_LCSV_[\tht] = HJ
\[\tht\] = H

// prefixed 't'

\# [ts]CS_*RCSV_NONSYLLABIC_ = TJ
\# [ts] = T

// base case

\[t\]CS_*RCSV_NONSYLLABIC_ = TJ
// NONSYLLABIC_LCSV_CS_*[t] = TJ
NONSYLLABIC_LCSV_CS_+[t] = TJ
NONSYLLABIC_LCSV_[t] = TJ
\[t\] = T

// start of test set for 't'

TEST t-uisce --> T UI SJ KJ @@
TEST t-éabhléidi --> TJ EE V LL 00 DJ II
TEST atfaidh --> A T IDH#
TEST titfidh --> TJ I TJ IDH#
TEST athdheanamh --> A GFJ EE NN AMH#
TEST REQUIRED
TEST REQUIRED
TEST REQUIRED
TEST REQUIRED
TEST REQUIRED
TEST REQUIRED
TEST REQUIRED
TEST meathfadh --> MJ A H ADH#
TEST rithfeadh --> RJ I XJ ADH#
TEST REQUIRED
TEST REQUIRED
875 TEST REQUIRED
876 TEST REQUIRED
877 TEST REQUIRED
878 TEST REQUIRED
879 TEST REQUIRED
880 TEST REQUIRED
881 TEST REQUIRED
882 TEST bláthra --> B LL AA R_D @@
883 TEST REQUIRED
884 TEST REQUIRED
885 TEST REQUIRED
886 TEST REQUIRED
887 TEST tharla --> H AA R LL @@
888 TEST thit --> HJ I TJ
889 TEST REQUIRED
890 TEST REQUIRED
891 TEST REQUIRED
892 TEST REQUIRED
893 TEST tseachtain --> TJ A X T @@ NJ
894 TEST tsagairt --> T A G @@ RJ TJ
895 TEST teann --> TJ A NN
896 TEST tit --> TJ I TJ
897 TEST togra --> T O G R @@
898 // end of testset for 't'
899
900 // base case
901 [v]CS_*RCSV_NONSYLLABIC_ = VJ
902 //NONSYLLABIC_LCSV_CS_*[v] = VJ
903 NONSYLLABIC_LCSV_CS_[v] = VJ
904 NONSYLLABIC_LCSV_[v] = VJ
905 [v] = V
906
907 // base case
908 [w] = V
909
910 [x-] = E KJ S
911 [x] = ZJ
912
913  [y] = GFJ
914
915  [z]CS_*RCSV_NONSYLLABIC_ = ZJ
916  //NONSYLLABIC_LCSV_CS_*[z] = ZJ
917  NONSYLLABIC_LCSV_CS_*+[z] = ZJ
918  NONSYLLABIC_LCSV_[z] = ZJ
919  [z] = Z
920
921  [-] =
922
923  ['] =
924
925  // random tests
926  TEST sadhbh --> S AI V
B.2 Global/global-units.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file lists the abstract inter-dialectal global units.

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie

A
AA
ABH
ADH#
AI
AÍO
AMH#
AO
AU
Letter-to-sound (LTS) rule files

29
30  B
31  BJ
32  D
33  DJ
34  DJZJ
35
36  E
37  EE
38  EIDH
39  EIGH
40
41  F
42  FJ
43  G
44  GF
45  GFJ
46  GJ
47  H
48  HJ
49
50  I
51  IΩ
52  IDH#
53  IGH#
54  II
55  IO
56  IODH#
57  ITHE#
58
59  K
60  KJ
61  LJ
62  LJ_D
63  LL
64  LL_D
65  LLJ
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>LLJ_D</td>
</tr>
<tr>
<td>67</td>
<td>M</td>
</tr>
<tr>
<td>68</td>
<td>M_D</td>
</tr>
<tr>
<td>69</td>
<td>MJ</td>
</tr>
<tr>
<td>70</td>
<td>MJ_D</td>
</tr>
<tr>
<td>71</td>
<td>NG</td>
</tr>
<tr>
<td>72</td>
<td>NG_G</td>
</tr>
<tr>
<td>73</td>
<td>NG_D</td>
</tr>
<tr>
<td>74</td>
<td>NGJ</td>
</tr>
<tr>
<td>75</td>
<td>NGJ_D</td>
</tr>
<tr>
<td>76</td>
<td>N</td>
</tr>
<tr>
<td>77</td>
<td>N_D</td>
</tr>
<tr>
<td>78</td>
<td>NJ</td>
</tr>
<tr>
<td>79</td>
<td>NJ_D</td>
</tr>
<tr>
<td>80</td>
<td>NN</td>
</tr>
<tr>
<td>81</td>
<td>NN_D</td>
</tr>
<tr>
<td>82</td>
<td>NNJ</td>
</tr>
<tr>
<td>83</td>
<td>NNJ_D</td>
</tr>
<tr>
<td>84</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>O</td>
</tr>
<tr>
<td>86</td>
<td>OBH</td>
</tr>
<tr>
<td>87</td>
<td>ODH</td>
</tr>
<tr>
<td>88</td>
<td>ÕDH#</td>
</tr>
<tr>
<td>89</td>
<td>OGH</td>
</tr>
<tr>
<td>90</td>
<td>OI</td>
</tr>
<tr>
<td>91</td>
<td>OIDH</td>
</tr>
<tr>
<td>92</td>
<td>OIGH</td>
</tr>
<tr>
<td>93</td>
<td>OMH</td>
</tr>
<tr>
<td>94</td>
<td>OO</td>
</tr>
<tr>
<td>95</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>P</td>
</tr>
<tr>
<td>97</td>
<td>PJ</td>
</tr>
<tr>
<td>98</td>
<td>R</td>
</tr>
<tr>
<td>99</td>
<td>R_D</td>
</tr>
<tr>
<td>100</td>
<td>RJ</td>
</tr>
<tr>
<td>101</td>
<td>RJ_D</td>
</tr>
<tr>
<td>102</td>
<td>RR</td>
</tr>
</tbody>
</table>
Letter-to-sound (LTS) rule files

103  RR_D
104  RRJ
105  RRJ_D
106  S
107  SJ
108  T
109  TJ
110
111  U
112  U@
113  UBH
114  UU
115
116  V
117  VJ
118  X
119  XJ
120  Z
121  ZJ
B.3 Global/lex-exceptions.txt

1 // This file is a component of a multi-dialect modular phonetiser for Irish.
2 // A phonetiser converts orthographic text to dialectal phonemic transcription.
3 // This phonetiser has two layers: one global, one local.
4 // At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
5 // At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
6 // The global layer consists of one global module.
7 // The local layer consists of one local module for each dialect implemented.
8 // Each module contains the following:
9 // (1) A set of rewrite (letter-to-sound) rules.
10 // (2) An exceptions lexicon.
11 // (3) A set of tests.
12
13 // This file is (part of) the global exceptions lexicon.
14
15 // Syntax of lexicon entry:
16 // <ORTHOGRAPHY>
17 // <TAB><TRANSRIPTION>
18
19 // Author: Brian Ó Raghallaigh
20 // oraghalb AG tcd PONC ie
21
22 eochair
23 1 O 0 X @@ RJ
24 seo
25 1 SJ O
26 deoch
27 1 DJ O X
B.4 Global/lex-function-words.txt

1 // This file is a component of a multi-dialect modular
2 // phonetiser for Irish.
3 // A phonetiser converts orthographic text to dialectal
4 // phonemic transcription.
5 // This phonetiser has two layers: one global, one local.
6 // At a global level, spelling is converted to an abstract
7 // inter-dialectal linguistic representation.
8 // At a local level, this abstract linguistic
9 // representation is mapped to a dialectal transcription.
10 // The global layer consists of one global module.
11 // The local layer consists of one local module for each
12 // dialect implemented.
13 // Each module contains the following:
14 // (1) A set of rewrite (letter-to-sound) rules.
15 // (2) An exceptions lexicon.
16 // (3) A set of tests.
17 // This file is (part of) the global exceptions lexicon.
18 // Syntax of lexicon entry:
19 // <ORTHOGRAPHY>
20 // <TAB><TRANSCRIPTION>
21 // Author: Brian Ó Raghallaigh
22 // oraghalb AG tcd PONC ie
23 a
24 0 @@
25 ab
26 0 @@ B
27 abhaile
28 0 @@ . 1 V A . 0 LJ @@
29 abhus
30 0 @@ . 1 V U S
abú
0 @@ . 1 B UU
acu
1 A . O K U
aduaidh
0 @@ . 1 D U@ GFJ
ag
0 E GJ
ágó
0 @@ . 1 G OQ
aici
1 E . O KJ I
aige
1 E . O GJ @@
air
1 E RJ
aisti
1 A SJ . O TJ I
amach
0 @@ . 1 M AX
amanathar
0 @@ . 1 M A . O N @@ . O H @@ R
amárcach
0 @@ . 1 M AA . O R @@ X
ambaiste
0 @@ M . 1 B A SJ . O TJ @@
amháin
0 @@ . 1 V AA NJ
amú
0 @@ . 1 M AU
amuigh
0 @@ . 1 M I GFJ
an
0 @@ NN
anall
0 @@ . 1 NN A LL
anallóid
Letter-to-sound (LTS) rule files

67  0 @@  .  1  NN  A  .  0  LL  O  O  D
68  aneas
69  0  @@  .  1  NNJ  A  S
70  aniar
71  0  @@  .  1  NNJ  I@  R
72  aníos
73  0  @@  .  1  NNJ  II  S
74  anocht
75  0  @@  .  1  NN  O  X  T
76  anoir
77  0  @@  .  1  NN  O  RJ
78  anóiritheas
79  0  @@  .  1  NN  O  O  .  0  RJ_D  @@  R
80  anois
81  0  @@  .  1  NN  I  SJ
82  annonn
83  0  @@  .  1  NN  O  NN
84  anseo
85  0  @@  NJ  .  1  SJ  O
86  ansin
87  0  @@  NJ  .  1  SJ  I  NN
88  anuas
89  0  @@  .  1  NN  U@  S
90  anuraidh
91  0  @@  .  1  NN  U  .  0  R  II
92  ar
93  0  @@  R
94  arae
95  0  @@  .  1  R  EE
96  araile
97  0  @@  .  1  R  I  .  0  LLJ  @@
98  arason
99  0  @@  .  1  R  II  NN
100  arb
101  1  @@  .  0  R  @@  B
102  arbh
103  1  @@  .  0  R  @@  V
Letter-to-sound (LTS) rule files

aréir
O @@ . 1 RJ EE RJ
ariamh
O @@ . 1 RJ I@ V
arís
O @@ . 1 RJ II SJ
astu
1 A S . 0 T U
ba
O B @@
chucu
1 X U . 0 K U
chuici
1 XJ I . 0 KJ I
cibé
0 KJ . 1 I BJ EE
de
0 DJ E
do
0 D 0
eatarthu
1 A . 0 T @@ . 0 R_D U
fadó
2 F A . 1 D 00
faraor
0 F @@ . 1 R II R
fúithi
1 F UU . 0 H I
fúthu
1 F UU . 0 H U
go
0 G @@
gur
O G @@
gura
0 G @@ . 0 R @@
gurab
Letter-to-sound (LTS) rule files

141 O G @@ . O R @@ B
142 gurb
143 O G @@ R B
144 gurbh
145 O G @@ . O R @@ V
146 i
147 0 I
148 in
149 0 I NJ
150 inné
151 0 @@ . 1 NNJ EE
152 inniu
153 0 @@ . 1 NNJ U
154 iontu
155 1 I NN . 0 T U
156 is
157 0 I S
158 isteach
159 0 @@ . 1 SJ TJ A H
160 istigh
161 0 @@ . 1 S T I GFJ
162 istoíche
163 0 @@ . 1 S T II . 0 XJ @@
164 laisteas
165 2 LL A SJ . 1 TJ A S
166 laistiar
167 2 LL A SJ . 1 TJ I@ R
168 laistigh
169 2 LL A SJ . 1 TJ I GFJ
170 laistíos
171 2 LL A SJ . 1 TJ II S
172 lasbhus
173 2 LL A S . 1 V U S
174 lasmuigh
175 2 LL A S . 1 M I GFJ
176 lasnairde
177 2 LL A S . 1 NN AA R . 0 DJ @@
178 lastall
179 2 LL A S . 1 T A LL
180 lastoir
181 2 LL A S . 1 T O RJ
182 lastuaidh
183 2 LL A S . 1 T U@ GFJ
184 lastuas
185 2 LL A S . 1 T U@ S
186 le
187 0 LJ @@
188 máguaird
189 2 M AA . 1 G U@ R DJ
190 mo
191 0 M @@
192 monuar
193 0 M @@ . 1 N N U@ R
194 mórthimpeall
195 2 M OO R . 1 HJ I MJ . O PJ @@ LL
196 na
197 0 N N @@
198 ochón
199 0 O . 1 X O N N
200 orthu
201 1 O . 0 R_D U
202 rompu
203 1 R O M . O P U
204 sa
205 0 S @@
206 san
207 0 S @@ N N
208 sna
209 0 S N N @@
210 tráthnóna
211 2 T R AA . 1 N N OO . 0 N N @@
212 tríothu
213 1 TJ RJ II . O H U
214 tríthi
Letter-to-sound (LTS) rule files

1 TJ RJ II . O H I
uimpi
1 I MJ . O PJ I
uirthi
1 E . O R_D I
umpu
1 U M . O P U
B.5 Global/lex-non-native.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file is (part of) the global exceptions lexicon.

// Syntax of lexicon entry:
// <ORTHOGRAPHY>
// <TAB><TRANSCRIPTION>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie

aeraidinimic
  1 E . 0 R @@ . 2 DJ I . 1 NNJ I . 0 MJ @@ KJ
aicéitiléin
  0 @@ . 1 KJ EE . 0 TJ @@ . 2 LLJ EE NNJ
aidréanailín
  0 @@ . 1 DJ RJ EE . 0 NN @@ . 2 LLJ II NNJ
aipindic
  2 A . 1 P I NNJ . 0 DJ @@ KJ
aipindicíteas
O A . 2 PJ I NNJ . O DJ @0 . 1 KJ II . O TJ @0 S
aláram
O A . 1 LL AA . O R @0 M
alúmanam
O @0 . 1 LL UU . O M @0 . O NN @0 M
ambassadóir
O A M . 1 B A . O S @0 . 2 D 00 RJ
amóinína
O A . 1 M 00 . O NNJ I0
anatamaíocht
O @0 . 1 NN A . O T @0 . O M I0 X T
antartach
2 A NN . 1 T A R . O T @0 X
asparagas
2 A S . 1 P A . O R @0 . O G @0 S
atlantach
2 A T . 1 LL A NN . O T @0 X
babún
2 B A . 1 B UU NN
banana
0 B @0 . 1 NN A . O NN @0
basár
0 B @0 . 1 S AA R
beagóinína
0 BJ @0 . 1 G 00 . O NNJ I0
cabheár
0 K A . 1 VJ AA R
canú
0 K @0 . 1 NN UU
cocatú
2 K O . 0 K @0 . 1 T UU
coimisinéir
2 K O . 1 MJ I . O SJ @0 . O NNJ E RJ
coimisiún
2 K O . 1 MJ I . O SJ U NN
coimisiúnta
Letter-to-sound (LTS) rule files

67 2 K 0 . 1 MJ I . 0 SJ U NN . 0 T @
68 coinbhinsiún
69 2 K 0 NNJ . 1 VJ I NNJ . 0 SJ U NN
70 coinbhinsiúnach
71 2 K 0 NNJ . 1 VJ I NNJ . 0 SJ U . 0 NN @ X
72 colaistéaról
73 2 K 0 . 1 LL A . 0 SJ TJ EE . 0 R O LL
74 créamatóiriam
75 2 KJ RJ EE . 0 M @ . 1 T OO . 0 RJ I M
76 cumá
77 0 K @ . 1 M AA
78 deirmititéas
79 2 DJ E . 0 RJ @ . 0 MJ @ . 1 TJ II . 0 TJ @ S
80 dibheán
81 0 DJ @ . 1 VJ AA NN
82 diméin
83 0 DJ @ . 1 MJ EE NNJ
84 dioplóma
85 2 DJ I . 1 P LL @O . 0 M @
86 éacuímeineach
87 2 EE . 0 K UU . 1 MJ EE . 0 NNJ @ X
88 éacuímeineachas
89 2 EE . 0 K UU . 1 MJ EE . 0 NNJ A . 0 H @ S
90 eindéimeach
91 2 I NNJ . 1 DJ EE . 0 MJ @ X
92 einríteas
93 2 E NNJ . 1 TJ RJ II . 0 TJ @ S
94 eipistil
95 2 E . 1 PJ I . 0 SJ TJ @ LLJ
96 gaistríteas
97 2 G A . 1 SJ TJ RJ II . 0 TJ @ S
98 geiréiniam
99 2 GJ E . 1 RJ EE . 0 NNJ I @ M
100 giomnáisiam
101 2 GJ I M . 1 NN AA . 0 SJ I @ M
102 giotár
103 0 GJ @ . 1 T AA R
goraille
2 G O . 1 R I . 0 LLJ @
guairille
2 G U@ . 1 RJ I . 0 LLJ @
guairilleach
2 G U@ . 1 RJ I . 0 LLJ @
hieana
2 H II . 1 EE . 0 NN @
hura
0 H @ . 1 R AA
imprisean
2 I MJ . 1 PJ RJ I . 0 SJ @ NN
impriseanachas
2 I MJ . 1 PJ RJ I . 0 SJ @ . 0 NN A . 0 H @ S
inleithscéil
0 I NNJ . 1 LLJ E . 2 SJ KJ EE LLJ
larainingíteas
2 LL A . 0 R @ NGJ . 1 GJ II . 0 TJ @ S
leoicéime
2 LL J @ . 1 KJ EE . 0 MJ @
lumbágó
2 LL U M . 1 B AA . 0 G @
macadam
0 M @ . 1 K A . 0 D @ M
mahagaine
0 M @ . 1 HA . 0 G @ . 0 NNJ @
maistíteas
2 M A . 1 SJ TJ II . 0 TJ @ S
maláire
2 M A . 1 LL AA . 0 RJ @
marún
0 M @ . 1 R UU NN
meiningíteas
2 MJ E . 0 NNJ @ NGJ . 1 GJ II . 0 TJ @ S
meireang
2 MJ E . 1 RJ A NG
míleoidéan
Letter-to-sound (LTS) rule files

141 0 MJ @@ . 1 LLJ OO . 0 DJ @@ NN
142 móiminteam
143 2 M O0 . 1 MJ I NNJ . 0 TJ @@ M
144 moinsíneoir
145 2 M O NNJ . 1 SJ II . 0 NNJ OO RJ
146 mósáic
147 2 M O0 . 1 S AA KJ
148 muinisean
149 2 M I . 1 NNJ I . 0 SJ @@ NN
150 nairciseas
151 2 NN A RJ . 1 KJ I . 0 SJ @@ S
152 néiríteas
153 2 NNJ EE . 1 RJ II . 0 TJ @@ S
154 niúmóíne
155 2 NNJ UU . 1 M O0 . 0 NNJ @@
156 nóibhéíne
157 2 NN O0 . 1 VJ EE . 0 NNJ @@
158 oilimpeach
159 2 O . 1 LLJ I MJ . 0 PJ @@ X
160 onamataípé
161 1 O . 0 NN @@ . 2 M A . 0 T @@ . 1 PJ EE
162 órang-útan
163 1 O0 . 0 R A NG . 1 UU . 0 T @@ NN
164 pampúta
165 2 P A M . 1 P UU . 0 T @@
166 peireatóiníteas
167 1 PJ E . 0 RJ @@ . 2 T O0 . 1 NNJ II . 0 TJ @@ S
168 peirspictíocht
169 2 PJ E RJ . 1 S PJ I KJ . 0 TJ I@ X T
170 peitriliam
171 2 PJ E . 1 TJ RJ I . 0 LLJ I@ M
172 pianó
173 2 PJ I . 1 A . 0 NN O0
174 pianódóir
175 2 PJ I . 1 A . 0 NN O0 . 0 D O RJ
176 plútóiníam
177 2 P LL UU . 1 T O0 . 0 NNJ I@ M
polaimiailíteas
1 P 0 . 0 LL @@ . 2 MJ I@ . 1 LLJ II . 0 TJ @@ S
potaisiam
2 P 0 . 1 T A . 0 SJ I@ M
proibhinseal
2 PR0 . 1 VJ I NNJ . 0 SJ @@ LL
racún
2 RA . 1 K UU NN
raíciteas
2 RA . 1 KJ II . 0 TJ @@ S
rosualt
2 R 0 . 1 S U@ LL T
salún
0 S @@ . 1 LL UU NN
sciaitice
2 SJ KJ II . 1 A . 0 TJ II . 0 KJ @@
seampú
2 SJ A M . 1 P UU
símeantaic
2 SJ EE . 1 MJ A NN . 0 T @@ KJ
sóraf
2 SJ I . 1 R AA F
sringe
0 SJ @@ . 1 RJ I NGJ . 0 GJ @@
sraigítí
2 S PA . 1 GJ I . 0 TJ II
saitistic
2 STA . 1 TJ I . 0 SJ TJ @@ KJ
saitistidil
2 STA . 1 TJ I . 0 SJ TJ U LLJ
tirramhacadam
2 TA . 0 R @@ . 0 V @@ . 1 K A . 0 D @@ M
tbac
0 T @@ . 1 B A K
tbacadóir
0 T @@ . 1 B A . 0 K @@ . 0 D O RJ
tinníteas
Letter-to-sound (LTS) rule files

215  2 T 0 . 1 NNJ II . 0 TJ @® S
tornádó
216  2 T 00 R . 1 NN AA . 0 D 00
217  úráiniam
218  2 UU . 1 R AA . 0 NNJ I® M
219  úránas
220  2 UU . 1 R AA . 0 NN @® S
221  útóipeach
222  2 UU . 1 T 00 . 0 PJ @® X
223  vearanda
224  0 VJ @® . 1 R A NN . 0 D @®
225  vióla
226  2 VJ II . 1 00 . 0 LL @®
B.6 Global/test.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.

// A phonetiser converts orthographic text to dialectal phonemic transcription.

// This phonetiser has two layers: one global, one local.

// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.

// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.

// The global layer consists of one global module.

// The local layer consists of one local module for each dialect implemented.

// Each module contains the following:

// (1) A set of rewrite (letter-to-sound) rules.

// (2) An exceptions lexicon.

// (3) A set of tests.

// This file contains global module tests.

// Test syntax: <INPUT><TAB><DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh

// oraghalb AG tcd PONC ie
B.7 Larchanuint/lts.lff

// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file contains local rules for the Lárfháinnt artificial Irish dialect.

// Variable syntax: <VARIABLE NAME><TAB>[<VALUE 1>|<VALUE 2>|...|[<VALUE N>]]
// Rule syntax: <INPUT> = <OUTPUT>
// Test syntax: TEST <INPUT> --> <DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie

// VARIABLES

// RULES AND TESTS

[AA] = aa
[ABH] = au
[ADH#] = @ x
[AGH] = ai
Letter-to-sound (LTS) rule files

29 \([\text{AI}O] = \text{ii}\)
30 \([\text{AMH#}] = \text{Q v}\)
31 \([\text{AI}] = \text{ai}\)
32 \([\text{AO}] = \text{ii}\)
33 \([\text{AU}] = \text{au}\)
34 \([\text{A}] = \text{a}\)
35
36 \([\text{EE}] = \text{ee}\)
37 \([\text{EIDH}] = \text{ai}\)
38 \([\text{EIGH}] = \text{ai}\)
39 \([\text{E}] = \text{e}\)
40
41 \([\text{IDH#}] = \text{ii}\)
42 \([\text{IGH#}] = \text{ii}\)
43 \([\text{ODH#}] = \text{ii X}\)
44 \([\text{THE#}] = \text{i h Q}\)
45 \([\text{HQ}] = \text{i Q}\)
46 \([\text{II}] = \text{ii}\)
47 \([\text{IO}] = \text{i}\)
48 \([\text{I}] = \text{i}\)
49
50 \([\text{JBH}] = \text{au}\)
51 \([\text{JDH#}] = \text{oo X}\)
52 \([\text{JDH}] = \text{au}\)
53 \([\text{JGH}] = \text{au}\)
54 \([\text{JIDH}] = \text{ai}\)
55 \([\text{JIGH}] = \text{ai}\)
56 \([\text{JI}] = \text{o}\)
57 \([\text{JMHI}] = \text{au}\)
58 \([\text{JO}] = \text{oo}\)
59 \([\text{J}] = \text{o}\)
60
61 \([\text{JBH}] \# = \text{u v}\)
62 \([\text{JBH}] = \text{Q v}\)
63 \([\text{JQ}] = \text{u Q}\)
64 \([\text{JI}] = \text{i}\)
65 \([\text{JU}] = \text{uu}\)
\[ u = [NN] \]
\[ f_u = [CNN] \]
\[ \eta u = [D^\sim NN] \]
\[ \eta [f_u = [D^\sim CNN] \]
\[ \tau = [\tau] \]
\[ f_\tau = [\tau\tau] \]
\[ \eta \tau = [D^\sim \tau\tau] \]
\[ \eta [f_\tau = [D^\sim \tau\tau\tau] \]
\[ \tau = [\tau\tau\tau] \]
\[ f_\tau = [\tau\tau\tau\tau] \]
\[ \eta \tau = [D^\sim \tau\tau\tau\tau] \]
\[ \eta [f_\tau = [D^\sim \tau\tau\tau\tau\tau\tau] \]
\[ \rho = [D] \]
\[ f_\rho = [\rho\rho] \]
\[ -p = [-D] \]
\[ \tau = [\tau\tau] \]
\[ f_\tau = [\tau\tau\tau] \]
\[ -\tau = [-\tau\tau] \]
\[ f z [f_\rho = [\rho\rho\rho] \]
\[ f s [f_\tau = [\tau\tau\tau]\tau\tau\tau] \]
\[ w = [W] \]
\[ f w = [\rho W] \]
\[ \eta w = [D^\sim W] \]
\[ \eta [f w = [D^\sim W\tau] \]
\[ M = [M] \]
\[ \Lambda = [\Lambda] \]
\[ f_\Lambda = [\Lambda\Lambda] \]
\[ \tau = [\tau\tau] \]
\[ f_\tau = [\tau\tau\tau] \]
\[ q = [B] \]
\[ f q = [B\beta] \]
\[ d = [D] \]
\[ f d = [D\tau] \]
\[ \emptyset = [\emptyset\emptyset] \]
\[ n = [\emptyset] \]

Letter-to-sound (LTS) rule files
Letter-to-sound (LTS) rule files

[103] [NJ_D] = nj h
[104] [N_D] = n h
[105] [NJ] = nj
[106] [N] = n
[107] [RRJ_D] = rj h
[108] [RJ_D] = rj h
[109] [RR_D] = r h
[110] [R_D] = r h
[111] [RRJ] = rj
[112] [RJ] = rj
[113] [RR] = r
[114] [R] = r
[115] [SJ] = sj
[116] [S] = s
[117] [ZJ] = zj
[118] [Z] = z
[119] [KJ] = kj
[120] [K] = k
[121] [GJ] = gj
[122] [G] = g
[123] [XJ] = xj
[124] [X] = x
[125] [GFJ] = gfj
[126] [GF] = gf
[127] [NGJ_D] = ngj h
[128] [NG_D] = ng h
[129] [NGJ] = ngj
[130] [NG] = ng
[131] [HJ] = h
[132] [H] = h
[133] [ ] =
[134] [O] = 0
[135] [1] = 1
[136] [2] = 2
[137] [.] = .
Letter-to-sound (LTS) rule files
B.8 Larchanuint/lex-exceptions.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file is the local exceptions lexicon for the Lárchanúint artificial Irish dialect.

// Syntax of lexicon entry:
// <ORTHOGRAPHY>
// <TAB><TRANSCRIPTION>

// Author: Brian Ó Raghallaigh
oraghalb AG tcd PONC ie

B.9 Larchanuint/test.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file contains tests for the Lárnáint artificial Irish dialect.

// Test syntax: <INPUT><TAB><DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie


abhus Ó v u s
ainm a nj Ó mj
anam a n Ó m
baile b a lj Ó
bán b aa n
Letter-to-sound (LTS) rule files

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>béal</td>
<td>bj</td>
<td>ee</td>
<td>l</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>bean</td>
<td>bj</td>
<td>an</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>bhí</td>
<td>vj</td>
<td>ii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>bhuí</td>
<td>v</td>
<td>ii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>bí</td>
<td>bj</td>
<td>ii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>bia</td>
<td>bj</td>
<td>i@</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>bó</td>
<td>b</td>
<td>oo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>bodhar</td>
<td>b au</td>
<td>r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>buail</td>
<td>b u@</td>
<td>lj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>buí</td>
<td>b</td>
<td>ii</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>cad</td>
<td>k a d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>cailín</td>
<td>k a</td>
<td>lj</td>
<td>ii</td>
<td>nj</td>
</tr>
<tr>
<td>39</td>
<td>cáis</td>
<td>k aa</td>
<td>sj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>cáis</td>
<td>k aa</td>
<td>s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>cead</td>
<td>kj</td>
<td>a</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>céad</td>
<td>kj</td>
<td>ee</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>ceist</td>
<td>kj</td>
<td>e sj</td>
<td>tj</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>ceol</td>
<td>kj</td>
<td>oo</td>
<td>l</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>chathair</td>
<td>x a h @</td>
<td>rj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>cheathair</td>
<td>xj</td>
<td>a h @</td>
<td>rj</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>ciú</td>
<td>kj</td>
<td>uu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>ciúin</td>
<td>kj</td>
<td>uu</td>
<td>nj</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>croí</td>
<td>k r i i</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>cú</td>
<td>k</td>
<td>uu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>cuid</td>
<td>k i dj</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>cúige</td>
<td>k uu</td>
<td>gj</td>
<td>@</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>curaí</td>
<td>k u r i i</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>dearg</td>
<td>dj</td>
<td>a r</td>
<td>@</td>
<td>g</td>
</tr>
<tr>
<td>55</td>
<td>deo</td>
<td>dj oo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>dhíol</td>
<td>gfj</td>
<td>ii</td>
<td>l</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>dó</td>
<td>d oo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>domhan</td>
<td>d au</td>
<td>n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>dorcha</td>
<td>d o r</td>
<td>@</td>
<td>x</td>
<td>@</td>
</tr>
<tr>
<td>60</td>
<td>dubh</td>
<td>d u v</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>duine</td>
<td>d i nj</td>
<td>@</td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>feoil</td>
<td>fj oo</td>
<td>lj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>fios</td>
<td>fj</td>
<td>i s</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Letter-to-sound (LTS) rule files

64 fliuch fj lj u x
65 fóill f oo lj
66 fuair f u@ rj
67 fuar f u@ r
68 gaeil g ee lj
69 gaeilge g ee lj gj @
70 gael g ee l
71 gall g a l
72 geall gj a l
73 ghaoil gf ii l
74 hata h a t @
75 im i mj
76 inné @ nj ee
77 ith i h
78 jab djzj a b
79 jíp djzj ii pj
80 leabhar lj au r
81 leigheas lj ai s
82 leon lj oo n
83 loinge l o ngj gj @
84 lón l oo n
85 longa l o ng g @
86 mac m a k
87 mála m aa l @
88 maoin m ii nj
89 mé mj ee
90 milis mj i lj @ sj
91 mín mj ii nj
92 mór m oo r
93 mothú m o h uu
94 naoi n ii
95 ní nj ii
96 óga oo g @
97 óige oo gj @
98 pacaí p a k ii
99 peacaí pj a k ii
100 radharc r ai r k
Letter-to-sound (LTS) rule files

101 riail r iØ lj
102 rua r uØ
103 saghas s ai s
104 saol s ii l
105 scoil s k o lj
106 seic sj e kj
107 seirbhís sj e rj @ vj ii sj
108 seo sj oo
109 shéamais h ee m @ sj
110 sí sj ii
111 siad sj iØ d
112 sin sj i nj
113 sláinte s l aa nj tj @
114 sleamhain sj lj au nj
115 solas s o l @ s
116 tá t aa
117 te tj e
118 thit h i tj
119 thug h u g
120 tí tj ii
121 tiubh tj u v
122 tseo tj oo
123 tú t uu
124 tuí t ii
125 veidhlín vj ai lj ii nj
126 xioleafón zj i lj @ f oo n
127 zú z uu
B.10 GaothDobhair/lts.lff

// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file contains local rules for the Gaoth Dobhair dialect of Irish.

// Variable syntax: <VARIABLE NAME><TAB>[<VALUE 1>|<VALUE 2>...]<VALUE N>]
// Rule syntax: <INPUT> = <OUTPUT>
// Test syntax: TEST <INPUT> --> <DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONG ie

// VARIABLES

// all global units
Letter-to-sound (LTS) rule files

25 V_ [AI|AU|IQ|UQ|II|IQ|I|EE|E|AA|A|UU|U|0|0|0|0|ABH|ADH#|AO|AMH#|AE|EIDH|EIGH|IDH#|IGH#|IODH#|ITHE#|OBH|ODH|ODH|OGH|OIH|OIH|O|UBH|UI]

26 // broad consonants
27 BC_ [P|B|F|V|M_D|T_|T|D_|D|LL_D|LL|L_D|L|NN_D|NN|N_D|N|
\RR_D|R_D|RR|R|S|Z|K|G|X|GF|NG_D|NG|H]

28 --stress marks
29 ANYSTRESS_ [0|1]
30
31 DNST_ [DJ|D|NNJ|NN|NJ|SJ|S|TJ|T]
32 MN_ [MJ|M|NJ|N|VJ|V]
33 XJ_ [HJ|H|XJ]
34 X_ [HJ|H|XJ|X]
35
36 // RULES AND TESTS
37
38 1 C_* [AA] X_ = a
39 C_* V_+ . 0 C_* [AA] = a
40 [AA] = aa
41 [ABH] = oo
42 [ADH#] = uu
43 [AGH] = ee
44 [AÍO] = iQ
45 [AMH#] = uu
46 [AI] DNST_ = ai
47 [AI] = ee
48 [AO] = ii
49 [AU] C_ # = ee
50 [AU] = oo
51 [A] XJ_ = ai
52 [A] = a
53
54 // máthair
55 TEST 1 M AA . 0 H @ @ R --. 1 m a . 0 h @ r
56 // amdan, urlár, sábháil, coimeád, oileán
57 TEST 1 A . 0 M @ @ . 0 D AA NN --> 1 a . 0 m @ . 0 d a nn
C_\* V_+ . 0 C_\* [EE] BC_ = a
C_\* V_+ . 0 C_\* [EE] = e
[EE] = ee
[EIDH] = ee
[EIGH] = ee
[E] X_ = ai
[E] = e

// buideil, sciléid, buicaeid
TEST 1 B UI . 0 DJ EE LJ --> 1 b i . 0 dj e lj
// coirnéal, buideál
TEST 1 B UI . 0 DJ EE LL --> 1 b i . 0 dj a ll

[IDH#] = @
[IGH#] = @
[IODH#] = ii . 0 uu
[ITHE#] = ii
[I@] = i@
[I@] = ii
[IO] = i
[I] = i

[OBH] = oo
[ÓDH#] = oo x
[ODH] = oo
[OGH] = ee
[OIDH] = ee
[OIGH] = ee
[OI] = i
[OMH] = oo
C_\* V_+ . 0 MN_ [OO] = o
C_\* V_+ . 0 C_\* [OO] = o_o
-- comhartha, bróna
[OO] AnyStress_\* MN_ = oo
-- mór, mó
Letter-to-sound (LTS) rule files

96 MN_ ANYSTRESS_* [00] = oo
97 -- óg, ór, bó, cófra
98 [00] = oo_o
99 [0] X_ = o_o
100 [0] = o

101 // bábóg, siopadóir, trioblóid, fuinneog, ainneoin
102 TEST 1 B AA . 0 B 00 G --> 1 b aa . 0 b o_o g
103
104 [UBH] = uu
105 [U@] = u@
106 [UI] = i
107
108 C_* V_+ . 0 C_* [UU] = u
109 [UU] = uu
110 [U] RJ = uu
111 [U] R = uu
112 [U] = u

113 // galún, séasúr, cosuíl
114 TEST 1 G A . 0 LL UU NN --> 1 g a . 0 ll u nn
115
116 [QQ] = @
117
118 [PJ] = pj
119 [P] = p
120 [BJ] = bj
121 [B] = b
122 [FJ] = fj
123 [F] = f
124 [VJ] = vj
125 [V] = v
126 [W] = w
127
128 [MJ_D] = mj_d
129 [M_D] = m_d
130 [MJ] = mj
131 [M] = m
132 [TJSJ] = tjsj
z = [Z] 691
fz = [fZ] 891
s = [S] 791
fs = [fS] 991
x = [R] 951
r = [RR] 291
fx = [fR] 391
p-1 = [D^{-1}] 191
p-2 = [D^{-2}] 1091
p-fx = [D^{-1}fR] 291

uu = [N] 721
fu = [fN] 821
uu = [NN] 721
fu = [fNN] 821
p-uu = [D^{-1}N] 141
p-fuu = [D^{-1}fNN] 241
p-fuu = [D^{-1}fNN] 341
p-fuu = [D^{-1}fNN] 441
p-fuu = [D^{-1}fNN] 541

III = [L] 181
fII = [fL] 281
p-II = [D^{-1}L] 941
p-fII = [D^{-1}fL] 841
p-fII = [D^{-1}fL] 741
p-fII = [D^{-1}fL] 641
p-fII = [D^{-1}fL] 541

p = [D] 691
fp = [fD] 891
-p = [-D] 131
t = [L] 191
fL = [fL] 291
-t = [-L] 141
fzfp = [fZfD] 331
Letter-to-sound (LTS) rule files

170 [KJ] = kj
171 [K] = k
172 [GJ] = gj
173 [G] = g
174 [XJ] = xj
175 [X] = h
176 [GFJ] = gfj
177 [GF] = gf
178 [NGJ_D] = ngj_d
179 [NG_D] = ng_d
180 [NGJ] = ngj
181 [NG] = ng
182 [HJ] = hj
183 [H] = h
184 [ ] =
185 [0] = 0
186 [1] = 1
187 [2] = 2
188 [.] = .
189
190 TEST R 00 NNJ --> r oo nnj
191 TEST R 00 VJ --> r oo vj
192 TEST R 00 T --> r oo_o t
193
B.11 GaothDobhair/lex-exceptions.txt

1 // This file is a component of a multi-dialect modular phonetiser for Irish.
2 // A phonetiser converts orthographic text to dialectal phonemic transcription.
3 // This phonetiser has two layers: one global, one local.
4 // At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
5 // At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
6 // The global layer consists of one global module.
7 // The local layer consists of one local module for each dialect implemented.
8 // Each module contains the following:
9 // (1) A set of rewrite (letter-to-sound) rules.
10 // (2) An exceptions lexicon.
11 // (3) A set of tests.
12
13 // This file is the local exceptions lexicon for the Gaith Dobhair dialect of Irish.
14
15 // Syntax of lexicon entry:
16 // <ORTHOGRAPHY>
17 // <TAB><TRANSCRIPTION>
18
19 // Author: Brian Ó Raghallaigh
20 // oraghalb AG tcd PONC ie
21
23
24 acra
25 1 a . 0 k @ . 0 rj @
26 aibhéis
27 1 a . 0 vj lj e sj
Letter-to-sound (LTS) rule files

28 aibí
29 1 a . 0 p ii
30 aicearra
31 1 a . 0 xj @ . 0 r @
32 aicearracht
33 1 a . 0 xj @ . 0 r @ x t
34 aingiallta
35 1 a ngj . 0 kj @ ll . 0 t @
36 áirithe
37 1 aa . 0 rj_d @ dj
38 áirneán
39 1 aa rj . 0 nnj a ll
40 aisteach
41 1 a . 0 sj tj ia h
42 aiteann
43 1 a . 0 tj @ . 0 nn a h
44 albanach
45 1 a . 0 ll @ b . 0 nn a h
46 amanna
47 1 a . 0 m @ . 0 nn ii
48 amharc
49 1 oo_o r k
50 anóirthear
51 0 a . 1 nn oo_o . 0 rj_d @ rj
52 anróiteach
53 1 aa . 0 r oo_o . 0 tj a h
54 argóint
55 1 a r . 0 g a lj
56 babhla
57 1 b oo lj
58 bail
59 1 b o lj
60 báitheadh
61 1 b aa . 0 xj uu
62 bascaed
63 1 b a . 0 s k aa dj
64 báthadh
Letter-to-sound (LTS) rule files

1 b aa . 0 h uu
beithíoch
1 bj e . 0 gfj a h
bídeach
1 b ii . 0 dj a h
bóitheach
1 b oo . 0 gfj a h
bronn
1 p r o nn
bronntanas
1 p r o nn . 0 t @ . 0 nn @ s
bua
1 b ua . 0 gfj @
buatais
1 b ua . 0 t @ . 0 sj @
buicéad
1 b u . 0 k a dj
caurán
1 k ee . 0 r a n
captaen
1 k a fj . 0 tj ii nn
cártta
1 k aa . 0 d @
casachtach
1 k a . 0 s @ x . 0 t ii
ceann
1 kj i nn
céanna
1 kj ia . 0 nn @
cé
1 kj ee gfj
chomh
1 g o
chuig
1 h i gj
cithréim
1 kj o . 0 r_d @ m
cithréimeán
1 kj o . 0 r_d @ . 0 m a n
clé
1 kj lj ii
clú
1 kj lj uu
cófra
1 k oo_o . 0 r_d @
cónra
1 k oo . 0 nn a rj
confach
1 k o . 0 nn @ . 0 f a h
contae
1 k o nn . 0 d a i
coinleach
1 k o nn . 0 ll a h
corr
1 k o_o r
cosán
1 k a . 0 s a nn
crampa
1 k r a m . 0 b ii
crua
1 k r ua gfj
cuardaigh
1 k ua r . 0 t ii
cúpla
1 k u . 0 p ll @
dada
1 d a . 0 d ii
dála
1 d aa ll . 0 t @
daingean
1 d ai . 0 @ nn
daíríre
1 d aa . 1 rj ii . 0 rj @ vj
dea-
Letter-to-sound (LTS) rule files

139  1 dj ee
140  deichniúr
141  1 dj e . 0 xj @ r
142  deifir
143  1 dj e . 0 fj rj @
144  dearóil
145  1 dj e . 0 llj o rj
146  deirfiúr
147  1 dj e . 0 rj @ . 0 fj @ r
148  deoch
149  1 dj o_o h
150  deoir
151  1 dj oo_o r
152  dhá
153  1 gfj aa
154  dheamhan
155  1 gfj oo nn
156  diaidh
157  1 dj ee gfj
158  difear
159  1 d i . 0 fj @ r
160  ding
161  1 dj i nnj
162  dóbháil
163  1 dj au lj
164  dlí
165  1 dj lj ii . 0 uu
166  dóbair
167  1 h o . 0 b @ rj
168  doirt
169  1 d oo rj tj
170  dosaen
171  1 d i . 0 sj ii nj
172  dreocilín
173  1 dj rj oo . 0 ll a nn
174  eagal
175  1 i . 0 g @ ll
éigin
1    i    .  0    n j    a    r t
eireaball
1    r    u    .  0    b    @    l l
eochair
1    o o    _ o    .  0    h    @    r j
falsean
1    f    a    .  0    s j    a    n n
faobhar
1    f    i i    .  0    v    @    r
faoi
1    f    a a
faoistin
1    f    i i    .  0    s j    @    .  0    d j    @
féach
1    f    j    i a    x
feadaíl
1    f j    a    .  0    d    @    .  0    l l    @    x
feadóg
1    f j    i    .  0    d j    o    g
feamainn
1    f j    a    .  0    m    n n    @    x
fearthainne
1    f j    a    .  0    r_d    @    .  0    n n    @
feirm
1    f a    .  0    r    @    m
feirmeoir
1    f a    .  0    r    .  0    m    o    r j
fill
1    p j    i l l j
foighne
1    f    e e    d j
freagra
1    f j    r j    a    .  0    g f j    @    r
furasta
1    f u    .  0    r    @    s t
gabháil
213 1 g o lj
214 gaeilge
215 1 g ee . 0 lj @ kj
216 gafa
217 1 g o . 0 tj @
218 gallda
219 1 g a ll . 0 t @
220 garraí
221 1 g a . 0 r uu
222 gealluínt
223 1 gj a ll . 0 s t @ nn
224 gearradh
225 1 gj aa . 0 r uu
226 geata
227 1 gj a f . 0 t @
228 giorria
229 1 gj a . 0 r ia
230 glúin
231 1 g ll uu nn
232 gnách
233 1 g r aa x
234 gnaíúil
235 1 g r ii . 0 v i lj
236 gnaoi
237 1 g r ii
238 gnó
239 1 g r ai . 0 hj @
240 gnóthach
241 1 g r ii . 0 hj a h
242 grág
243 1 g r aa gj
244 gráinne
245 1 g r ai nj
246 iargúlta
247 1 ii . 0 r @ . 0 g uu ll . 0 t @
248 iarraidh
249 1 gfj i . 0 r ii
Letter-to-sound (LTS) rule files

250  iníon
251     1 nj ii nn
252  idù
253     1 uu lj
254  leaba
255     1 llj a . 0 b ii
256  lear
257     1 lj ee r
258  leointe
259     1 ll oo nn . 0 t ii
260  luchóg
261     1 ll o . 0 x oo g
262  madra
263     1 m a . 0 d uu
264  máistir
265     1 m ai . 0 sj tj @ rj
266  máistreáis
267     1 m ai . 0 sj tj rj a s
268  mala
269     1 m a . 0 ll ii
270  marú
271     1 m a . 0 r @ . 0 v @
272  mearbháin
273     1 mj i r . 0 ll a nn
274  móra
275     1 m o . 0 r @
276  mothaigh
277     1 m ai . 0 hj @
278  muidne
279     2 m i . 1 dj i . 0 nnj @
280  muinchille
281     1 m i . 0 xj i . 0 llj @
282  műineadh
283     1 m uu . 0 nn uu
284  naonúr
285     1 nn ii . 0 @ r
286  níos
Letter-to-sound (LTS) rule files

1 nn @ s
nuacht
1 nn ua . 0 gfj a r t
oidhre
1 ee . 0 rj @
oidreacht
1 ee . 0 rj ia r t
oiread
1 o . 0 r @ d
oscaíl
1 f o . 0 s k @ lj
piopa
1 p ii . 0 p @
piséeog
1 pj i . 0 sj lj o g
poitín
1 p oo_o . 0 tj ii nj
portán
1 p aa . 0 t a nn
post
1 p o . 0 s t @
punt
1 p u nn . 0 t @
rámh
1 r aa . 0 v @
ramhar
1 r a . 0 v @
riamh
0 @ . 1 r j ia v
roimh
0 @ . 1 r i . 0 v j @
saoithídil
1 s ii . 0 f @ lj
scaip
1 s k a b
scabhóg
1 s k a . 0 v a nn
Letter-to-sound (LTS) rule files

324 scian
325  1 sj kj i nn
326
327 scith
328  1 sj kj rj ii . 0 sj tj @
329 scoth
330  1 s k ai hj
331 seic
332  1 tj e kj
333 seisreach
334  1 sj e . 0 sj @ . 0 rj a h
335 seo
336  1 sj o_o
337 sil
338  1 s i lj
339 siúcra
340  1 sj uu . 0 r_d a
341 siúd
342  1 sj o d
343 smoineamh
344  1 s m ii . 0 tj @ v
345 soir
346  1 s i r
347 sonraíoch
348  1 s oo nn . 0 r ia h
349 sórt
350  1 sj oo r t
351 spionn
352  1 s pj i n
353 spuaic
354  1 s p ee kj
355 stiúg
356  1 sj tj uu k
357 stiúgtha
358  1 sj tj uu . 0 k ii
359 suaimhneach
360  1 s ua . 0 nnj a h
361 tanaí
Letter-to-sound (LTS) rule files

1 t a . 0 nn @
taobh
1 t ii vj
tapa
1 t a . 0 p @
tarnocht
1 t aa r . 0 nn a r t
teanga
1 t a ng . 0 ii
thar
1 h a . 0 rj @
tine
1 t j i . 0 nj ii
teilgean
1 t j l j i . 0 gj @ nn
tóirneach
1 t aa r j . 0 nj a h
tosach
1 t o . 0 sj a h
tosaigh
1 t o . 0 sj ii
tré
1 t r a a gfj
trua
1 t r u a . 0 gfj @
tua
1 t u a gfj
tuigbheáil
1 t i gj . 0 vj a lj
tuilleadh
1 t j i . 0 llj uu
ubh
1 i vj
ubhágán
1 i . 0 vj @ . 0 g a nn
úd
1 a . 0 d ii
Letter-to-sound (LTS) rule files

398  uile
399    1 i . 0 lj @ gj
400  úll
401    1 uu . 0 ll @
402  veain
403    1 vj a nn
404  veist
405    1 vj e . 0 sj tj @
Letter-to-sound (LTS) rule files

B.12 GaorthDobhair/test.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file contains tests for the Gaoth Dobhair dialect of Irish.

// Test syntax: <INPUT><TAB><DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie

abhainn 1 oo nnj
ádh 1 aa
aidhm 1 ee mj
aighneas 1 ai . 0 nj @ s
airde 1 aa rj . 0 dj @
áireamh 1 aa . 0 rj uu
áireamhán 1 aa . 0 rj @ . 0 v a nn
áiseanna 1 aa . 0 sj @ . 0 nn @
ait 1 a tj
29 aithris 1 a . 0 rj_d @ sj
30 amharc 1 oo r k
31 ard 1 aa r d
32 at 1 a t
33 atfaidh 1 a . 0 t @
34 áthas 1 a . 0 h @ s
35 athdhéanamh 1 a . 0 gfj ee . 0 nn uu
36 bainseó 1 b a nj . 0 sj o_o
37 bantaboc 1 b a nn . 0 t @ . 0 b @ k
38 bantaboic 1 b a nn . 0 t @ . 0 b @ kj
39 beairic 1 bj a . 0 rj @ kj
40 bhfánne 1 v aa . 0 nnj @
41 bhfianaise 1 vj i@ . 0 nn @ . 0 sj @
42 bhris 1 vj rj i sj
43 bhrostaigh 1 v r o . 0 s t @
44 bhuel 1 v e ll
45 bláthra 11 b l l a . 0 r_d @
46 bodhair 1 b oo_o rj
47 bodhar 1 b oo r
48 bogha 1 b oo
49 boird 1 b oo rj dj
50 borb 1 b o . 0 r @ b
51 bord 1 b oo r d
52 bos 1 b o s
53 bpáistí 1 b a a . 0 sj tj ii
54 bpéisteanna 1 bj ee . 0 sj tj @ . 0 nn @
55 brian 1 bj rj i@ nn
56 broghais 1 b r e e sj
57 bróna 1 b r oo . 0 nn @
58 bruíne 1 b r ii nj @
59 bruíon 1 b r ii nn
60 bualadh 1 b u@ . 0 ll uu
61 burdún 1 b uu r . 0 d u nn
62 bus 1 b u s
63 cait 1 k a tj
64 cat 1 k a t
65 cead 1 kj a d
Letter-to-sound (LTS) rule files

ceadharlach 1 kj au r . 0 ll @ x
ceannaíodh 1 kj a . 0 nn ii . 0 uu
cеard 1 kj aa r d
ceird 1 kj ee rj dj
cеoil 1 kj oo lj
ceol 1 kj oo_o ll
cheadaigh 1 xj a . 0 d @
cheannaigh 1 xj a . 0 nn @
cheannódh 1 xj a . 0 nn oo x
ciumhais 1 kj uu sj
clibfeadh 1 kj lj i . 0 pj uu
clibfidh 1 kj lj i . 0 pj @
clibthe 1 kj lj i . 0 pj @
coibhín 1 k oi . 0 vj ii nj
comhair 1 k oo_o rj
comhar 1 k oo r
coid 1 k i dj
cumhacht 1 k uu x t
cur 1 k u r
déanamh 1 dj ee . 0 nn uu
déis 1 dj e sj
díbhfeadh 1 dj ii . 0 fj uu
díbhfidh 1 dj ii . 0 fj @
diúltaithe 1 dj uu ll . 0 t ii
dumhach 1 d uu x
dúnfaidh 1 d uu . 0 nn_d @
é 1 ee
eamhain 1 ee nj
eán 1 ee nn
eas 1 a s
éiníní 1 ee . 0 nj ii . 0 nj ii
eitpheil 1 e tj . 0 fj e lj
eoghain 1 oo nj
eoghan 1 oo nn
feadhain 1 fj ee nj
feidhm 1 fj ee mj
Letter-to-sound (LTS) rule files

103  fios 1 fj io s
104  fiuch 1 fj ux
105  foighid 1 f ee dj
106  gaoil 1 g ii lj
107  gaol 1 g ii ll
108  gardai 1 g aa r. 0 d ii
109  giuirléid 1 gj uu rj. 0 lj e dj
110  guird 1 g uu rj dj
111  imithe 1 i. 0 mj ii
112  leathbhosca 1 lj a. 0 v @ s k @
113  leigheas 1 lj ee s
114  leighis 1 lj ee sj
115  leódhas 1 lj oo_o s
116  lobhadh 1 ll oo uu
117  lobhair 1 ll oo_o rj
118  lobhar 1 ll oo r
119  meathfadh 1 mj ai. 0 h uu
120  cíche 1 ii. 0 xj @
121  oídhe 1 ee. 0 @
122  oighhear 1 ee r
123  óil 1 oo lj
124  ól 1 oo ll
125  riall 1 rj i@ lj
126  rithfeadh 1 rj i. 0 xj uu
127  sadhbh 1 s ee vj
128  saghas 1 s ai s
129  saolaíodh 1 s ii. 0 ll ii. 0 uu
130  sciath 1 sj kj i@
131  scríobhfadh 1 sj kj rj ii. 0 f uu
132  scríobhfaidh 1 sj kj rj ii. 0 f @
133  scuabfadh 1 s k u@. 0 p uu
134  scuabfaidh 1 s k u@. 0 p @
135  scuabtha 1 s k u@. 0 p @
136  seabhac 1 sj oo k
137  seabhaic 1 sj oo kj
138  seán 1 sj aa nn
139  searbh 1 sj a. 0 r @ v
Letter-to-sound (LTS) rule files

seinneadh 1 sj e . 0 nnj uu
seinnfidh 1 sj e . 0 nnj_d @
seirbhís 1 sj e . 0 rj @ . 0 vj ii sj
seirbiach 1 sj e . 0 rj @ . 0 bj i@ x
sheáin 1 xj aa nj
sí 1 sj ii
siad 1 sj i@ d
siadhail 1 sj i@ lj
sin 1 sj i nj
síos 1 sj ii s
siúil 1 sj uu lj
siúl 1 sj uu ll
sleamhnán 1 sj lj au . 0 nn a nn
sodair 1 s o d @ rj
sodar 1 s o . 0 d @ r
soir 1 s i rj
subh 1 s uu
súil 1 s uu lj
súl 1 s uu ll
t-éabhlóidí 1 tj ee . 0 v ll o_o . 0 dj ii
teaghaí 1 t a sj
teaghásán 1 tj a i . 0 s a nn
teaghlach 1 tj a i . 0 ll @ x
teann 1 tj a nn
tharla 1 h aa r . 0 ll @
thit 1 hj i tj
tit 1 tj i tj
titfidh 1 tj i . 0 tj @
tiubh 1 tj ubh
todhsaí 1 t oo . 0 x ii
toghadh 1 t ee . 0 uu
togra 1 t o . 0 g r @
tsagairt 1 t a . 0 g @ rj tj
tseachtain 1 tj a x . 0 t @ nj
t-usc 1 t i . 0 sj kj @
uaine 1 u@ . 0 nj @
uan 1 u@ nn
Letter-to-sound (LTS) rule files

177 uathðil  ı  u@ . 0 u lj
178 uile  ı  i lj @
B.13 Conamara/lts.lff

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file contains local rules for the Conamara dialect of Irish.

// Variable syntax: <VARIABLE NAME><TAB>[<VALUE 1>|<VALUE 2>...|<VALUE N>]

// Rule syntax: <INPUT> = <OUTPUT>

// Test syntax: TEST <INPUT> --> <DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie

// VARIABLES

RST_ [RR|R|S|T]
NJ_ [MJ|NGJ|NJ]

// RULES AND TESTS
Letter-to-sound (LTS) rule files

[AA] = aa
[ABH] = au
[ADH#] = @ x
[AGH] = ai
[AfO] = i@
[AMH#] = @
[AI] = ai
[AO] = ii
[AU] = au
[A] = a

// cáis, abháinn, bhrisfeadh, laghad, safocht
// déanamh, tadhg, saol, samhail, cad
TEST 1 K AA S --> 1 k aa s
TEST 1 ABH NNJ --> 1 au nnj
TEST 1 VJ RJ I . O SJ ADH# --> 1 vj rj i . O sj @ x
TEST 1 LL AI D --> 1 ll ai d
TEST 1 S Ao x T --> 1 s i@ x t
TEST 1 DJ EE NN AMH# --> 1 dj ee nn @
TEST 1 T AI G --> 1 t ai g
TEST 1 S AO LL --> 1 s ii ll
TEST 1 S AU LJ --> 1 s au lj
TEST 1 K A D --> 1 k a d

[EE] = ee
[EIDH] = ai
[EIGH] = ai
[E] = e

// céim, feidhm, leigheas, deis
TEST 1 KJ EE MJ --> 1 kj ee mj
TEST 1 FJ EIDH MJ --> 1 fj ai mj
TEST 1 LJ EIGH S --> 1 lj ai s
TEST 1 DJ E SJ --> 1 dj e sj

[IDH#] = @
Letter-to-sound (LTS) rule files

65 [IGH#] = @
66 [IODH#] = ii . 0 @ v
67 [ITHE#] = ii
68 [IQ] = iQ
69 [II] = ii
70 [IO] RST_ = i
71 [IO] X = o
72 [IO] = u
73 [I] = i
74
75 // brisfidh, cheannaigh, ceannáiodh, ceannaithe
76 // iarr, sín, fios, sliocht, global, sin
77 TEST 1 BJ RJ I . 0 SJ IDH# --> 1 bj rj i . 0 sj @
78 TEST 1 XJ A . 0 NN IGH# --> 1 xj a . 0 nn @
79 TEST 1 KJ A . 0 NN IODH# --> 1 kj a . 0 nn ii . 0 @ v
80 TEST 1 KJ A . 0 NN ITHE# --> 1 kj a . 0 nn ii
81 TEST 1 IQ RR --> 1 i@ r
82 TEST 1 SJ II NJ --> 1 sj ii nj
83 TEST 1 FJ IO S --> 1 fj i s
84 TEST 1 SJ LJ IO X T --> 1 sj lj o x t
85 TEST 1 GJ IO . 0 B @@ LL --> 1 gj u . 0 b @ ll
86 TEST 1 SJ I NJ --> 1 sj i nj
87
88 [OBH] = au
89 [ODH#] = oo x
90 [ODH] = au
91 [OGH] = au
92 [OIDH] = ai
93 [OIGH] = ai
94 [OI] XJ . 0 TJ = o
95 [OI] XJ TJ = o
96 [OI] R TJ = o
97 [OI] . 0 NJ_ = i
98 [OI] NJ_ = i
99 [OI] = e
100 [OMH] = au
101 [O0] = oo
Letter-to-sound (LTS) rule files

\[O\] = o

// lobhar, ceannódh, bodhar, toghchán, oidhreacht, oighear
// boichte, goirt, gloine, oirthear, domhan, mór, bog

TEST 1 LL O BH R --> 1 ll au r
TEST 1 K J A . O NN ÓDH# --> 1 kj a . O nn oo x
TEST 1 B O DH R --> 1 b au r
TEST 1 T OGH . O X AA NN --> 1 t au . O x aa nn
TEST 1 O IDH . O RJ A X T --> 1 ai . O rj a x t
TEST 1 O IGH R --> 1 ai r
TEST 1 B O I X J . O TJ @@ --> 1 b o xj . O tj @
TEST 1 G O I R TJ --> 1 g o r tj
TEST 1 G LL O I . O NJ @@ --> 1 g ll i . O nj @
TEST 1 O I . O R_D @@ R --> 1 e . O r_d @ r
TEST 1 D O MH NN --> 1 d au nn
TEST 1 M OO R --> 1 m oo r
TEST 1 B O G --> 1 b o g

\[UBH\] # = u v
\[UBH\] = @ v
\[U@\] = u@
\[UI\] = u
\[UU\] = uu
\[U\] = u

// ubhagán, bua, cuid, cúr, cur
TEST 1 UBH . O @@ . O G AA NN --> 1 @ v . O @ . O g aa nn
TEST 1 B U@ --> 1 b u@
TEST 1 K UI DJ --> 1 k u dj
TEST 1 K UU R --> 1 k uu r
TEST 1 K U R --> 1 k u r

\[@@\] = @

\[PJ\] = pj
\[P\] = p
\[BJ\] = bj
Letter-to-sound (LTS) rule files

176 \[[R\_D]\] = r\_d
177 \[[RRJ]\] = rj
178 \[[RJ]\] = rj
179 \[[RR]\] = r
180 \[[R]\] = r
181 \[[SJ]\] = sj
182 \[[S]\] = s
183 \[[ZJ]\] = zj
184 \[[Z]\] = z
185 \[[KJ]\] = kj
186 \[[K]\] = k
187 \[[GJ]\] = gj
188 \[[G]\] = g
189 \[[XJ]\] = xj
190 \[[X]\] = x
191 \[[GFJ]\] \# =
192 \[[GFJ]\] = gfj
193 \[[GF]\] = gf
194 \[[NGJ\_D]\] = ngj\_d
195 \[[NG\_D]\] = ng\_d
196 \[[NGJ]\] = ngj
197 \[[NG]\] = ng
198 \[[HJ]\] = h
199 \[[H]\] = h

200 TEST 1 KJ A . 0 NN 00 GFJ --> 1 kj a . 0 nn 00

201 \[\] =

202 \[[0]\] = 0
203 \[[1]\] = 1
204 \[[2]\] = 2
205 \[[.]\] = .
B.14  Conamara/lex-exceptions.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file is the local exceptions lexicon for the Conamara dialect of Irish.

// Syntax of lexicon entry:
// <ORTHOGRAPHY>
// <TAB><TRANSCRIPTION>

// Author: Brian Ó Raghallaigh
// oraghbalb AG tcd PONC ie


beach
1 mj a x
bearrán
1 bj i . 0 r aa nn
28 bolgam
29 1 b l l o . 0 g @ m
30 chaon
31 1 x e e n n
32 custaiméir
33 1 k u . 0 s j l j @ . 0 m j e e . 0 r @
34 diabhal
35 1 d j a u . 0 @ l l
36 dheamhan
37 1 g f j u u n n
38 drochmhúinte
39 1 d r o . 1 x u u n j . 0 t j @
40 foscadh
41 1 f a . 0 s k @
42 foscúil
43 1 f a . 0 s k u u l j
44 gabháil
45 1 g o o l j
46 gal
47 1 g a l j
48 gnó
49 1 g r o . 0 h @
50 gnóigh
51 1 g r u u
52 meirfean
53 1 m j e . 0 r j @ . 0 f j @ n n
54 racán
55 1 r u . 0 k a a n n
56 scáfánta
57 1 s g u . 0 f a a n n . 0 t @
58 suim
59 1 s ii m j
60 triail
61 1 t r i i . 0 a a l j
62 trust
63 1 t r u s d
B.15 Conamara/test.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.
// This file contains tests for the Conamara dialect of Irish.

// Test syntax: <INPUT><TAB><DESIRED OUTPUT>
// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie
B.16 CorcaDhuibhne/lts.lff

1 // This file is a component of a multi-dialect modular phonetiser for Irish.
2 // A phonetiser converts orthographic text to dialectal phonemic transcription.
3 // This phonetiser has two layers: one global, one local.
4 // At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
5 // At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
6 // The global layer consists of one global module.
7 // The local layer consists of one local module for each dialect implemented.
8 // Each module contains the following:
9 // (1) A set of rewrite (letter-to-sound) rules.
10 // (2) An exceptions lexicon.
11 // (3) A set of tests.
12
13 // This file contains local rules for the Corca Dhuibhne dialect of Irish.
14
15 // Variable syntax: <VARIABLE NAME><TAB>[<VALUE 1>|<VALUE 2>]. . . |<VALUE N>]
16 // Rule syntax: <INPUT> = <OUTPUT>
17 // Test syntax: TEST <INPUT> --> <DESIRED OUTPUT>
18
19 // Author: Brian Ó Raghallaigh
20 // oraghalb AG tcd PONG ie
21
22 // VARIABLES
23
24 // broad consonants
25 BC_ [P|B|F|V|M_D|M_T_|T_|D_|D|LL_D|LL|L_D|L|NN_D|NN|N_D|N]
26 RR_D|R_D|RR|R|S|Z|K|G|X|GF|NG_D|NG|H]
27 LLJ_ [LLJ|NNJ|RRJ|MJ]
Letter-to-sound (LTS) rule files

```
27  LL_ [LL|NN|RR|M]
28
29  // RULES AND TESTS
30
31  [AA] = aa
32  [ABH] = au
33  [ADH#] = @
34  [AGH] = ai
35  [AIO] = i@
36  [AMH#] = @ v
37  [AI] = ai
38  [AO] = ee
39  [AU] = au
40  [A] LLJ_ # = ai
41  [A] LL_ # = au
42  [A] = a
43
44  // fas, abhainn, saghad, safocht, deanamh, tadhg
45  // saol, samhail, caill, crann, cad
46  TEST 1 F AA S --> 1 f aa s
47  TEST 1 ABH NNJ --> 1 au nj
48  TEST 1 S AGH D --> 1 s ai d
49  TEST 1 S AIO X T --> 1 s i@ x t
50  TEST 1 DJ EE . O NN AMH# --> 1 dj ee . O n @ v
51  TEST 1 T AI G --> 1 t ai g
52  TEST 1 S AO LL --> 1 s ee l
53  TEST 1 S AU LJ --> 1 s au lj
54  TEST 1 K A LLJ --> 1 k ai lj
55  TEST 1 K R A NN --> 1 k r au n
56  TEST 1 K A D --> 1 k a d
57
58  [EE] BC_ = i@
59  [EE] = ee
60  [EIDH] # = e gj
61  [EIDH] = ai
62  [EIGH] = ai
63  [E] = e
```
// óan, sé, beidh, feidhm, leigheas, deis
TEST 1 EE NN --> 1 i@ n
TEST 1 SJ EE --> 1 sj ee
TEST 1 B EIDH --> 1 b e gj
TEST 1 FJ EIDH MJ --> 1 fj ai mj
TEST 1 LJ EIGH S --> 1 lj ai s
TEST 1 DJ E SJ --> 1 dj e sj

// post-lexical rules required to deal with exceptions to next two rules
[IDH#] = @ gj
[IGH#] = @ gj
[ÍODH#] = i@ x
[ITHE#] = ò . 0 h @
[IΩ] = iΩ
[II] = ii
[IO] = i
[I] LLJ_ # = ii
[I] = i

// brisfidh, cheannaigh, ceannáidh, ceannaithe
// iarr, sín, fios, im, sin
TEST 1 BJ RJ I . 0 SJ IDH# --> 1 bj rj i . 0 sj @ gj
TEST 1 XJ A . 0 NN IGH# --> 1 xj a . 0 n @ gj
TEST 1 KJ A . 0 NN ÍODH# --> 1 kj a . 0 n i@ x
TEST 1 KJ A . 0 NN ÍTHER# --> 1 kj a . 0 n @ . 0 h @
TEST 1 IΩ RR --> 1 i@ r
TEST 1 SJ II NJ --> 1 sj ii nj
TEST 1 FJ IO S --> 1 fj i s
TEST 1 I MJ --> 1 ii mj
TEST 1 SJ I NJ --> 1 sj i nj

[OBH] = au
[ÓDH#] = oo x
[ODH] = au
[OOGH] = au
Letter-to-sound (LTS) rule files

[OIDH] = ai
[OIGH] = ai
[OI] LLJ_ # = ii
[OI] = i
[OMH] = au
[00] = oo
[0] LL_ # = au
[0] = o

// lobbhe, ceannódh, bodharch, toghchán, oidhreacht
// oighcare, coill, coille, móir, poll, polladh
TEST 1 LL OBH R --> 1 l au r
TEST 1 KJ A . 0 NN ÓDH# --> 1 kj a . 0 n oo x
TEST 1 B ODH R --> 1 b au r
TEST 1 T OGH . 0 X AA NN --> 1 t au . 0 x aa n
TEST 1 OIDH . 0 RJ A X T --> 1 ai . 0 rj a x t
TEST 1 OIGH R --> 1 ai r
TEST 1 K OI LLJ --> 1 k ii lj
TEST 1 K OI LLJ @@ --> 1 k i lj @
TEST 1 D OMH NN --> 1 d au n
TEST 1 M 00 R --> 1 m oo r
TEST 1 P 0 LL --> 1 p au l
TEST 1 P 0 . 0 LL ADH# --> 1 p o . 0 l @ x

[UBH] # = u v
[UBH] = @ v
[U@] = u@
[UI] LLJ_ = ii
[UI] = i
[UU] = uu
[U] = u

// ubhagán, bua, suim, cuid, cúr, cur
TEST 1 UBH . 0 @@ . 0 G AA NN --> 1 @ v . 0 @ . 0 g aa n
TEST 1 B U@ --> 1 b u@
TEST 1 S UI MJ --> 1 s ii mj
TEST 1 K UI DJ --> 1 k i dj
Letter-to-sound (LTS) rule files

137 TEST 1 K UU R --> 1 k uu r
138 TEST 1 K U R --> 1 k u r
139
140 [@@] = @
141
142 [PJ] = pj
143 [P] = p
144 [BJ] = bj
145 [B] = b
146 [FJ] = fj
147 [F] = f
148 [VJ] = vj
149 [V] = v
150 [W] = w
151 [MJ_D] = mj_d
152 [M_D] = m_d
153 [MJ] = mj
154 [M] = m
155 [TJSJ] = tjsj
156 [DJZJ] = djzj
157 [T_] = t_
158 [TJ] = tj
159 [T] = t
160 [D_] = d_
161 [DJ] = dj
162 [D] = d
163 [LLJ_D] = lj_d
164 [LL_D] = l_d
165 [LLJ] = lj
166 [LL] = l
167 [LJ_D] = lj_d
168 [L_D] = l_d
169 [LJ] = lj
170 [L] = l
171 [NNJ_D] = nj_d
172 [NN_D] = n_d
173 [NNJ] = nj
Letter-to-sound (LTS) rule files

[NN] = n
[NJ_D] = nj_d
[N_D] = n_d
[NJ] = nj
[N] = n
[RRJ_D] = rj_d
[RJ_D] = rj_d
[RR_D] = r_d
[R_D] = r_d
[RRJ] = rj
[RJ] = rj
[RR] = r
[R] = r
[SJ] = sj
[S] = s
[ZJ] = zj
[Z] = z
[KJ] = kj
[K] = k
[GJ] = gj
[G] = g
[XJ] = xj
[X] = x
[GFJ] # =
[GFJ] = gfj
[GF] = gf
[NGJ_D] = ngj h
[NG_D] = ng h
[NGJ] # = nj
[NGJ] = ngj
[NG] = ng
[HJ] = h
[H] = h

// fairsing, scilling
TEST 1 F A RJ . O SJ @@ NGJ --> 1 f a rj . O sj @ nj
[ ] =

[0] = 0

[1] = 1

[2] = 2

[..] = .
B.17 CorcaDhuibhne/lex-exceptions.txt

1 // This file is a component of a multi-dialect modular phonetiser for Irish.
2 // A phonetiser converts orthographic text to dialectal phonemic transcription.
3 // This phonetiser has two layers: one global, one local.
4 // At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
5 // At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
6 // The global layer consists of one global module.
7 // The local layer consists of one local module for each dialect implemented.
8 // Each module contains the following:
9 // (1) A set of rewrite (letter-to-sound) rules.
10 // (2) An exceptions lexicon.
11 // (3) A set of tests.
12 // This file is the local exceptions lexicon for the Corca Dhuibhne dialect of Irish.
13
14 // Syntax of lexicon entry:
15 // <ORTHOGRAPHY>
16 // <TAB><TRANSCRIPTION>
17
18 // Author: Brian Ó Raghallaigh
19 // oraghalyb AG tcd PONC ie
20
22
23 abláil
24 1 a . 0 b @ . 0 l aa lj
25 achasán
26 0 @ . 0 s @ . 1 x aa n
Letter-to-sound (LTS) rule files

28 aer
29 1 ee rj
30 aibreán
31 0 @ . 1 b r a a n
32 aireach
33 0 @ . 1 rj a x
34 amadán
35 0 @ . 0 m @ . 1 d a a n
36 amárách
37 0 @ . 1 m aa . 0 rj @ x
38 amharc
39 1 a . 0 v @ r k
40 amhrán
41 0 uu . 1 r aa n
42 anáil
43 0 @ . 1 nj aa l
44 annamh
45 1 a . 0 n @
46 anois
47 0 @ . 1 nj i sj
48 anráith
49 1 a . 0 n @ . 0 rj_d @
50 anseo
51 0 @ n . 1 s o
52 ansin
53 0 @ n . 1 s u n
54 ansiúd
55 0 @ n . 1 s u u d
56 arú
57 0 @ . 0 r @ . 1 v u u
58 ascaill
59 1 o . 0 s k @ lj
60 banaltra
61 1 b a . 0 n @ r_d . 0 l @
62 beach
63 1 mj a x
64 blaincéad
1 p l a n gj . 2 kj ee d
boladh
1 b l a
bollóg
1 b l o o g
bos
1 b a s
bradán
0 b u r . 1 d a a n
braillín
0 b a . 0 r @ . 1 l j i i nj
buicéad
1 b o . 0 k @ d j
buíoch
1 b e e x
caincín
1 k i i nj . 0 k j i i nj
canta
1 k a u n . 0 t @
carbhat
1 k a . 0 r @ . 2 v a t
céannaithé
1 k j n j a . 0 h @
ceant
1 k j a n j t j
céirtlín
0 k j e r j _ d . 1 l j i i nj
céobhrán
1 k j o o . 0 v @ . 0 r a a n
ciotrúnta
0 k j @ . 0 t @ . 1 r u u n . 0 t @
ciumhais
1 k j o . 0 s @
cliathán
1 k j l j i i . 0 h a a n
clipthe
1 k j l j i . 0 p @ . 0 h @
102 cluiche
103 1 k l o . 0 h @
104 cluthar
105 1 k l u . 0 h @ rj
106 coimhthioch
107 1 k ii . 0 h a x
B.18 CorcaDhuibhne/test.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file contains tests for the Corca Dhuibhne dialect of Irish.

// Test syntax: <INPUT><TAB><DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie
B.19 Template/lts.lff

// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file contains local rules for the <DIALECT REGION> Irish dialect.

// Variable syntax: <VARIABLE NAME><TAB>[<VALUE 1>|<VALUE 2>|...|<VALUE N>]
// Rule syntax: <INPUT> = <OUTPUT>
// Test syntax: TEST <INPUT> --> <DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie

// VARIABLES

// RULES AND TESTS

[AA] = aa
[ABH] = au
[ADH#] = ø x
[AGH] = ai
[AI0] = ii
[AMH#] = @ v
[AI] = ai
[A0] = ii
[AU] = au
[A] = a
[EE] = ee
[EIDH] = ai
[EIGH] = ai
[E] = e
[IDH#] = ii
[IGH#] = ii
[IODH#] = ii x
[ITHE#] = i h @
[IQ] = i@
[II] = ii
[IO] = i
[I] = i
[OBH] = au
[ODH#] = oo x
[ODH] = au
[OGH] = au
[OIDH] = ai
[OIGH] = ai
[OI] = o
[OH#] = au
[OO] = oo
[O] = o
[UBH] = u v
[UBH] = @ v
[U@] = u@
[UI] = i
[UU] = uu
<table>
<thead>
<tr>
<th>Line</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>([U] = u)</td>
</tr>
<tr>
<td>67</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>([@] = @)</td>
</tr>
<tr>
<td>69</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>([PJ] = pj)</td>
</tr>
<tr>
<td>71</td>
<td>([P] = p)</td>
</tr>
<tr>
<td>72</td>
<td>([BJ] = bj)</td>
</tr>
<tr>
<td>73</td>
<td>([B] = b)</td>
</tr>
<tr>
<td>74</td>
<td>([FJ] = fj)</td>
</tr>
<tr>
<td>75</td>
<td>([F] = f)</td>
</tr>
<tr>
<td>76</td>
<td>([VJ] = vj)</td>
</tr>
<tr>
<td>77</td>
<td>([V] = v)</td>
</tr>
<tr>
<td>78</td>
<td>([W] = w)</td>
</tr>
<tr>
<td>79</td>
<td>([MJ-D] = mj\ _h)</td>
</tr>
<tr>
<td>80</td>
<td>([M-D] = m\ _h)</td>
</tr>
<tr>
<td>81</td>
<td>([MJ] = mj)</td>
</tr>
<tr>
<td>82</td>
<td>([M] = m)</td>
</tr>
<tr>
<td>83</td>
<td>([TJSJ] = tjsj)</td>
</tr>
<tr>
<td>84</td>
<td>([DJZJ] = djzj)</td>
</tr>
<tr>
<td>85</td>
<td>([T__] = t__])</td>
</tr>
<tr>
<td>86</td>
<td>([TJ] = tj)</td>
</tr>
<tr>
<td>87</td>
<td>([T] = t)</td>
</tr>
<tr>
<td>88</td>
<td>([D__] = d__])</td>
</tr>
<tr>
<td>89</td>
<td>([DJ] = dj)</td>
</tr>
<tr>
<td>90</td>
<td>([D] = d)</td>
</tr>
<tr>
<td>91</td>
<td>([LLJ-D] = lj\ _h)</td>
</tr>
<tr>
<td>92</td>
<td>([LL-D] = l\ _h)</td>
</tr>
<tr>
<td>93</td>
<td>([LLJ] = lj)</td>
</tr>
<tr>
<td>94</td>
<td>([LL] = l)</td>
</tr>
<tr>
<td>95</td>
<td>([LJ-D] = lj\ _h)</td>
</tr>
<tr>
<td>96</td>
<td>([L-D] = l\ _h)</td>
</tr>
<tr>
<td>97</td>
<td>([LJ] = lj)</td>
</tr>
<tr>
<td>98</td>
<td>([L] = l)</td>
</tr>
<tr>
<td>99</td>
<td>([NNJ-D] = nj\ _h)</td>
</tr>
<tr>
<td>100</td>
<td>([NN-D] = n\ _h)</td>
</tr>
<tr>
<td>101</td>
<td>([NNJ] = nj)</td>
</tr>
<tr>
<td>102</td>
<td>([NN] = n)</td>
</tr>
</tbody>
</table>
[NJ_D] = nj h
[N_D] = n h
[NJ] = nj
[N] = n
[RRJ_D] = rj h
[RJ_D] = rj h
[RR_D] = r h
[R_D] = r h
[RRJ] = rj
[RJ] = rj
[RR] = r
[R] = r
[SJ] = sj
[S] = s
[ZJ] = zj
[Z] = z
[KJ] = kj
[K] = k
[GJ] = gj
[G] = g
[XJ] = xj
[X] = x
[GFJ] = gfj
[GF] = gf
[NGJ_D] = ngj h
[NG_D] = ng h
[NGJ] = ngj
[NG] = ng
[HJ] = h
[H] = h
[
] = 

[0] = 0
[1] = 1
[2] = 2
[.] = .
Letter-to-sound (LTS) rule files
B.20 Template/lex-exceptions.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file is the local exceptions lexicon for the <DIALECT REGION> Irish dialect.

// Syntax of lexicon entry:

// <ORTHOGRAPHY>
// <TAB><TRANSCRIPTION>

// Author: Brian Ó Raghallaigh
// oraghanb AG tcd PONC ie
B.21 Template/test.txt

// This file is a component of a multi-dialect modular phonetiser for Irish.
// A phonetiser converts orthographic text to dialectal phonemic transcription.
// This phonetiser has two layers: one global, one local.
// At a global level, spelling is converted to an abstract inter-dialectal linguistic representation.
// At a local level, this abstract linguistic representation is mapped to a dialectal transcription.
// The global layer consists of one global module.
// The local layer consists of one local module for each dialect implemented.
// Each module contains the following:
// (1) A set of rewrite (letter-to-sound) rules.
// (2) An exceptions lexicon.
// (3) A set of tests.

// This file contains tests for the <DIALECT REGION> Irish dialect.

// Test syntax: <INPUT><TAB><DESIRED OUTPUT>

// Author: Brian Ó Raghallaigh
// oraghalb AG tcd PONC ie
Appendix C

The Lárchanúint
<table>
<thead>
<tr>
<th>Phonemes</th>
<th>Allophones</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labials</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/b/</td>
<td>[b̃] or [b̃]</td>
<td>buf, bó</td>
</tr>
<tr>
<td>/b'/</td>
<td>[b] or [b̃]</td>
<td>bí, beo</td>
</tr>
<tr>
<td>/p/</td>
<td>[p̃] or [p̃]</td>
<td>pacaí</td>
</tr>
<tr>
<td>/p'/</td>
<td>[p] or [p̃]</td>
<td>peacaí</td>
</tr>
<tr>
<td>/v/</td>
<td>[ṽ] or [ṽ]</td>
<td>an-bhuf</td>
</tr>
<tr>
<td>/v'/</td>
<td>[v] or [ṽ]</td>
<td>bhí</td>
</tr>
<tr>
<td>/f/</td>
<td>[f̃] or [f̃]</td>
<td>go fóill</td>
</tr>
<tr>
<td>/f'/</td>
<td>[f] or [f̃]</td>
<td>feoil</td>
</tr>
<tr>
<td>/m/</td>
<td>[m̃] or [m̃]</td>
<td>maoin</td>
</tr>
<tr>
<td>/m'/</td>
<td>[m] or [m̃]</td>
<td>mín</td>
</tr>
<tr>
<td>Phonemes</td>
<td>Allophones</td>
<td>Examples</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td><strong>Coronals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/d/</td>
<td>[d̪]</td>
<td>dó</td>
</tr>
<tr>
<td>/dʼ/</td>
<td>[d̠]</td>
<td>deo</td>
</tr>
<tr>
<td>/t/</td>
<td>[t̪]</td>
<td>tuí</td>
</tr>
<tr>
<td>/tʼ/</td>
<td>[t̠]</td>
<td>tí</td>
</tr>
<tr>
<td>/l/</td>
<td>[l̪]</td>
<td>Gael, lón</td>
</tr>
<tr>
<td>/lʼ/</td>
<td>[l̠]</td>
<td>Gaeil, leon</td>
</tr>
<tr>
<td>/n/</td>
<td>[n̪]</td>
<td>anam, naoi</td>
</tr>
<tr>
<td>/nʼ/</td>
<td>[n̠]</td>
<td>ainm, ní</td>
</tr>
<tr>
<td>/r/</td>
<td>[ɾ̪] or [ɾ̠]</td>
<td>fuar</td>
</tr>
<tr>
<td>/rʼ/</td>
<td>[ɾ̠]</td>
<td>fuair</td>
</tr>
<tr>
<td>/z/</td>
<td>[z̪]</td>
<td>zú</td>
</tr>
<tr>
<td>/zʼ/</td>
<td>[z̠]</td>
<td>xioleafón</td>
</tr>
<tr>
<td>/s/</td>
<td>[s̪]</td>
<td>cás</td>
</tr>
<tr>
<td>/sʼ/</td>
<td>[ʃ̪] or [ʃ̠]</td>
<td>cáis</td>
</tr>
<tr>
<td>/dʼzʼ/</td>
<td>[ʧ̪]</td>
<td>jab, jíp</td>
</tr>
<tr>
<td>/tʼsʼ/</td>
<td>[ʤ̪]</td>
<td>tseo</td>
</tr>
<tr>
<td>Phonemes</td>
<td>Allophones</td>
<td>Examples</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Dorsals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/g/</td>
<td>[ɡ]</td>
<td>gall, óga</td>
</tr>
<tr>
<td>/gˈ/</td>
<td>[j]</td>
<td>geall, óige</td>
</tr>
<tr>
<td>/k/</td>
<td>[k]</td>
<td>cú, cad</td>
</tr>
<tr>
<td>/kˈ/</td>
<td>[c]</td>
<td>ciú, cead</td>
</tr>
<tr>
<td>/y/</td>
<td>[ɣ]</td>
<td>mo ghaol</td>
</tr>
<tr>
<td>/yˈ/</td>
<td>[i] or [j]</td>
<td>dhíol</td>
</tr>
<tr>
<td>/x/</td>
<td>[ʃ] or [x]</td>
<td>an chathair</td>
</tr>
<tr>
<td>/xˈ/</td>
<td>[s]</td>
<td>faoi cheathair</td>
</tr>
<tr>
<td>/ŋ/</td>
<td>[ŋ]</td>
<td>(na) longa</td>
</tr>
<tr>
<td>/ŋˈ/</td>
<td>[n]</td>
<td>(ar bord) loinge</td>
</tr>
<tr>
<td>Irish Glottals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/h/</td>
<td>[h]</td>
<td>hata, thit, a Shéamais</td>
</tr>
<tr>
<td>Phonemes</td>
<td>Allophones</td>
<td>Examples</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td><strong>Front vowels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/i:/</td>
<td>[i:]</td>
<td>sí, mín, cailín</td>
</tr>
<tr>
<td></td>
<td>[i:]</td>
<td>buí, croí, naoi, saol</td>
</tr>
<tr>
<td>/i/</td>
<td>[i] or [i]</td>
<td>ithub, sin, im</td>
</tr>
<tr>
<td></td>
<td>[i] or [i]</td>
<td>cuid, duine, fios</td>
</tr>
<tr>
<td>/eː/</td>
<td>[ɛː] or [ɛː]</td>
<td>mé, béal, céad</td>
</tr>
<tr>
<td></td>
<td>[ɛː]</td>
<td>Gaeilge, Fine Gael</td>
</tr>
<tr>
<td>/e/</td>
<td>[e] or [ɛ] or [ɛ]</td>
<td>ceist, te, seic</td>
</tr>
<tr>
<td><strong>Low vowels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/aː/</td>
<td>[æː] or [aː] or [ɑː]</td>
<td>tá, bán, sláinte</td>
</tr>
<tr>
<td>/a/</td>
<td>[æ] or [a] or [ɑ]</td>
<td>ainm, bean, mac</td>
</tr>
<tr>
<td>Phonemes</td>
<td>Allophones</td>
<td>Examples</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>Back vowels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/u:/</td>
<td>[u:] or [u:]</td>
<td>tú, Cúige Mumhan, ciúin</td>
</tr>
<tr>
<td>/u/</td>
<td>[u] or [u]</td>
<td>dubh, thug, fliuch, tiubh</td>
</tr>
<tr>
<td>/o:/</td>
<td>[o:] or [o:]</td>
<td>bó, ceol, móir</td>
</tr>
<tr>
<td>/o/</td>
<td>[o] or [o] or [ɔ]</td>
<td>mothú, scoil, seo, curáí, murdar</td>
</tr>
<tr>
<td><strong>Centralised vowels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/a/</td>
<td>[a]</td>
<td>abhus, mála, solas, dearag, dorcha</td>
</tr>
<tr>
<td></td>
<td>[i]</td>
<td>baile, inné, milis, ainm, seirbhís</td>
</tr>
<tr>
<td><strong>Diphthongs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/ai/</td>
<td>[ai] or [æi]</td>
<td>radharc, saghas, veidhlín, leigheas</td>
</tr>
<tr>
<td></td>
<td>[iæi]</td>
<td></td>
</tr>
<tr>
<td>/au/</td>
<td>[au] or [əu]</td>
<td>samhradh, domhan, bodhar</td>
</tr>
<tr>
<td></td>
<td>[əu]</td>
<td>leabhar, sleamhain</td>
</tr>
<tr>
<td>/iə/</td>
<td>[iə] or [iə] or [iə]</td>
<td>bia, siad, riail</td>
</tr>
<tr>
<td>/uiə/</td>
<td>[uə] or [uə] or [uə]</td>
<td>rua, fuar, buail</td>
</tr>
</tbody>
</table>